# AIED 2015:

# 17th International Conference on Artificial Intelligence in Education

Workshop Proceedings

Edited by:

Jesus Boticario
*aDeNu Research Group, UNED, Spain*

Kasia Muldner
*Carleton University, Canada*

June 22 – 26, 2015

Madrid, Spain

# Preface

The supplementary proceedings of the workshops held in conjunction with AIED 2015, the seventeenth International Conference on Artificial Intelligence in Education, June 22-26, 2015, Madrid, Spain, are organized as a set of volumes - a separate one for each workshop.

The set contains the proceedings of the following workshops:

**Volume 1: Sixth International Workshop on Culturally-Aware Tutoring Systems (CATS)**
*Ma Mercedes T. Rodrigo, Emmanuel G. Blanchard, Amy Ogan, Isabela Gasparini*

**Volume 2: Intelligent Support in Exploratory and Open-ended Learning Environments;**
**Learning Analytics for Project Based and Experiential Learning Scenarios**
*Manolis Mavrikis, Gautam Biswas, Sergio Gutierrez-Santos, Toby Dragon, Rose Luckin, Daniel Spikol, James Segedy*

**Volume 3: Fourth Workshop on Intelligent Support for Learning in Groups (ISLG)**
*Ilya Goldin ,Roberto Martinez-Maldonado,Erin Walker, Rohit Kumar ,Jihie Kim*

**Volume 4: Workshop on Les Contes du Mariage: Should AI stay married to Ed?**
*Kaska Porayska-Pomsta, Gord McCalla, Benedict du Boulay*

**Volume 5: Second Workshop on Simulated Learners**
*John Champaign and Gord McCalla*

**Volume 6: Workshop on Developing a Generalized Intelligent Framework for Tutoring (GIFT): Informing Design through a Community of Practice**
*Benjamin Goldberg, Robert Sottilare, Anne Sinatra, Keith Brawner, Scott Ososky*

**Volume 7: International Workshop on Affect, Meta-Affect, Data and Learning (AMADL 2015)**
*Genaro Rebolledo-Mendez, Manolis Mavrikis, Olga C. Santos, Benedict du Boulay, Beate Grawemeyer, Rafael Rojano-Cáceres*

While the main conference program presents an overview of the latest mature work in the field, the AIED2015 workshops are designed to provide an opportunity for in-depth discussion of current and emerging topics of interest to the AIED community. The workshops are intended to provide an informal interactive setting for participants to address current technical and research issues related to the area of Artificial Intelligence in Education and to present, discuss, and explore their new ideas and work in progress.

All workshop papers have been reviewed by committees of leading international researchers. We would like to thank each of the workshop organizers, including the program committees and additional reviewers for their efforts in the preparation and organization of the workshops.

June, 2015
Jesus Boticario & Kasia Muldner

# Sixth International Workshop on Culturally-Aware Tutoring Systems (CATS2015)

*held in conjunction with*

Seventeenth International Conference on Artificial Intelligence in Education (AIED 2015)

Monday, June 22, 2015
Madrid, Spain


**Workshop Co-Chairs:**

Ma Mercedes T. Rodrigo[1], Emmanuel G. Blanchard[2], Amy Ogan[3], Isabela Gasparini[4]

[1]*Ateneo de Manila University, Philippines*
[2]*IDÛ Interactive, Montreal, Canada*
[3]*Carnegie Mellon University. USA*
[4]*University of Santa Catarina State (UDESC), Brazil*


http://cats-ws.org

# Table of Contents

# Preface

Culture has a profound effect on the way people interact with, react to, think and feel about knowledge, symbols, situations, etc. Yet it is underestimated in AIED research. Most of the currently influential learning systems have indeed been created by and for developed world contexts and with Western cultural perspectives in mind. However in recent years, more and more opportunities to design, develop, and deploy educational software for and in different contexts have emerged. This state of affairs naturally leads to broader questions. What features of culture are important to consider in the design process? Can software designed and developed in a specific cultural context transfer to other parts of the world and remain effective? The answers to these questions remain unclear although a growing body of research suggests that the use of AIED systems across cultural contexts results in variations of the knowledge acquisition process.

Over the last seven years, Culturally-Aware Tutoring Systems (CATS) workshops have been organized in conjunction with ITS2008, AIED2009, ITS2010, AIED2013, and ITS2014. The series is a venue for researchers to reflect on the universality of their work. CATS2015 thus proposes to discuss culture and AIED from five perspectives:

1. Developing both pedagogical strategies and system infrastructure mechanisms that incorporate cultural features to enculturate AIED systems;
2. Designing acquisition-oriented CATS, i.e. AIED systems to teach cultural knowledge and intercultural skills;
3. Designing adaptation-oriented CATS, i.e. AIED systems that can be personalized overtly or automatically based on users' cultural profiles;
4. Considering human features that are connected with the learning process, and that are culturally-sensitive, e.g. affect, behavior, cognition, or motivation; and
5. Considering cultural biases in the AIED research cycle.

In addition to describing the current state of the art in these domains, the workshop engages participants in working to expand the reach of AIED research to a greater global audience, including those disadvantaged due to a lack of resources or other obstacles.

Overseeing the quality of CATS2015 papers was a program committee of 37 members from Asia, Europe, North America, and South America. The program committee members were well-versed in AIED, culture, technology, and other relevant fields. The committee selected 4 full papers and 1 short paper for inclusion in this year's workshop.

We thank all the program committee members and authors for contributing their time and expertise to making CATS2015 possible. We also thank the Workshop Chairs and the Organizing Committee of AIED2015 for including CATS in this year's conference.

Ma Mercedes T. Rodrigo, Emmanuel G. Blanchard, Amy Ogan, & Isabela Gasparini
The CATS2015 Co-Chairs

# Leveraging Comparisons between Cultural Frameworks: Preliminary Investigations of the MAUOC Ontological Ecology

Phaedra Mohammed[1], Emmanuel G. Blanchard[2]

[1]Department of Computing and Information Technology, The University of the West Indies, St. Augustine, Trinidad and Tobago
phaedra.mohammed@gmail.com

[2]IDÛ Interactive Inc.
Montreal, Canada
ebl@idu-interactive.com

**Abstract.** Many theoretical cultural frameworks have been proposed in the literature. For comparisons and critiques of these frameworks to make sense, community members have to assign similar-enough meanings to the terms that they use when interacting. This entails overcoming the challenge of dealing with the imprecise and interpretable definitions conveyed in frameworks due to the use of common language. The MAUOC Ontological Ecology (MOE) approach offers a strategy for dealing with this through reinterpretation of all cultural frameworks along a singular, common conceptual baseline. In this way, a far more cohesive, consistent, and controlled representation of cultural frameworks becomes available compared to just common language descriptions. The purpose of this paper is to clarify the MOE methodology, and report initial efforts into practically applying it to the Hofstede cultural framework.

**Keywords:** Culture, Heavyweight Ontology, Systematic Methodology, Hofstede Framework

## 1. Introduction

Culture is a key phenomenon in many academic disciplines such as psychology, anthropology, sociology, education, philosophy, and therefore has been studied from diverse perspectives. Consequently, many theoretical frameworks have been proposed, each with specific purposes as endorsed by different research communities. These frameworks are mostly described with common language terms which disguise the complexity and philosophical nuances within. For these reasons and others, frameworks are frequently prone to misinterpretation, and disagreements are common when conflicting claims are made regarding particular frameworks. A common source of dispute is the use of the same terminology across frameworks which may or may not refer to the same conceptualization, such as *Individualism* and *Collectivism* in the GLOBE and Hofstede frameworks [4].

As an emerging interdisciplinary field, research on Culturally-Aware Tutoring Systems (CATS) is driven by scholars with different profiles, both in terms of cultural

backgrounds and expertise. This rich diversity places the CATS community in a unique position to properly tackle the techno-cultural objectives it has assigned to itself. However, the variety of existing cultural frameworks and the lack of time for many community members to deeply understand them creates challenges for cumulating research efforts and findings. Indeed, for comparisons and critiques to make sense, community members have to assign similar-enough meanings to the terms that they use when interacting. This is one way of overcoming the challenge of dealing with the imprecise and interpretable definitions conveyed in frameworks due to the use of common language.

The More Advanced Upper Ontology of Culture (MAUOC) aims to identify conceptual building blocks of the cultural domain, and it has several potential applications for CATS. The one that is considered in this paper is the possibility it offers for reinterpretation of all cultural frameworks along a singular, common conceptual baseline. In this way, a far more cohesive, consistent, and controlled representation of cultural frameworks becomes available compared to just common language descriptions. This would in turn promote objective comparisons between frameworks, and enhance interoperability between research efforts. Before this can be done, a structured, scientific methodology is necessary. One such strategy has been theorized and presented in [3]. It is referred to as the MAUOC Ontological Ecology (MOE) approach, and the purpose of this paper is to clarify this methodology, and report initial efforts into practically applying it to the challenges articulated earlier.

The remainder of the paper is organized as follows. Section 2 presents a justification for the choice of heavyweight ontology engineering as the basis for this research, and briefly describes the development processes behind MAUOC and the MOE approach which motivate the systematic methodology taken in the paper. Section 3 goes into the specifics of this methodology, briefly describes the Hofstede cultural framework, and gives insight regarding why this framework was chosen for analysis. The section then provides illustrative examples arising from the preliminary analysis of the Hofstede framework using the MOE approach, along with a brief discussion of each example. Section 4 discusses what is to be learnt from this preliminary investigation and identifies the limitations of the work so far. The paper concludes in Section 5 with future plans for the investigation.

## 2. Ontological Grounding of our Analytical Process

### 2.1 A Heavyweight Ontology Initiative

Heavyweight ontology engineering is strongly connected to the original philosophical meaning of ‘ontology’. Whereas heavyweight and other (lightweight) ontologies look similar to non-specialists (simply put, they could be seen as a set of concepts/constructs interconnected with relations), the critical difference lies in the way heavyweight vs lightweight ontologies assign identities to these concepts/constructs and relations. Authors of lightweight ontologies commonly refer to a ‘rule of thumbs’ approach: they may look for, and accept a definition that makes sense to them in the context of the specific application(s) they have in mind, and according to their personal experience. This obviously limits its applicability while bringing risks of per-

sonal and socio-cultural biases. Heavyweight ontologies on the other hand must not target a specific application, but rather aim to capture the true essence of a domain or task (as in philosophy). A definition obtained following proper heavyweight ontological analyses can thus be reapplied in any situation related to the domain of interest.

Eventually, distinctions between heavyweight and lightweight ontologies are largely ignored by non-specialists. This is a major issue since these ontologies have very different properties. However, the purpose of this paper is not to reflect upon this point, and readers are invited to look at [8] for clarifications. Overall, if heavyweight ontologies are innately superior from a conceptual perspective, they have a major drawback: they are far more complex and consequently require more expertise and development time before being considered to be sufficiently stable for use. But for ontology specialists, these difficulties are overshadowed by the breadth of applicability and the subsequent interoperability that heavyweight ontologies allow once stable-enough. We therefore adopt a heavyweight ontological approach because capturing the philosophical essence of cultural frameworks requires careful, precise definitions that can bridge the operational data/solutions produced by different disciplines [3].

### 2.2 From MAUOC to MOE: Two Phases in Framework Reinterpretations

Initiated in 2008 [1], MAUOC is a heavyweight ontology initiative. Rather than describing MAUOC itself, which is prohibitive in this paper due to space constraints (see [3] for an overview), we will now make a brief presentation of MAUOC's development process. This is essential for understanding the remainder of the paper because it forms the basis for the systematic methodology described in the next section. The process has several objectives:

- Distinguishing 'natural concepts' (i.e. conceptual units which exist inherently in nature. See [8]) from 'constructs' (i.e. artificial conceptual units defined in the context of a framework to better carry out its message, connect with a user community, and/or facilitate its adoption and use) for the cultural domain,
- Providing precise definitions for natural concepts by figuring out their essential parts and properties. These features are 'essential' because the removal of one of them leads instances to be classifiable in more than one definition. In the same time, a proper definition has to respect Okham's razor principle, i.e. the simplest definition is always the best one.

The development process of MAUOC can thus be decomposed into five steps:

1. Acquiring a deep understanding of several cultural frameworks representing different schools of thought and disciplines
2. Identifying major framework terms as 'natural concept' candidates
3. Classifying the ideas behind these terms as trans-framework or framework-specific into a more restricted ensemble of 'natural concept' candidates while discarding those that are too specific or not innately cultural
4. Eliciting ontology-grade definitions for the remaining 'natural concept' candidates and their relations, and testing if the resulting ecology of concepts allows for expressing any cultural situations and issues that may arise
5. Iteratively repeating one or more of the previous steps if d) has failed, because this would mean that the current version of the ontology is incomplete, and/or includes inappropriately-defined elements.

In the course of its development, MAUOC has thus been revised many times before reaching the first version thought to be stable-enough [3]. Yet, one cannot be certain that the current version of MAUOC will not be challenged by cultural issues to be tested in the future. Developing MAUOC is both a top-bottom and bottom-up process that attempts to identify cultural building blocks by cross-analysing various frameworks. Now that a stable-enough version has been proposed, the MAUOC Ontological Ecology (MOE) aims to further this initiative by following a bottom-up approach where ontological translations of cultural frameworks will be designed and grounded on these building blocks. In other words, the goal of MOE is not to state what frameworks should or should not say, but rather to achieve clearer and more precise formulations of what they already intend to say.

Figure 1 presents a simplified view of MAUOC and MOE processes. Note that YAMATO is a top ontology, on which MAUOC is grounded (see [9]).

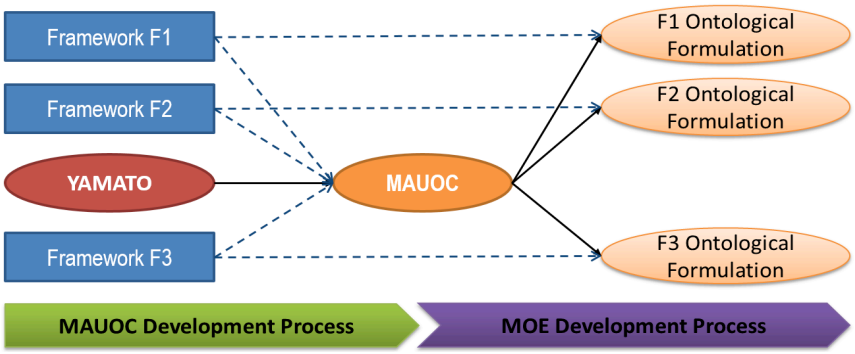



**Figure 1**. A Simplified View of the MAUOC and MOE Development Processes.

## 3. Applying the MOE Approach to Hofstede's Framework

### 3.1 A Systematic Methodology

The systematic methodology described in this section is framework-independent and therefore it can be applied to any cultural framework for which intercultural comparisons are desired using the MOE approach. It is important to note that this process first requires the perspective of external reviewers who have no connection to the particular framework being studied in order to guard against bias [2]. This is crucial since the analysis deals with matters of interpretation and comparison of meanings. At this early stage, only the two authors of the paper are solely involved in the process. Both authors are independent of the cultural framework to which the methodology is being applied and both have different cultural backgrounds which provide an additional layer for guarding against bias.

a) Identify major references for the cultural framework within the literature. Here, sources may include books, journal articles, or conference papers where the overarching quality is the frequency of reference.
b) Identify key terms and several corresponding quoted definitions within these references, by authors of the framework and/or the representative user com-

munity. Key terms, for our purposes, refer to words or phrases which define essential features or ideas that contribute towards the major theoretical underpinnings of a cultural framework.

c) Highlight any discrepancies, consistencies, and/or differences (if any) in the quoted definitions for the key terms. Two levels of analysis are performed in this step: Terminological analysis - which asks whether the definition is consistent over time from a grammatical and a lexical perspective, and Conceptual/ontological analysis - which asks whether the definition is precise enough. Consistency refers the number of changes in the grammatical and lexical structure across the quoted definitions, and it is used to assess whether those changes may alter the meaning in the definitions over time. Precision refers to the self-explanatory nature of expression used in the quoted definition, and the extent to which that expression is potentially subject to interpretations amongst readers.
d) Determine whether a coherent, durable definition can be extracted for each key term. In this step, a key term would still be expressed in common language, but it would now be ontology-ready. In other words, the term would have a logical and consistent structure that is made up of several other conceptualizations that fit together precisely.
e) Consult with experts of the cultural framework to assess the validity of the extracted definitions in keeping with the intended 'spirit' of the framework. If necessary, the definitions would be refined or modified to eventually come to a consensual definition that satisfies both the experts and reviewers while still remaining ontology-ready.
f) Interpret and convert the resulting common language, consensual definitions to MAUOC-grade formulations, using logical representations such as mathematical notations or those originating from HOZO.

Our approach currently focuses on achieving 'heavyweight ontology'-grade definitions for constructs articulated in various cultural frameworks, and as such it only partially reflects the vision stated in MOE. Subsequent and interleaved steps are thus required to clearly state relations and dependencies between these construct definitions in order to achieve true MAUOC-grounded ontologies.

### 3.2 Primer on Hofstede's Framework

The Hofstede cultural framework was chosen as the starting point in this research for several reasons. Firstly, it is the most popular one used in intercultural research as evidenced by the large body of work using the framework for theoretical and practical reference. Due to over 30 years of study, it is also one of the best documented and consequently one of the most attacked and critiqued of the available frameworks. This rich body of work and the clear evolution that naturally has taken place in the framework due to intense scrutiny, further provides a good distribution of terms upon which to test our methodology.

A brief description of the Hofstede framework is necessary at this point in order to give readers a sense of what the framework is about. The Hofstede framework takes an empirical, generalized approach towards studying cultural differences. It focuses on the identification of dimensions of national culture which were originally: Power Distance, Individualism, Masculinity, and Uncertainty Avoidance [5]. Since then, two

more dimensions have been added to the framework: Long Term Orientation and Indulgence vs Restraint [7]. These dimensions are used to score and classify countries according to how members of those societies cope with problems and concerns that are basic to all human societies [7]. Using these scores and statistical relationships between the dimensions, the framework quantified the differences reported across 40 countries originally in 1980. The data set has since been extended to 107 countries [7]. Country clusters were used to account for cultural observations about behaviour which may apply at various levels (national, regional, individual). Table 1 shows definitions of the six Hofstede dimensions, as well as scores for three countries.

**Table 1.** Hofstede Dimensions and Country Scores for Three Sample Countries

| ***Hofstede's Dimension*** | ***Dimension Description*** | ***U.S.A.*** | ***Spain*** | ***Japan*** |
|---|---|---|---|---|
| *Power Distance* | The degree to which the less powerful members of a society accept and expect power to be distributed unequally | 40 | 57 | 54 |
| *Individualism* | Preference for a loosely-knit social framework | 91 | 51 | 46 |
| *Masculinity* | Preference for achievement, material rewards, assertiveness over modesty, cooperation, caring | 62 | 42 | 95 |
| *Uncertainty Avoidance* | The degree to which members of a society feel uncomfortable with uncertainty and ambiguity | 46 | 86 | 92 |
| *Long Term Orientation* | The degree to which a society maintains links with its own past while dealing with challenges of the present and future | 26 | 48 | 88 |
| *Indulgence vs Restraint* | The degree to which a society allows relatively free gratification of basic and natural human drives over suppression and regulation with strict social norms | 68 | 44 | 42 |

### 3.3 Illustrative Examples and Analyses

In applying the MOE systematic methodology to the Hofstede framework, three reference sources [5, 6, 7] were selected. These three refer to some of the most commonly cited sources of the framework, and together they cover over 30 years of the framework's evolution: the original source in 1980, the currently most cited source from 2001, and the most recent source in 2010. To illustrate part of the process, only 6 framework-specific terms were selected for analysis and presentation in this paper due to space constraints. The 6 key terms were chosen since they are core terms for the Hofstede framework (and most other frameworks), they test different situations in the methodology, and they are commonly used in the user community. These terms

are considered according to their meaning in the scope of the Hofstede's framework. Hence there must be no confusion between some of these constructs (e.g. *value,* or *dimensions)* and heavyweight ontology concepts using the same labels (see [9]).

Table 2 below shows the directly quoted definitions (if present) extracted for each key term from each source. Summarized, unquoted descriptions are provided if there were no formal definitions found for a given key term. The sources [5, 6, 7] are referred to as 1), 2), and 3) respectively. It should be noted that only the first three steps of the systematic methodology were carried out on the Hofstede framework in this paper.

**Table 2.** Six Key Terms in Hofstede's Framework and their Representative Definitions in Reference Sources from 1980, 2001, and 2010.

| ***Key Terms*** | ***Key Term Definitions from Hofstede Sources*** |
|---|---|
| *Value* | 1) "*A value is a broad tendency to prefer certain states of affairs over others*." (1980, p.19)<br>2) "*A value is a broad tendency to prefer certain states of affairs over others*." (2001, p.9)<br>3) "*Values are broad tendencies to prefer certain states of affairs over others*." (2010, p.9) |
| *Culture* | 1) "*The collective programming of the mind which distinguishes the member of one human group from another*." (1980, p.25)<br>2) "*The collective programming of the mind that distinguishes the members of one group or category of people from another*." (2001, p.9)<br>3) "*The collective programming of the mind that distinguishes the members of one group or category of people from others*." (2010, p.6) |
| *Dimension* | 1) Empirically verifiable, independent phenomena (behaviours of individuals or situations, institutions, or organizations) on which cultures can be meaningfully ordered. (1980, p.36)<br>2) A dimension is described by two possible extremes which can be seen as ideal types. "*A dimension is rooted in a basic problem which all societies have to cope, but on which their answers vary*." (2001, p.28-29)<br>3) "*A dimension is an aspect of a culture that can be measured relative to other cultures*." A dimension groups together a number of phenomena in a society that were empirically found to occur in combination. (2010, p.31) |
| *Individualism* | *1) "... the relationship between the individual and the collectivity which prevails in a given society*." (1980)<br>*2) "... the relationship between the individual and the collectivity that prevails in a given society*." (2001, p.209). "*Individualism stands for a society in which the ties between individuals are loose: Everyone is expected to look after her/his immediate family only*." (2001, p.225)<br>*3)* "*Individualism pertains to the societies in which the ties* |

| | |
|---|---|
| | *between individuals are loose: everyone is expected to look after him- or herself and his or her immediate family*." (2010, p.92) |
| *Collectivism* | 1) No formal definition in the 1980 source.<br>2) "*Collectivism stands for a society in which people from birth onwards are integrated into strong, cohesive in-groups, which throughout people's lifetime continue to protect them in exchange for unquestioning loyalty*." (2001, p.225). "*Collectivism is the degree to which individuals are supposed to remain integrated into groups usually around the family*." (2001, p. xx)<br>3) "*Collectivism pertains to societies in which people from birth onward are integrated into strong, cohesive in-groups, which throughout people's lifetime continue to protect them in exchange for unquestioning loyalty*." (2010, p.92) |
| *IDV Dimension* | 1) "*It describes the relationship between the individual and the collectivity which prevails in a given society*." (1980)<br>2) "*It describes the relationship between the individual and the collectivity that prevails in a given society*." (2001, p.209) "*Individualism versus collectivism is related to the integration of individuals into primary groups*." (2001, p. 29). The IDV dimension is defined also by combining the Individualism and Collectivism definitions from 2) above.(2001, p.225)<br>3) The IDV Dimension is defined by combining the Individualism and Collectivism definitions from 3) above. (2010, p.92) |

**Value.** Terminologically, the definition of value is cohesive from 1980 to 2010 with one grammatical change in 2010. The grammatical change, i.e. pluralisation, does not affect the meaning of the definition so it is cohesive from this perspective. However it is ontologically since inner terms leave room for interpretation (*state of affairs*, *broad tendency* – what do they refer to? Are these to be understood from a group, individual, or both levels?).

**Culture.** The definition is terminologically-inconsistent due to changes between 1980 and 2001 from *member* to *members*, and *one human group* to *one group or category of people*, and from *another* to *others* in 2010. In all of the definitions, comparisons are made between A and B, but the nature of A and B changes with each evolution of the definition. This has ontological implications for the cardinality of the comparisons namely a shift from a one-to-one comparison between two individuals in 1980 to a many-to-many comparison across individuals from two groups in 2001 to a broader comparison between not just two groups but amongst many groups in 2010. There are also imprecise inner terms: *collective programming of the mind* and *human group*.

**Dimension.** The first plain definition for dimension is found in the 2010 source. The term was used and described in 1980 and 2001 across a few pages, however neither source provides a precise definition; the salient parts are summarised in Table 2. Terminologically, there is no cohesion amongst the descriptions. Ontologically, the lack of more than one plain definition provides more room for interpretation. The 2001 quote is imprecise since inner terms (*rooted on*, *basic problem*) are subject to

interpretation, whereas *society* is not clearly defined. The 2010 quote is also ontologically imprecise due to interpretable inner terms such as *aspect*, and *culture*. The measurable property of a dimension is however coherently and consistently articulated across all three sources.

**Individualism.** The quotes are terminologically cohesive for the first part between 1980 and 2001. The additional section added in 2001 is not cohesive with 1980, and not consistent with the 2010 due to two evolutions: society to societies and immediate family only to him or herself and his or her immediate family. Ontologically, there is a change in cardinality as in the culture definition, and the inner terms are imprecise in 1980 (*relationship*), and imprecise and subjective in both 2001 and 2010 (*ties, loose*).

**Collectivism.** Terminologically there is limited cohesion with no formal definition in 1980, and one evolution between the common quotes in 2001 and 2010: society changes to societies. Ontologically, the definitions in 2001 and 2010 are imprecise due to inner terms requiring further explanations (*strong, cohesive in-groups, society, protect* - from what, why, and by whom? -, *unquestioning loyalty* - allegiance to whom?, forced or voluntary? -).

**IDV (Individualism-Collectivism) Dimension.** The quotes from 1980 and the first part of 2001 are terminologically cohesive but ontologically imprecise due to inner terms requiring further definition (*relationship, collectivity)*. The quotes from the second part of 2001 and that of 2010 have the same outcome as the individualism and collectivism analyses above.

## 4. Discussion

The analysis in the previous section should not be construed as a criticism or praise of the Hofstede framework, nor should it be seen as an effort to create our own definitions for key terms. Rather, the intention is to raise awareness of the possible interpretations of the framework's core terms which can have wide-reaching implications for CATS research especially if misunderstanding and oversights are not cleared up. Contradictions from incorrect usage of framework term can lead to wrong conclusions in educational applications, and cascade dangerously in culturally-aware contexts. The goal is therefore to understand the cultural framework and confirm whether existing definitions are prone to significant misunderstandings.

At this point we cannot say that the MOE methodology is fully validated yet since the research is still in its early stages. More work is needed, and naturally there are limitations. Only three quotes were used for each term and we agree that more and deeper reflection is needed for each term in order to solidify the analysis. In addition, quotes were sourced from material written by authors of the framework only. User community quotes can help identify further misunderstandings as well as consensus from a broader perspective, and should be investigated as well. Finally, only the first three steps of the MOE systematic methodology were carried out on the Hofstede framework. Despite this, clear risks of misinterpretation were identified for key term definitions in the framework in these early, simple stages. As ontology-ready definitions are extracted and validated through consultation with experts of the cultural framework, the systematic process hopefully will reveal weaknesses in the MOE

approach as well as provide additional validation of the soundness of existing concepts in MAUOC. For example, if a definition requires particular concepts that should have been defined in MAUOC, the missing concepts can be added to strengthen the ontology. If successful, this investigation will then create a baseline for analysing other existing cultural frameworks, and produce further validation of MAUOC as a deep ontological model of culture. Folk-based validation of definitions could also provide practical insight since ontologies, both lightweight and heavyweight, require a community of users. This type of validation however needs to be moderated since reliance on inexperienced users can lead to the design of a folksonomy. It is nonetheless still useful to be considered for future work.

## 5. Conclusion and Future Research

Derived from the MAUOC Ontological Ecology (MOE) approach, this paper presented a systematic methodology for overcoming the challenge of dealing with the imprecise and interpretable definitions conveyed in cultural frameworks due to the use of common language. Preliminary analysis of the Hofstede framework, using the MOE approach, indicates that the methodology is holding up. The next steps involve analysis of more Hofstede framework key terms, such as national culture, and country score for examples, and figuring out whether ontology-ready definitions are possible for the quoted definitions collected thus far in consultation with framework experts.

## References


1. Blanchard, E.G., Mizoguchi R.: Designing culturally-aware tutoring systems: Toward an upper ontology of culture. In E. Blanchard, and D. Allard (eds.), Proc. Culturally Aware Tutoring Systems, 23-34. (2008)
2. Blanchard, E.G.: Is it adequate to model the socio-cultural dimension of e-learners by informing a fixed set of personal criteria? In Proc. 12th IEEE International Conference on Advanced Learning Technologies. 388-392. USA: IEEE Computer Society. (2012)
3. Blanchard, E.G., Mizoguchi R.: Designing Culturally-Aware Tutoring Systems with MAUOC, the More Advanced Upper Ontology of Culture. *Research and Practice in Technology Enhanced Learning* 9(1): 41-69. (2014)
4. Brewer, P., S. Venaik.: Individualism-Collectivism in Hofstede and GLOBE. *Journal of International Business Studies* 42: 436-445. (2011)
5. Hofstede, G.: Culture's Consequences: International Differences in Work-Related Values. Beverly Hills, CA: Sage. (1980)
6. Hofstede, G.: Cultures' Consequences: Comparing Values, Behaviours, Institutions, and Organizations Across Nations. Thousand Oaks, CA: Sage Publications Inc. (2001)
7. Hofstede, G., G-J. Hofstede, M. Minkov.: Cultures and Organizations: Software of the Mind: Intercultural Cooperations and Its Importance for Survival. NY: McGraw-Hill. (2010)
8. Mizoguchi, R.: Tutorial on ontological engineering - part 1: Introduction to ontological engineering. *New Generation Computing* 21(4): 365–384. (2003)
9. Mizoguchi, R.: YAMATO: Yet Another More Advanced Top-level Ontology, *Proceedings of the Sixth Australasian Ontology Workshop Adelaide (AOW2010)*, 1-16. (2010)

# Exploring Power Distance, Classroom Activity, and the International Classroom Through Personal Informatics

David Gerritsen, John Zimmerman, Amy Ogan

Human-Computer Interaction Institute,
Carnegie Mellon University, Pittsburgh, USA
{dgerrits, johnz, aeo}@cs.cmu.edu

**Abstract.** Research shows the benefits of active learning in American college classrooms. International graduate students in American universities may face difficulties in teaching students with different cultural dispositions. The current research uses power distance to explore cultural juxtapositions in classrooms and personal informatics design to propose an adaptive system for cultural acquisition. The work shows that even though instructors are aware of the distinctly Western value of speaking up in class, they do not employ it in their own classes. They show surprise at the amount of time they spend lecturing, but they express ambivalence about the importance of vocal contributions from the students. We describe a technical system design that supports the development of cultural fluency by providing ITAs with feedback such as visualizations of time spent lecturing and suggestions for strategy selection in culturally challenging scenarios. The system would reflect changes in classroom activity over time as a way for TAs to reflect on their own professional development.

**Keywords:** Power distance, international teaching assistants, classroom activity, personal informatics

## 1 Introduction

Research in the learning sciences has recently produced an explosion of experimental evidence that college students benefit from less lecture and more student activity. This evidence exists even for content-heavy science, technology, engineering, and mathematics (STEM) classes where instructors have traditionally emphasized the importance of covering and memorizing facts rather than exploring, curating, and constructing knowledge. Most of these studies have taken place in American classrooms and have not addressed questions of cultural dimensions of learning and teaching. Meanwhile, the number of international graduate students teaching introductory STEM classes in American universities continues to grow. These students tend not to have experienced the cultural shift toward active learning and its concomitant decrease in social distance to figures of authority that is familiar to most students from the U.S. This can lead to challenges for international graduate students in the U.S. when they are required to teach American students.

The CATS community has a history of developing systems to improve education and cultural awareness. We build on this line of research by focusing on new design methods that frame the instructor as both the learner and the agent of change in the classroom. Using methods from *Personal Informatics* (PI), we explore the state of international teaching assistants (ITAs) leading STEM classes in an American university, and propose a system that potentially simplifies the implementation of active learning in order to more fully engage students.

PI is an approach to behavior change and maintenance that gathers user data and generates digital artifacts for reflection, such as visualizations of change toward a behavioral goal. Very little research has looked at its value in education, and none has attempted to use it to better understanding culture. It incorporates methods of contextual design and development that may be valuable in improving educational outcomes while investigating culturally adaptive interactions.

To assess the feasibility of this line of research and development, we carried out several overlapping activities: classroom observation of ITAs in action in order to understand the context need for adaptive instruments, surveys and interviews in order to understand how ITAs might make sense of classroom behavior, and data visualization feedback for ITAs in order to understand and explore the potential interface for a PI system. Finally we constructed and evaluated a prototype classroom detection system to investigate if we could sense relevant behaviors.

We confirmed that ITAs' knew of the cultural value of classroom activity, yet their recitations were almost completely based on lecture, with little student participation. They were open to more classroom activity, but with some reservations. They shared an interest in monitoring their teaching behaviors and aligning their performance with expert models. Also, our technical system functioned with 85% accuracy. We propose that these findings support further investigation of PI methods for investigating and supporting the acquisition of cultural fluency in unfamiliar educational contexts.

## 2 Background

Several decades of research in U.S. higher education has produced a wealth of studies showing the benefits of active learning compared to passive lecture and fact memorization [1, 2, 3, 4]. These studies have investigated and advocated active learning tactics such as *think-pair-share* and *cooperative learning*, showing that students improve academically, socially, and psychologically [1, 4]. Like most education research, the studies tend not to include considerations of cultural dimensions of learning. Cultural dimensions of instructors and learners in American universities are poorly understood. Given the evidence that different cultures have different valuations of student activity in the classroom [5, 6], the call for increased student participation may create a tension when it fails to address how international instructors perceive and value active learning practices. This situation deserves attention as the number of international graduate students teaching STEM classes in the U.S. continues to grow [7].

One way to orient the conversation about cultural differences in praxis is to frame it in terms of power distance [8, 9]. Higher and lower national indices of power dis-

tance (PDI) attempt to describe the level of deference that individuals express toward members of higher and lower social status. Given the long history of measurable social distance between Asian students and American instructors [5, 6], power distance is a reasonable construct with which to study classroom practices. It seems to have a direct mapping to the differences students exhibit as a function of cultural orientation to learning [9]. A low PDI score of 40 in the U.S., compared to 77 and 80 in India and China [8], may partially explain these students' general tendencies to speak or remain silent when they attend American university classes, regardless of how well they know the material [6].

This distance is becoming increasingly important to address. International enrollment to American graduate schools has grown since 2005, with the most recent report showing a 17% jump in enrollment to engineering schools and a 40% increase in graduate students from India [7]. These students often fund their education by teaching small classes that act as a supplement to large introductory STEM courses. These small classes, normally called *recitations*, allow groups of undergraduates from a large class to review course material and interact more closely with each other and an expert instructor.

Although many states require ITAs to pass an oral proficiency exam before teaching, there is little support for developing cultural fluency (or even general teaching skills). In other domains, such as health and finance, PI has recently emerged as a technique for motivating changes in behavior [10–13] with only a small investment of time or conscious effort on the part of the user. It is a new class of socio-technical system based on self-monitoring through data visualization [14]. The process helps motivate people to make new decisions by increasing their awareness of behaviors that are normally obscure and hard to observe, such as encouraging more activity by showing people a record of how much (or how little) they move throughout the day. That awareness is a critical step in the process of making changes [12]. These systems have gained popularity due to advances in wearable technology and smartphones. Current PI systems can track a user's number of steps [10], hours and quality of sleep [15], levels of glucose in relation to food intake [16], consumption of non-renewable goods [17], and many more important activities that are hard to monitor without technological assistance.

Research investigating how people use and make sense of PI systems produced a five-stage model of behavior change that applies to a large number of general cases [14]. The model (Preparation, Collection, Integration, Reflection, and Action) describes the types of data users collect, the integration of data collection and reflection into a daily routine, and the transition from reflection to goal setting. The framework provides a list of barriers and design recommendations for each stage. Researchers have recently proposed that incorporating this framework into adaptive training systems may improve classroom interactions [18], but only one project has evaluated such an application. The Live Interest Meter is a PI system that tracks student engagement through a mobile app and provides data visualization to the instructor. It shows the potential to increase audience engagement and instructor responsiveness [19], but at the cost of increased cognitive demand by relying on live manual data input. Our system advances this work by investigating automatic detection of the

presence of classroom features that may indicate enhanced learning, such as peer-to-peer interaction and student participation, both of which have been shown to correlate with students' critical thinking in American universities [20], and both of which would likely be difficult for cultural non-natives to enact in their classrooms [21]. Additional strategies for involving students include the use of student names, asking students to elaborate on ideas, and asking deep questions [22].

AIED work has addressed professional development for teachers by means of student tracking and data visualization [23, 24], but these systems have focused on online learning or blended classrooms, and did not offer instructors guidance on how to enact change in a live classroom. Other systems have attempted to visualize student participation (e.g., [25, 26]), but these have been deployed to support students' own self-reflection rather than to support the instructor, and only in online applications where participation can be tracked through clickstream data.

In our work, we advance the state of the art by focusing on the instructor as the primary agent of change. We focus on student participation in class as an achievable goal that is likely to provide academic benefits to students and cultural fluency for ITAs. The current stage of the work includes classroom observations and iterative phases of design for the adaptive system. Specifically, we wanted to answer the following research questions:

1. Do ITAs from a culture with a high PDI encourage active classrooms?
2. Are ITAs open to adapting their teaching style to an unfamiliar cultural context?
3. Are ITAs open to using PI to set and reflect on goals for their teaching?
4. Can we easily and inexpensively sense and create visualizations of classroom activity in terms of TA and student interactions?

## 3 Method

To answer the research questions, we recruited 5 ITAs, observed them teaching, issued surveys, conducted interviews, and showed them visualizations of their classroom data. We also developed a prototype technical system to detect instructor talk, student talk, and silence.

The TAs were all from India, male, and in their mid-twenties. India has a relatively high PDI (77) compared to the U.S. (40). Each TA had similar levels of teaching experience and content knowledge. None of them had received pedagogical training by the institution or the professor in charge of the course. We observed six to seven sessions of each TA's weekly course, a sophomore level computer science recitation, for a total of 32 sessions. We logged behaviors that would adduce attempts to engage active learning. We inferred activity from frequency and duration of student talk, as opposed to TA talk and silence. We logged the time and locus of all spoken contributions in order to extrapolate episodes of discussion vs. passive lecture.

We surveyed and interviewed the ITAs about their teaching experiences in and perspectives on American classrooms. The survey collected theoretical orientations toward cultural dimensions of learning via items such as demographics, definitions of terms (e.g., "classroom contribution"), and perceived locus of responsibility for learn-

ing (e.g., instructor, student, or a combination). We met with each TA three times during the semester (totaling 2.5 – 4 hours per TA) to discuss their survey responses, their perspectives on and motivations for teaching, and to explore their own teaching behaviors with data visualizations.

The data visualizations were initial sketches of what might exist in a PI system. These were meant as a probe for discussion that allowed TAs to reflect on the behaviors they most wanted to capture and view. This is a common technique in the design of new computing systems when there are no design patterns or social conventions to inform the design space [i.e., 27]. We gathered reactions to the visualizations, and redesigned them after each round of feedback. We also probed TAs on their willingness to try new teaching techniques, such as praising students, using students' names, encouraging elaboration, and asking difficult questions. To analyze the results we transcribed the interviews and iteratively searched for areas of strong agreement and disagreement amongst the participants' comments.

Finally, we developed an initial prototype system for a feasibility study, following a typical user-centered design process. We synthesized a set of system needs from the observations and interviews and proposed a minimal set of detection requirements. We developed a prototype system with two Microsoft Kinects and tested it with 20 students and a 60-minute lecture that included various kinds of classroom talk. We hand-coded the audio data with discrete categories of *instructor talk*, *no talk*, and *student talk*. Periods when students talked simultaneously were coded as *student talk*. We tested these categories against the Kinect's angle detection, confidence calculation, and audio amplitude, i.e., whether or not the device picked up sound and if so, where in the room it originated.

## 4 Findings

Exploring the presence of classroom activity, we observed that ITAs conducted nearly all recitation sections as lectures covering a subset of slides from the most recent primary course lecture. Instructor talk dominated the class, taking up 91.97% of class time (SD=3.6%). Student talk took up only 5.25% of class time on average (SD=2.3%), and the length of their contributions averaged 6.2 seconds (Median=3.4, SD=12.6). The most common prompt for student participation was to ask the class, "Do you have any questions?" The resulting patterns of speech were as follows:

1. TA-talk | silence | TA-talk
2. TA-talk | silence | Student-talk | TA-talk
3. TA-talk | silence | Student-talk | Student-talk

TAs were the first to speak after 85% of their pauses (SD=.088) (pattern 1). 13% of the time (SD=.088) students responded, followed by the TA again (pattern 2). These student contributions were typically brief. 2% of the time (SD .02) a different student contribution followed immediately from a prior student (pattern 3).

Student-student interactions were rare. From an active learning perspective, these interactions are useful as students build on each other’s ideas. These conversations

were typically animated discussions of the course content that took place in the few minutes before class began. TAs usually called a stop to such interactions in order to begin the lecture, and over the course of the semester most students stopped talking as soon as the TA entered the room. This matched an overall pattern of decreasing student talk (and attendance) for most classes over the semester.

ITAs did express that student participation was important to them, but they defined participation as *students asking or answering questions*. They used that information for diagnosis. TA-2: "*If you … don't answer [a question asked by the instructor] there is no way for a teacher to know whether you are understanding what he is teaching or what is going on*." Nevertheless, the TAs made lecturing their priority, and student questions were a distraction from this goal. TA-5: "*Maybe I might want to involve their participation a bit more than what it is, but I also fear by doing so [that I won't] be able to complete the contents*."

To explore ITAs' positions on the cultural dimensions of the American classroom, we asked about their explanation for student silence (pattern 1). They speculated that students already understood the content, only had specific questions about their own work, feared appearing dumb, or that they would rather check with peers. When asked how one might increase participation, there were two types of response: ask students if they have questions (TA-1: "*Probably I should ask more times if they have questions*."), and push student to respond to recall questions (TA-3: "*I'll say … at least take a guess … I'm sure that one of them will say something*.").

Viewing visualizations of their teaching helped to assess the TAs' stance toward adopting new cultural strategies. At times these graphs triggered immediate motivation for change. When TA-1 saw he talked 99% of the time in the preceding class (Fig. 1), he shared that an interactive class was important to him and that he wanted to include the students more. Yet when he later viewed four weeks of data revealing that he never spoke less than 95% of the time (Fig. 2), he became frustrated with the students. "*I would prefer if the class had more [student participation]. I keep asking if there are any questions, but no one speaks so, I cannot help this one*."

We probed TAs about their attitudes toward culturally specific strategies for teacher-student interaction. TAs generally agreed that lengthening the pause after asking students a question might be useful and expressed a familiarity with the idea. They showed interest in the tactic of pausing after a student stops talking, and were surprised that it might be valuable. When asked about asking students to elaborate, they expressed skepticism, sharing that students should only elaborate when the instructor does not understand them. We probed them on asking students deep questions from course content as opposed to simple recall questions. This met with mixed reactions. Most worried that asking hard questions would reduce the time needed to cover the material, and all were reluctant to slow down class. TA-5 described his technique of asking content questions in order to highlight important concepts, but only when the questions could be answered rapidly.

We raised the idea of calling on students by their name and of praising their contributions as approaches to create a supportive environment for student participation. Most TAs agreed that these ideas would help students feel valued and might improve their confidence in the learning process, but none of them were willing to employ

these techniques. They worried that they might call a student by the wrong name and feel embarrassed, or that calling on a student directly might make them feel picked on. TA-2 shared that calling on specific students would point out that the student had not been speaking and that this might generate shame.

After looking at many visualizations of their classroom behaviors, including talk time, distribution of student participation, number of unique speakers per class, proportions of each event type per class, changes in rates across multiple classes, timelines of event types, and more, almost all TAs expressed an interest in eliciting more student talk, but each spoke about wanting explicit goals for different behaviors. How much is the *right* amount of student and TA talk? How long should the TA wait after asking a question? Are enough of the students participating? Most also asked how their individual data compared to the other TAs in the course. They were all open to the idea of using a PI system to empirically answer these kinds of questions.

Finally, as a first technical step towards a PI system, we built a prototype detector for speaker events meant to identify three states of classroom discourse that would indicate interesting patterns of events when viewed in sequence: (i) instructor speaking (in front of class), (ii) student speaking (from seats), and (iii) no one speaking for at least one second. Researchers have previously had success using microphone arrays for speaker localization [e.g. 28], a process that triangulates the angle of a noise source in relation to microphones placed in a line (the array). We chose to use the Microsoft Kinect, an inexpensive commodity device with a robust microphone array, a developers' kit, and a support community for software development.

In our 60-minute test of various kinds of classroom talk, we evaluated the accuracy of a single Kinect on one side of a classroom and the inclusion of a second Kinect at the front of the room facing the students. We used a *Nominal Logistic Fit for Categories* test (JMP V.10.0) with standard output from the device (angle detection and confidence), and were able to discriminate between students and the instructor with high accuracy (Table 1). We expanded the test to also detect silence by including average amplitude for each second of recorded audio as an input variable. This reduced accuracy overall, but much of that loss was amended by the inclusion of a second Kinect.

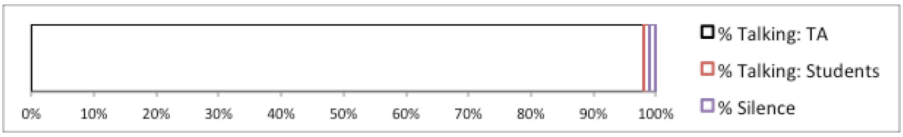



**Fig. 1.** TA-1's first day of recorded data.

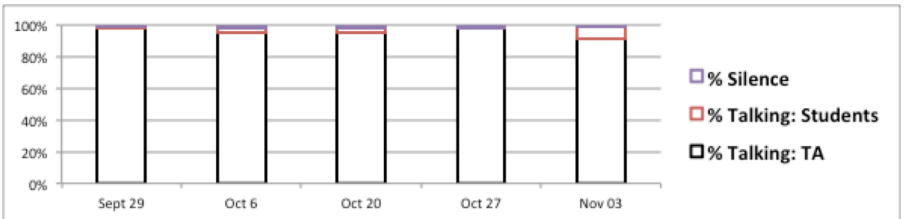



**Fig. 2.** Aggregate talk time for TA-1 across four classes.

**Table 1.** Accuracy of Kinects detecting instructor talk, student talk, and no talk

| | 1 Kinect | 2 Kinects |
|---|---|---|
| Student/Instructor | 94.78% | 95.36% |
| Student/Instructor/Silent | 77.70% | 85.44% |

## 5 Discussion and Conclusion

Our research explored classroom activity in a university STEM course taught by graduate students from a country with a PDI higher than the host country. We used design methods from PI to better understand the perspectives of ITAs who teach in an unfamiliar cultural context. This process led to the development of a prototype system for identifying levels of classroom activity based on speech events that could indicate higher order discourse phenomena. Our findings suggest that ITAs and their students may benefit from an adaptive feedback system built on measuring levels of classroom activity, and that international instructors would be open to using such a system.

ITAs were open to varying degrees of active learning techniques in their own classrooms. Some were easy for them to imagine using (e.g., pausing after students talk), and others were harder to accept (e.g., asking for elaboration). They showed reluctance to decrease the amount of time spent "covering" critical course material, yet they all valued when students got involved in the lecture. These tensions are clues that an adaptive system for cultural training may need do more than measure and report on behavior, but also provide scaffolding for implementing relatively low-cost active learning strategies, such as *think-pair-share*. The next step would be to assess the user's knowledge and stance toward different contextual behaviors and provide individualized instruction and adding more advanced scaffolding prompts as the TA becomes ready for them. Future research would need to navigate this complex space. To refine the detection system further and more easily differentiate between user states, it would be possible to include machine learning and more factors than we currently use, such as Kinect error rates, classroom details, pitch fluctuations and filters, and so on. With more tuning the system might identify individual speakers, leading to reflection opportunities based on individual student speaking patterns. Turn detection at this level could point out disproportionate properties of classroom talk, such as a group of dominant speakers.

There are aspects of the classroom that the proposed system would not be able to detect. ITAs were curious about whether they had lectured for "too long." They made reasonable requests, such as seeing when they had made a "good" explanation, or if students understood the material. A fully operational PI system would necessarily need supplemental human input to provide such feedback, which is already standard practice in current systems: much like annotating the quality of a recent jog when using a fitness-tracking app, our proposed system could request post-class assessments from students or the TA. Some TAs remarked that it would be a simple procedure to personally label the broad topic of the class, or the context of specific pauses throughout the lecture if they were able to review the data and access the audio. Alt-

hough previous PI systems have not explored user input this deeply, such interactions would be possible to implement, and may be critical for system design.

Our study only observed one genre of recitation, but there are many others. It is critically important to assess how much the observed behaviors in this study were an artifact of culture, context, or simply being new to teaching. In our current work we are performing additional observations of a broad selection of classroom contexts taught by students from many different cultural backgrounds in order to assist in making these distinctions.

Research in professional development for teachers might note that our work did not address the quality of interactions, but only quantity and abstract patterns of discourse. As a first step, we argue that any increase in student talk would more closely align with the cultural context of the U.S. classroom, although in the future quality may prove to be a critical area of investigation. Currently, however, the space of cultural acquisition for graduate students and the professional development of novice instructors is under-investigated, and thus this early work makes a contribution.

The implications of this research are important in their potential to address the lack of research in supporting the cultural fluency of ITAs in a challenging new environment. Our work shows preliminary evidence that PI could be an approach to support reflection on classroom dynamics and an opportunity to adaptively expand an instructor's set of pedagogical tools. The impact of the work points to a better experience for international graduate students and potentially better learning for their students.

## 6 Acknowledgements

This work is supported in part by Carnegie Mellon University's Program in Interdisciplinary Education Research (PIER) funded by grant number R305B090023 from the US Department of Education.

## 7 References

1. Faust, J. L., Paulson, D. R.: Active learning in the college classroom. Journal on Excellence in College Teaching, 9(2), 3–24 (1998)
2. Michael, J.: Where's the evidence that active learning works? Advances in Physiology Education, 30(4), 159–67 (2006)
3. Lujan, H. L., DiCarlo, S. E.: Too much teaching, not enough learning: what is the solution? Advances in Physiology Education, 30(1), 17–22 (2006)
4. Freeman, S., Eddy, S.L., McDonough, M., Smith, M.K., Okoroafor, N., Jordt, H., Wenderoth, M.P.: Active learning increases student performance in science, engineering, and mathematics. Proc. Natl. Acad. Sci. U. S. A. 111, 8410–5 (2014)
5. Nisbett, R.: The Geography of Thought: How Asians and Westerners Think Differently... and Why. Simon and Schuster. (2010)
6. Liu, J.: Asian students' classroom communication patterns in US universities: An emic perspective. Greenwood Publishing Group. (2001)
7. Allum, J. R.: Findings from the 2013 CGS International Graduate Admissions Survey Phase III : Final Offers of Admission and Enrollment, 20036 (2013)

8. Hofstede, G., Hofstede, G. J., & Minkov, M.: Cultures and organizations: Software of the mind: 3rd edition. McGraw-Hill Professional. (2010)
9. Wang, M.: Designing online courses that effectively engage learners from diverse cultural backgrounds. British Journal of Educational Technology, 38(2), 294–311 (2007)
10. Lin, J.J., Mamykina, L., Lindtner, S., Delajoux, G., Strub, H.B.: Fish'n'Steps: Encouraging physical activity with an interactive computer game. UbiComp 2006. pp. 261–278. Springer-Verlag (2006)
11. Hsieh, G., Li, I., Dey, A., Forlizzi, J., Hudson, S.E.: Using visualizations to increase compliance in experience sampling. Proc. UbiComp '08. 164 (2008)
12. Li, I., Dey, A.K., Forlizzi, J.: Using context to reveal factors that affect physical activity. ACM Trans. Comput. Interact. 19, 1–21 (2012)
13. Levy, L., Jones, B., Robertson, S., Price, E.D.: Health Mashups: Presenting Statistical Patterns between Wellbeing Data and Context in Natural Language to Promote Behavior Change. ACM Trans. Comput. Interact. 20, (2013)
14. Li, I., Dey, A., Forlizzi, J.: A stage-based model of personal informatics systems. CHI 2010. pp. 557–566. ACM, Atlanta, Georgia, USA (2010)
15. Kay, M., Choe, E.K., Shepherd, J., … Kientz, J.A.: Lullaby: A Capture & Access System for Understanding the Sleep Environment. UbiComp '12. ACM, Pittsburgh, USA (2012)
16. Mamykina, L., Mynatt, E.D., Kaufman, D.R.: Investigating health management practices of individuals with diabetes. Proc. CHI '06. 927 (2006)
17. He, H.A., Greenberg, S., Huang, E.M.: One size does not fit all: Applying the transtheoretical model to energy feedback technology design. CHI 2010. 927–936 (2010)
18. Muller, R., Long, Y., Koedinger, K.R.: Personal informatics for self-regulated learning. In: Personal Informatics in Practice: Improving Quality of Life Through Data — CHI 2012 Workshop. CHI, Austin, TX (2012)
19. Rivera-Pelayo, V., Munk, J., Zacharias, V., Braun, S.: Live interest meter: Learning from quantified feedback in mass lectures. Proc. 3rd Int. Conf. on Learn. Anal. and Knowledge. 23–27. ACM, Leuven, Belgium (2013)
20. Smith, D.G.: College classroom interactions and critical thinking. J. Educ. Psychol. 69, 180–190 (1977)
21. Gerritsen, D., Zimmerman, J., & Ogan, A.: A theoretical approach to the development of critical incidents for cultural training. 5th Int. Workshop on Culturally-Aware Tutoring Systems. 51–60 (2014)
22. Nunn, C. E.: Discussion in the college classroom: Triangulating observational and survey results. J. of High. Educ., 243–266 (1996)
23. Mazza, R., Milani, C.: Exploring Usage Analysis in Learning Systems: Gaining Insights From Visualisations. AIED'05 Work. Usage Anal. Learn. Syst. 65–72 (2005)
24. Scheuer, O., Zinn, C.: How did the e-learning session go? The Student Inspector. Proc. Conf. Artif. Intell. Educ. 487–494 (2007)
25. Janssen, J., Erkens, G., Kanselaar, G., Jaspers, J.: Visualization of participation: Does it contribute to successful computer-supported collaborative learning? Comput. Educ. 49, 1037–1065 (2007)
26. Smith, M.K., Vinson, E.L., Smith, J. a., Lewin, J.D., Stetzer, M.R.: A Campus-Wide Study of STEM Courses: New Perspectives on Teaching Practices and Perceptions. Cell Biol. Educ. 13, 624–635 (2014)
27. Davidoff, S., Lee, M., Dey, A., Zimmerman, J.: Rapidly Exploring Application Design Through Speed Dating. UbiComp 2007. 429–446 (2007)
28. Valin, J., Rouat, J., Dominic, L.: Robust Sound Source Localization Using a Microphone Array on a Mobile Robot. Proc. IEEE/RSJ (IROS); p. 1128–1233. (2003)

# Culture-Oriented Factors in the Implementation of Intelligent Tutoring Systems in Chile

Ignacio Casas[1], Patricia Fernandez[1], Marcia Barrera[1], Amy Ogan[2]

[1]School of Engineering, Universidad Católica de Chile, Santiago, CHILE
[2]Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, USA
{icasas, pifernac, mbarrerb}@uc.cl, aeo@andrew.cmu.edu

**Abstract.** With the aim of assessing the use of intelligent tutoring technology for math teaching in Chilean public schools, an experimental study was performed in the period 2013-2014. Although it was a successful experience in terms of number of participants and learning outcomes, it was not achieved without a number of difficulties which could be explained by focusing on the cultural challenges encountered in the endeavor. In this paper we explore the impact of cultural dimensions such as: organizational strategies and structure; organizational culture; pedagogical processes, human resources, and technology deployment. We characterize each one of these aspects by means of a qualitative study of the implementation process, involving tasks such as planning and technical support, class observations, interviews, and support to teachers in the classroom and lab. As a result, we propose a Diagnostic Chart which could help in the identification of pre-conditions to be solved at an earlier stage of the implementation phase.

**Keywords:** Intelligent tutoring experimentation; teaching strategies; country-specific developments; evaluation of CAL systems

## 1 Introduction

We describe a qualitative study focused on cultural issues encountered in the implementation of intelligent tutoring technology for Chilean public middle schools (5th to 8th grade in a K-12 system)[1]. The experimentation was carried out during two academic years (2013, 2014) and one of its objectives was to understand the challenges faced by teachers, students and authorities when engaged in the change of their teach-

[1] By implementation we refer to the complex endeavor of introducing new strategies and technology into the teaching-learning processes. This includes development and adaptation of software tools, planning, training, demos, on-line and field support.

ing-learning strategies by means of intelligent tutors[2]. The long range vision is to improve math learning in public education for underserved populations.

Based on the literature and the experimentations' findings, we have identified culture-oriented critical factors to be dealt with when implementing an intelligent tutoring system environment in the math class. From this characterization we construct a Diagnostic Chart which could help identifying pre-conditions to be solved at an earlier stage of the implementation process.

The implementation endeavor includes the development of a pedagogical framework that, considering scarce technological resources, takes advantage of personalized student-centered activities in the computer lab and collaborative-constructivist strategies in the classroom. Even though the ultimate goal has been to improve math learning among students, the core methodology has focused on the teachers: they provided training for teachers and implemented teaching support tools. In the training courses, the new technology-based strategies were socialized, situated and adapted to local contexts. We wanted to make sure teachers felt motivated and are willing participants-leaders of the required change process. After training, we provide constant support and follow-up of the implementation in the classroom and lab.

The focus is on the tools and support activities needed by teachers to adequately implement the new technology-enhanced teaching strategies. This involves substantial change in the teacher's attitude, motivations, activities, and plans. The teachers need training, time and support for studying and planning the new classroom-lab strategies. It involves major changes in planning, instructional design and the teaching processes itself; it is a complex task. We have identified that once the basic technology issues are resolved (computer labs with one functioning PC for each student, reliable local area networks, client software correctly installed, sufficient Internet access to the servers, and effective technical support), there are several cultural-organizational drawbacks that work against a successful implementation. Most teachers complain about the extra effort required for the process.

To understand the particularities associated with setting up a class on an intelligent tutoring environment, we first describe the technology and its strategies.

### 1.1 Cognitive Tutor Technology

Following the theoretical principles developed by Anderson [1], [2], a personalized digital learning system known as a Cognitive Tutor (CT) was built at Carnegie Mellon University and is maintained and operated by Carnegie Learning Inc.[3] In this software, each student has a personalized "problem-solving" space, with just-in-time feedback and detailed tracking of his or her progress [3]. CT follows a personalized self-paced approach, allowing students to sequentially tackle progressively more difficult tasks. It tracks students' progress in real time as they answer questions, ask for

[2] We acknowledge the generous support of district-municipality authorities, school principals and teachers together with funding from the Inter-American Development Bank (grant ATN/KK-11117-RS) and CONICYT-Chile (project FONDEF-D10i1286).

[3] Cognitive tutoring technology is a trademark property of Carnegie Learning Inc.

help and solve problems. It provides personalized feedback and hints when errors are made in key points [4].

Cognitive tutors have shown considerable potential, and evidence in the literature indicates that they are effective in improving mathematics and science problem-solving skills [5], [6]. Specific mathematics cognitive tutors have been used in large school systems (primary/secondary level) in the United States, including Los Angeles and Chicago, as well as in rural areas [7].

### 1.2 Cognitive Tutor Strategies

The main objective of the CT software is to provide each student with a unique, enriched environment where he/she can interact with the system by solving specific problems. Multiple graphical representations can be explored by the student for creative thinking practice [8], [9], [10].

The software presents a problem and the student is requested to work towards the solution. Instead of jumping to the final answer, the software provides step-by-step scaffolding [11]. This divide-&-conquer strategy asks specific questions, from easier to more complex, so that the student can advance at his/her own pace in the solution of the problem.

The first question in each problem presented to the student is always related to the appropriate reading of the problem narrative. The next questions (posed by the software) guide the student towards the solution of the problem[4].

The student gets feedback (positive or negative points in a roster of skills to be achieved) whenever he/she answers questions within a problem. This immediate feedback is continuously represented via a "skill-o-meter" in the interface of the tutor [12]. Based on the "skill-o-meter" we have developed a web-based tool that provides teachers with a complete view of students' progress, both at an individual and full class scale. The teacher knows at any time where individual students are standing and thus can give them reinforcement on topics of struggle [13].

## 2 Experimental Study

The broad objective of the study is to understand how the culture-oriented challenges, that may be an obstacle for the implementation of an intelligent tutoring system in schools, can be characterized to detect deal-breaker barriers at an early stage of the implementation. We state that dealing with these obstacles is a condition sine qua non to successfully engage teachers, school authorities and students in an intelligent tutoring environment, hence the importance of achieving this goal.

The key questions are: Which are the culture-oriented challenges that can be identified during the experimentation? Which are the critical factors that can be deduced from the cultural challenges? Are there verifiable achievement indicators that can be

[4] There is extensive literature with thorough description of CT technology ([2], [4], [6], [7], [8]).

linked to those challenges? How can these indicators be arranged into an evaluation instrument to be used as a guideline for teachers and school authorities in the process of setting up an intelligent tutoring implementation?

### 2.1 Methodology

Building from experiences in USA, the Chilean initiative seeks an important innovation: the definition and application of new teaching strategies that, based on the CT technology, are adapted to the local educational context. This starts with the negotiation of change strategies with the district and school authorities. It follows with the involvement of teachers in training and instructional design blended-courses (90% of work is on-line) based on the CT. It culminates with the implementation of the technology-supported strategies in the math classroom.

At an early stage, we decided to work with public Chilean schools (totally or partially dependent of Municipalities) which enroll the largest percentages of vulnerable students and present the lowest learning results. These are the students with most diminished education opportunities explained by the lack of household economic resources. Once the schools were selected and authorities had committed their support, we provided training for teachers to engage them in the new strategies and technologies. Teacher involvement was the most critical issue in the implementation plan. The training goal was to achieve high motivation and strong commitment of the teachers towards the new technology-based strategies. However, a common denominator that plays against this goal is a dramatic lack of time for innovations on the part of the teachers. We also checked the technological infrastructure at the schools, providing support and solutions when needed[5].

In addition to the definition of the pedagogical strategies, we took an English version of the software content and, considering cultural and contextual differences, transformed it into a Spanish version. Even though the underlying theory and structure of the software tool remains the same as in the English version, contents and exercises were localized to the local culture. Finally, we have conducted activities to collect the data needed for constructing the Diagnostic Chart.

### 2.2 The Sample (2013-2014 Implementation)

In general, the selection of the participating districts was a difficult process. It is obvious that without full support and involvement of the district authorities, implementation was impractical. There were some initially invited districts that were necessary to discard due to their lack of real involvement. All schools within a district were invited to participate, but only a few of them decided to experiment with the CT technology.

During the implementation process, a number of treatment schools dropped out for different reasons: problems with infrastructure, lack of involvement in training, reluc-

[5] Even though the technological infrastructure of public schools in Chile is generally adequate, in some cases we needed to provide local servers and networks due to low connectivity.

tance toward teaching changes, and lack of support from school authorities. Due to the training process most participating teachers were enthusiastic and willing to adopt the new strategies and technology. Some teachers (about 20% of initial participants) didn't have enough time to complete the training. The later ones constituted drop-outs from the implementation and in some cases the school as a whole could not participate. Table 1 shows the total number of participants separated by geographic location (Villarrica is mainly a rural area.)

**Table 1.** Total number of participants by geographic location

| | Schools | Teachers | Courses | Students |
|---|---|---|---|---|
| Santiago | 17 | 36 | 76 | 2915 |
| Villarrica | 5 | 7 | 14 | 340 |
| Others | 4 | 6 | 8 | 95 |
| TOTAL | 26 | 49 | 98 | 3350 |

### 2.3 Culture-Oriented Challenges

Culture-oriented challenges continue to be a significant obstacle in the adoption of new technologies for the classroom and lab as means of improving teaching practices [14]. Based on the literature and best practices in industry [15], in our experimentation we have identified a number of these challenges, which rise up as significant barriers to be dealt with in the implementation of intelligent tutors[6]. We have grouped them in 5 categories or dimensions: (1) Pedagogical processes (teaching & learning); (2) Organizational strategies and structure; (3) Organizational culture (teacher's attitudes towards change and technology); (4) Human resources (teachers' skills and knowledge; student attitudes); (5) Technology acquisition and deployment.

A characterization of these dimensions can be obtained by a series of questions to be answered during the study (i.e., observations, interviews, empirical data analysis), as follows.

(1) **Pedagogical processes (teaching & learning):** Are the actual teaching processes adequate for improved learning? Are these processes student-centered or teacher-centered? Is the technology used to innovate (and improve) the teaching process or just to micro-improve a specific task (i.e., projectors for lectures, e-books for reading)?

(2) **Organizational Strategies and Structure:** Are the organization's structures and strategies adequate to motivate, lead and perform effective changes in the teaching processes? Is it feasible to implement changes in the classroom? Do authorities facilitate resources (equipment, time for training, planning, and implementation) to involved teachers?

[6] We focus here on "organizational" culture as opposed to "ethnical" culture. Notwithstanding, there are organizational issues that may be influenced by the local culture, such as dealing with scarce resources, poor planning and assessment, social unrest, vulnerable student communities, etc.

(3) **Organizational Culture:** Are teachers comfortable-satisfied with the actual pedagogical strategies? Are they committed to introduce changes for improvement? Using the CT technology, was it possible to change the classroom-lab processes? Were the resources assigned (by school authorities) adequate? Were there other critical factors? Do teachers perceive that the resources and support for innovation are adequate?

(4) **Human Resources:** Is the teacher's level of proficiency in the domain (math) adequate for teaching? Do teachers master the features present in the CT technology? Are the teachers confident on the contributions of technology for improved learning? Are they confident on the CT technology? What is the student's attitude towards learning, technology and math?

(5) **Technology Acquisition and Deployment:** Are there enough computers in the lab for a "one computer per student" strategy? Are there enough local area networks (e.g., Wi-Fi) to support the use of the new technology? Is there a sound Internet connection and Web services? Does the school have appropriate technical support?

# 3 Results and Discussion

Using assessment instruments such as interviews and surveys, during the experimentation we have identified specific factors for each dimension of culture-oriented challenges. These factors can be evaluated by means of achievement indicators. The set of dimensions, factors and achievement indicators provide a coherent characterization of culture-oriented challenges found in our study. What follows is a brief description of factors and indicators for each dimension.

## 3.1 Factors and Achievements for Culture-Oriented Dimensions

As shown in Table 2, within the "Pedagogical Processes" dimension we have identified two factors: teaching strategies and teaching tools.

**Table 2.** Factors and Indicators for Pedagogical Processes (Dimension 1)

| Factor | Achievement Indicator |
|---|---|
| Teaching strategies | Facilitates a student-centered process v/s teacher-centered. |
| Teaching tools | Use of technology tools<br>Use of other resources in the classroom (hands-on material, etc.) |

The "Organizational Strategies and Structure" dimension addresses school's organizational structure and strategies for teaching-learning innovations. In this matter, school's authorities have the main saying; they should be motivators and promoters of transformations in the classroom. If authorities are open to changes, it is necessary to verify the feasibility of these transformations. Table 3 summarizes factors and achievement indicators for this dimension.

**Table 3.** Factors and Indicators for Organizational Strategies and Structures (Dim. 2)

| Factor | Achievement Indicator |
|---|---|
| Authorities motivated towards changes | Interested in innovative pedagogical activities (with or without technology).<br>Comfortable with current teaching strategies.<br>Encourages teachers towards changes.<br>Values the use of technology for teaching-learning.<br>Positive evaluation of CT as a new learning strategies |
| Feasibility of implementation | Facilitates pedagogical innovations in the school.<br>Facilitates the use of technology in the classroom. |
| Resources for teacher | Provides enough time for planning activities.<br>Provides extra time for training activities.<br>Provides enough time for implementation.<br>Encourages school community involvement in innovation.<br>Provides resources. |

As part of the third dimension, organizational culture of a school, teachers are the most important agents of change and innovation in the classroom. Table 4 shows factors and achievement indicators for this dimension.

**Table 4.** Factors and Indicators for Organizational Culture (Dimension 3)

| Factor | Achievement Indicator |
|---|---|
| Teacher's motivation towards change | Open to innovative pedagogical activities (with or without technology).<br>Performs innovative pedagogical activities (with or without technology).<br>Feels pleased about current teaching strategies.<br>Encourages other teachers towards changes.<br>Values the use of technology and CT for teaching. |
| Feasibility of implementation in the school | There is enough time for re-planning learning activities.<br>There is enough time for attending training sessions.<br>There is enough time to carry out the implementation.<br>The school community is engage and supportive towards innovation.<br>There are resources to carry out the innovation activities. |
| Training in new contents, methods and tools | Interest in training.<br>Suggests training opportunities to his or her colleagues and school authorities.<br>Participates in training sessions (school authorities initiative)<br>Participates in training sessions (personal initiative) |

Within the "Human Resources" dimension, we consider teachers and students as shown in Table 5.

**Table 5.** Factors and Achievement Indicators for Human Resources (Dim. 4)

| Factor | Achievement Indicator |
|---|---|
| Teacher's tech skills | Mastering technology, at a user level: Internet, desktop tools. |
| Teacher's attitude towards technology | Introduction of technology into the annual or semester class planning<br>Positive opinion towards the use of technology for teaching. |
| Teacher's self-perception towards math | Self-confidence on knowledge for domain area.<br>Masters the learning objectives of the grade he/she teaches. |

| Factor | Achievement Indicator |
|---|---|
| Teacher's mastery level of CT software (technology and contents) | Check the lessons in "student" mode.<br>Identifies fundamental strategies present in the CT software<br>Understands CT methodology for problem solving and scaffolding |
| Teacher's confidence with technology based strategies | Self-confidence on his/her technology skills<br>Comfort level regarding technology |
| Student's attitude towards technology | Interested in carrying out activities using technology<br>Positive opinion towards the use of CT in the math classroom<br>High level of comfort in using CT for math learning |
| Student's attitude towards math | Improved perception about math after using the CT technology |

Factors and achievement indicators for the "Technology Acquisition and Deployment" dimension are shown in Table 6.

**Table 6.** Factors and Indicators for Technology Acquisition and Deployment (Dim. 5)

| Factor | Achievement Indicator |
|---|---|
| Computers availability | Feasibility for adapting a one-computer-per-student strategy. |
| Internet connection and local networks | Sufficient Internet access and local area networks for full deployment of one-computer-per-student in a class. |
| Technical support | Permanent technical support staff for the lab.<br>Lab administrator present during lab sessions. |
| Exclusive dedication of technical resources | Technical resources used exclusively for educational purposes (as opposed to administrative). |

### 3.2 Diagnostic Chart

Following the dimensions, factors and indicators presented in the previous section, we have constructed a Diagnostic Chart of culture-oriented factors. With this tool we can pin-point those issues that seriously impact or endanger the feasibility of the implementation. Even though the chart is a result of our experimentation, it could be used in future studies to identify pre-conditions to be solved at an earlier stage of an intelligent tutoring endeavor.

**Table 7.** Diagnostic Chart Application: Critical Factors for Drop-Out Schools

| School | Culture-Oriented Factors that Constrained the Implementation |
|---|---|
| School 1 | **Dim 2:** Authorities (school principal and academic coordinator) were not motivated towards changing the actual teaching methodology.<br>**Dim 4:** Lack of technological skills among teachers. |
| School 2 | **Dim 2:** Authorities (school principal and academic coordinator) were not motivated towards changing the actual teaching methodology.<br>**Dim 5:** No enough computers; lack of a reliable Internet connection; lack of technical support. |
| School 3 | **Dim 2:** Authorities (school principal and academic coordinator) were not motivated towards changing the actual teaching methodology.<br>**Dim 5:** Lack of a reliable Internet connection and local networks. |
| School 4 | **Dim 5:** Lack of a reliable Internet connection and local networks. |
| School 5 | **Dim 3, 4:** Teachers not open to change. Teachers do not value the use of CT technology. |

We have used the diagnostic chart to assess the results of the experimentation with 26 schools in urban and rural areas. Out of 26 participating schools, 5 of them showed culture-oriented issues that endangered the implementation effort (resulting in drop-outs). These drop-outs and related inhibiting factors are shown in Table 7.

It could be inferred from Table 7 that the most frequent culture-oriented inhibitors (in our experimentation) are the ones related to "Technology Acquisition and Deployment" (Dim. 5), "Organizational Strategies and Structures" (Dim. 2) and "Human Resources" (Dim. 4).

## 4 Conclusions

Based on the analysis of culture-oriented factors encountered during our experimentation, we have constructed an instrument that helps identifying schools likely to drop out from an intelligent tutoring endeavor. Although the sample size is relatively small (5 out of 26 schools drop-out), observations in the field clearly highlight those factors which are critical in the implementation.

Cultural factors that had more impact on our experimentation (diminishing though the feasibility of implementation) are, in order of importance:

- Innovation is not facilitated by school authorities; no interest on innovative technologies.
- Lack of adequate Internet connection and local area networks.
- Lack of positive attitude towards changes (authorities and teachers).
- Teacher's claim that there are not enough resources to implement.

It can be noticed that there were no cultural issues related to students. According to our surveys and interviews, all drop-outs were due to problems with infrastructure, reluctance toward teaching changes, and lack of support from school authorities. Despite the sense that change was difficult for the teachers and administration, the fact that 100% of non-drop-out teachers and authorities want to continue using the CT technology in the future is an encouraging result that shows motivation and willingness to change once the value of the new technology is established.

## 5 References


1. Anderson, J.R. (2002). Spanning seven orders of magnitude: A challenge for cognitive modeling. *Cognitive Science 26,* 85-112.
2. Koedinger, K., Anderson, J.R., Hadley, W., & Mark, M.A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education, 8,* 30-43.
3. Ritter, S. (2011). *The research behind the Carnegie Learning math series* [Online]. Available: http://www.carnegielearning.com.
4. Koedinger, K., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review 19,* 239-264.

5. Koedinger, K., & Corbett A. T. (2006). Cognitive tutors: Technology bringing learning science to the classroom. In *The Cambridge Handbook of the Learning Sciences* (pp. 61-78). Cambridge University Press.
6. Ritter, S., Anderson, J.R., Koedinger, K., & Corbett, A. T. (2007). The cognitive tutor: applied research in mathematics education. *Psychonomics Bulletin & Review 14(2)*, 249-255.
7. Arroyo, I., Woolf, B., Royer, J.M., Tai, M., & English, S. (2010). Improving math learning through intelligent tutoring and basic skills training. In Aleven, Kay and Mostow (Eds.), *Intelligent Tutoring Systems Part I* (pp. 423-432), *Lecture Notes In Computer Science 6094*.
8. Feenstra, L., Aleven, V., Rummel, N., & Taatgen, N. (2010). Multiple interactive representations for fraction learning. *Proceedings 10th Int. Conference on Intelligent Tutoring Systems (ITS), (*pp. 221-223).
9. Rau, M., Aleven, V., & Rummel, N. (2009). Intelligent tutoring systems with multiple representations and self-explanation prompts support learning of fractions. *Proceedings 14th Int. Conference on Artificial Intelligence in Education (AIED),* (pp. 441-448).
10. Wegerif, R., MacLaren, B.M., Chamrada, M., Scheuer, O., Mansour, N., Miksatko, J. & Williams, M. (2010). Exploring creative thinking in graphically mediated synchronous dialogues. *Computers & Education 54(3),* 613-621.
11. Luckin, R. (2008). The learner centric ecology of resources; a framework for using technology to scaffold learning. *Computers & Education 50,* 449-462.
12. Mitrovic, A., Ohlsson, S., Barrow, D.K. (2013). The effect of positive feedback in a constraint-based intelligent tutoring system. *Computers & Education 60,* 264-272.
13. Schnaubert, L., Andres, E., Narciss, S., Eichelmann, A., Goguadze, G., & Mellis, E. (2011). Student behavior in error-correction-tasks and its relation to perception of competence. In *Towards Ubiquitous Learning Proceedings of the 6th European Conference on Technology Enhanced Learning* (pp 370-383), Berlin: Springer.
14. Casas I., Imbrogno J., Ochoa S., Ogan A. "Cultural Factors in the Implementation and Use of an Intelligent Tutoring System in Latin America", Proceedings AACE E-Learn 2014 World Conference on E-Learning, New Orleans, USA, Oct. 2014.
15. Laube, D.R., Zammuto, R.F., Editors. "Business Driven Information Technology", Stanford University Press, 2003.

# More Fun in the Philippines? Factors Affecting Transfer of Western Field Methods to One Developing World Context

Juan Miguel L. Andres[1], Ma. Mercedes T. Rodrigo[1], Jessica O. Sugay[1], Michelle P. Banawan[1,2], Yancy Vance M. Paredes[1,2], Josephine S. Dela Cruz[3], Thelma D. Palaoag[3]

[1] Ateneo de Manila University, Quezon City, Philippines
[2] Ateneo de Davao University, Davao City, Philippines
[3] University of the Cordilleras, Baguio City, Philippines
{mandres, mrodrigo, jsugay}@ateneo.edu
{mpbanawan, yvmparedes}@addu.edu.ph
{delacruzpen, tpalaoag}@gmail.com

**Abstract.** This paper presents some of the challenges encountered by a field research team when deploying an educational game for Physics. These included problems with site infrastructure and institutional support, logistical challenges, compliance with ethics requirements, launch delays, and student inattention or misunderstanding of directions. The paper shares these experiences with the wider community to help fellow researchers prepare, should they decide to conduct field studies in the Philippines.

**Keywords:** intelligent tutoring systems · research methods · field study· Physics Playground

## 1 Introduction

In 2012, two experienced human-computer interaction researchers said, "Fieldwork takes you to strange locations to meet new people. Despite the best-laid plans, surprises will happen and some amount of mayhem will ensue [5]." Nowhere is this more true than during attempts to transfer software or field methods from a developed country to the developing world. Because the software or field methods are usually designed in and for developed countries, the assumptions made during the design process and the circumstances surrounding deployment vary, sometimes extremely, from ground conditions in other countries. When describing the deployment of an American intelligent tutoring system in Brazil, Ogan and colleagues [4] found that most students had no computers in their homes, that teachers had little to no technology expertise and were not familiar with ways in which computers could be used for education. On a technical level, schools had a limited number of computers for student use and the ones that were available were often riddled with viruses. Other barriers discussed extensively in [2] include data costs, Internet reliability, the availability

and reliability of electricity, and localization of content in terms of both culture and language.

Since 2006, the Ateneo Laboratory for the Learning Sciences (ALLS) has been conducting field studies in different schools all over the Philippines. In [6], key members of ALLS documented five of the challenges of transferring Western educational software and study methods to the Philippines. As in both [2] and [4], [6] observed that the overall level of technology adoption for education was generally low and that technology infrastructure was generally limited. [6] further added that school support, while essential, was not always easy to obtain. Students were culturally conditioned to be respectful of authority, therefore the presence of observers sometimes had an effect on behavior. Finally, typhoons are common occurrences in the Philippines. In one field experiment, they disrupted data gathering and introduced a possible confound: post-traumatic stress.

The goal of this paper is to present the challenges that confronted another ALLS research team during a more recent study. The goal of the paper is to describe additional considerations that researchers should take into account when planning field studies.

"It's More Fun in the Philippines" is the country's official tourism tagline, which presents how otherwise mundane activities such as commuting (as seen in Fig. 1) are more fun in the country by highlighting places, activities, and artifacts that are uniquely Filipino.


**Fig. 1.** Example poster of the "It's More Fun in the Philippines" tourism campaign.

## 2 Description of the Field Study

Data from 180 students was collected over three weeks from January to February 2015 in three schools (Sites A, B, and C) in different regions of the Philippines. The goals of the study were to assess the persistence and affect of students using an educational game for Physics, and to determine any differences among the different region-

al groups. The subsections that follow describe the methods and materials used to these ends.

### 2.1 Learning Environment

Data was gathered from students using Newton's Playground (now Physics Playground, PP). PP is a computer game for physics that was designed to help secondary school students understand qualitative physics. Qualitative physics is a nonverbal, conceptual understanding of how the physical world operates [7].

PP is a two-dimensional computer-based game that requires the player to guide a green ball to a red balloon. Two example levels are shown in Fig.1. PP has 74 levels that require the player to guide a green ball to a red balloon. The game presents these levels divided into eight different playgrounds. The player achieves this goal by drawing agents (ramps, pendulums, springboards, or levers) or by nudging the ball to the left or right by clicking on it. The moment the objects are drawn, they behave according to the law of gravity and Newton's 3 laws of motion [7].

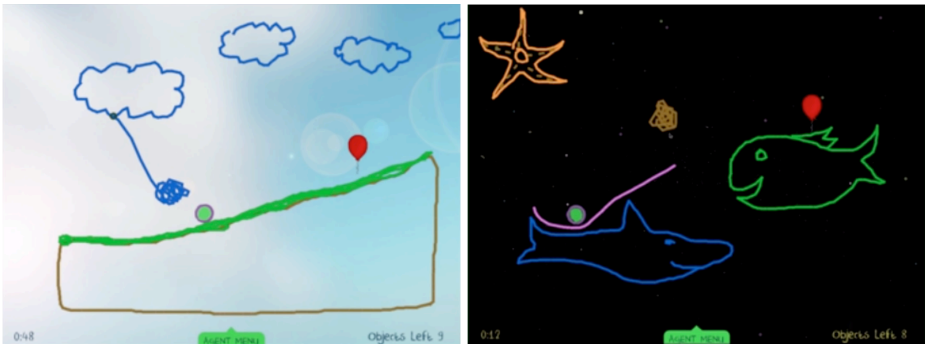


**Fig. 2.** Example PP levels.

A ramp is any line drawn that helps to guide a ball in motion. A ramp is useful when a ball must travel over a hole. A lever rotates around a fixed point, usually called a fulcrum or pivot point. Levers are useful when a player wants to move the ball vertically. A swinging pendulum directs an impulse tangent to its direction of motion. The pendulum is useful when the player wants to exert a horizontal force. A springboard stores elastic potential energy provided by a falling weight. Springboards are useful when the player wants to move the ball vertically. In Fig. 2, the level on the left requires a pendulum, and the level on the right requires a lever.

During gameplay, PP automatically generates log files. Each level a student plays creates a corresponding log file, which tracks every interaction the student has with the game in terms of particular counts and times for selected features of gameplay.

### 2.2 Participants

Data were gathered from 180 students in the Philippines, equally divided among three geographical locations in the country: 60 eighth grade students from Baguio City, 60 tenth grade students from Cebu City, and 60 eighth grade students from Davao City.

### 2.3 The Observation Protocol

The Baker-Rodrigo-Ocumpaugh Monitoring Protocol (BROMP) is a protocol for quantitative field observations of student affect and engagement-related behavior, described in detail in [3]. The affective states observed within Physics Playground in this study were engaged concentration, confusion, frustration, boredom, happiness, delight, and curiosity. The affective categories were drawn from [1].

Participants were divided equally among the two to three BROMP-certified observers present per session. Students were observed in 5 to 8 second intervals through each site's respective observation period, resulting in at least one observation per student per minute. If the student exhibited two or more distinct states during his or her respective observation period, the observers only coded the first state.

The observers recorded their observations using the Human Affect Recording Tool, or HART. HART is an Android application developed to guide researchers in conducting quantitative field observations according to BROMP, and facilitate synchronization of BROMP data with educational software log data.

### 2.4 Data Collection Methods

Before playing PP, students completed a demographics sheet and a 16-item multiple-choice pretest. Students then played the game for a certain period of time (i.e., 90 minutes in Site A, 75 minutes in Site B, and 30 minutes in Site C), during which the trained BROMP observers coded student affect and behavior on the HART application. After completing gameplay, participants completed a 16-item multiple-choice posttest. The pretest and posttest were designed to assess knowledge of physics concepts, and have been used in previous studies involving PP [7].

## 3 Challenges Encountered

Poverty is intrinsic to the Philippine situation, and as such, the adoption of information and communication technologies (ICTs) in the classrooms of the Philippines has been slow and marred by hindrances and limitations. Of the 46,000 public schools run by the country's Department of Education (DepEd), for example, about 8,000 have no power, and even more have no connectivity. There also exists a tremendous need for ICT integration in pre- and in-service teacher training in order to gain appreciation for the use of technology in the curriculum and in the classroom.

As in [6], infrastructure and institutional support remained challenging. This field study also introduced new challenges in terms of logistics, compliance with ethics

requirements, launch delays, and student inattention or misunderstanding of directions.

### 3.1 Infrastructure

In preparation for data gathering, arrangements were made with on-site counterparts to have the software installed and tested prior to the arrival of the research team. PP requires several peripherals in order to launch smoothly. An error thrown by any of these necessary components can cause faulty data capture, which can result in having to throw out gathered data, or cause the game not to run at all. The three main components necessary for PP to run are 1) the software itself, 2) a steady Internet connection not blocked by a firewall or proxy, and 3) a webcam to record the participants' facial expressions.

A previous research project outside of this project's scope already required the team to install and debug the system in the past. Hence, the research team had solutions to problems encountered before. Site A, however, experienced problems with the installation of the software and hardware drivers, which required around three hours of debugging possible conflicts in the computer laboratory's system configurations, including webcam driver incompatibilities and the unstable Internet connection. PP had been running smoothly on one machine, but continued to encounter launch errors on every other machine in the computer laboratory. The team eventually found that the machines were configured to use a virtual environment, which was causing conflicts with the PP software installation and webcam drivers. Once the virtual environment was disabled, PP ran smoothly.

PP's Internet connection posed a technical challenge. The Internet connection was essential for the game's timing functionality to run smoothly. The timing functionality's main purpose is to synchronize all interaction events with Internet time, allowing for a unified set of timestamps for all the participants, as well as for the BROMP coders. Having to synchronize multiple data sources (including human-recorded data) into a single time-stream is a challenge all on its own; having to deal with time inconsistencies in the process makes the task much harder, and the resulting analyses less accurate.

This timing functionality on PP can be turned off optionally (though it is not advised), requiring the research team to take note of session start times manually. Computer labs are usually protected by firewalls and proxies, and as such, the research team had made it a point to request for a firewall exception and for proxies to be disabled a week before data gathering. The research team had to disable the timing functionality of the software in Site B because the administration would not allow addition of a firewall exception for the timeserver. Another solution to this issue could be the use of a local time server.

Another critical issue of PP is that, in order to ensure that the interaction logs and video files are properly saved to secondary storage, the software must exit cleanly. On several occasions, the research team observed that the software did not exit properly. This was consistently experienced in Site B, wherein the software had to be forcefully

terminated before log files could be retrieved from temporary folders. Conversely, the problem was only encountered on two occasions in Site A, and never in Site C.

### 3.2 Institutional Support

Institutional support, in this case, refers to the willingness of the institution to participate in the study and their readiness to make adjustments to accommodate the arrangements required to properly conduct the study. These adjustments include, but are not limited to, scheduling of the experiment and access to the computer laboratories and the students.

The research team received some resistance from the school administration in Site B. Consent forms had been distributed to participants a week prior to data gathering, but had not been collected at the time of the research team's arrival. This caused concerns about research methods and scheduling, which ultimately led to the delay in system configuration and installation. School officials did not allow the local ground team to begin software and hardware installation until two days before the beginning of data gathering. Fortunately, installation and launching in Site B ran smoothly, and data gathering was able to proceed as scheduled.

The study was designed to be conducted over a period of three hours, allotting 30 minutes each for both the pretest and posttest, as well as delays in arrival and about 90 to 120 minutes of interaction with the software. Site B allotted only two hours for each session, including buffers for delay in arrival, introductions, and the administration of the pretest and the posttest. As a result, students were only able to interact with the software for 70-75 minutes per session.

Site C posed the most limitations in the schedule for data gathering. Instead of the prescribed three-hour period, each session was only allotted about 90 minutes, including the delayed arrival of the participants and the administration of the pretest and the posttest. To maximize the allotted time, PP was launched on each system before the participants arrived, which minimized the problems usually encountered when launching the software. As a result, students only interacted with the software for 30-45 minutes.

The final component of the study's design was the administration of a delayed posttest. Local teams in each site were instructed to administer a posttest exactly one week after a participant's interaction with the software. Due to the limited time, restricted by the school's schedule of activities as they were already on their final weeks of the semester, the delayed posttest was not administered to participants in Site C.

### 3.3 Logistics

Two local high schools took part in the study in Site A. Students here needed to travel from their high school campuses to the site where the study was conducted. School A had asked the research team to arrange for transportation of their participating students one week ahead of data gathering: from their high school to the data gathering venue, and vice versa once the session was over. As a result, members of the team were able to commission transportation for the 30 students coming from School A.

Conversely, School B instructed their students to proceed to the venue on their own. Because students had to manage their own transportation and because their commute was not properly managed, more than half of the time allotted (i.e., about an hour and a half) for the data gathering session was spent waiting for the participants to arrive. The delay caused the research team to shorten the interaction time with the software. For the succeeding groups of students from School B, the research team hired a shuttle service to transport the students to the venue in order to ensure timely arrival.

### 3.4 Compliance with Ethics Requirements

In line with university's guidelines on ethical research, the team was required to prepare and collect informed consent forms from each participant and his/her parents. While the study's data collection methods were non-invasive, the requirement applied to this study because interacting with the software required capturing the participant's face on video throughout the session.

Although arrangements were made with the partner schools in advance, only School A in Site A was able to distribute and collect the consent forms prior to the scheduled data gathering sessions. In effect, counterparts in Site A collected the consent forms from School B after the study was conducted, then sent the forms to the research team via courier. Similarly, counterparts in Site B also collected the consent forms one week after the study was conducted, and scanned copies were electronically sent to the research team.

Site C, being the last leg in the data gathering push, presented the most difficulty as their school year was already coming to a close. A week after data gathering had concluded, the research team's main counterpart in Site C said that, with the limited time and schedule constraints, it was going to be impossible to distribute and collect the consent forms. The team reached out instead to another member of the local team in Site C, and only after explaining the gravity of the situation and offering to compensate whoever can get it done was the request obliged. Consent forms were distributed, collected, and mailed back to the research team via courier within a week after contracting help.

### 3.5 Launch Delays

When launching PP, a number of technical problems sometimes occur. Most frequently, if the Internet is unstable when the game is launched, an error message will pop up saying that the game was unable to connect to the timeserver. Launching the game again usually resolves this issue. If the problem persists, however, the team had to resort to disabling the timing functionality of that specific machine.

Another frequent error that occurs has to do with the webcam malfunctioning. Previous experience with the webcam and its connection to PP has shown that when other applications on the machine are using the webcam, it was likely to malfunction when PP was launched. As a result, the research team usually quit all webcam-related software before launching PP. Webcam-related errors popped up on several occasions

in Site A and Site B. Quitting and launching the game again usually resolves the problem as well.

Also, in order to better manage webcam software, the research team had its own set of webcams, which they install onsite immediately before data gathering. In Site C, however, because the school's officials wanted all students in each of the three participating classes to take part in the study, the research team had to use the built-in webcams of the site's machines. These built-in webcams had built-in webcam software that would pop up every time PP was launched. Because data gathering in Site C was already very limited time-wise, the research team resolved to launching the game before students arrived in order to address all launch delays before the session began.

### 3.6. Inattention to Directions

Not listening, reading, understanding, and paying attention to instructions also contributed to delays in gameplay. Because the timeserver synchronizes all student interactions in its logs with Internet time, it is important that all participants in each session begin at the same time. Once PP is launched, participants are asked to input a username (which is provided to them upon arrival), and to press OK. Participants will then be presented another screen to read, shown in Fig. 3, telling participants to wait for the moderator's go signal before pressing OK again. Clicking this OK button launches the game and begins the logging sequence.

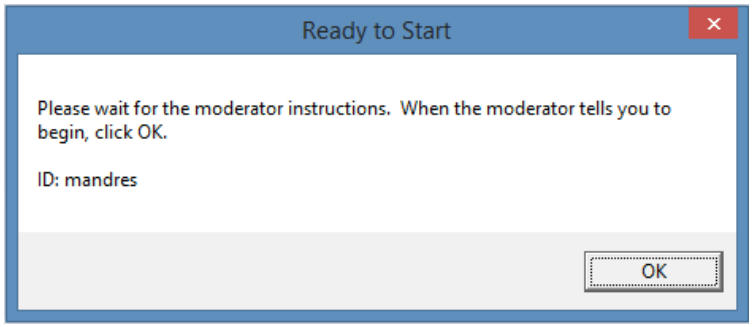



**Fig. 3.** Instruction screen telling participants to wait.

Participants are given the instructions to wait both verbally through the moderating member of the research team, and in writing through the pop-up screen in Fig. 3. Despite these, however, members of the research team have had to quit a game that was launched prematurely about two times every session. Once everyone is back on this screen and waiting for the go signal, participants are instructed to press OK, after which they are presented with a tutorial on how to play the game.

This tutorial ends with a string of text, instructing the students to "hit ESCAPE and select 'Quit'," as shown in Fig. 4. The research team noticed that almost half the participant population in each session gets stuck on this screen, possibly waiting for an "ESCAPE" button to come up on screen, as opposed to tapping the Escape button on the keyboard, which in turn brings the menu up, and allows the participants to

click “Quit”, which then brings them back to the game’s main screen where they can choose what level they want to play.

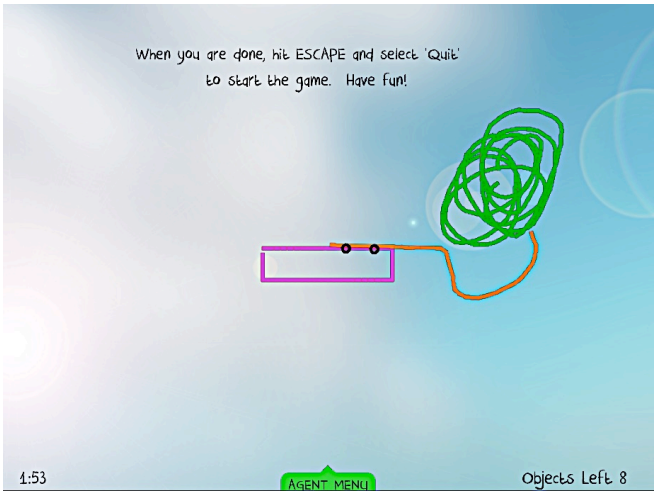



**Fig. 4.** Hit ESCAPE instruction.

## 4 Discussion and Conclusions

In an extensive literature review, [2] regards the Philippines as a significant producer of intelligent tutoring systems research outside of high-income nations. This finding implies an openness to new technology as well as commitment of Filipino researchers to collaborate with their counterparts abroad and to shepherd the deployment and study of technology use to improve educational institutions. However, many factors on the ground prevent adoption of these technologies. This paper describes some of the challenges that a Philippine team had to overcome to gather data from three local sites.

Infrastructure and institutional support were major roadblocks in the research method’s smooth implementation. The learning environment used had three main components: the software itself, a stable Internet connection not blocked by a firewall or proxy, and a webcam. Any error produced by any of these three components results in faulty log capture, which eventually leads to data being thrown out. Having to ensure that each component runs without error in three separate data gathering sites in a country where education is only beginning to embrace the use of ICTs was the study’s biggest hurdle to overcome. On top of this, resistance from and miscommunication with school administrators had caused the delay of both hardware/software setup and compliance with ethics requirements. The other challenges encountered during the study’s execution were transportation arrangement, launch delays, and the students’ inattention to directions.

All these challenges taken into consideration, there were some lessons learned in the process. In terms of dealing with institutional support and ethics compliance, start-

ing the process early of arranging for data gathering schedules and the efficient distribution and collection of ethical consent forms. Avoiding the conduct of studies towards the end of the school year will give both the researchers and the partner institutions more time to fix issues that may have arisen during research execution. In terms of research execution itself, controlling transportation to and from the data gathering sites will ensure the participants' timely arrival, which is important especially when you are given only a certain number of hours for the session.

For educational technology adoption to widen, researchers must continue to plan for and address these challenges, and to share these experiences with the wider community to inform like-minded researchers about what to expect when conducting fieldwork in the Philippines.

**Acknowledgements.** We would like to thank the Bro. Robbie Paraan, S.J., Cecilie Villacrucis, and the officials at University of the Cordilleras, Bakakeng National High School, Ateneo de Davao University, Sacred Heart School – Ateneo de Cebu, Drs. Valerie Shute, Matthew Ventura, and Matthew Small of Florida State University for collaborating with us. This study was made possible through a grant from the Philippines' Department of Science and Technology Philippine Council for Industry, Energy and Emerging Technology Research and Development entitled "Stealth assessment of student conscientiousness, cognitive-affective states, and learning using Newton's Playground."

## 5 References

1. D'Mello, S. K., Craig, S. D., Witherspoon, A., McDaniel, B., & Graesser, A. (2005). Integrating affect sensors in an intelligent tutoring system. In Proceedings of the Workshop on Affective Interactions: The computer in the affective loop workshop, International conference on intelligent user interfaces (pp. 7- 13). New York: Association for Computing Machinery.
2. Nye, B. D. (2014). Intelligent tutoring systems by and for the developing world: a review of trends and approaches for educational technology in a global context. *International Journal of Artificial Intelligence in Education*, 1-27.
3. Ocumpaugh, J., Baker, R.S.J.d., Rodrigo, M.M.T. (2012) Baker-Rodrigo Observation Method Protocol (BROMP) 1.0. Training Manual version 1.0. Technical Report. New York, NY: EdLab. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.
4. Ogan, A., Walker, E., Baker, R.S.J.d., de Carvalho, A., Laurentino, T., Rebolledo-Mendez, G., Castro, M.J. (2012) Collaboration in Cognitive Tutor Use in Latin America: Field Study and Design Recommendations. *Proceedings of ACM SIGCHI: Computer-Human Interaction*, 1381-1390.
5. Portigal, S., & Norvaisas, J. (2012). Never eat anything raw: fieldwork lessons from the pros. *interactions*, *19*(4), 10-12.
6. Rodrigo, M. M. T., Sugay, J. O., Agapito, J., & Reyes, S. (2014). Challenges to Transferring Western Field Research Materials and Methods to a Developing World Context. *Research & Practice in Technology Enhanced Learning*, *9*(1).
7. Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and Learning of Qualitative Physics in Newton's Playground. The Journal of Educational Research, 106(6), 423-430.

# Investigating the Impact of Designing and Implementing Culturally Aligned Technological Systems on Educators' Ideologies

Samantha Finkelstein

Human Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, USA
{slfink}@cs.cmu.edu

**Abstract.** Culturally sensitive educational technologies may be able to help improve under-represented students' learning and engagement when they are deployed in the classroom. However, there may be challenges integrating these systems into the classroom when the cultural components they incorporate are heavily stigmatized in contemporary society. In this on-going work, we are using an action research approach to investigate how involving teachers in the design of these technologies may not only affect the effectiveness of these interventions on students, but also teachers' own ideologies surrounding the targeted stigmatized cultural components.

**Keywords:** dialect, classrooms, teachers, culture, AAVE, action research

## 1 Introduction

The pervasive achievement gap between Euro-American and African American students is perpetuated by challenging and inter-related factors, including access to resources, socio-economic status, and racism (and vestiges of racism) in contemporary society [1]. One common manifestation of these vestiges of racism is a *deficit perspective* within the classroom, where the school system views certain aspects of a student's cultural background as a *challenge to overcome* rather than an *asset to leverage* [2]. For example, many African American students come into school as speakers of a non-standard dialect of English called African American Vernacular English (AAVE), which is rarely represented, or even accepted, within the classroom. Despite that AAVE has great cultural importance for its speakers and linguists regard AAVE as valid and grammatically consistent, it is common practice for educators to criticize or even shame students for their use of this dialect [3], such as by saying that they are speaking incorrectly, or even that they sound like they belong on the streets. However, some evidence suggests that when non-Standard English speakers are allowed to use their primary dialect within the classroom or when this dialect is represented in learning materials, students may improve on their task performance, academic engagement, self-efficacy, and even their use of Standard English [4, 5, 6]. While this evidence is promising, standard teacher training programs rarely incorporate enough background in language variation to prepare teachers for methods of incorporating students' dialect diversity into the classroom. For this reason, some researchers have proposed that culturally adaptive educational technologies may be a productive way

for students to gain access to learning materials that may best support their learning [7, 8, 9].

Despite the potential promise of these systems, a notable challenge in the design of culturally adaptive classroom technologies is ensuring that they work with, and not against, the teacher. There is substantial evidence that teachers may be hesitant to incorporate classroom interventions that expose their students to stigmatized cultural behaviors such as non-standard dialect use. This is often due to lack of appropriate teacher training about cultural variation, misconceptions about the role of non-standard dialect use in their students' lives, and concern that they might accidentally cause offense and put their job at risk. As interventions are less likely to be successful if teachers do not believe that the systems are helping them meet their own goals [10], this may make even the most well-designed educational technologies unusable in real classroom settings. In this work, we are investigating how an *action research* (AR) approach may be used to both design technologies that best meet teachers' needs, while also helping them develop more progressive and positive ideologies about cultural variation. By action research, we refer to the cyclical process of researchers working alongside community partners (in this case, educators) to create knowledge by *learning through action* – taking steps, reflecting on the outcomes, and iterating together [11]. In AR, the researcher works alongside the community partners to open up productive lines of communication and facilitate activities expected to create change, rather than as a distanced observer of subjects. This method will allow us to work alongside educators to quickly iterate on different ways of incorporating a technology that can use AAVE into the classroom. This will help us understand what social and scientific impacts these interventions may have on the classroom culture, as well as investigate how this collaborative design process itself impacts teachers' ideologies about their students.

## 2 Previous Work on Culturally Aligned Technologies

Over the past two decades, there have been a small but notable number of educational technologies that have considered how to align to students' underrepresented cultural backgrounds. These projects demonstrate some of the potential scope for the impact culturally-aligned technologies may be able to have on students. For example, Pinkard's work on literacy learning for young African American students resulted in two systems, Rappin' Reader and Say Say Oh Playmate, which leveraged students' culturally-based knowledge of rhythm patterns and clap sequences to acquire early literacy components through writing rap lyrics [7]. Rap lyrics were also applied in Gilbert's African American Distributed Multiple Learning Styles System (AADMLSS) program, which is an intelligent tutoring system that additionally uses gaming components to allow students to practice math word problems where explanations are provided via rap lyrics that use AAVE features [8]. Other educational technologies have began exploring the potential impact of dialect congruence on students' performance in other non-standard dialects, such as Mohammad's Trinbago Adventures for Caribbean students, where students are allowed to customize the

amount of dialect features they hear (and other cultural references) within the system [9]. Each of these systems has demonstrated success with the underrepresented population they had targeted, including both academic performance and student engagement. However, the teachers' response to these systems, and the potential impact that the deployment of these systems in the classroom had on the teachers over time, was either not performed or not reported.

There have also been a small number of investigations that examine the impact of simply manipulating only the dialect used in a system. For example, in our own previous work, we have found that when AAVE-speaking 3rd grade students were exposed to a system that provided them with identical science examples in either Standard English or AAVE, students demonstrated an average of two standard deviations improvement on the quality of their own science reasoning when they heard the example in AAVE [12]. However, in follow-up interviews with teachers, we found that they would be very uncomfortable with deploying such a system to their students in the future, regardless of the potential learning benefits. The impact of a German non-standard dialect was also investigated with German adults using a virtual agent who either spoke in Standard or Non-Standard German, finding that participants aligned their own dialect to match that of the agent, but that the Non-Standard agent was viewed as more likable [13]. In our current work, we are performing a similar analysis, and investigating how 3rd grade AAVE-speaking students' language use, self-efficacy, language ideologies, and science achievement is impacted by a virtual agent who either exclusively speaks Standard English or code-switches between Standard English and AAVE based on context over the course of six weeks. Previous work with this virtual agent, Alex, found that even during one session with the character, students switched between dialect features based on context along with the agent – even though they did not perform this type of code-switching with their teachers [14].

## 3 Educational Interventions to Impact Teacher Ideologies

Our previous research (in preparation) has found that teachers would be very hesitant to expose their students to AAVE via an educational technology, regardless of the potential learning benefits to students. This is consistent with what other researchers have found about integrating non-technical curricula into the classroom. However, research suggests that if teachers feel that an educational technology is working to support their overall goals, it is possible that teachers may experience a *pedagogical evolution* [10], whereby the technologies in their classrooms may support and structure class activities that the educator previously did not think possible. The challenge, then, is identifying methods for integrating these technological systems into a classroom in a way that is able to work *with,* rather than *against,* educators.

To address this problem, some designers of non-virtual curricula have found it effective to host professional development workshops with teachers to help teach them about linguistic variation [4, 15]. When paired with this knowledge, teachers become able to not just host the intervention within their classroom (such as is often the case with technologies), but also become active facilitators of the learning activities with

their students. In fact, there is additionally evidence that when teachers have the opportunity to teach a pre-packaged learning activity involving linguistic variation to their students themselves, they develop a stronger positive change in their own ideologies compared to teachers who only attend professional development workshops [4]. These findings support the potential positive impact of action research on influencing teachers' ideologies, as action research involves many of these components, such as professional development discussions facilitated by researchers, reflection with other peer educators, and implementation of curricula within the classroom.

## 4 Investigating the impact of culturally aligned systems

The goal of this work is to employ AR approaches with urban elementary school teachers to promote a positive change in the often-negative classroom culture surrounding students from linguistically-diverse backgrounds. To do this, our approach will involve a combination of professional development workshops surrounding language variation, group reflection discussions about what learning goals they feel are important for their students to know regarding language variation, and hands-on activities to develop classroom activities to meet some of those identified learning goals. The classroom activities will involve the use of Alex, a virtual peer character capable of communicating to students about different science activities and some other social topics (e.g., video games) in either Standard English or AAVE (described above). Because one of the noted reasons that many teachers avoid talking about AAVE with students is many do not identify as speakers of this dialect, a system that is able to demonstrate dialect differences as a peer to the students may be a productive platform for helping to introduce this discussion. We additionally argue that providing educators with an existing technology that can be deployed differently in the context of different classroom activities may allow us to more efficiently iterate new ideas into the classroom.

In this planned work, we will work with approximately ten educators between two and four times a month for a full semester to facilitate and participate in these discussions and lesson plan design sessions. We will aim for teachers to deploy a new classroom activity surrounding the virtual character in the classroom approximately twice a month throughout the semester. We expect a large variation in the sorts of activities teachers design, for example, ranging from using the technology as part of a guided class discussion and worksheet, to a hands-on group activity where students are asked to make the character speak differently in different situations. The researchers and each of the teachers will observe how the students interact with the class activity, and bring their observations to the group discussion the following week. This discussion will spark teachers' iterations on their next class activity.

We will perform pre- and post-intervention measures including meta-linguistic awareness, language ideology, and dialect use for both teachers and students. These quantitative measures will be paired with qualitative measures of how different activities promoted different sorts of student interactions and responses and the types of interactions students and teachers shared throughout the lesson. We are currently

performing a pilot analysis of this process with three elementary school teachers at a local, urban 100% African American charter school to help prepare us for the upcoming semester-long study. Through this pilot and the full-length study, we aim to gain a better understanding of how culturally-aligned educational technologies, and the collaborative process of designing them with teachers, may impact the classroom culture in ways that support positive social change.

**Acknowledgments.** Many smiles to the ArticuLab, HCII, and the Graduate Training Grant # R305B090023 from the US Department of Education (IES).

## 5 References

1. Ogbu, J. U. (2003). Black American students in an affluent suburb: A study of academic disengagement. Routledge.
2. Atkinson, J. L. Are We Creating the Achievement Gap? Examining How Deficit Mentalities Influence Indigenous Science Curriculum Choices. Cultural Studies and Environmentalism, 439-446. (2010)
3. Lippi-Green, R. (1997). *English with an accent: Language, ideology, and discrimination in the United States*. Psychology Press.
4. Sweetland, J. (2006). Teaching writing in the African American classroom: A sociolinguistic approach. Stanford University.
5. Webb, L., & Webb, P. (2008). Introducing discussion into multilingual mathematics classrooms: An issue of code switching?. Pythagoras: Teaching and learning mathematics in multilingual classrooms: Special Issue 67, 26-32.
6. Wheeler, R. (2006). "What do we do about student grammar—all those missing -ed's and -s's?" Using comparison and contrast to teach Standard English in dialectally diverse classrooms. English Teaching: Practice and Critique, 5(1), 16–33.
7. Pinkard, N. (2001). Rappin'Reader and Say Say Oh Playmate: Using children's childhood songs as literacy scaffolds in computer-based learning environments. *Journal of Educational Computing Research*, *25*(1), 17-34.
8. Gilbert, J. E., Arbuthnot, K., Hood, S., Grant, M. M., West, M. L., McMillian, Y., ... & Eugene, W. (2008). Teaching Algebra Using Culturally Relevant Virtual Instructors. *IJVR*, *7*(1), 21-30.
9. Mohammed, P., & Mohan, P. (2011). Integrating culture into digital learning environments: studies using cultural educational games. *The Caribbean Teaching Scholar*, *1*(1).
10. Pajares, M. (1992) Teachers' Beliefs and Educational Research: cleaning up a messy construct, Review of Educational Research, 62, pp. 307-332.
11. Hayes, G. R. (2011). The relationship of action research to human-computer interaction. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *18*(3), 15.
12. Finkelstein, S., Yarzebinski, E., Vaughn, C., Ogan, A., & Cassell, J. (2013, January). The effects of culturally congruent educational technologies on student achievement. In *Artificial Intelligence in Education* (pp. 493-502). Springer Berlin Heidelberg.
13. Kühne, V., Rosenthal-von der Pütten, A. M., & Krämer, N. C. (2013, January). Using Linguistic Alignment to Enhance Learning Experience with Pedagogical Agents: The Special Case of Dialect. In *Intelligent Virtual Agents* (pp. 149-158). Springer Berlin Heidelberg.
14. Rader, E., Echelbarger, M., & Cassell, J. (2011, May). Brick by brick: iterating interventions to bridge the achievement gap with virtual peers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2971-2974). ACM.

15. Reaser, J., & Adger, C. (2007). Developing language awareness materials for nonlinguists: Lessons learned from the Do you speak American? curriculum development project. Language and Linguistics Compass, 1(3), 155-167.
16. Yocum, K. (1996) Teacher-centered Staff Development for Integrating Technology into Classrooms, Technology Horizons in Education, 24(4), pp. 88-91

# Intelligent Support in Exploratory and Open-ended Learning Environments

# Learning Analytics for Project Based and Experiential Learning Scenarios

*held in conjunction with*

Seventeenth International Conference on Artificial Intelligence in Education (AIED 2015)

Monday, June 22, 2015
Madrid, Spain

**Workshop Co-Chairs:**

Manolis Mavrikis[1], Gautam Biswas[2], Sergio Gutierrez-Santos[3], Toby Dragon[4], Rose Luckin[1], Daniel Spikol[5], James Segedy[2]

[1]*London Knowledge Lab, University College, London*
[2]*Electrical Engineering and Computer Science, Vanderbilt University*
[3]*London Knowledge Lab, School of Computer Science, Birkbeck*
[4]*Department of Computer Science, Ithaca College, US*
[5]*Faculty of Technology and Society, Malmo University*

http://link.lkl.ac.uk/iseole15
https://aied15learninganalytics.wordpress.com

# Table of Contents

# Preface

By encouraging interaction, exploration and experimentation in environments that directly represent the domain to the learner, Exploratory Learning Environments (ELE) adhere to constructivist theories of learning that emphasize learners' control to construct their own understanding. More generally, Open-ended Learning Environments (OLEs) offer students opportunities to take part in authentic and complex problem-solving and inquiry learning activities. These environments provide learning context and a set of tools to support learners while they engage in many activities, including (i) seeking and acquiring knowledge and information, (ii) applying that information to a problem-solving context, (iii) assessing the quality of the constructed solution, (iv) evaluating and reflecting on the overall approach, and (v) assessing and enacting cognitive and metacognitive processes.

However, there are several factors that prevent appropriate learning within ELEs or OLEs. The structure of the activity sequences and the level of support by teachers, peers, technologies are crucial determinants of learning. This is particularly true in domains where knowledge is not a directly observable outcome of a situation under exploration (e.g. simulators) but is externalized by cognitive tools in the environment. There is a wealth of learning sciences literature about support for learning in exploratory environments, but developing the technology to support these still faces several impressive challenges that the community is only beginning to address.

At the same time the migration of technology from the desktop to the wider learning environment provides the opportunity to collect data about learners' interactions with a greater bandwidth of learning resources. Smart phones, tablets and technologies embedded in the fabric of the environment are now commonplace in educational settings. In parallel with these developments, there has been great progress in developing techniques to analyse learning interactions through the large amount of data that is generated by these various systems. This kind of learning analytics offers the potential for novel feedback and scaffolding to support project-based and experiential learning that involves physical computing projects and other hands-on type projects.

The papers submitted to this workshop address various aspects of the above-listed issues, which are all at the heart of the AIED community's interest.

Summarizing the papers in brief, Chase et al. and Mazziotti et al. focus mostly on the *design and evaluation of exploratory learning environments*. Chase et al. in particular describe the design of an ELE to support invention activities, inspired by a model of naturalistic teacher guidance. Mazziotti et al. present a pedagogical intervention model that selects and sequences exploratory learning activities and structured practice activities. Four papers focus more on the *tools, algorithms and approaches* behind the implementation of intelligent support in ELEs. Karkalas et al. evaluate requirements and present a prototype for learning analytics for constructionist mathematical e-books. Segedy and Biswas use coherence analysis to provide measures of the quality

of students' problem-solving processes. Silva et al. propose an automatic rating system to assess students and to sequence activities. Harpstead et al. demonstrate a method of accelerating model development for both knowledge and skills by applying a concept formation algorithm.

Lastly, two papers focus specifically on *Learning analytics for project based and experiential learning scenarios.* Luckin et al. present an analysis framework for project-based learning situations that involve the use of technology. Spikol et al. present the design of a visual-based programming language for physical computing and mobile tools to invite learners to actively document and reflect on their projects in a way that creates possibilities of intelligent support and learning analytics.

This workshop builds on the previous work from several editions of the Intelligent Support in Exploratory Environments workshop, and the Scaffolding in Open-Ended Learning Environments in AIED 2013. The format of the workshop is based on a question-oriented organisation around open problems raised by the papers accepted for the workshop. It also includes a posters and hands-on interactive session for participants to present prototypes and get or provide feedback. Our website (http://link.lkl.ac.uk/iseole15) provides more information as well as the current and previous proceedings.

Manolis Mavrikis, Gautam Biswas, Sergio Gutierrez-Santos, Toby Dragon, Rose Luckin, Daniel Spikol, James Seged

Workshop Co-Chairs

# The design of an exploratory learning environment to support Invention

Catherine C. Chase, Jenna Marks, Deena Bernett, and Vincent Aleven

[1]Teachers College, Columbia University, New York, United States
(chase, jnm2146, dlb2175) @tc.columbia.edu
[2]Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, United States
aleven@cs.cmu.edu

**Abstract**. We describe the design of the Invention Coach, an intelligent, exploratory learning environment (ELE) to support Invention, an exploratory learning activity. Our design is based on a two-pronged approach. Our own study of naturalistic teacher guidance for paper-based Invention uncovered phases in the Invention process. Prior research on the mechanisms of learning with Invention activities revealed specific instructional strategies. These two sources informed the design of the guidance offered by the Invention Coach. To our knowledge, this is the first design of a guided environment for Invention activities inspired by a model of naturalistic teacher guidance. Our work offers insight into styles of guidance that could apply to other exploratory learning environments.

**Keywords**: intelligent learning environment, human tutoring, exploratory learning, intelligent tutors

## 1 Introduction

While exploratory tasks support the constructivist nature of learning and have the potential to enhance 21st century skills, there is broad agreement that learners need guidance in their exploration [1]. But what kind of guidance will help learners to engage in productive exploration without eliminating the exploratory nature of the task? Designers of exploratory learning environments have investigated this question through various lenses – types of learner feedback [2, 3], "cognitive tools" for inquiry [4], and participation structures [5]. We explore the question of effective guidance for exploration in the context of an exploratory learning task called Invention, where learners invent their own formulas to describe scientific phenomena. We are now in the process of developing an intelligent, exploratory learning environment (ELE) called the Invention Coach, which scaffolds students through the Invention process.

Invention is an exploratory task that invites students to engage with deep, conceptual ideas by analyzing a set of data [6]. Students are asked to invent an expression of an underlying structure that runs throughout a set of contrasting cases. Cases are examples of phenomena with predesigned contrasts that highlight key features, provid-

ing students with clues to the abstract, underlying concepts. After exploring the cases and inventing their own structures, students are told the canonical structures, through traditional expositions (lecture, reading). Prior work suggests that Invention creates "a time for telling," preparing students to appreciate the "mathematical work" of equations [6] or "function of tools for solving relevant problems" [7].

Figure 1 shows an Invention task our computerized Invention Coach is designed to support. In this "Crowded Clowns" task, students are asked to invent a numerical "index" to describe how crowded the clowns are in each set of buses. Though students do not realize it, they are inventing the equation for density (d=m/v, where density is the number of objects crowded into a space). Most students initially attempt to describe crowdedness using a single feature – the number of clowns. They do not realize that crowdedness must consider two features related in a *ratio* structure (e.g. #clowns ÷ #boxes). The six buses in Figure 1 are contrasting cases designed to highlight the critical features of "crowdedness." For example, by contrasting cases A1 and B1 (see Figure 1), which both have 3 clowns but different-sized buses, students may notice that clowns alone cannot account for crowdedness, and space must be considered as well. Through an iterative process of generating and evaluating their inventions, students begin to realize that a workable solution must involve both features in some kind of relational structure. While many students do not produce the correct formula, the invention process prepares them to learn from a later lecture on ratio structures, which is the targeted content of our instruction.

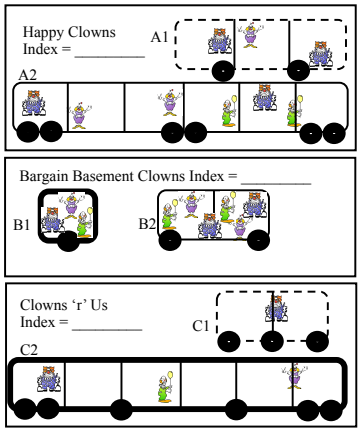



**Fig 1**. Invention task, adapted from Schwartz et al., 2011.

Invention activities are very successful in supporting transfer. In several studies, Invention has been more effective than traditional instruction at enhancing transfer and deep learning in science and math domains, both with adolescents and adults [6, 8, 9, 10]. But in most studies, students need subtle guidance from a teacher to engage in productive invention. In a move towards scaling up, we are developing a computer-based Invention Coach that will ultimately provide adaptive guidance as students

engage in Invention. Through the design of the Invention Coach, we also explore what types of guidance are most effective in scaffolding an exploratory task.The most applicable related work comes from Roll, Aleven, and Koedinger [11], who developed an ELE for Invention activities in statistics. The learning environment we propose will share some characteristics with their Invention Lab but will differ in a fundamental way. While Roll et al.'s technology was developed through rational analysis of the task and empirical study of components of the Invention process, our Invention Coach is modeled on guidance from a human teacher.

To develop the Invention Coach, we are following a multi-phase approach of formal empirical research interspersed with design cycles and informal user testing. We began with a study of naturalistic human teachers' guidance of Invention and a review of the literature on learning with Invention. In the following section, we briefly review the results of both. We then describe the design of our current Invention Coach, focusing on the pedagogical elements of our design rather than the technical aspects underlying it. We are now in the process of implementing a Wizard-of-Oz version of the Coach, though we plan to build a fully adaptive system in the future.

## 2 A Two-pronged Approach to Design

The design of the Invention Coach was driven by a combination of our own empirical work and prior research and theory on Invention. Our study of naturalistic teacher guidance demonstrated the process of Invention by explicating the various subgoals teacher-student pairs tackle as they work towards a solution. The specific instructional strategies embedded in our Coach were drawn from research and theories on the mechanisms that make Invention a successful instructional paradigm.

Our analysis of naturalistic teacher guidance uncovered a process model of guided Invention with four phases [12]. In the "understand the problem" phase, teachers explained the task goal and constraints to students who were confused by the ill-defined goal of inventing an "index." In the "notice features" phase, teachers guided students to notice key features they often overlooked (most often bus size) or to think conceptually about what "crowdedness" means. In the "produce and reflect on an Invention" phase, students generated their numerical index and teachers helped them evaluate whether it was correct. There was also a "math calculation" phase, in which teachers and students worked to simplify and manipulate fractions or count key features. Informally, we noted that phases were not completed in a linear fashion; teacher-student pairs moved back-and-forth between them. As a result, our initial prototype Invention Coach supports each phase, without prescribing a specific phase order.

While the study of naturalistic tutor guidance revealed the subgoals of solving an Invention problem, specific instructional strategies were derived largely from the existing literature on Invention. Instructional strategies were designed to scaffold three core components of the Invention paradigm: noticing deep features of a domain, monitoring errors, and withholding direct feedback. First, noticing deep features of a domain is a critical step for problem-solving success. For instance, novices often focus on the surface features of a problem while experts focus on the deep principles that underlie a problem solution [13]. An effective way to help novice learners notice

key features is to have them compare and contrast example cases that explicate the features [7]. Our carefully designed contrasting cases systematically differ on key features, so that certain pair-wise comparisons reveal the necessity of considering a not-so-obvious feature. Second, Invention helps learners to identify gaps in their understanding, which they can then seek to fill in later expository instruction [14]. Through the process of monitoring and reflecting on their solution attempts, learners often come to see that their invention is inadequate. When they later receive a lecture on the canonical problem solution, they are prepared to understand how it avoids the errors they made in their own solution attempts. We scaffold monitoring by encouraging learners to explain their solutions. Related work on self-explanation suggests that it strongly enhances metacognitive monitoring [15]. A third critical component of Invention is that giving away the answer or showing students how to solve the problem cuts off learners' exploration and hinders their ability to notice and monitor [9]. Thus, instead of providing direct right/wrong feedback and elaborative explanatory feedback, our system exposes inconsistencies in the learner's solution. In sum, the three instructional strategies our system employs are (1) encouraging learners to contrast cases (2) inviting learners to explain their solutions and (3) providing feedback that exposes inconsistencies in a learner's solution.

## 3 Design of Invention Coach Prototype

Our research findings along with prior work on Invention informed the design of the Invention Coach. We designed instructional components corresponding to each phase of the Invention process model derived from our study. Additionally, some components scaffold students as they engage in the core learning mechanisms of the Invention paradigm. Our initial prototype was designed to be operated by a "Wizard-of-Oz" (the experimenter), who can launch the student into instructional components in any order, based on her assessment of the student's current knowledge state. While we ultimately plan to build a fully adaptive Invention Coach, the Oz configuration allows for flexible application of process phases across students. Perhaps more importantly, the Oz configuration will help us identify the trigger conditions for each type of coach guidance. We are now in the throes of building our first prototype Invention Coach. We are using the Cognitive Tutor Authoring Tools (CTAT, [16]) to build our ILE as an example-tracing tutor with additional custom programming.

In our Invention Coach, the student is initially left to work independently on his invention. During this independent work time, students typically inspect the cases provided and begin entering potential index numbers for each case. Students can also click the "rules tab" to re-read the rules that their index must follow, the "calculator tab" to display an on-screen calculator, the "notepad" tab to display an on-screen notepad, or the "help" or "submit" buttons to request feedback from Oz. Oz only provides guidance in response to the student's request for feedback, or whenever the student has been working uninterrupted for five minutes.

There are two types of guidance that Oz can provide: modules and hints. A module is a short exchange between the computer and student focused on a particular subgoal.

For example, our "ranking module" (Figure 2A) asks students to rank the bus companies from most to least crowded. After the student ranks the companies, the system automatically provides feedback and, if needed, additional scaffolding. Once the student has successfully ranked the companies, the module ends, and the student is left to work independently again. Hints represent the second type of guidance Oz can provide. Hints are much simpler than modules, consisting of a single text bubble displayed to the student. The system provides largely high-level hints with broad suggestions and never gives a "bottom-out" hint, which would give away the answer.

Each of the instructional components included in the Invention Coach was designed to guide students through one of the four process phases revealed in our analysis of teacher guidance (Table 1). Most components employ one of three instructional strategies that support the mechanisms of learning with Invention: encouraging students to contrast cases, inviting students to explain their solutions, and provide feedback that exposes inconsistencies in the student's inventions.

**Table 1.** Invention Process Model, Instructional Strategies, and Instructional Components

| Process Phases | Process Description | Instructional Strategy | Instructional Component |
|---|---|---|---|
| Understand the Problem | Explain or describe task goal and constraints | Expose inconsistencies | Rule-related hints<br>Rules tab |
| Notice Features | Notice key features of the underlying structure (e.g. #objects, space) | Contrast cases | Ranking module<br>Feature Contrast module |
| Produce and Reflect on an Invention | Generate a solution (e.g. index) and evaluate its correctness | Explain solution | Tell-Me-How module |
| Math Calculation | Simplify/manipulate fractions | -- | Calculator |

The two instructional components that help students through the "understand the problem" phase are the "rules tab" and the rule-related hints. Rule-related hints provide feedback exposing inconsistencies in students' inventions. For instance, if a student's invention is not generalizable and only works for specific cases, Oz can provide the following hint: "Don't forget: you have to use the exact same method to find the index for each bus!"

The Invention Coach also supports the "notice and understand features" phase of the Invention process via the "ranking" (described above) and "feature contrast" modules. Ranking the buses from most to least crowded helps students think about why some companies are more crowded than others, which starts to focus them on the features that determine crowdedness. In the "feature contrast" module (Figure 2C), Oz can select two specific buses to contrast. The student is then asked to note which features make one bus more crowded than the other. For example, Oz could ask the student to contrast cases A1 and C2 in Figure 1. Since the number of clowns is held constant across the cases while space changes, the student may begin to notice that clowns alone cannot account for crowdedness, the feature of bus size is important too.

Both "ranking" and "feature contrast" modules employ the instructional strategy of comparing and contrasting cases, to scaffold learners in noticing key features of the problem space.

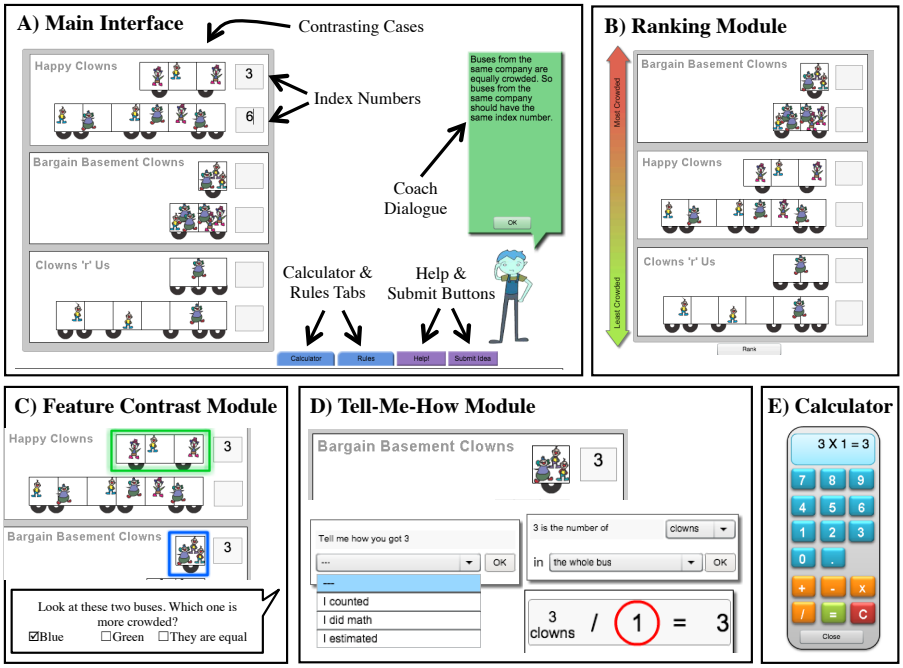



Fig. 2. Prototype Interface and Modules.

The backbone of the Invention Coach is the "tell-me-how" module (Figure 2D), where students are asked to enter and explain their inventions. This serves to recreate the "produce and reflect on an invention" phase of the process while encouraging students to monitor their own errors. In this module, students explain how they arrived at their answer (by selecting whether they "counted," "estimated," or "used math"). Students who indicate that they "counted" are further prompted to identify what exactly they counted, while students who "used math" must then use a calculator feature to show how they derived their answers. Students are never provided with direct right/wrong feedback on their solutions. Instead, the tell-me-how module encourages students to explain how they arrived at their solutions, right or wrong. We hope that in the process of explaining their answers, and connecting the math to referents in the cases, students will begin to reflect on their answers and identify gaps in their own understanding. Another key function of this module is to help Oz (and eventually the fully adaptive system) understand how a student generated her index so it can determine appropriate feedback.

Finally, to enable the math calculation phase of the Invention process, students are provided with a calculator (Figure 2E). In our study of naturalistic teacher guidance, many students had difficulty engaging in simple math (e.g. 6 divided by 3), and a

large proportion of teacher talk focused on math calculations such as simplifying fractions. The calculator enables students to off-load some of this challenging calculation work and instead focus on the larger concepts behind the math. The "calculator" tab is available in the main interface for students to call up at any time during the task. A calculator is also part of the "tell-me-how" module as described above.

Throughout the phases of the Invention process, the Coach's feedback points out inconsistencies in students' problem solutions. Instead of providing right/wrong or elaborative feedback when students create an incorrect invention, the Coach presents information to contradict the wrong invention. For instance, the Coach may remind the student that their Invention must generalize to all cases or that it must account for two cases that have the same crowdedness. The coach may also present pairs of cases that directly contradict the student. For instance, if a student believes that an irrelevant feature is important, the Coach will show two cases where the irrelevant feature varies but crowdedness does not. This type of feedback enables students to explore on their own, while encouraging them to self-monitor errors and "see" deep features.

In our current design, several components of the Invention Coach must be selected by Oz, while some intelligence is built into the system. The Oz selects whether to respond to a request for feedback by launching a student into a module (e.g. feature contrast, tell-me-how, or ranking) or by giving a single hint, adapting the path through the Invention space based on each student's individual needs. However, once inside a module, the system largely controls the interaction by selecting appropriate feedback and prompting the student to take action.

## 4 Discussion and Conclusion

We have described the design of a computer-based Invention Coach, which was inspired by a study of naturalistic teacher guidance of paper-based Invention and by prior research on the mechanisms behind Invention. The Invention Coach contains instructional components to address each phase in the Invention process, which can be adaptively selected. The system employs three instructional strategies that target key mechanisms in learning from Invention: contrasting cases, self-explanation of problem solutions, and feedback that exposes inconsistencies in students' solutions. While we are currently implementing a Wizard-of-Oz version of the Invention Coach, we ultimately aim to develop a fully adaptive system.

This work contributes more broadly to work on Invention and exploratory learning environments. To the best of our knowledge, the work presented here is the first design of a guided environment for Invention activities that is based on a model of naturalistic teacher guidance. Our design offers insight into possible strategies and phases of guidance that could be more broadly applicable in other exploratory learning environments and tasks. Specifically, if the Invention Coach we've built proves successful, it would argue that unguided exploration can be augmented by guidance that highlights inconsistencies in student work, contrasts cases to make relevant features salient, and invites students to explain their solutions. These forms of guidance may prove especially useful for developers who wish to retain the emphasis on active pro-

cessing and construction of ideas inherent in exploratory learning environments, while avoiding the pitfall of unproductive aimless exploration [2, 3].

## References


1. Mayer, R. E.: Should There Be a Three-strikes Rule Against Pure Discovery Learning?. American Psychologist, 59(1), 14-19 (2004)
2. Mavrikis, M., Gutierrez-Santos, S., Geraniou, E., & Noss, R. (2013). Design requirements, student perception indicators and validation metrics for intelligent exploratory learning environments. *Personal and ubiquitous computing*, *17*(8), 1605-1620.
3. McLaren, B. M., Timms, M., Weihnacht, D., & Brenner, D. Exploring the Assistance Dilemma in an Inquiry Learning Environment for Evolution Theory. In the Proceedings of the Workshop of Intelligent Support for Exploratory Environments 2012 (ITS 2012).
4. De Jong, T.: Scaffolds for Scientific Discovery Learning. Handling Complexity in Learning Environments: Research and Theory, 107-128 (2006)
5. Kapur, M., & Bielaczyc, K.: Designing for Productive Failure. Journal of the Learning Sciences, 21(1), 45-83 (2012)
6. Schwartz, D. L., Martin, T.: Inventing to Prepare for Future Learning: The Hidden Efficiency of Encouraging Original Student Production in Statistics Instruction. Cognition and Instruction 22(2), 129-184 (2004)
7. Bransford, J. D., Franks, J.J., Vye, N.J., Sherwood, R.D.: New Approaches to Instruction: Because Wisdom Can't be Told. Similarity and analogical reasoning, 470 (1989)
8. Schwartz, D. L., Bransford, J. D.: A Time for Telling. Cognition and Instruction 16(4), 475-5223 (1998)
9. Schwartz, D. L., Chase, C. C., Oppezzo, M. A., Chin, D. B.: Practicing Versus Inventing with Contrasting Cases: The Effects of Telling First on Learning and Transfer. Journal of Educational Psychology 103(4), 759 (2011)
10. Shemwell, J. T., Chase, C. C., Schwartz, D. L.: Seeking the General Explanation: A Test of Inductive Activities for Learning and Transfer. Journal of Research in Science Teaching 52(1), 58-83 (2015)
11. Roll, I., Aleven, V., Koedinger, K. R.: The Invention Lab: Using a Hybrid of Model Tracing and Constraint-based Modeling to Offer Intelligent Support in Inquiry Environments. In Intelligent Tutoring Systems, pp. 115-124. Springer, Heidelberg (2010)
12. Chase, C.C., Marks, J., Bernett, D., Bradley, M., & Aleven, V.: Towards the Development of an Invention Coach: A Naturalistic Study of Teacher Guidance for an Exploratory Learning Task. In C. Conati, N. Heffernan, A Mitrovic, & F. Verdejo (Eds.), Proceedings of the 17th International Conference on Artificial Intelligence in Education (2015)
13. Chi, M. T., Feltovich, P. J., & Glaser, R.: Categorization and Representation of Physics Problems by Experts and Novices*. Cognitive science,5(2), 121-152 (1981)
14. Loibl, K., & Rummel, N.: Knowing What You Don't Know Makes Failure Productive. Learning and Instruction, 34, 74-85 (2014)
15. Chi, M. T., Leeuw, N., Chiu, M. H., & LaVancher, C.: Eliciting self-explanations improves understanding. Cognitive science, 18(3), 439-477 (1994).
16. Aleven, V., McLaren, B. M., Sewall, J., Koedinger, K. R.: A New Paradigm for Intelligent Tutoring Systems: Example-Tracing Tutors. International Journal of Artificial Intelligence in Education 19(2), 105-154 (2009)

# Discovering knowledge models in an open-ended educational game using concept formation

Erik Harpstead, Christopher J MacLellan, Vincent Aleven

Human Computer Interaction Institute, Carnegie Mellon University,
5000 Forbes Avenue, Pittsburgh, PA, 15232, USA
{eharpste, cmaclell, aleven}@cs.cmu.edu

**Abstract.** Developing models of the knowledge and skills being exercised in a task is an important component of the design of any instructional environment. Developing these models is a labor intensive process. When working in exploratory and open-ended environments (EOLEs) the difficulty of building a knowledge model is amplified by the amount of freedom afforded to learners within the environment. In this paper we demonstrate a way of accelerating the model development process by applying a concept formation algorithm called TRESTLE. This approach takes structural representations of problem states and integrates them into a hierarchical categorization, which can be used to assign concept labels to states at different grain sizes. We show that when applied to an open-ended educational game, knowledge models developed from concept labels using this process show a better fit to student data than basic hand-authored models. This work demonstrates that it is possible to use machine learning to automatically acquire a knowledge component model from student data in open-ended tasks.

## 1 Introduction

When designing intelligent instructional support in educational learning environments it is important to have a model of the skills and knowledge employed during problem solving. A common approach to modeling skills in intelligent tutoring systems (ITSs) is knowledge component (KC) modeling [1]. In the KLI Framework a KC is "an acquired unit of cognitive function or structure that can be inferred from performance on a set of related tasks" [2]. A KC model is a mapping of each problem-solving step in a particular educational environment to the skills necessary to solve that step. KC models can be used in pedagogical software to drive feedback and hints, guide problem selection [3], and inform redesign of the interface [4].

While KC models are useful for a number of purposes in the development of intelligent software they take significant effort to develop. The process of creating a KC model often employs elements of empirical and theoretical task analyses [1], soliciting expert feedback and rationally constructing the skills used in a task. When working in exploratory and open-ended environments (EOLEs) this process is aggravated by the freedom learners experience in these environments. It can be assumed that as the space learners are allowed to explore grows, so too must a KC model grow to

continue to provide useful support and feedback to learners. In addition to providing large spaces for exploration, EOLEs often contain more complex representations of domains making it more difficult to articulate the rules defining the applicability of a given KC.

To address the challenges of KC model creation in EOLEs we have developed a novel method for generating new KC models based only on problem states taken from the learning environment. Our approach uses a form of automated model discovery that employs a concept formation algorithm called TRESTLE [5]. This algorithm creates a hierarchical categorization tree based on training examples, which can then be used to label problem states at various grain sizes. The algorithm is designed to handle messy, mixed representations of data, making it ideal for application to EOLEs. It has previously been shown to create clusters similar to humans [5]. In this paper, we show how the conceptual patterns learned by TRESTLE can be used to discover new KC models in the open-ended educational game *RumbleBlocks* [6]. Finally, we conclude with a brief discussion of the implications of this approach and detail how we plan to expand it in future work.

## 2 The TRESTLE Algorithm

TRESTLE [5] is an incremental concept formation algorithm that creates a hierarchical categorization tree from a set of structured instances. In this section we briefly describe the algorithm's major structures and categorization procedure for more details see [5][1].

The TRESTLE algorithm produces a categorization tree and functions over a set of instances, each described by a set of attribute-value pairs. Instance attributes can have nominal, numeric, or component values that have their own sub-attributes and values. When integrating a new instance TRESTLE proceeds through 3 major steps:

1. Partial Matching, which renames instance attributes to align with the algorithm's current domain understanding
2. Flattening, which converts structured attributes to unstructured ones, while preserving structural information.
3. Categorization, which incorporates the instance into the knowledge base.

TRESTLE's knowledge base is an evolving category structure being built from training examples and is organized into a hierarchical tree of concepts. In building its tree, the algorithm optimizes for a heuristic called category utility, which is similar to maximizing for the expected number of correct guesses that a given concept could make about the attribute-values of a given instance. During categorization new instances are sorted into the tree. At each node in the categorization tree TRESTLE considers 4 different operations and performs whichever one would result in the highest category utility: (1) adding the instance to the best child, (2) creating a new node

[1] A reference implementation is available at: https://github.com/cmaclell/concept_formation

for the instance, (3) merging the best 2 nodes and adding the instance to the result, or (4) splitting the best node by promoting its children to be children of the current node.

After categorizing an instance into its knowledge base, TRESTLE returns a concept label for the instance. Since concepts in TRESTLE are organized in a hierarchical tree, the cluster labels returned from categorization can be generalized if more coarse clusters are desired. At the coarsest, i.e. the root of the tree, everything is considered to be the same concept, while at the most specific, i.e. the leaves of the tree, everything is considered to be unique.

To arrive at a KC label for a step, the problem state in which the step too place is categorized and label is generated based on the returned concept and a desired depth. For a given depth model the state is categorized down the TRESTLE tree. Once the state reaches the desired depth the current concept's label is returned. If the state reaches a leaf of the tree before reaching the desired depth, then the label of the deepest node is used instead. When generating KC models this allows for the creation of multiple model variants that consider the domain at different levels of granularity (see **Fig. 1**).

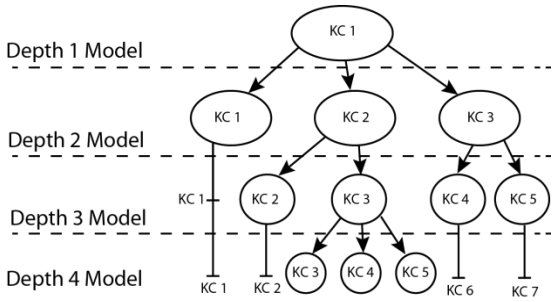



**Fig. 1.** A diagram of how KC labels are attributed to problem states based on their categorization in the TRESTLE tree for a given depth mode.

## *3 RumbleBlocks*

To demonstrate how TRESTLE can be used to aid in the process of KC modeling we introduce *RumbleBlocks* [6], an open-ended educational game. *RumbleBlocks* is a physics game designed to teach children (ages 5-8) three basic concepts of structural stability and balance: (1) objects with wider bases are more stable, (2) objects that are symmetrical are more stable, and (3) objects with lower centers of mass are more stable.

In the game, players are tasked with building a tower out of blocks to help a stranded alien power their spaceship (see **Fig. 2**). The tower must be tall enough to reach the alien and cover a series of energy orbs that power the spaceship. Once players have finished building their tower they place the spaceship on top, which triggers an earthquake. If, after the earthquake, the ship is still on top of the tower, then the player has succeeded and advances on to the next level, otherwise they must try the level again.

**Fig. 2.** A screenshot of *RumbleBlocks*

Each level in *RumbleBlocks* is designed to emphasize one of the three key concepts of stability. This emphasis is accomplished through the placement of energy orbs, the target zone for the spaceship, and the palette of available blocks. While each level is targeted at a particular principle, there is a wide range of variance in the kinds of solutions players design to in-game challenges. Our previous analysis found that there are several levels where less than 10% of students actually used the solution envisioned by the game's designers [7]. The variance in player behavior demonstrates the open-endedness of the game as well as highlights the challenge inherent in defining KC models to measure learning in the game.

## 4 KC Model Discovery in *RumbleBlocks*

To evaluate the application of TRESTLE to the KC modeling process we used it to discover a set of new KC models in *RumbleBlocks*. For comparison we also created a "hand built" KC model meant to capture the original design intent behind the game. This model labels each level in the game with the principle it is designed to emphasize. Since the first 5 levels of the game are primarily a mechanical tutorial for the game rather than instructional levels dealing with physics principles, we relabeled these levels with an "Intro" KC, resulting in a hand-built model with 4 KCs.

For this first demonstration of the use of TRESTLE to generate KC models we chose to focus on a broad definition of a step as solving an entire level of *RumbleBlocks*. This is in keeping when Van Lehn *et al.*'s definition of a step as "the smallest possible correct entry that a student can make" [8] because, in its current form, *RumbleBlocks* only provides correctness feedback to players at the end of a level. In this context a step is then considered in terms of the initial level state given to the player to construct a solution in and evaluated based on their final construction. The state representation used for training TRESTLE contained the positions of each of the energy orbs, the target position of the spaceship, and the available number of each block type. The resulting categorization tree, based on the initial state data from *Rumble-*

*Blocks'* 47 levels, was 7 levels deep giving us 7 candidate KC models each with differing levels of granularity.

To evaluate relative appropriateness of different candidate KC models we used the tool suite provided by DataShop [9]. In particular, we used AFM [10], a specialized form of logistic regression that fits a given KC model to student log data. The resulting regression model can be used to assess the fit of a particular KC to the real student data. DataShop provides several model fit statistics to compare KC models: AIC and BIC, both standard model fit statistics that penalize for model complexity and Cross Validated Root Mean Square Error (CV-RMSE) using 3-fold cross validation with different stratification schemes (i.e. student, item and un-stratified).

The data we use in our evaluation comes from a formative evaluation of the game with 174 players in the target demographic. Players were allowed to play the game for two 20-minute sessions.

The model fit estimates for the 7 Trestle-based models and the original Principle (i.e., hand-built) model can be seen in **Table 1**. In general, more fine grained models tend to fit the data better. The TRES-Depth7 model is preferred according to AIC and both item-stratified and un-stratified RMSE. This would suggest that an appropriate model for initial states in *RumbleBlocks* is one that treats all levels as nearly unique from each other.

**Table 1.** Fit statistics for each KC model. Cross Validated Root Mean Square Errors (CV-RMSE) are based on 3 fold cross validation using different forms of stratification.

| *Model* | *KCs* | *AIC* | *BIC* | *CV-RMSE (student)* | *CV-RMSE (item)* | *CV-RMSE (none)* |
|---|---|---|---|---|---|---|
| Principle | 4 | 6560.73 | **8544.74** | .3856 | .3883 | .3869 |
| TRES-Depth1 | 1 | 6828.35 | 8771.45 | .3924 | .3948 | *NA* |
| TRES-Depth2 | 5 | 6737.21 | 8734.85 | .3899 | .3921 | .3923 |
| TRES-Depth3 | 14 | 6661.67 | 8782.03 | .3878 | .3904 | .3915 |
| TRES-Depth4 | 24 | 6530.78 | 8787.50 | .3845 | .3853 | .3855 |
| TRES-Depth5 | 32 | 6350.50 | 8716.31 | .3794 | .3821 | .3826 |
| TRES-Depth6 | 39 | 6152.75 | 8614.01 | **.3734** | .3761 | .3739 |
| TRES-Depth7 | 41 | **6152.28** | 8640.81 | .3736 | **.3754** | **.3732** |

## 5 Discussion

We can see from the results that KC models based on depth cuts of a TRESTLE categorization tree better fit student data than a model based on the original design of the game in terms of AIC and cross-validation. According to these statistics, we find that a more specific KC model better fits student data than more general models. This would make it appear that there is little transfer going on within the game. However, this is likely due to our unit of analysis. An approach that employs a more fine grained definition of a correct step (e.g., steps defined at the transaction level) might reach a different conclusion with regards to transfer because there is likely to be some

common application of knowledge components used across building towers in different levels.

The approach presented here deals with concept granularity at a holistic level. By this we mean that all KCs in a model are being considered at the same depth of the concept tree. There is some evidence that suggests human learners will employ concepts at different levels of granularity based on their expertise [11]. It is possible that the most appropriate KC model uses a combination of specific and general concepts depending on the context of the task at hand. Rather than creating KC labels as uniform cuts of a concept hierarchy, where concepts all exist at the same depth, we could instead start all problem states at their coarsest label and iteratively split concept nodes into more specific labels. After each split the resulting KC model could be tested for fit using student data until an optimal model is found. This is similar to the Learning Factors Analysis search algorithm [12] but it would not require human developed models as seeds. Exploring this process is something we look forward to in future work.

Our current analysis defined steps to be the complete solution to each level. This follows with Van Lehn *et al.*'s definition of a step in KC analysis as the smallest amount of action that a student can perform correctly [8]. This definition still assumes that all possible solutions to a level exercise the same skill, which may not be the case in practice. One way of going beyond this assumption in analyzing *RumbleBlocks* is to create a TRESTLE model based on the solutions players make to each in-game level rather than the initial conditions of the level. Such an approach would allow for analysis according to different kinds of solutions rather than the constraints under which problem solving took place. One issue with taking into account the content of students' solutions is how to handle the assignment of KC labels when there are multiple valid solutions to a level, as is the case with *RumbleBlocks* [7]. In the case of correct solutions it is simple to state that each unique correct solution embodies the use of a different KC. When looking at incorrect solutions, however, the question of attribution becomes more difficult as it is hard to know which of the possible correct approaches the student failed to execute correctly. A standard modeling approach would assign an incorrect step with the labels of all possible correct solutions; using a variant to AFM's statistical formula to allow for the disjunction of KCs [10]. A TRESTLE based approach could go beyond this by categorizing incorrect solutions into a knowledge base trained on correct solutions and assigning a KC label based on which correct solution the error most closely resembles. This is similar to the approach taken by Rivers and Koedinger to create next step feedback in programing tasks [3] but has the potential to be domain general. Exploring this approach to KC modeling with TRESTLE remains a topic of our future work.

Ideally, we would like to go beyond the final state definition of a step to a transaction-level model. Having a full transaction-level model would allow for the inclusion of targeted feedback to players while they are playing rather than providing feedback only at the end of building. Additionally, more detailed understanding of player problem solving could better inform adaptive sequencing. The challenge in taking this approach in *RumbleBlocks* is that that evaluation of player performance is currently only performed at the end of a level. This creates similar correctness attribution chal-

lenges in deciding whether a particular build step is a good or bad example of a given concept. Again we could turn to TRESTLE to aid in this analysis by having it perform categorization on whole solution paths rather than final solutions. There are several open questions with this analysis in terms of how best to represent a solution path for categorization but we hope to resolve these issues in future work.

## 6 Conclusion

This paper presents a preliminary use of TRESTLE as a way to discover new KC models in an open-ended game. The models automatically discovered by TRESTLE better fit student data than one hand-built to capture the design intent of the game. This demonstrates the promise of concept formation based approaches to KC model creation. In future work we plan to further explore the implications of TRESTLE-based KC models including discovering transaction-level models and exploring models that capture mixed grain sizes. We hope other researchers can find utility in these methods and apply them to their own exploratory and open-ended environments.

## 7 Acknowledgement

We would like to thank the developers of *RumbleBlocks* and our colleagues who performed the evaluation that provided our data. This work was supported in part by the DARPA ENGAGE research program under ONR Contract Number N00014-12-C-0284 and by a Graduate Training Grant awarded to Carnegie Mellon University by the Department of Education # R305B090023.

## 8 References

1. Aleven, V., Koedinger, K.R.: Knowledge Component Approaches to Learner Modeling. In: Sottilare, R.A., Graesser, A., Hu, X., and Holden, H. (eds.) Design Recommendations for Intelligent Tutoring Systems: Volume 1 - Learner Modeling. pp. 165–182. U.S. Army Resarch Laboratory (2013).
2. Koedinger, K.R., Corbett, A.T., Perfetti, C.: The Knowledge-Learning-Instruction Framework: Bridging the Science-Practice Chasm to Enhance Robust Student Learning. Cogn. Sci. 36, 757–98 (2012).
3. Rivers, K., Koedinger, K.R.: Automating Hint Generation with Solution Space Path Construction. Proc. ITS 2014. pp. 329–339 (2014).
4. Koedinger, K.R., Stamper, J.C., Mclaughlin, E.A., Nixon, T.: Using Data-Driven Discovery of Better Student Models to Improve Student Learning. Proc. AIED 2013. pp. 421–430 (2013).
5. Maclellan, C.J., Harpstead, E., Aleven, V., Koedinger, K.R.: TRESTLE: Incremental Learning in Structured Domains using Partial Matching and Categorization. Proceedings of the 3rd Annual Conference on Advances in Cognitive Systems - ACS 2015 (2015).
6. Christel, M.G., Stevens, S.M., Maher, B.S., Brice, S., Champer, M., Jayapalan, L., Chen, Q., Jin, J., Hausmann, D., Bastida, N., Zhang, X., Aleven, V., Koedinger, K., Chase, C.,

Harpstead, E., Lomas, D.: RumbleBlocks: Teaching science concepts to young children through a unity game. Proc. CGAMES 2012. pp. 162–166. IEEE (2012).
7. Harpstead, E., Maclellan, C.J., Koedinger, K.R., Aleven, V., Dow, S.P., Myers, B.A.: Investigating the Solution Space of an Open-Ended Educational Game Using Conceptual Feature Extraction. Proc. EDM 2013. pp. 51–58 (2013).
8. Vanlehn, K., Koedinger, K.R., Skogsholm, A., Nwaigwe, A., Hausmann, R.G.M., Weinstein, A., Billings, B.: What's in a Step? Toward General, Abstract Representations of Tutoring System Log Data. Proceedings of the 11th International Conference on User Modeling - UM 2007. pp. 455–459. Springer Berlin Heidelberg (2007).
9. Koedinger, K.R., Baker, R.S.J. d, Cunningham, K., Skogsholm, A., Leber, B., Stamper, J.: A Data Repository for the EDM community: The PSLC DataShop. In: Romero, C., Ventura, S., Pechenizkiy, M., and Baker, R.S.J. d. (eds.) Handbook of Educational Data Mining. pp. 43–55 (2010).
10. Cen, H., Koedinger, K., Junker, B.: Comparing Two IRT models for conjunctive skills. Proc. ITS 2008. pp. 796–798 (2008).
11. Fisher, D.: A Computational Account of Basic Level and Typicality Effects. Proc. 1988 Nat. Conf. Artificial Intelligence (AAAI'88). pp. 233–238 (1988).
12. Cen, H., Koedinger, K., Junker, B.: Learning Factors Analysis – A General Method for Cognitive Model Evaluation and Improvement. Proceedings of the 8th International Confernence on Intelligent Tutoring Systems - ITS 2006. pp. 164–175 (2006).

# Towards Configurable Learning Analytics for Constructionist Mathematical e-Books

Sokratis Karkalas, Christian Bokhove, Patricia Charlton, and Manolis Mavrikis

London Knowledge Lab, Institute of Education, University College London, UK
s.karkalas,c.bokhove,m.mavrikis,p.charlton@ioe.ac.uk

**Abstract.** This paper presents emerging requirements for learning analytics on interactive mathematical e-books and a framework that can be used for the seamless integration of complex learning objects with e-book platforms. We describe the opportunities that this approach opens up regarding interoperability and configurability of learning analytics and intelligent support. The framework is generic and can be used for any type of system with similar requirements. In this paper we present a case that covers configuration of learning analytics for teachers and intelligent support for students in constructionist mathematical e-books.

## 1 Introduction

The emergence of authoring software for e-books means that digital books with text, images and other interactive elements are increasingly being used on personal computers and other electronic devices for educational purposes. However, most of these e-books are simple transformations of traditional textbooks into a digital format and do not take advantage of the dynamic and computational affordances offered by this emerging technology. The MCSquared project[1] is investigating whether the affordances of state-of-the-art e-books can be exploited to support the learning of abstract mathematical concepts. We are looking into the design of highly interactive constructionist e-book widgets, and exploring their potential for providing learners with opportunities to construct mathematical artefacts in order to engage creatively with mathematical problems.

Within this context the increase of both process and product data collected provides unprecedented opportunities for knowledge discovery through state-of-the-art data analysis and visualization techniques. However, despite the fact that in the past two decades intelligent technology has become increasingly feasible, the power of these methods has not reached its full potential in education. For example, although it is now possible for intelligent pedagogical agents to monitor learners' interactions within educational applications and provide individualised support, only a handful of intelligent tools are employed in practice, yet they are tied to particular instructional approaches, domains and context.

[1] The Mathematical Creativity Squared project is funded by the EU, under FP7 ICT-2013.8.1 Project #610467. For more details see http://www.mc2-project.eu

We believe that one of the reasons that the promises of ubiquitous, individualised and adaptive technology has had a very small impact in education is that learning environments are often rigid and limited to specific learning contexts and pedagogical approaches. Our previous research [6, 7] and that of others (e.g., [9]) has primarily enabled the rapid revision and management of content. In line with previous research in the field (e.g. [8]), our vision is that teachers and educational organizations will be able to also mould the nature and type of support provided to a learner (cf. [2, 10]) and the information they want to glean from their interaction. Then the unrealised potential of the technology could begin to be exploited.

This paper presents our preliminary efforts towards this vision: a prototype where e-book pages and the widgets that they contain can be configured. First, we present below a set of emerging requirements for Learning Analytics in the context of constructionist mathematical e-books.

## 2 Emerging Requirements for Configurable Learning Analytics

With the advent of data science and analytics in general, there are several 'analytics' tools that have appeared. While we have looked into a large subset of them, we cannot review them all in detail here. However, we have been unable to find a tool that focuses on providing information from constructionist, exploratory mathematical environments (with the exception of our previous work in [3] where we also review related work in more detail).

In the context of commercial e-books in particular publishers and authors are interested in (and to some extent only have access to) high level information such the number of pages read, average reading times, exit rates and other details that reveal reading patterns that can correlate with, for example, sales figures. However, from an educational point of view teachers, designers and even students require a more in-depth analysis of learners' interaction with the e-books.

The MCSquared project comprises four Communities of Interest (COI) across 4 EU countries (France, Greece, Spain, UK) and engaged their members in requirements elicitation and stakeholders' analysis. Through several face-to-face workshops and sustained online interaction and communication between members of the COI we have identified many scenarios in which e-books are being used in teaching and other requirements of learning analytics tools and data visualisation that are emerging.

Digital resources like e-books are being used either directly in the classroom or in 'blended' learning scenarios (e.g. for practice exercises at home) or in a 'flipped' learning model where students read and interact with the e-book content online (e.g. at home) and complete other parts of the e-book in the classroom with the help of other students or the teacher. So neither context can be excluded. We present below high-level categories of the themes around which requirements have emerged:

– Usage and other book-level descriptive statistics

- the order of pages
- time spent on each page/activity
- how quickly students read a page
- the percentage of coverage of particular pages from a book

– Structured answer and related descriptive statistics
  - Student answers and performance in structured questions
  - Number of attempts to answer a question
  - Repeated wrong answers across students

– Constructionist Analytics
  - Constructionist descriptive statistics (i.e. number of objects constructed, moved, deleted, etc.)
  - Data regarding construction operations (achievements of key 'landmarks')
  - Specific patterns of interaction within a widget

While the first and second category of data analytics are interesting in their own right, we are focusing mostly on the third type of data that we refer to as 'deep' analytics of constructionist e-books for learning. This is particularly interesting because it goes beyond the 'low-hanging fruit' of descriptive statistics (which, in principle, are technically and conceptually well understood) and looks into extracting some meaningful information that could support decision making. Constructionist analytics opens up the door to real-time formative and summative assessment (as discussed in [1]). In our previous work, we found that even a simple traffic-light system could satisfy the teacher's need for finding out which students are progressing satisfactorily towards completing the task and which ones may be in difficulty [3].

In addition, a requirement across all the types of analysis mentioned above, is the availability of a generic, interoperable framework that enables configurability of learning analytics and intelligent support. We present a prototype of this in the next section.

## 3 Prototype

In this section, using an example of an e-book page, we demonstrate a basic but complete integration scenario. The page is part of a mathematics e-book developed by the Greek COI and features a learning activity developed in Geogebra. The page is integrated in a prototype that has a local in-memory database that stores data generated from the student activities and a rule-based reasoner that provides real-time intelligent support to the students (fig. 1). The purpose of the activity is to get the student select an appropriate combination of variables in order to get both parts of the ladder to the same level. Converging the two parts can then display a single heart at the top (join the two halves). All of these heterogeneous components are pluggable widgets that operate in their own secure environment (sandbox). They are hosted in their own domains and they are executed concurrently without interfering with one another. Integration with the host page takes place through a lightweight set of mediator

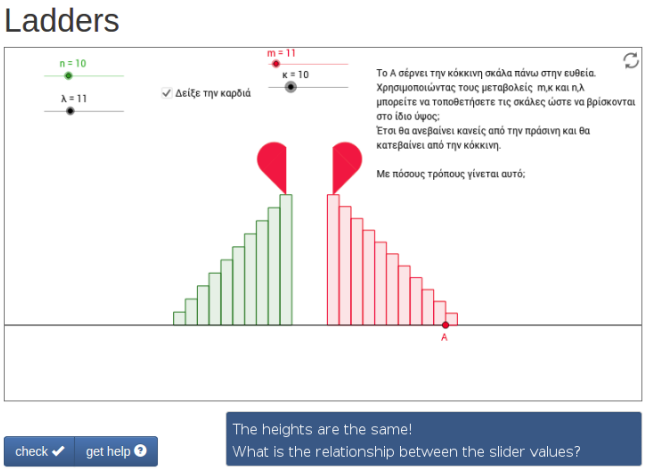



**Fig. 1.** The 'Ladders' Activity from Geogebra Tube

wrappers that enable full two-way communication over a simple and common interface. Each widget is allowed to expose its own functionality (or part of it) and make it available to the platform through a wrapper interface. This scheme allows better performance (multithreading), security (sandboxing), controllable interoperability (widget interface exposed through the wrapper) and seamless integration (common wrapper interface) [5].

This e-book page demonstrates an example of an activity that offers real-time intelligent support to students through visual controls and real-time formative and summative feedback to teachers through graphs. The activity widget offers interactivity through sliders and a checkbox. As the student interacts with the widget, action indicators are generated and sent to the page. The platform populates the local (in-memory) database which in turn incrementally synchronises with the back-end database through REST [2] web-service endpoints (fig. 2). These updates are asynchronous for better performance. The local database serves as a buffer for data that needs to be immediately available and thus enables fast and more reliable responses. The local data is then sent to the rule-based reasoner for processing. If the reasoner identifies a case that justifies a discreet intervention, a message is displayed in the textbox and/or some visual indicator is presented in the activity frame (heart). The latter presupposes that messages are sent to the activity widget through the platform. This process may also be initiated by the student. If the student asks for help or wants the system to evaluate the work that has been submitted so far, then the reasoner responds with an appropriate message in the textbox. In parallel, the data generated from both the activity and the reasoner is sent to the database.

[2] http://en.wikipedia.org/wiki/Representational_state_transfer

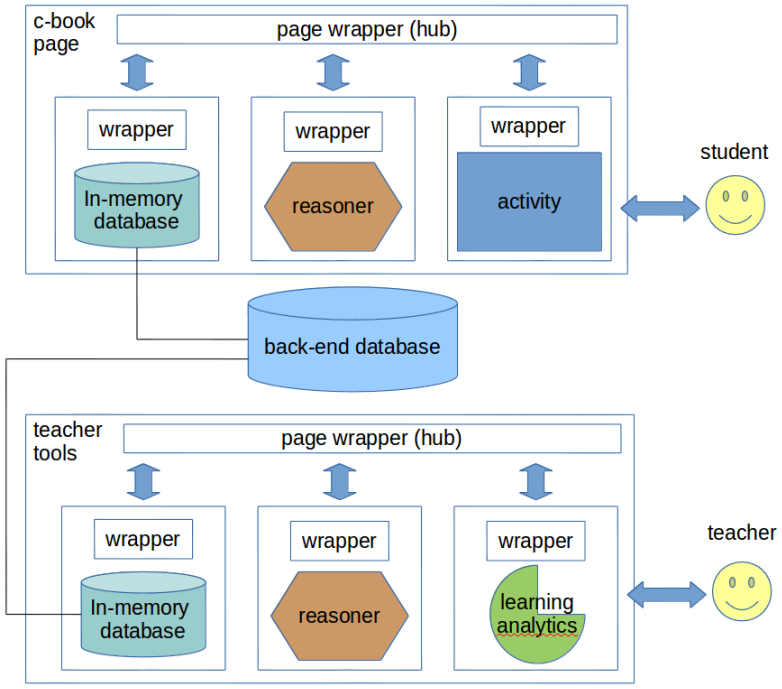



**Fig. 2.** The Architecture

The page that hosts the teacher tools has a similar structure. It also contains a local database widget and a reasoner. As the back-end database gets updated with student actions and reasoner findings, the local database incrementally retrieves the changes. Some of this data is used as a direct feed to other widgets that host learning analytics visualisations. In this particular example we have visualisations that measure student activity and performance (fig. 4). Both measurements are presented as histograms and provide real-time feedback to the teacher. The first visualisation measures what has been used in the activity and how much. For example the elements n, m, k and  are numeric variables that correspond to sliders in the construction. The visualisation shows which of these sliders and how many times have been used by the student. The second visualisation presents a comparative measurement of effort and levels of achievement. Some of the local data is then processed by the reasoner and new data may be inserted into the database. This data may be used to populate other visualisations or provide some intelligent support to the teacher.

The teacher tool is both an authoring and a monitoring environment. The teacher has the ability to dynamically configure the system to log actions performed by specific widget elements. Widget instances can be dynamically inserted into the authoring environment in the same way they can be integrated with a c-book page. The widget communicates with its host through the wrap-

pers and makes available its internal structure to it. The metadata extracted from the widget is then used by the host to dynamically construct an authoring graphical user interface that is presented to the teacher (fig. 3). The teacher can then select the widget elements deemed necessary to log their actions. This information is sent to the database along with the id of the c-book the widget belongs to. When the widget is invoked in a c-book, the page uses this information to dynamically register event handlers in the widget in order to intercept student actions for the selected elements.

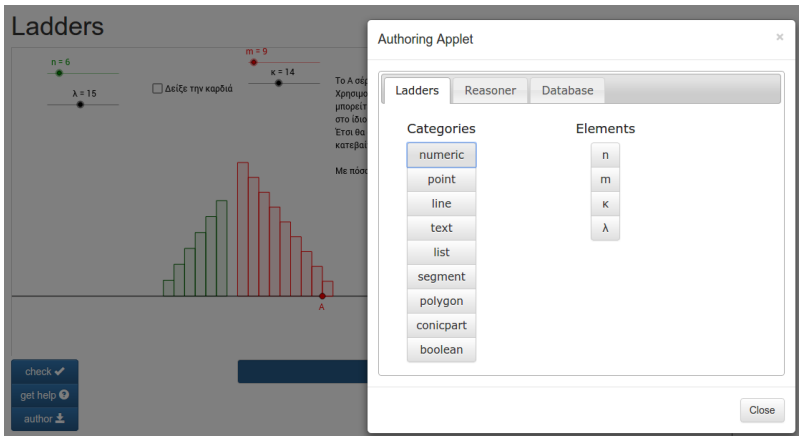



**Fig. 3.** Authoring Applet for the Teacher

## 4 Conclusion and Future Work

In this paper we presented a prototype authoring environment that enables configurable learning analytics and intelligent support in educational e-books. The specific example used in this presentation focuses on constructionist mathematical learning activities and the configuration of appropriate analytics for them. The system has been implemented and used by members of COIs and preliminary results show that it meets its original design objectives. It can be used effectively for rapid integration of learning objects and dynamic configuration of learning analytics and intelligent support. The next step is to specify how this data will be processed by the reasoner in order to provide effective support to the students. This part requires the use of a rule editor by a domain expert. Preliminary work towards this aspect has been undertaken in [4].

A distinguishing characteristic of the prototype presented here is the ability to dynamically generate user interfaces that enable the configuration of learning analytics on heterogeneous learning objects. Heterogeneity is hidden behind the mediator wrappers. A possible future enhancement would be to analyse a number

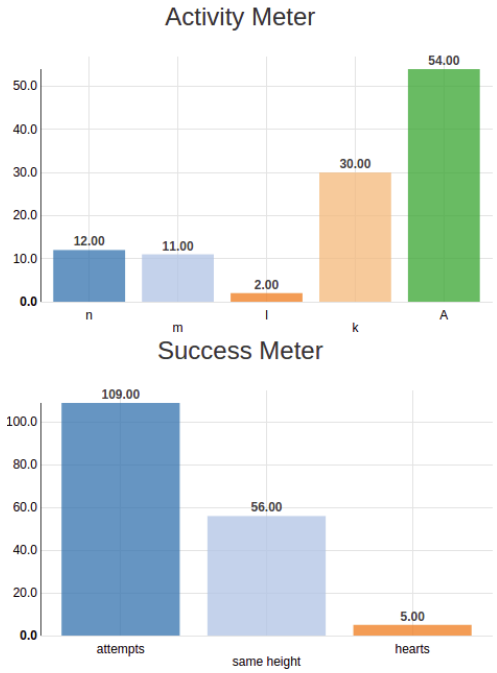



**Fig. 4.** Teacher Visualisations. In this example the x axis represents the different aspects that the teacher selected to log and the y-axis the number of logged cases.

of representative learning objects and create a learning component description language that can be used as a standard description of the construction that represents an activity. This language could then be used to semantically enhance the component in the wrapper in a standardised way.

## References

1. Berland, M., Baker, R., Blikstein, P.: Educational data mining and learning analytics: Applications to constructionist research. Technology, Knowledge, and Learning (19), 205–220 (2014)
2. Bokhove, C.: Implementing feedback in a digital tool for symbol sense. International Journal for Technology in Mathematics Education 17(3), 121–126 (2010)
3. Gutierrez-Santos, S., Mavrikis, M., Geraniou, E., Poulovassilis, A.: Usage scenarios and evaluation of teacher assistance tools for exploratory learning environments. Computers and Education (in press)
4. Karkalas, S., Gutierrez-Santos, S.: Enhanced javascript learning using code quality tools and a rule-based system in the flip exploratory learning environment. In: Advanced Learning Technologies (ICALT), 2014 IEEE 14th International Conference on. pp. 84–88. IEEE (2014)
5. Karkalas, S., Mavrikis, M., Charlton, P.: Turning web content into learning content. a lightweight integration and interoperability technique (2015), under review

6. Mavrikis, M., González Palomo, A.: Mathematical, interactive exercise generation from static documents. Electronic Notes in Theoretical Computer Science 93, 183–201 (Feb 2004), http://www.lkl.ac.uk/manolis/pubs/pdfs/mkm-interactive-04.pdf
7. Mavrikis, M., Maciocia, A.: Wallis: a web-based ILE for science and engineering students studying mathematics. three years on. (2006), http://webalt.math.helsinki.fi/webalt2006
8. Murray, T., Blessing, S., Ainsworth, S.: Authoring Tools for Advanced Technology Learning Environments: Toward cost-effective adaptive, interactive, and intelligent educational software. Springer (2003)
9. Pardo, A., Fisteus, J.A., Kloos, C.D.: A distributed collaborative system for flexible learning content production and management. Journal of Research and Practice in Information Technology 44(2), 203 (2012)
10. Sangwin, C.: Automating the marking of core calculus and algebra: eight years on. In: Robinson, M., Challis, N., , Thomlinson, M. (eds.) Maths at University, pp. 135–139. More Maths Grads (2010)

# Analysing Project Based Learning Scenarios to inform the design of Learning Analytics: Learning from related concepts

Rose Luckin, Manolis Mavrikis, Katerina Avramides, Mutlu Cukurova

London Knowledge Lab, UCL Institute of Education, University College London

**Abstract.** Project Based Learning is a complex concept that is related to Problem Based Learning and Collaborative Problem Solving. These latter concepts are well represented in the literature by models and frameworks that can usefully be adapted to develop a framework for the analysis of Project Based Learning. We present such a framework that has been designed for learning situations that involve the use of technology. This technology can be used to capture data about learners' interactions as well as to support their learning. We suggest that this data can be combined with data collated by human observers and analysed using the framework.

## Introduction

The literature on Project Based Learning is complex with many related concepts, for example: Practice Based Learning, Problem Based Learning, Collaborative Problem Solving and Inquiry Learning. In this paper we explore the frameworks for two of these concepts: Problem Based Learning (PBL) and Collaborative Problem Solving (CPS) in an attempt to identify a framework for the analysis of Project Based Learning activities to inform the design of Learning Analytics. We have selected these two concepts, because they are well supported by existing models and frameworks.

### 1.1 Problem Based Learning

Problem based approaches encourage learners to become actively engaged in meaningful real-world problems that often require practical as well as intellectual activity. The premise is that the students who participate in a PBL approach will learn through solving problems together and then reflecting upon their experience (Barrows and Tamblyn, 1980). Problem-based approaches to learning (PBL) are not new, they date back to the early 20th century in the work of Dewey (1938) for example (Hmelo-Silver, 2004). Whilst they were initially part of medical education and law schools; they have recently gained more popularity with educators in schools and universities for teaching STEM subjects. A key element of PBL is that the students work collaboratively, learning from each other and solving the problem together. The teacher's

role is that of facilitator, but the students are very much self-directed. The PBL approach therefore requires that participating students have good collaborative skills and sufficient metacognitive awareness to steer them through the problem space in a manner that enables their learning. As a result the potential outcomes for the students are not merely cognitive in terms of their increased understanding of the subject matter of the problem, but also there are advances in the transferable twenty first century skills of communication, collaboration and critical thinking.

Hmelo-Silver (2004) uses a stepwise model to describe the PBL process from the teacher's perspective (see Figure 1). Students start by identifying relevant facts about the problem, which increases their understanding and enables them to generate their hypotheses about potential solutions. The teacher or potentially a more able peer helps the student to recognize what are referred to as knowledge deficiencies that will become the goals of their self-directed study. Once these knowledge deficiencies have been addressed the student can re-evaluate their hypotheses and learn through a process of reflection and application.

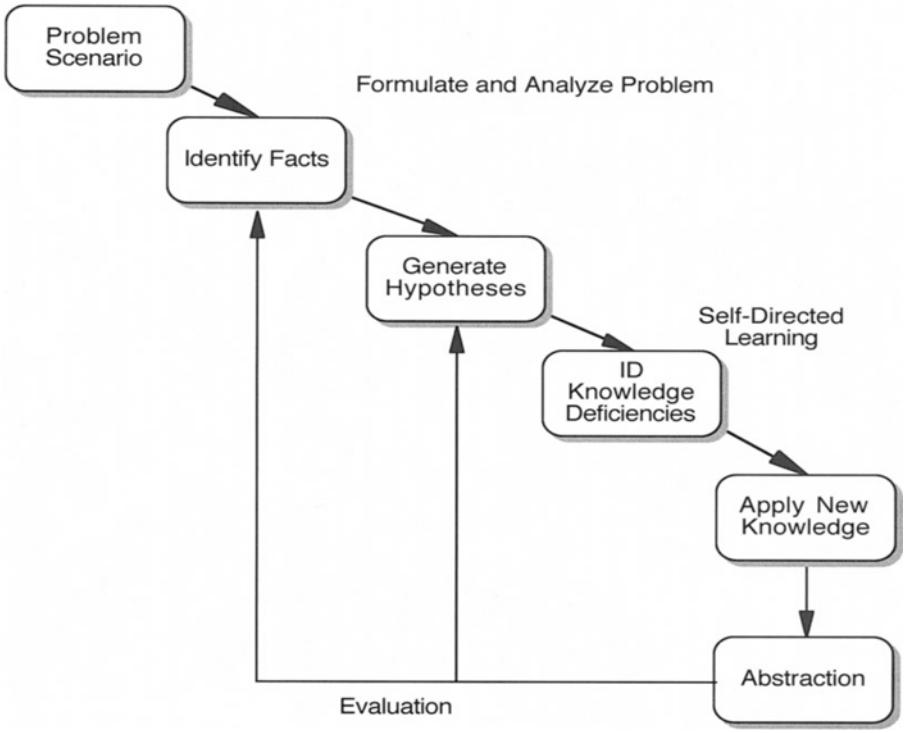



**Fig. 1.** PBL Tutorial Model (Hmelo-Silver, 2004)

### 1.2 Collaborative Problem Solving

More recently, and in preparation for the 2015 PISA assessments, the OECD has developed a framework for the assessment of collaborative problem solving (CPS) that is complementary to the traditional PBL approach outlined above (OECD, 2013). The OECD defines CPS as:

> Collaborative problem solving competency is the capacity of an individual to effectively engage in a process whereby two or more agents attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills and efforts to reach that solution.
>
> (OECD, 2013, p.6)

There are three core competencies that are fundamental to this definition of CPS:

1. Establishing and maintaining shared understanding;
2. Taking appropriate action to solve the problem;
3. Establishing and maintaining team organisation.

These are combined with a set of problem solving competencies that are similar to those outlined by Hmelo-Silver (2004), although there is no explicit reference to knowledge deficiencies. This is not surprising because the PBL model is one of tuition, whereas the OECD CPS model is one of assessment:

1. Exploring and Understanding
2. Representing and Formulating
3. Planning and Executing
4. Monitoring and Reflecting

The OECD framework for CPS also includes three further elements:

1. Three conceptual dimensions for the assessment of problem solving. These are the problem context, the nature of the problem situation, and the problem solving process;
2. Two aspects of the problem solving context: the setting (whether or not it is based on technology) and the focus (whether it is personal or social);
3. Two problem presentation types: static problem situations in which the information about the problem situation is complete, and interactive problem situations, where it is necessary for the problem solver to explore the problem situation in order to obtain additional information.

These additional elements highlight the complexity of CPS activities and are pulled together in Figure 2 below.

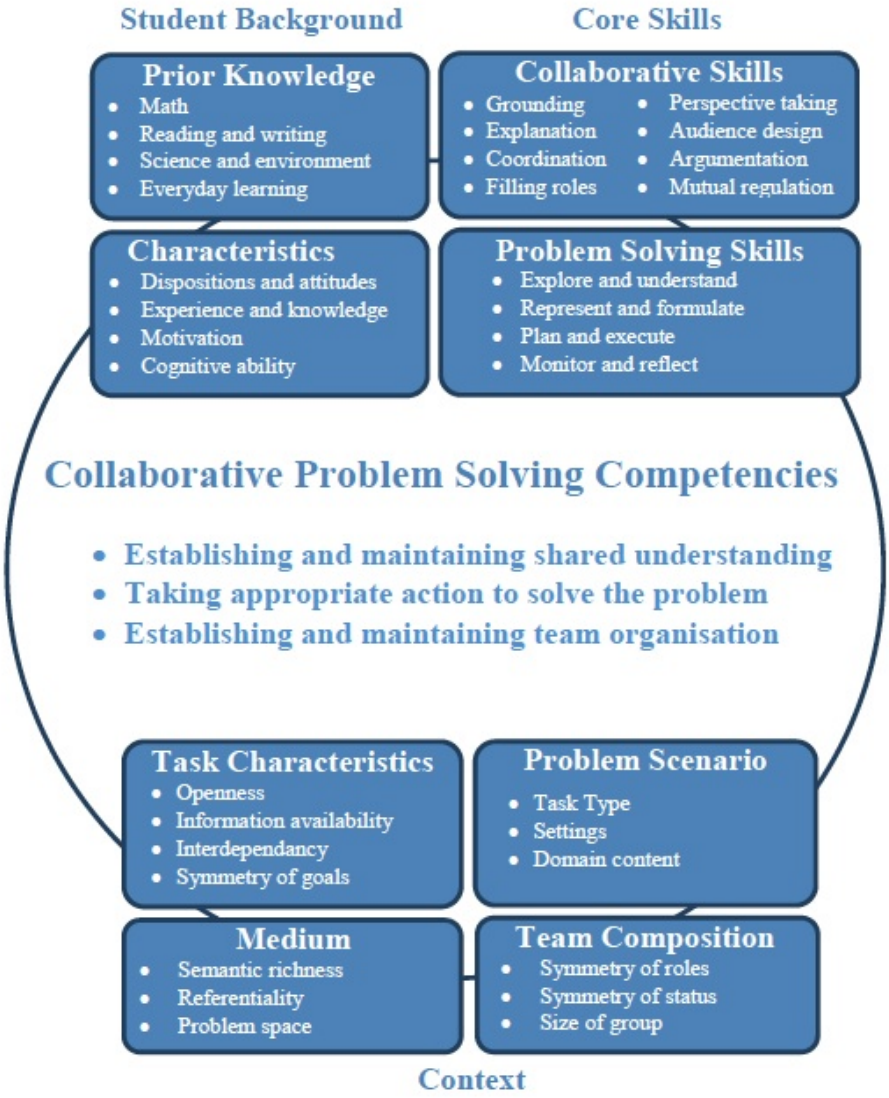



**Fig. 2.** Overview of factors and processes for Collaborative Problem Solving in PISA 2015

In addition to this overview the four problem solving processes and the three major collaborative problem solving competencies are merged to form a matrix of specific

skills, see Table 1. In the resulting matrix, the skills have associated actions, processes, and strategies. These specify what it means for the student to be competent.

| | (1) Establishing and maintaining shared understanding | (2) Taking appropriate action to solve the problem | (3) Establishing and maintaining team organisation |
|---|---|---|---|
| **(A) Exploring and Understanding** | (A1) Discovering perspectives and abilities of team members | (A2) Discovering the type of collaborative interaction to solve the problem, along with goals | (A3) Understanding roles to solve problem |
| **(B) Representing and Formulating** | (B1) Building a shared representation and negotiating the meaning of the problem (common ground) | (B2) Identifying and describing tasks to be completed | (B3) Describe roles and team organisation (communication protocol/rules of engagement) |
| **(C) Planning and Executing** | (C1) Communicating with team members about the actions to be/ being performed | (C2) Enacting plans | (C3) Following rules of engagement, (e.g., prompting other team members to perform their tasks.) |
| **(D) Monitoring and Reflecting** | (D1) Monitoring and repairing the shared understanding | (D2) Monitoring results of actions and evaluating success in solving the problem | D3) Monitoring, providing feedback and adapting the team organisation and roles |

**Table 1.** Matrix of Collaborative Problem Solving skills for PISA 2015

## Learning from Problem Based and Collaborative Problem Solving

The type of matrix in Fig. 1 has the potential for use when analyzing data of collaborative activity, but for a PBL approach, the missing component of knowledge deficiency requires attention. In Table 2, we add the PBL tutorial stages to the matrix to address this limitation. In this way we combine a tuition model with an evaluation model and in so doing address both aspects of the teaching learning process.

| | (1) Establishing and maintaining shared understanding | (2) Taking appropriate action to solve the problem | (3) Establishing and maintaining team organisation |
|---|---|---|---|
| **(A) Identifying facts** | (A1) Discovering perspectives and abilities of team members, making knowledge explicit | (A2) Discovering the type of collaborative interaction to solve the problem, along with goals | (A3) Understanding roles to solve problem |
| **(B) Representing and Formulating** | (B1) Building a shared representation and negotiating the meaning of the problem (common ground) | (B2) Identifying and describing tasks to be completed | (B3) Describe roles and team organisation (communication protocol/rules of engagement) |
| **(C) Generating Hypotheses** | (C1) Critically analysing the problem representation | (C2) Generating and Communicating potential solution paths | (C3) Present Hypothesis, encourage feedback from others and offer feedback on others' hypotheses |
| **(D) Planning and Executing** | (D1) Communicating with team members about the actions to be/ being performed | (D2) Enacting plans | (D3) Following rules of engagement, (e.g., prompting other team members to perform their tasks.) |
| **(E) Identifying Knowledge and Skill Deficiencies** | (E1) Comparing the team's knowledge and skills with the proposed actions | (E2) Identifying and specifying individual deficiencies | (E3) Identifying and specifying team deficiencies |
| **(F) Monitoring, Reflecting and Applying** | (F1) Monitoring and repairing the shared understanding | (F2) Monitoring results of actions and evaluating success in solving the problem | (F3) Monitoring, providing feedback and adapting the team organisation and roles |

**Table 2.** Combined Matrix that merges PBL and CPS concepts adapted from PISA 2015

Each of the 18 cells can be associated with different levels of learner proficiency. For example;

**Low** — the student responds to or generates information that has little relevance to the task.

**Medium** — the student responds to most requests for information and prompts for action, and generally selects actions that contribute to achieving group goals.

**High** — the student responds to requests for information and prompts for action, and selects actions that contribute to achieving group goals (OECD, 2013).

The contents of the cells C1 to C3 and E1 to E3 have been generated by the authors informed by Hmelo-Silver (2004).

## Final Remarks and Further Research

Frameworks such as this offer a flexible approach to the analysis of data collected from project based learning scenarios. This analysis may be that completed by humans as we strive to understand whether and how learning happens, but could it also be useful for data collected and analysed by machine? It needs to be acknowledged that PBL activity may not be captured completely through technology and that there will be aspects of the activity that take place away from any current technology. It may therefore be necessary for any analytics to use a combination of human and machine generated data. Our next steps are to test the framework empirically with a project based data set and to consider what appropriate learning analytic requirements might be extracted. At the workshop we will bring some examples of data and associated analysis to support further discussion of the framework.

## Acknowledgements

This work is co-funded by the European Union under the Practice-based Experiential Learning Analytics Research And Support (PELARS) STREP Project of the FP7-ICT-2013-11 Objective ICT-2013.8.2 Technology-enhanced Learnin #619738. See http://www.pelars-project.eu/

## References

1. Barrows, H. S., and Tamblyn, R. (1980). Problem-Based Learning: An Approach to Medical Education, Springer, New York.
2. Dewey, J. (1938). Experience and Education, Macmillan, New York.
3. Hmelo-Silver, C. E. (2004) Problem-Based Learning: What and How Do Students Learn? Educational Psychology Review, Vol. 16, No. 3.
4. OECD (2013) Draft Collaborative Problem Solving Framework http://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf

# Robust student knowledge: Adapting to individual student needs as they explore the concepts and practice the procedures of fractions

Claudia Mazziotti[1], Wayne Holmes[2], Michael Wiedmann[1], Katharina Loibl[3], Nikol Rummel[1], Manolis Mavrikis[2], Alice Hansen[2], Beate Grawemeyer[4]

[1] Ruhr-University Bochum, Germany
{claudia.mazziotti,michael.wiedmann,nikol.rummel}@rub.de
[2] UCL Institute of Education, University College London, UK
{w.holmes, m.marvrikis, a.hansen}@lkl.ac.uk
[3]University of Education Freiburg, Germany
loibl@ph-freiburg.de
[4]Birkbeck College, University of London, UK
beate@dcs.bbk.ac.uk

**Abstract.**
Robust knowledge consists of both conceptual *and* procedural knowledge. In order to address both types of knowledge, offering students opportunities to explore target concepts in an exploratory learning environment (ELE) is insufficient. Instead, we need to combine exploratory learning environments, to support students acquisition of conceptual knowledge, with more structured learning environments that allow students to practice problem-solving procedures step-by-step, to support students' acquisition of procedural knowledge. However, how best to combine both kinds of learning environments and thus both types of learning activities is an open question. We have developed a pedagogical intervention model that selects and sequences learning activities, exploratory learning activities and structured practice activities, that are appropriate for the individual learner. Technically, our intervention model is implemented as a rule-based system in a learning platform about fractions. The model's decision-making process relies on the detection of each individual student's level of challenge (i.e. whether they were under-, appropriately or over-challenged by the previous learning activity). Thus, our model adapts flexibly to each individual student's needs and provides them with a unique sequence of learning activities. Our formative evaluation trials suggest that single components of the intervention model, such as the ELE, mostly achieve their aims. The interplay between the different components of the intervention model (i.e. the outcomes of sequencing and selecting exploratory and structured practice activities) is currently being evaluated.

# 1 Introduction

Exploratory Learning Environments (ELEs), that include intelligent support, facilitate constructivist learning by offering opportunities for student self-determined exploration of a virtual environment [1]. The exploration of an ELE allows for sense-making activities which in turn promote the student's conceptual knowledge [2]. However, when integrating ELEs into the classroom, conceptual knowledge alone is insufficient. We need to move beyond this and enable students to achieve *robust* knowledge. Robust knowledge is deep, connected and comprehensive knowledge about a domain that lasts over time, accelerates future learning, transfers easily to new situations and is thus a very desirable learning goal [2–4]. It consists of both conceptual knowledge (understanding 'why') *and* procedural knowledge (knowing 'how') [5]. Thus, in addition to exploratory learning opportunities, we also need to provide students with learning opportunities that foster procedural knowledge [5] – opportunities for practicing problem-solving procedures, in structured learning environments such as that offered by some Intelligent Tutoring Systems (ITSs) [2] [6].

While prior work in the learning sciences and educational technology has mostly focused on fostering *either* procedural knowledge with structured practice activities (SPA) within ITSs *or* conceptual knowledge with exploratory learning activities (ELA) within ELEs, we aim to extend the existing literature by combining both types of learning activities – exploratory and structured – in order to help students acquire robust knowledge. This novel approach, combining both types of learning activities in one learning environment, also exploits the fact that conceptual and procedural knowledge evolve both iteratively and simultaneously [5].

Here, we report on a pedagogical intervention model (Figure 1), that specifies how to intelligently combine and sequence both ELA and SPA in order to promote complete robust knowledge. In doing so, we followed a theory and a data driven approach and thus iteratively improved our pedagogical model [7]. For example, our pedagogical intervention model builds on the cognitive psychology literature and, as such, is domain-neutral and thus transferable to other domains. However, as learning always depends on a target domain, the model also builds on previous work in the field of mathematics education, particularly fractions learning. The intervention model focuses on the individual student's level of challenge (categorized as either under-, appropriately or over-challenged) and selects the next learning activity accordingly. The model further specifies when students should receive cognitive support, so called task-dependent-support (TDS) , and emotional support, so called task-independent-support (TIS) [8]. The technical implementation of the intervention model is based on a rule-based system that, in order to determine each individual student's level of challenge, evaluates various input indicators (for example the student's response to the activity and the amount of feedback the system has provided).

A speech-enabled learning platform about fractions represents our intervention model and is embedded in the larger context of the 7$^{th}$ grant European research project "iTalk2Learn". In the following sections, we explain the rationale behind the intervention model in more detail, in particular describing how we determine each student's level of challenge, and we finish by discussing future steps.

## 2 The pedagogical intervention model

When combining ELA and SPA, the first question we have to address is which should come first? We argue that students should first start with an ELA rather than an SPA. The benefits of beginning with an ELA are evident in findings from Kapur [9]. He was able to show that students who started with an ill-structured task (*cf.* ELA) and continued with a well-structured task (*cf.* SPA) gained significantly more conceptual knowledge than students learning in the reverse order. This research was extended by Kapur in his work on Productive Failure [10] which replicated the finding that exploring concepts first fosters conceptual knowledge without hampering the acquisition of procedural knowledge. The choice to start with an ELA was also rooted in a domain-specific reason. From more than 20 years of research, the Rational Number Project [11] elicited four essential beliefs about how best to support students learning fractions [12]. One of these essential beliefs is that "teaching materials for fractions should focus on the development of conceptual knowledge prior to formal work with symbols and algorithms" [13].

The next question to be addressed when combining ELAs and SPAs is what activity comes after the initial ELA? The answer depends on the individual student's level of challenge. Students who are over-challenged with the initial ELA should continue with another less challenging ELA, in order to prevent them applying rules without prior reasoning [14]. On the other hand, students who are under-challenged should be given a more challenging ELA, in order to extend their learning. Finally, for students who are appropriately challenged by the ELA, switching from the exploratory to a structured activity is useful because the acquisition of conceptual and procedural knowledge mutually depend upon each other: changes in one type of knowledge lead to changes in the other type of knowledge which in turn lead to changes in the first type [5]. For example, when a student is appropriately challenged by an ELA, an SPA that is mapped to the ELA allows the student to elaborate and consolidate the conceptual knowledge that was acquired during the ELA.

A third question to be addressed is once a student has engaged with a SPA, what activity comes next? In light of ACT-R theory [15] and the power law of practice [16] students should be provided with more than a single SPA because they need sufficient practice in order to become fluent in the application of a problem-solving procedure. Accordingly, the student should engage with more than a single SPA. In addition to providing students with opportunities to become fluent with a given procedure, we also aim to facilitate students' flexible retrieval of different procedures by providing them with interleaved practice of SPAs, rather than simple blocked practice [17, 18]. However, once the student has become fluent with a given procedure, then additional practice does not lead to better learning [17]. Therefore, students are switched back to the ELE. In this way, the student starts a new learning cycle, which (in the context of our project) is embedded in a particular coarse grain goal of fractions learning (e.g. equivalence of fractions). Here again, depending on the student's level of challenge, the new learning cycle focuses either on the same coarse grain goal, and thus provides the student with additional learning opportunities for that goal, or moves to another coarse grain goal (e.g. adding fractions).

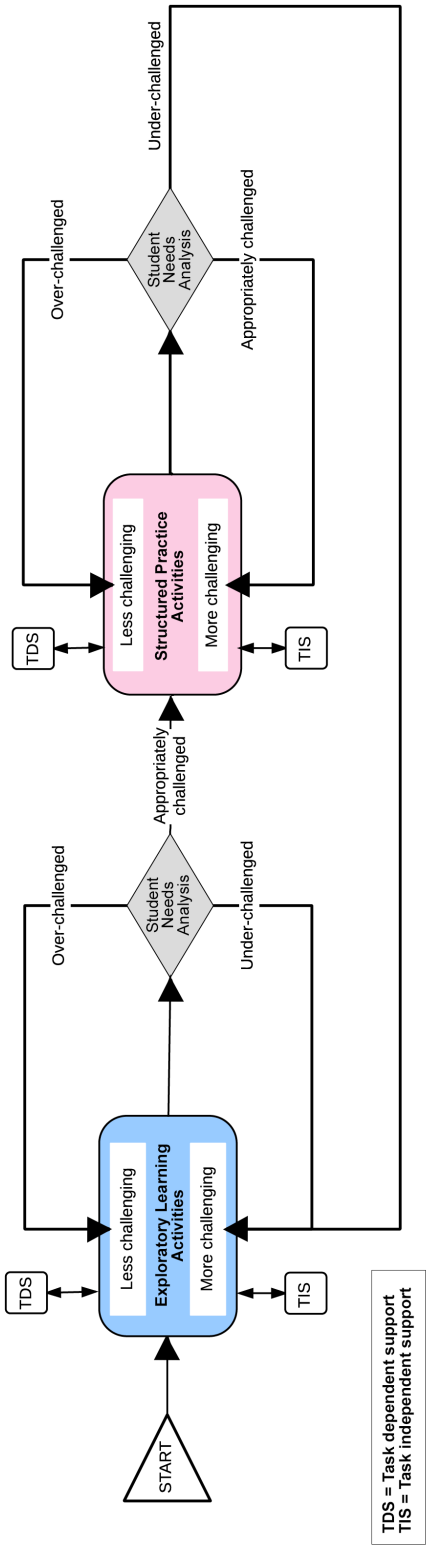



Figure 1: The pedagogical intervention model.

## 3 Determining a student's level of challenge

Determining a student's current level of challenge is a complex affair, because it is a function of characteristics both of the student *and* of the activity. For example, an ELA is likely to be less challenging for a student with high prior knowledge than for another student with low prior knowledge. Based on our pedagogical intervention model and a student model (i.e. considering the various input variables) the analytical engine (that we call the Students Needs Analysis or SNA) determines the student's level of challenge and thus the learner's appropriate next activity (i.e. output decision). For example, the SNA draws on the student's response to previous activities and to the current activity (using as a proxy the amount of task-dependent support, TDS [19], and the amount of task-independent support, TIS [8], delivered by the system), and the affective state inferred from the student's speech. Combining all these various inputs, each of which is assigned a weighting based on expert pedagogy, provides the SNA with a level of redundancy: a decision about the next appropriate activity can still be reached even if one of the inputs does not give any useful information or gives contradictory information.

### 3.1 Student Needs Analysis for exploratory learning activities

After each ELA, the SNA determines whether the student was under-, appropriately or over-challenged, based on the following input variables:

- the student's response to the current activity (using as a proxy the amount of TDS and TIS delivered by the system);
- the student's affect state inferred from prosodic cues in the student's speech;
- the student's affect state inferred from their screen and mouse behavior.

Based on these data, the SNA makes an output decision, selecting the next activity that is appropriate for the learner. If, for example, the system has had to deliver a large amount of TDS and the student's affective state has been calculated as *frustrated*, the SNA will determine that the student was over-challenged by the ELA and will sequence to a less challenging ELA. If, on the other hand, few TDS prompts have been delivered and the student's affect is inferred from speech to be *bored*, the SNA will determine that the student was under-challenged by the ELA and will sequence to a more challenging ELA.

Finally, if the SNA infers the student is appropriately challenged (for example, if there has been a minimal number of TDS and the affect has been categorized as *enjoyment*), the SNA switches to the structured practice environment. To ensure that students are provided opportunities to build upon and consolidate their conceptual knowledge, by applying it during structured practice, the SPA are mapped as closely as possible to the just-explored ELA. The close mapping of activities also aims to keep the individual student in their zone of proximal development, that is "the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving

under adult [or an Intelligent Tutor's] guidance, or in collaboration with more capable peers" [20].

### 3.2 Student Needs Analysis for structured practice activities

After students have completed a SPA, the SNA determines what the next activity should be based on the following input indicators (a future implementation will also take account of the number of SPAs the student has completed and the time taken):

- performance prediction, based on a machine-learning model that uses a student's past activity performance to predict future activity performance [21];
- the student's affect state inferred from prosodic cues;
- the TIS previously delivered.

Here, again, the SNA determines whether the student is under-, appropriately or over-challenged. If the SNA detects that a student was over-challenged by a SPA and the student's affect is categorized as *frustrated*, the SNA will deliver a less challenging SPA. By providing over-challenged students with a less challenging SPA we aim to enable the student to become fluent with a less challenging procedure, before re-exposing him to the more challenging procedure that they had not managed before. On the other hand, if the SNA detects that the student is appropriately challenged, he will be assigned a more challenging SPA. A machine-learning-based performance prediction model is used to determine how challenging activities are to the student. It takes into account data about the student's performance on previous tasks and data from other students working on these tasks from a historic dataset. Finally, if the SNA detects that the student is under-challenged, the SNA will switch back to the ELE and will assign a new ELA that is more challenging than the last ELA that they explored.

## 4 Summary and outlook

Our intervention model, currently implemented within the context of learning fractions, combines exploratory learning activities (ELA) with structured practice activities (SPA) according to each individual student's level of challenge, in order to achieve robust knowledge. In addition to the adaptive selection of the next activity, our intervention model also provides adaptive support in the form of TDS and TIS during each learning activity. Accordingly, students are provided with both cognitive and emotional support as they learn about fractions. Although our intervention model evolved within the domain of fractions learning, it is transferable to other domains as the rationale behind the intervention model is domain-neutral.

Repeated formative evaluation trials across the UK and Germany have tested the effectiveness of all the separate components of the intervention model. For example, various Wizard-of-Oz studies have delivered first empirical evidence that our ELE and its TDS supports students' exploratory behavior and fosters their conceptual understanding of fractions. Meanwhile, the interplay between different components of

the intervention model is currently being evaluated. To test the effectiveness of the intervention model we have created different versions of our learning platform. For example, in two quasi-experimental studies in the UK and Germany, we are comparing a full version of the learning platform representing our intervention model with a version that is without the ELE (but has all the other components). We expect differential effects in terms of students' knowledge acquisition (full version, complete robust knowledge, vs. the version without the ELE, procedural knowledge only) and user experiences. The initial results of these evaluation studies will be presented during the AIED workshop.

Once the learning platform is evaluated we will intensify our effort to facilitate the use of the platform for teachers by providing guidelines about how best to prepare for students' interaction with the platform. Additionally, for when working with the platform in class, we aim to provide teachers with a tool (e.g., a teacher dashboard) which will allow them to monitor individual student's use of the learning platform [22]. A further promising approach would be to enable students to learn collaboratively with the platform, as collaborative learning might further support students exploratory behavior and hence additionally support students' learning. From a more technical perspective, our next step is to develop a Bayesian network that is able to predict more precisely the learner-appropriate next activity. However, this first requires the collection of training data for the network from our current rule-based implementation of the SNA.

**Acknowledgments.** The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 318051 - iTalk2Learn project. This publication reflects only the authors' views and the Union is not liable for any use that may be made of the information contained therein.

## References

1. Noss, R. & Hoyles, C. Windows on Mathematical Meanings: Learning Cultures and Computers. Dordrect: Kluwer Academic Publishers. (1996)
2. Koedinger, K.R., Corbett, A.T., Perfetti, C.: The knowledge-learning-instruction framework: bridging the science-practice chasm to enhance robust student learning. Cognitive science 36, 757–798 (2012)
3. Richey, J.E., Nokes-Malach, T.J.: Comparing Four Instructional Techniques for Promoting Robust Knowledge. Educ Psychol Rev 27, 181–218 (2015)
4. Koedinger, K.R., Aleven, V.: Exploring the Assistance Dilemma in Experiments with Cognitive Tutors. Educ Psychol Rev 19, 239–264 (2007)
5. Rittle-Johnson, B., Siegler, R.S., Alibali, M.W.: Developing conceptual understanding and procedural skill in mathematics: An iterative process. Journal of Educational Psychology 93, 346–362 (2001)

6. van Lehn, K.: The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. Educational Psychologist 4, 197–211 (2011)
7. Cobb, P., Confrey, J., diSessa, A., Lehrer, R., Schauble, L.: Design Experiments in Educational Research. Educational Researcher 32, 9–13 (2003)
8. Grawemeyer, B., Holmes, W., Gutiérrez-Santos, S., Hansen, A., Loibl, K., Mavrikis, M. and Brdiczka, O.: Light-Bulb Moment? Towards Adaptive Presentation of Feedback based on Students' Affective State // IUI '15. Proceedings of the 20th International Conference on Intelligent User Interfaces ; March 29 - April 1, 2015, Atlanta, Georgia, USA, 400–404 (2014)
9. Kapur, M.: Productive Failure. Cognition and Instruction 26, 379–424 (2008)
10. Kapur, M.: Productive failure in learning the concept of variance. Instr Sci 40, 651–672 (2012)
11. Cramer, K., Behr, M., T., P., & Lesh, R. (1997). Rational Number Project: Fraction Lessons for the Middle Grades - Level 1. Dubuque Iowa: Kendall/Hunt Publishing Co. (1997)
12. Cramer, K.A., Post, T.R., delMas, R.C.: Initial Fraction Learning by Fourth- and Fifth-Grade Students: A Comparison of the Effects of Using Commercial Curricula with the Effects of Using the Rational Number Project Curriculum. Journal for Research in Mathematics Education 33, 111 (2002)
13. Cramer, K., & Henry, A. (2002). Using Manipulative Models to Build Number Sense for Addition of Fractions. In B.H. Littwiller & G.W. Bright (Eds.), Making Sense of Fractions, Ratios and Proportions: 2002 Yearbook (pp. 41-48). Reston, VA: National Council of Teachers of Mathematics. (2002)
14. Skemp, R. (1976). Relational Understanding and Instrumental Understanding Mathematics Teaching, 77, 20–26. (1976)
15. Anderson, J.R., Lebiere, C.: The atomic components of thought. Lawrence Erlbaum Associates., Mahwah, NJ (1998)
16. Newell, A.Hillsdale, NJ: Erlbaum & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), Cognitive skills and their acquisition (pp. 1-55). (1981)
17. Rohrer, D., Taylor, K.: The effects of overlearning and distributed practise on the retention of mathematics knowledge. Appl. Cognit. Psychol. 20, 1209–1224 (2006)
18. Rau, M.A., Aleven, V., Rummel, N.: Interleaved practice in multi-dimensional learning tasks: Which dimension should we interleave? Learning and Instruction 23, 98–114 (2013)
19. Holmes, W., Mavrikis, M. Hansen, A., Grawemeyer, B.: Purpose and Level of Feedback in an Exploratory Learning Environment for Fractions. In: Proceedings of AIED (2015)
20. Vygotskiĭ, L.S., Cole, M.: Mind in society. The development of higher psychological processes. Harvard University Press, Cambridge (1978)
21. Schatten, C., Schmidt-Thieme, L.: Adaptive Content Sequencing without Domain Information. In: Proceedings of the 6th International Conference on Computer Supported Education (www.csedu.org), Springer (2014)

22. Dillenbourg, P., Jerman, P.: Technology for classroom orchestration. In: Khine, M.S. , Saleh, I.M. (ed.) New science of learning: Cognition, computers and collaboration in education, pp. 525–552. Springer (2010)

# Adapting Collaboratively by Ranking Solution Difficulty: an Appraisal of the Teacher-Learner Dynamics in an Exploratory Environment

Rômulo C. Silva[1,2], Alexandre I. Direne[2], Diego Marczal[3], Paulo R. B. Guimarães[2], Ângelo S. Cabral[2], and Bruno F. Camargo[2]

[1] Western University of Paraná (UNIOESTE)
[2] Federal University of Paraná
[3] Federal Technological University of Paraná
{romulocesarsilva,dmarczal,guimaraes.prb,angeloscabral}@gmail.com
alexd@inf.ufpr.br
brunofilla_camargo@hotmail.com

**Abstract.** The work approaches theoretical and implementation issues of a framework aimed at supporting human knowledge acquisition of mathematical concepts. We argue that the problem solving tasks to be carried out by a learner should be ordered according to the matching of two parameters: (1) human skill level and (2) solution difficulty. Both are formally defined here as algebraic expressions based on fundamental principles derived from extensive consultations with experts in pedagogy and cognition. Our general definition of skill level is a rating-based measure that resembles the ones of game mastery scales. Likewise, the solution difficulty includes valuations based on a calibration method that computes mistakes and successes of learners' attempts to deal with the problem. The framework is instantiated by implemented software tools for the domain of logarithmic properties. Finally, we draw conclusions about the suitability of the claims based on a four-highschool-class experiment.

**Keywords:** rating, exercises calibration, Intelligent Tutoring Systems

## 1 Introduction

The student's expertise is usually developed by solving exercises that require a set of assessed skills. This is done in both conventional education schools and when applying advanced learning technologies, such as Intelligent Tutoring Systems (ITS). Normally, human teachers detect students' misconceptions when marking tests and exercises. Depending on how much the answer of a question departs from its correct version, two students that missed the same question could be scored different grades for that specific question.

Another aspect that can be used to compose the score is how difficult the question is. The difficulty degree of a question can be measured by the number of students that have skipped or made a mistake in that question. Thus, a student

who finds the correct answer of a question that many missed, probably has more skills than others and the score should reflect that. Conversely, a student who makes a mistake in a question that many were successful to answer, might possess fewer skills. Therefore, when posing questions to a student, it's desirable that an ITS calibrates the difficulties of such questions properly in order to match them against the expertise level of the student.

The student models have become a key element in ITS, supporting the development of individual help and detecting off-task behaviour [1]. The more recent approaches of student displacement behaviour from what is expected are influenced by the other students' behaviour. In this sense, a larger sampling of learners should provide better automatic assessments of a specific learner.

In the construction of student models, an important issue is weather just one or multiple skills will be considered. Some of the proposed models are based on the IRT (Item Response Theory), which is a classical model in psicometrics that assumes that success in every item of a test is determinated by one ability, named $\theta$, referred to as latent trait.

Another desirable aspect in ITS is predicting or prospecting if a learner will be able to answer a question correctly or not before it is actually showed to him or her. This feature allows the exercises to be presented according to the student's skills or rating.

## 2 Literature Review

Champaign and Cohen propose an algorithm [3] for content sequencing that selects the appropriate learning object to present to a student, based on previous learning experiences of like-minded users. The granularity of sequencing is on the LO level, not exercises or issues. A limitation of the work is that the algorithm was validated only by using simulated students.

Ravi and Sosnovsky [14] propose a calibration method for solution difficulty in ITS based on applying data mining techniques to a student's interaction log. Using the classical bayesian Knowledge Tracing (KT) method [5], the probability that a student has acquired a skill is calculated on the basis of a tentative sequence of exercises for which the soluctions involve a given concept. The logged events are grouped by exercises and classified according to the student's skills. All the data generated by the process is then used to match the sigmoid curve of IRT to connect different students using the standard clustering algorithm k-means.

Schatten and Schmidt-Thieme [15] present the Vygotski Policy Sequencer (VPS), based on the concept of Zone of Proximal Development devised by Vygotski. In this approach, the matrix factorization, which is a method for predicting user rating, is combined with a sequencing policy. This is done in order to select at each time step the content according to the predicted score.

Clement *et al.* [4] propose two algorithms for the tutoring model of ITS. The first, named RiARiT (Right Activity at Right Time), is based on multi-arm bandit techniques [2] such that each activity involves different skills, referred to as

Knowledge Components (KCs). The student model is a generalization of the one used in the bayesian KT method, representing the student's competence level ($c_i$) by a Real number in the range [0..1]. Furthermore, a reward representing the learning progress is defined by the difference between required KC and $c_i$. The second algorithm, ZPDES (Zone of Proximal Development and Empirical Success) [4] is a modified version of RiARiT where the calculation of the reward is changed in order to remove the dependence of the student's estimated competence level. The reward becomes a measure of how the success rate is increasing, providing a more predictive choice of activities.

Guzmán and Conejo [10] propose a cognitive assessment model based on IRT for ITS that calibrates the items of a topic (or concept). The method of item calibration is based on the kernel smoothing statistical technique that requires a reduced number of prior students sessions compared to conventional methods. In their approach, each possible answer has a characteristic curve that expresses the probability that a student with a certain knowledge level will more than likely select this answer.

There are several works about rating prediction techniques. Desmarais *et al.* [7] presented a comparative study between different linear models of student skill based on matrix factorization, IRT model and the k-nearest-neighbours approach. The linear models based on matrix factorization make predictions using a subset of the observed performance data for each student to predict the remaining subset, and measure the prediction accuracy. For other works, see [9], [6] and [16].

## 3 Automatic Calculation of Rating

Rating systems are frequently used in games to measure the players skills and to rank them. Usually, the rating is a number in a range $[minRank, maxRank]$ such that it is very unlikely that a player falls on the extremes. Inspired by game rating systems and taking the performance of other learners, this study proposes Equation 1 to assess iteratively a student's ability.

The following guidelines were adopted: *(1)* each question is scored a difficulty degree with a value in the range [0..10] and the student is rated in the range [1..10] to express his or her expertise level in the subject matter; *(2)* the easier the question, the greater the likelyhood that students will answer it correctly (in this case, a student's rating should have just a small increase if he or she enters the correct answer and should have a large decrease in the case of failure); *(3)* students that are successful in the first attempt to solve a question are scored a higher increment in their expertise level compared to those who need several attempts; *(4)* skipped questions are considered wrong.

Consider Equation 1.The details of its parameters are as follows:

$$R_J^q = R_J^{q-1} + Ak_1\alpha(10 - \frac{9T_J^q}{T_{med}^q}) - Ek_2\beta \times 10\frac{T_J^q}{T_{med}^q} \quad (1)$$

– $R_J^q$: student $J$'s rating after answering question $q$;

- $R_J^{q-1}$: previous student $J$'s rating. $R_J^0 = 5.5$ (initial rating);
- $A = 1$ and $E = 0$ if the student is successful in answering $q$, otherwise $A = 0$ and $E = 1$;
- $T_J^q$: number of unsuccessful attempts of student $J$ to answer question $q$;
- $T_{med}^q$: median of wrong attempts on question $q$ during classroom time;
- $N_a^q$: number of students that were successful in answering question $q$;
- $N_e^q$: number of students that were unsuccessful in answering question $q$;
- $\alpha = \frac{1}{N_a^q}$: weight factor to increase rating;
- $\beta = \frac{1}{N_e^q}$: weight factor to decrease rating;
- $k_1$ and $k_2$: multiplier factors of rating increase and decrease, respectively, calculated by $k_1 = 1 - \frac{R_J^{q-1}}{10}$ and $k_2 = \frac{R_J^{q-1}-1}{10}$.

Furthermore, $10 - \frac{9T_J^q}{T_{med}^q}$ and $10\frac{T_J^q}{T_{med}^q}$ represent the score of student $J$ in question $q$ in case the answer is correct and incorrect, respectively. There is no limit to the number of attempts $T_J^q$ a student can make to answer a question. However, if there are more than 10 trials, then 10 is taken as the maximum value for calculation purposes. Factors $k_1$ and $k_2$ avoid results of the expression in Equation 1 to reach upper and lower bounds of the range [1..10].

Using only the number of attempts and considering that the student usually tries until he or she gets the correct answer, the difficulty degree of a question $q$ can be defined by Equation 2 and its parameters as follows:

$$D^q = \frac{\sum_{J=0}^{J=n} T_J^q}{N_e^q + N_a^q} \tag{2}$$

- $D^q$: difficulty degree of the question $q$ after an exercise session;
- $T_J^q$: number of unsuccessful attempts of student $J$ to answer question $q$. If the number of attempts is greater than 10 trials, then 10 is taken as $T_J^q$;
- $N_e^q$ and $N_a^q$ are the same as in Equation 1

## 4 The ADAPTFARMA environment

The ADAPTFARMA (Adaptive Authoring Tool for Remediation of errors with Mobile Learning) prototype software tool is a modified version of FARMA[12], an authoring shell for building mathematical learning objects. In ADAPTFARMA, a learning object (LO) consists of a sequence of exercises following their introduction. The introduction is the theoretical part of a LO where concepts are defined through text, images, sounds and videos. The ADAPTFARMA implementation was carried out aiming its use on the web, either through personal computers or mobile devices.

To build an introduction and its corresponding exercise statements, ADAPTFARMA offers a WYSIWYG (What you See Is What You Get) interface, similar to those of highly interactive word processors. The teacher defines the number of questions related to each exercise. For each question, the teacher-author must

set a reference solution, which is the correct response to the question. ADAPTFARMA allows arithmetic and algebraic expressions to be entered as the reference solution. Under the learner's functioning mode, the tool deals automatically with the equivalence between the learners response and the reference solution.

A feature of ADAPTFARMA is the capability of backtracking the teacher to the exact context in which the learner made a mistake. This gives the opportunity to the teacher to identify the wrong steps performed by the learner and, thus, deal with the causes of the error accordingly. In addition, ADAPTFARMA allows the teacher to view a learner's complete interaction with the tool in a chronological order, in the form of a timeline. The teacher can make a closer monitoring of problem solution from other classrooms, as long as system permission is given through the collaboration mechanisms.

Likewise, learners can backtrack to the context of any of their right or wrong answers in order to reflect about their own solution steps. Additionally, on the collaborative side, it is possible for the teacher to carry out a review of students' responses and then provide them with non-automatic feedback, which can be done by exchanging remote messages through the system.

## 5 Algorithm for Exercises Sequencing

An important aspect in ITS is how the exercises should be sequenced after they are calibrated in order to match them to the expertise level of the student. At the beginning, the system doesn't have any information about the student. We propose an algorithm for sequencing exercises to be shown in ascending order of difficulty, combined with a mechanism similar to numerical interpolation:

- a minimal sequence of exercises is defined such that always begins with the easiest exercise and finishes with the most difficult one;
- the intermediate level exercises in the minimal sequence are distributed evenly among the easiest and most difficult exercises such that the number of exercises is $\left\lceil \frac{n}{stepsize} \right\rceil$ where $n$ is the total of exercises and the *stepsize* refers to the number of exercises that may be skipped when the student is successful. The *stepsize* can be set by the LO's author;
- the exercises are presented in the minimal sequence order;
- the number of attempts is limited to the average number of attempts obtained in the calibration phase. When the number of attempts is exceeded, the next exercise presented to the student is of a mid range difficulty considering the last exercise correctly answered and the current one.

For example, consider a LO with 30 exercises in ascending order of difficulty $[e_1, e_2, ..., e_{30}]$ and $stepsize = 4$. The minimal sequence of exercises will be $min\_seq =< e_1, e_5, e_9, e_{13}, e_{17}, e_{21}, e_{25}, e_{29}, e_{30} >$, and the exercises will be presented to the student in that order at first. For example, if the student misses $e_9$ until the attempts are over, then $e_7$ (of mid range difficulty between $e_5$ e $e_9$) is presented. Unlike the calibration phase, the student cannot skip exercises and if he/she continually misses the correct answer, the presentation becomes sequential.

## 6 Experiment

In order to evaluate the learning effectiveness of the four sequencing strategies, we carried out an experiment with four different classes of highschool students, aging fifteen to seventeen. The same LO about logarithms was applied to all four classes. It was created with the ADAPTFARMA enviroment to include thirty exercises. For each class, the LO was applied with a different sequencing method to order the exercises as follows:

- class A: random sequencing method (RSM);
- class B: teacher-defined sequencing method (TSM);
- class C: difficulty-biased sequencing method (DSM), where the difficulty degree was calculated by Equation 2 using outcome data from the calibration phase of class A;
- Class D: adaptive sequencing method (ASM), using the algorithm described in the previous section

The same pre- and post-tests were applied to all four classes. Students who did not participate in any step have been excluded from the analysis, resulting 119 participants. For the RSM, TSM and DSM methods, there was no limit to the solution attempts while in ASM, the average of attempts in class A was used. The Shapiro-Wilk test was applied to all samples to check for normality. Because only the DSM data passed the normality test (p-value = 0.0827), the pairwise T Student test was applied to it (p-value = 0.532). For the other three, the choice was the Wilcoxon test in order to evaluate the individual sequencing methods. The p-value of RSM, TSM and ASM were 0.0007, $< 0.0001$ and 0.0037, respectively. All methods, except for DSM, had a significant increase in scores.

The ANOVA method was applied to the pre-test data that showed normality whereas the Kruskal-Wallis, to the others, both to the post-test and to the average difference between pre- and post-tests. The results indicate that there is no significant difference among the four classes in the pre-test scores (p-value = 0.2539). However, there is significant difference in the post-test scores (p-value = 0.00579) and in the average difference between pre- and post-tests scores (p-value = 0.0307), suggesting that RSM, TSM and ASM led to better student performance than DSM. Besides, student performances among the three (RSM, TSM and ASM) were similar. Surprisingly, RSM led to the best performance while DSM, to the worst. This contradicts quite a large proportion of literature research on pedagogic practice, machine-led [8] or otherwise, for developing problem solving skills. Some reasons might explain such a phenomenon:

- the problem-statement ordering is a relevant issue that should be whatched more carefully to verify the influence of tacit knowledge contained in the textual organization of the statement;
- the lack of significant differences between RSM, TSM and ASM is also supported by evidence based on past research findings [11, 13];

– the DSM may have connected some sort of subject matters that caused an increase in the cognitive load, resulting in problem solutions that diverted from the correct ones;
– although most students have participated in the experiment, only the scores of pre- and post-tests accounted for the final student score in the official school records.

## 7 Conclusion and Future Work

Usually the student's expertise is developed by solving exercises that require a set of assessed skills, including ITS. We proposed an automatic rating system that can be used as an additional tool to assess students. Depending on the number of attempts and the difficulty degree of a question, students can get different scores for the same question.

Also, we proposed an algorithm, referred as ASM, for sequencing exercises that uses difficulty degree combined with a mechanism similar to numerical interpolation. It composes the ADAPTFARMA environment, a web authoring tool with WYSIWYG interface for creating and executing LOs. Taking advantage of it is very easy to change the strategy for exercises sequencing, we carried out a four-highschool-class experiment to test different sequences strategies: RSM, TSM, DSM and ASM. Only DSM had not a significant increase in the students' scores and the RSM had the best performance, demonstrating that problem-statement ordering is a relevant issue that should be researched more carefully in the near future. The ASM had also better performance compared to DSM.

Future research concentrates in adding new features to FARMA in two ways. Firstly, we are working in a deeper approach to user adaptation that includes more dimensions than just the matching between problem difficulty and student skill. One such new feature will be a function for generating problem statements based on teacher-defined problem statement parameters. Secondly, on the interface side, more interaction modes will be available to improve collaboration tasks for monitoring student performance progress.

## References


1. Ryan S.J. D. Baker, Adam B. Goldstein, and Neil T. Heffernan. Detecting the Moment of Learning. *LNCS*, 6094(PART I):25–34, 2010. Springer-Verlag Berlin Heidelberg.
2. Sébastin Bubeck and Nicolò Cesa-Bianchi. Regret analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.
3. John Champaign and Robin Cohen. A Model for Content Sequencing in Intelligent Tutorign Systems Based on the Ecological Approach and Its Validation Through Simulated Students. pages 486–491. Association for the Advancement of Artificial Intelligence (AAAI), 2010.

4. Benjamin Clement, Didier Roy, and Pierre-Yves Oudeyer. Online Optimization of Teaching Sequences with Multi-Armed Bandits. In Pardos Z. Mavrikis M. McLaren B.M. Stamper, J., editor, *Proceedings of the 7th International Conference on Educational Data Mining*, pages 269–272, 2014.
5. Albert T. Corbett and John R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278, 1994.
6. Maunendra Sankar Desarkar and Sudeshna Sarkar. Rating prediction using preference relations based matrix factorization. In Kalina Yacef Eelco Herder and Stephan Weibelzahl, editors, *Workshop and Poster Proceedings of the 20th Conference on User Modeling, Adaptation, and Personalization (UMAP2012)*, volume 872. http://http://ceur-ws.org/, Jul 2012.
7. Michel C. Desmarais, Rhouma Naceur, and Behzad Beheshti. Linear models of student skills for static data. In Kalina Yacef Eelco Herder and Stephan Weibelzahl, editors, *Workshop and Poster Proceedings of the 20th Conference on User Modeling, Adaptation, and Personalization (UMAP2012)*, volume 872. http://http://ceur-ws.org/, Jul 2012.
8. Alexandre Direne. Authoring intelligent systems for teaching visual concepts. *International Journal of Artificial Intelligence in Education*, 1(4):3–14, 1990.
9. Lucas Drumond, Nguyen Thai-Nghe, Tomáš Horváth, and Lars Schmidt-Thieme. Factorization techniques for student performance classification and ranking. In Kalina Yacef Eelco Herder and Stephan Weibelzahl, editors, *Workshop and Poster Proceedings of the 20th Conference on User Modeling, Adaptation, and Personalization (UMAP2012)*, volume 872. http://http://ceur-ws.org/, Jul 2012.
10. Eduardo Guzmán and Ricardo Conejo. Towards efficient item calibration in adaptive testing. In Liliana Ardissono, Paul Brna, and Antonija Mitrovic, editors, *User Modeling 2005*, volume 3538 of *Lecture Notes in Computer Science*, pages 402–406. Springer Berlin Heidelberg, 2005.
11. N. Major and H. Reichgelt. COCA: A shell for intelligent tutoring systems. In *Proc. of the International Conference on Intelligent Tutoring Systems (ITS92)*, pages 523–530. Springer, 1992.
12. Diego Marczal and Alexandre Direne. Farma: Uma ferramenta de autoria para objetos de aprendizagem de conceitos matemáticos. In *Anais do Simpósio Brasileiro de Informática na Educação*, volume 23, 2012.
13. Antonija Mitrovic. An intelligent SQL tutor on the web. *International Journal of Artificial Intelligence in Education*, 13(3):173–197, 2003.
14. Gautham Adithya Ravi and Sergey Sosnovsky. Exercise difficulty Calibration Based on Student Log Mining. In F. Mödritscher, V. Luengo, E. Lai-Chong Law, and U. Hoppe, editors, *Proceedings of DAILE'13: Workshop on Data Analysis and Interpretation for Learning Environments*, Villard-de-Lans (France), Janeiro 2013.
15. Carlotta Schatten and Lars Schmidt-Thieme. Adaptive Content Sequencing without Domain Information. *6th International Conference on Computer based Education*, April 2014.
16. Avi Segal, Ziv Katzir, Kobi Gal, Guy Shani, and Bracha Shapira. EduRank: A Collaborative Filtering Approach to Personalization in E-learning. In Pardos Z. Mavrikis M. McLaren B.M. Stamper, J., editor, *Proceedings of the 7th International Conference on Educational Data Mining*, pages 68–75, 2014.

# Towards Using Coherence Analysis to Scaffold Students in Open-Ended Learning Environments

James. R. Segedy & Gautam Biswas

Institute of Software Integrated Systems, Department of Electrical Engineering and Computer Science, Vanderbilt University, 1025 16th Avenue South, Nashville, TN, 37212, U.S.A.
{james.segedy,john.s.kinnebrew,gautam.biswas}@vanderbilt.edu

**Abstract.** Scaffolding students in open-ended learning environments (OELEs) is a difficult challenge. The open-ended nature of OELEs allows students to simultaneously pursue, modify, and abandon any of a large number of both short-term and long-term approaches to completing their tasks. To overcome these challenges, we have recently developed *coherence analysis*, which focuses on students' ability to *interpret* and *apply* the information available in the OELE. This approach has yielded valuable dividends: by characterizing students according to the coherence of their behavior, teachers and researchers have access to easily-calculated, intuitive, and actionable measures of the *quality* of students' problem-solving processes. The next step in this line of research is to develop a framework for using coherence analysis to adaptively scaffold students in OELEs. In this paper, we present our initial ideas for this work and propose guidelines for the construction of a scaffolding framework.

**Keywords:** Open-ended learning environments, metacognition, coherence analysis, scaffold

## 1 Introduction

Open-ended computer-based learning environments (OELEs) [1-2] are learner-centered; they present students with a challenging problem-solving task, information resources, and tools for completing the task. Students must use the resources and tools to construct and verify problem solutions, and in this process learn about the problem domain and develop their general problem-solving abilities. In OELEs, students have to distribute their time and effort between exploring and organizing their knowledge, creating and testing hypotheses, and using their learned knowledge to create solutions. Since there are no prescribed solution steps, students may have to discover the solution process over several hours. For example, learners may be given the following:

> *Use the provided simulation software to investigate which properties relate to the distance that a ball will travel when rolled down a ramp, and then use what you learn to design a wheelchair ramp for a community center.*

Whereas OELEs support a constructivist approach to learning, they also place significant cognitive demands on learners. To solve problems, students must simultaneously wrestle with their emerging understanding of complex topics, develop and utilize skills to support their learning, and employ *self-regulated learning* (SRL) processes to manage the open-ended nature of the task. SRL is a theory of learning that describes how learners actively set goals, create plans for achieving those goals, continually monitor their progress, and revise their plans when necessary to continue to make progress [3]. As such, OELEs can *prepare students for future learning* [4] by developing their ability to independently investigate and develop solutions for complex open-ended problems.

However, research with OELEs has produced mixed results. While some students with higher levels of prior knowledge and SRL skills show large learning gains as a result of using OELEs, many of their less capable counterparts experience significant confusion and frustration [5-7]. Research examining the activity patterns of those students indicates that they typically make ineffective, suboptimal learning choices when they independently work toward completing open-ended tasks [7-10].

The strong self-regulatory component of OELEs makes them an ideal environment for studying SRL. The open-ended nature of the environment forces students to make choices about how to proceed, and these choices reveal information about students' understanding of: (i) the problem domain; (ii) the problem-solving task; and (iii) strategies for solving the problem. By studying these choices, we can gain a better understanding of how students regulate their learning and how best to design scaffolds to support students who struggle to succeed.

Recently, we have introduced *coherence analysis* (CA) [11], a technique for studying students' problem-solving behaviors in OELEs. CA analyzes learners' behaviors in terms of their demonstrated ability to seek out, interpret, and apply information encountered while working in the OELE. By characterizing behaviors in this manner, CA provides insight into students' problem-solving strategies as well as the extent to which they understand the nuances of the learning and problem solving tasks they are currently completing.

In this paper, we present an overview of our findings with coherence analysis as applied to the *Betty's Brain* OELE (REF) and present our plans on extending this research. Our goal with CA is to empower both human and virtual tutors to more powerfully support students as they learn complex open-ended problem solving.

## 2 Betty's Brain

*Betty's Brain* [11] presents the task of teaching a virtual agent, Betty, about a science phenomenon (e.g., climate change) by constructing a causal map that represents that phenomenon as a set of entities connected by directed links representing causal relationships. Once taught, Betty can use the map to answer causal questions. The goal for students is to construct a causal map that matches an expert model of the domain.

In *Betty's Brain*, students acquire domain knowledge by reading resources that include descriptions of scientific processes (*e.g.*, shivering) and information pertaining

to each concept that appears in the expert map (*e.g.*, friction). As students read, they need to identify causal relations such as "*skeletal muscle contractions create friction in the body.*" Students can then apply this information by adding the entities to the map and creating a causal link between them (which "teaches" the information to Betty). Learners are provided with the list of concepts, and link definitions may be either increase (+) or decrease (-).

Learners can assess their causal map by asking Betty to answer questions and explain her answers. To answer questions, Betty applies qualitative reasoning to the causal map (*e.g.*, *the question said that the hypothalamus response increases. This causes skin contraction to increase. The increase in skin contraction causes…*). After Betty answers a question, learners can ask Mr. Davis, another pedagogical agent that serves as the student's mentor, to evaluate her answer. If Betty's answer and explanation match the expert model (*i.e.*, in answering the question, both maps utilize the same causal links), then Betty's answer is correct.

Learners can also have Betty take *quizzes* (by answering sets of questions). Quiz questions are selected dynamically by comparing Betty's current causal map to the expert map such that a portion of the chosen questions, in proportion to the completeness of the current map, will be answered correctly by Betty. The rest of her quiz answers will be incorrect or incomplete, helping the student identify areas for correction or further exploration. When Betty answers a question correctly, students know that the links she used to answer that question are correct. Otherwise, they know that at least one of the links she used to answer the question is incorrect. Students may keep track of correct links by annotating them as such.

## 3 Coherence Analysis

The Coherence Analysis (CA) approach analyzes learners' behaviors by combining information from sequences of student actions to produce measures of *action coherence*. CA interprets students' behaviors in terms of the information they encounter in the OELE and whether or not this information is utilized during subsequent actions. When students take actions that put them into contact with information that can help them improve their current solution, they have *generated potential* that should *motivate future actions*. The assumption is that if students can recognize relevant information in the resources and quiz results, then they should act on that information. If they do not act on information that they encountered previously, CA assumes that they did not recognize or understand the relevance of that information. This may stem from incomplete or incorrect understanding of the domain under study, the learning task, and/or strategies for completing the learning task. Additionally, when students add to or edit their problem solution when they have not encountered any information that could motivate that edit, CA assumes that they are guessing[1]. These two notions come together in the definition of action coherence:

[1] Students may be applying their prior knowledge, but the assumption is that they are novices to the domain and should verify their prior knowledge during learning.

*Two ordered actions ($x \rightarrow y$) taken by a student in an OELE are **action coherent** if the second action, $y$, is based on information generated by the first action, $x$. In this case, $x$ provides **support** for $y$, and $y$ is **supported** by $x$. Should a learner execute $x$ without subsequently executing $y$, the learner has created **unused potential** in relation to $y$. Note that actions $x$ and $y$ need not be consecutive.*

CA assumes that learners with higher levels of action coherence possess stronger metacognitive knowledge and task understanding. Thus, these learners will perform a larger proportion of supported actions and take advantage of a larger proportion of the potential that their actions generate. In the analyses performed to date, we have incorporated the following coherence relations:

- Accessing a resource page that discusses two concepts *provides support for* adding, removing, or editing a causal link that connects those concepts.
- Viewing assessment information (usually quiz results) that proves that a specific causal link is correct *provides support for* adding that causal link to the map (if not present) and annotating it as being correct (if not annotated).
- Viewing assessment information (usually quiz results) that proves that a specific causal link is incorrect *provides support for* deleting it from the map (if present).

Using these coherence relations, we derived six primary measures describing students' problem solving processes:

1. *Edit Frequency*: The number of causal link edits and annotations made by the student per minute on the system.
2. *Unsupported edit percentage*: the percentage of causal link edits and annotations not supported by information encountered within 5 minutes of the edit/annotation.
3. *Information viewing time*: the amount of time spent viewing either the science resources or Betty's graded answers. *Information viewing percentage* is the percentage of the student's time on the system classified as *information viewing time*.
4. *Potential generation time*: the amount of *information viewing time* spent viewing information that could support causal map edits that would improve the map. To calculate this, we annotated each hypertext resource page with information about the concepts and links discussed on that page. *Potential generation percentage* is the percentage of *information viewing time* classified as *potential generation time*.
5. *Used potential time*: the amount of *potential generation time* associated with information viewing that both occurs within a prior five minute window of and also supports an ensuing causal map edit. *Used potential percentage* is the percentage of *potential generation time* classified as *used potential time*.
6. *Disengaged time*: the sum of all periods of time, at least five minutes long, during which the student neither viewed a source of information for at least 30 seconds nor edited the map. *Disengaged percentage* is the percentage of the student's time on the system classified as disengaged time.

Metrics one and two capture the quantity and quality of a student's causal link edits and annotations, where supported edits and annotations are considered to be of higher quality. Metrics three, four, and five capture the quantity and quality of the student's time viewing either the resources or Betty's graded answers. These metrics speak to the student's ability to seek and identify information that may help them build or refine their map (potential generation percentage) and then utilize information from those pages in future map editing activities (used potential percentage). Metric 6 represents periods of time during which the learner is not measurably engaged with the system.

### 3.1 Summary of Findings with Coherence Analysis

Coherence analysis has proved to be a valuable tool for understanding how students learn as they solve open-ended problems. Thus far, we have investigated it with one group of 98 6$^{th}$-grade students (11 year olds). Thus, we interpret our findings with cautious optimism. We have identified the following relationships:

- *CA predicts learning and performance*: in general, students with higher levels of coherent behaviors have shown significantly higher levels of success in teaching Betty. Moreover, these learners have shown a better understanding of the science domain they were learning [11].
- *Prior skill levels predict CA*: students who were better able to identify causal links in abstract text passages (*e.g.*, A decrease in Ticks leads to an increase in Tacks) exhibited higher levels of coherence while using *Betty's Brain* [11].
- *CA identifies common problem solving profiles across students*: we clustered students by describing them with the six CA metrics described above, and we identified five common profiles among students: researchers and careful editors; strategic experimenters; confused guessers; disengaged students; engaged and efficient students. Interestingly, there were few differences in learning and performance among the clusters. Engaged and efficient students showed higher learning and performance than the other clusters, but there were not any other meaningful differences, suggesting that CA allows us to understand how different learning approaches lead to similar learning outcomes [11].
- *CA identifies common day-to-day problem solving profiles and transitions among them*: we clustered students as before, but this time the unit of analysis was a single day of using the system instead of the entire time using the system. We found a set of behavior profiles quite similar to those identified in the previous analysis. In analyzing day-to-day transitions, we found that many students performed fairly consistently while several other students performed inconsistently (that is, they have days of high coherence and days of low coherence). We also identified common transitions among days, which allowed us to find a potentially *at-risk* behavior profile. Students who behave like researchers and careful editors are far more likely than chance to transition to confused or disengaged behavior in subsequent days [12].

## 4 An Initial Coherence-Based Scaffolding Framework

Given the previous findings with CA, we aim to utilize the power of the analysis in real time as students use the system in order to detect non-coherent behavior, diagnose the cause of it, and take steps to support students in overcoming the difficulties they are experiencing. The core idea behind CA is that when students work in OELEs, they have two primary sets of tasks: *information seeking tasks* related to identifying and interpreting important information and *information application tasks* related to applying that information to improving the problem solution. All coherence metrics are based on identifying relationships between activities related to these two sets of tasks. By analyzing student behaviors with CA, we can identify problems related to information seeking and information application.

### 4.1 Diagnosing Problems with CA Metrics

The initial framework for diagnosing problems using CA metrics appears in Figure 1. This framework maps CA metrics to the problems they may indicate. For example, low levels of potential generation indicate that the learner is spending a large portion of their information viewing time on non-helpful information. This indicates that they may be struggling to identify relevant vs. non-relevant information in the environment. Problems with information seeking may also manifest as high levels of unused potential (*i.e.*, not applying viewed information), a high proportion of unsupported edits, and a low rate of editing the solution. Problems with information application are indicated by high unused potential and a low rate of editing the solution.

CA metrics may also be used to identify behaviors associated with effort avoidance. Specifically, low levels of information viewing, a low rate of editing the solution, a high unsupported edit percentage, and high levels of disengagement indicate that the learner may be purposefully avoiding effort. This may be due to a number of reasons, including low self-efficacy and low skill understandings.

Using this framework, our initial plan for using CA to scaffold students is as follows:

1. Observe the student for a period of time (*e.g.*, 10 minutes) and calculate their coherence metrics for that period. Identify any problematic behaviors (*e.g.*, high unused potential).
2. Form hypotheses about the sources of these behaviors. This involves looking at the combination of problematic behaviors observed and the student's previous activities in the system. For example, if the problematic behaviors are high unused potential and a low editing rate, the system may hypothesize that the student is struggling to apply information.

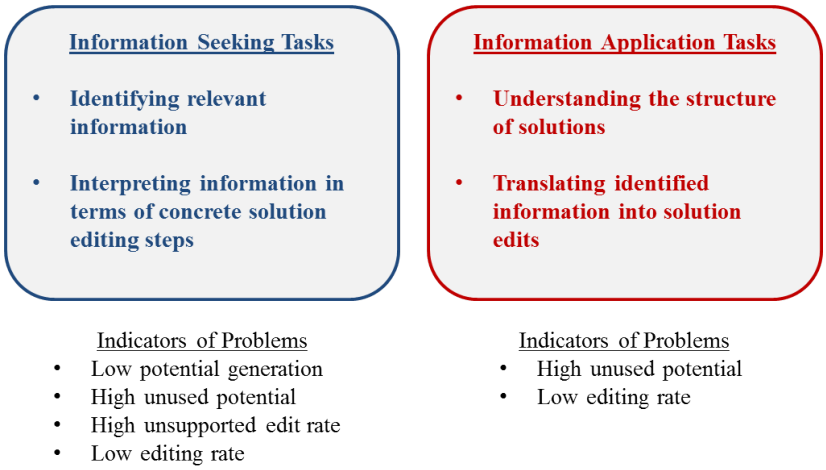



**Fig. 1.** Initial Problem Diagnosis Framework

3. Perform active diagnosis of the student to resolve competing hypotheses and gain additional information. For example, if the student has a high unsupported edit rate, this may be due to effort avoidance or a misunderstanding related to information seeking. The system can have the student answer questions and complete short problems in order to gain additional evidence as to which of these is the actual problem.
4. Once the system is confident that the student is struggling to understand something, it can use *guided practice scaffolds* [13] to help the student learn the knowledge and skills that they are missing or about which they are confused. Throughout guided practice, the system should provide encouragement, feedback, and scaffolding. It should also reinforce the relevance of the targeted knowledge and skills to the primary problem solving task, problem solving in general, and academic success.
5. If the system is confident that the student is exhibiting effort avoidance, then it should offer to help the student. If the behavior continues after the offer (and potential scaffolding related to that offer), then the system should provide guided practice scaffolds on the important knowledge and skills they need to understand to be successful. Hopefully, the student's abilities will improve during guided practice, and that will re-engage them with the learning task. As in the previous step, the system should provide the student with encouragement, feedback, and scaffolding and it should reinforce the relevance of the targeted knowledge and skills.

## 5 Conclusion

In this paper, we have provided an overview of *coherence analysis* (CA), an analysis approach that provides insight into how students behavior in open-ended computer-based learning environments (OELEs). Additionally, we have presented an initial

scaffolding framework that describes how CA might be leveraged to provide adaptive scaffolds to students who are struggling. As we move forward, we will continue developing this scaffolding framework, build it into *Betty's Brain*, and test its effectiveness with students.

**Acknowledgements**. This work has been supported by Institute of Education Sciences CASL Grant #R305A120186.

## References

1. Land, S., Hannafin, M., Oliver, K.: Student-centered learning environments: Foundations, assumptions and design. In: D. Jonassen & S. Land (eds.) Theoretical Foundations of Learning Environments, pp. 3-25. Routledge, New York, NY (2012)
2. Segedy, J.R., Biswas, G., Sulcer, B.: A model-based behavior analysis approach for open-ended environments. The Journal of Educational Technology & Society, 17(1), 272-282 (2014)
3. Zimmerman, B., Schunk, D. (eds.): Handbook of Self-Regulation of Learning and Performance. Routledge , New York, NY (2011)
4. Bransford, J., Schwartz, D.: Rethinking transfer: A simple proposal with multiple implications. Review of Research in Education, 24(1), 61-101 (1999)
5. Azevedo, R., Witherspoon, A.: Self-regulated learning with hypermedia. In: Hacker, D., Dunloskey, J., Graesser, A. (eds.), Handbook of Metacognition in Education, pp. 319-339. Taylor and Francis (2009)
6. Hacker, D., Dunloskey, J., Graesser, A. (eds.), Handbook of Metacognition in Education, pp. 319-339. Taylor and Francis (2009)
7. Kinnebrew, J.S., Loretz, K., Biswas, G.: A contextualized, differential sequence mining method to derive students' learning behavior patterns. Journal of Educational Data Mining, 5(1), 190-219 (2013)
8. Land, S.M.: Cognitive requirements for learning with open-ended learning environments. Educational Technology Research and Development, 48, 61-78 (2000)
9. Mayer, R.E.: Should there be a three-strikes rule against pure discovery learning?, American Psychologist, 59, 14-19 (2004)
10. Sabourin, J., Shores, L., Mott, B., Lester, J.: Understanding and predicting student self-regulated learning strategies in game-based environments. International Journal of Artificial Intelligence in Education, 23, 94-114 (2013)
11. Segedy, J.R., Kinnebrew, J.S., Biswas, G.: Using coherence analysis to characterize self-regulated learning behaviours in open-ended learning environments. Journal of Learning Analytics (in press)
12. Segedy, J.R., Biswas, G.: Coherence over time: Understanding day-to-day changes in students' open-ended problem solving behaviors. In: Proceedings of the 17th International Conference on Artificial Intelligence in Education (in press)
13. Segedy, J., Biswas, G., Blackstock, E., Jenkins, A.: Guided skill practice as an adaptive scaffolding strategy in open-ended learning environments. In: Proceedings of the 16th International Conference on Artificial Intelligence in Education. pp. 532-541. Springer (2013)

# Design Strategies for developing a Visual Platform for Physical Computing with Mobile Tools for Project Documentation and Reflection

Daniel Spikol[1], Nils Ehrenberg[1], David Cuartielles[2], and Janosch Zbick[3]

[1] Malmö University, Malmö, 211 19 Sweden
daniel.spikol@mah.se
[2] Arduino Verkstad, Malmö, 211 19 Sweden
[3] Linnæus University, Växjö, 351 95 Sweden

**Abstract.** This poster discusses work on the design of a visual-based programming language for physical computing and mobile tools for the learners to actively document and reflect on their projects. These are parts of a European project that is investigating how to generate, analyze, use and provide feedback from analytics derived from hands-on learning activities. Our aim is to raise a discussion about how learning analytics, intelligence, and the role of learners' documenting their work can provide richer opportunities for supporting learning and teaching.

**Keywords:** learning analytics, human factors and interface design, prototyping

## 1 Introduction

Educators, researchers, business leaders, and politicians are working to initiate new modes of education to provide 21st Century skills that focus on the following: creativity, innovation, critical thinking, problem solving, communication, and collaboration [4]. Recently, researchers and practitioners have provided strong cases for the value of hands-on activities like digital fabrication than could be part of the toolbox to bring powerful ideas, literacies, and expressive tools to learners [1]. This poster presents on-going work in the Practice-based Experiential Learning Analytics Research And Support project (PELARS) that aims to generate, analyze, use and provide feedback for analytics derived from hands-on these project-based learning activities. The focus of the PELARS project activities is on learning and making things with physical computing that provide learners with opportunities to build and experiment with tangible technologies and digital fabrication. One of the key research aims of the PELARS project can be summarised as: *How can physical learning environments that use hands-on digital fabrication technologies be better designed for ambient and active data collection for learning analytics?* The project addresses three different learning contexts (university interaction design, engineering courses, and high school science) across multiple settings in Europe. The goals of the project are first to

define learning (skills, knowledge, competencies) that is developing, and how we can assess it in the frame of learning analytics. Then to determine what elements of this learning we can capture by designing the physical environment and activities around digital fabrication technologies. Then to identify what patterns of data we collect can tell us about learning, collaboration and how the system can help support the learning activities.

The PELARS project approach has been to develop an intelligent system for collecting activity data (moving image-based and embedded sensing) for diverse learning analytics (data-mining, reasoning, visualisation) with active user-generated material from practice-based and experiential activities. This rich range of data is used to create learning analytics tools for learners and teachers that range from assessment to exploring intelligent tutoring. The PELARS system carries forwards the ideas of knowledge communities and inquiry [7] and provide conceptualising, representing, and analysing distributed interaction [8]. However, there are multiple challenges for designing learning analytics and intelligent support for these types of tangible activities. Learning situations in these contexts include open-ended projects, small group work, and the use of physical computing components that require construction and programming. Therefore, these types of activities present difficulties for collecting meaningful data for learning analytics.

This poster specifically discusses our work on the development of a visually based programming platform for the physical computing hardware and the mobile tools for the learners to actively document and reflect on their projects. These two parts of the PELARS project provide opportunities for discussion on the relationships between intelligent support, active learner engagement, and analytics. Our aim for the workshop is to raise a discussion about how learning analytics, intelligence, and the role of learner documenting their work can provide richer opportunities for supporting learning.

## 2 Methodological Approach

The PELARS project has a design-centric approach that includes the use of low-fidelity prototyping and "wizard of oz" scenarios [5]. These methods that include paper prototypes and technology sketches to investigate how to find the best way to get the design right [2]. The goal of the two cases below is to investigate how we can better understand the needs of the users. The need to develop a visually based programming experience to support students and supply data for analysis and lack of student documentation were identified as challenges through literature and own contextual user research in the project.

For the visual programming platform, a kit was created that contained foam core versions of hardware blocks with strings and pins to act as the cables to connect them. A small magnetic board with paper-based magnets acted as the computer screen that represented what blocks were connected. A set of simple tasks were provided to pairs of testers (recruited students from Interaction Design and Computer Science) while one of the researchers acted as the computer

in a "wizard of oz" scenario. Figure 1 illustrates how the students connected the hardware blocks of sensors and actuators (the paper blocks and strings) on the table and then the researcher put on a magnetic board (computer screen) the associated blocks (printed magnets). The researcher acted as the wizard representing the smart system that recognised which blocks the students had connected and represented the computer screen showing the visual programming interface. This prototyping system allowed the teams to discuss and adjust the inputs to generate the hypothetical outcome for the different tasks.

**Fig. 1.** Visual programming platform prototyping

For the mobile reporting part of the PELARS system we adopted a web-based system developed by colleagues [9] that allowed us to create a series of forms that could be accessed by students in an Interaction Design course where they have a 4 week block in physical computing. The students needed to fill in three forms, the first form asks them to briefly describe their plan for solving the task, the second form allows them to document their progress with text and photos, and the final form asks them to reflect on the outcome, did the project succeed as planned. Figure 2 shows the different screens of the mobile system. The intention of our prototyping effort has been to explore the similarities between practice-, problem-, and inquiry-based learning [3] and the challenges in student self-documentation practices in physical computing.

In addition to the forms, the students were also asked to complete a lightweight pre-survey and post-survey to evaluate the usability of the mobile documentation tool. The surveys were inspired by Read and MacFarlane's [6] work on surveys for children in computer interaction and designed to take a few minutes to fill out. The survey results were intended to supplement the submissions received through the mobile system. The pre-survey intended to cover their general ex-

perience with documenting their work and the post-survey their views on the usefulness of the tool. Additionally, the pairs of students were interviewed in a semi-structured after the prototyping session.

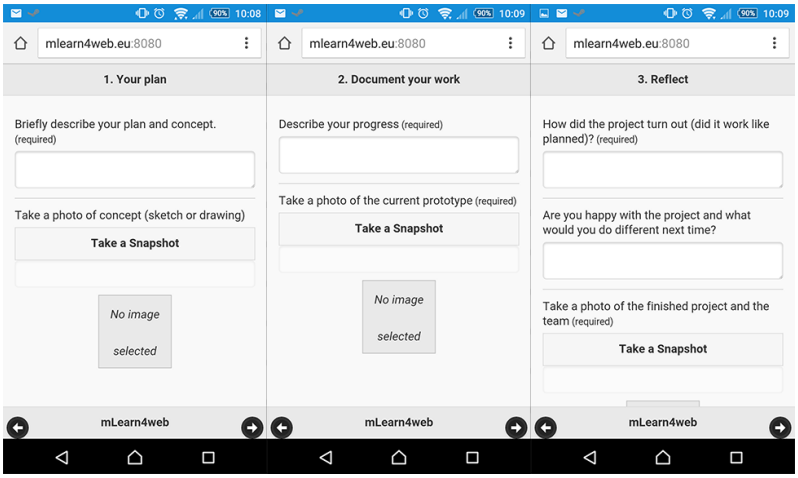



**Fig. 2.** Mobile system screen captures

## 3 Initial Results

### 3.1 Visual Programming

The initial results for the visual programming platform points towards the less experienced programmers finding the visual programming system easier for solving the different tasks. The less experienced students were more open to exploring how to solve the open-ended tasks. While the experienced programmers were frustrated by their perceived limitations of the system, for example not being able to code a loop statement to blink an LED. During the post activity interview, the experienced programmers did however see the system as useful both for learning programming but also for communicating ideas in a prototype stage. Importantly to note, that these perceptions may reflect that design students are more used to open-ended tasks and familiar with throw-away prototyping.

In some cases, the designers worked with more experienced programmers and in these cases communication between the team members helped the programmer shift metaphors to a more visual style of programming. After the initial tasks the more experienced programmers felt they had a better understanding of the concept. Additionally, in the follow-up interviews, they expressed that they liked the idea of visual coding, but primarily saw it as a teaching tool or a communication tool rather than something that they would use to build their projects.

### 3.2 Documentation Tools

The mobile tools initially seemed to have the right balance of short text entries and the uploading of rich media. The aim was to allow the students to plan easily, document and reflect via smartphones or laptops. Our initial findings suggest that the structure of planning, documenting the process and then reflecting on the project was utilised by the students. The students reported in the post-survey that it is easy to forget to document, to ignore it, or do it later. While the submitted documentation captured the students progress, it was also often submitted the day after or when they were finishing their work, rather than at the end of each session. Our thoughts for these results are that students faced the combination of not seeing the relevance of documenting the projects was important and not having practiced documenting their work.

Students reported in the post survey that the usability of the system needs to be improved. For example, they pointed out that the system did not let them go back to add, or amend their documentation. The need for better clarity what happens with the data after they submit it could help with the students. Connecting the documenting tools to their normal work practices and digital tools, like blogs or online portfolios need to be explored. Additionally while documenting some students appeared were frustrated when submitting as a group. The data shows that when students used a personal device they choose to submit individually. This suggests that the group submissions are useful, the students desire to submit individual reports as well.

## 4 Discussion

We feel that that the low-fidelity and sketching the technology for the PELARS project are important means to design better intelligent support while engaging with the needs of the different users. The PELARS project has been influenced by inquiry-science learning. However, the nature of making and solving problems with physical computing in interaction design courses can be more dynamic and open-ended than more traditional classwork. Prototyping both the programming interface and the documentation tool as parts of the same project, rather than as separate entities gives a broader design approach. This allows us to explore different aspects of the learning environment and test out ideas in pseudo-real world situations. One of the design goals is to support the visual programming activities with intelligent tutoring and means for teachers and students to analyse of time how they programmed and built the different projects. Additionally, the documentation tool provides a different perspective to the ambient data collection and a process framework for the learning activity. We feel that using these different design approaches provides us with a means to explore the complexity of project-based experiential learning scenarios.

**Acknowledgments.** The Project is a European Union Research Project for Small or Medium-scale Focused Research Project (STREP) part of the FP7-

ICT-2013-11 Objective ICT-2013.8.2 Technology-enhanced Learning. The Grant Agreement number is 619738. http://www.pelars-project.eu/.

## References

1. Blikstein, P.: Multimodal learning analytics. In: Proceedings of the third international conference on learning analytics and knowledge. pp. 102–106. ACM (2013)
2. Buxton, B.: Sketching user experiences: getting the design right and the right design (interactive technologies). Burlington, MA: Morgan Kaufmann (2007)
3. Hmelo-Silver, C.E.: Problem-based learning: What and how do students learn? Educational psychology review 16(3), 235–266 (2004)
4. Lai, E.R., Viering, M.: Assessing 21st century skills: Integrating research findings. In: annual meeting of the National Council on Measurement in Education, Vancouver, BC, Canada (2012)
5. Mavrikis, M., Dragon, T., McLaren, B.M.: The design of wizard-of-oz studies to support students learning to learn together. In: Proceedings of the International Workshop on Intelligent Support in Exploratory Environments: Exploring, Collaborating, and Learning Together (2012)
6. Read, J.C., MacFarlane, S.: Using the fun toolkit and other survey methods to gather opinions in child computer interaction. In: Proceedings of the 2006 conference on Interaction design and children. pp. 81–88. ACM (2006)
7. Slotta, J.D., Tissenbaum, M., Lui, M.: Orchestrating of complex inquiry: Three roles for learning analytics in a smart classroom infrastructure. In: Proceedings of the Third International Conference on Learning Analytics and Knowledge. pp. 270–274. ACM (2013)
8. Suthers, D.D., Dwyer, N., Medina, R., Vatrapu, R.: A framework for conceptualizing, representing, and analyzing distributed interaction. International Journal of Computer-Supported Collaborative Learning 5(1), 5–42 (2010)
9. Zbick, J., Jansen, M., Milrad, M.: Towards a web-based framework to support end-user programming of mobile learning activities. In: Advanced Learning Technologies (ICALT), 2014 IEEE 14th International Conference on. pp. 204–208. IEEE (2014)

# ISLG 2015
# Fourth Workshop on
# Intelligent Support for Learning in Groups

*held in conjunction with*

Seventeenth International Conference on
Artificial Intelligence in Education (AIED 2015)

Monday, June 22, 2015
Madrid, Spain

**Workshop Co-Chairs:**

Ilya Goldin[1] ,Roberto Martinez-Maldonado[2] ,Erin Walker[3] ,
Rohit Kumar[4] ,Jihie Kim[5]

[1]*Center for Digital Data, Analytics, & Adaptive Learning, USA*
[2]*School of Information Technologies, University of Sydney, Australia*
[3]*Computing, Informatics, and Decision Systems Engineering, Arizona State University, USA*
[4]*Raytheon BBN Technologies, USA*
[5]*Software R&D Center, Samsung Electronics, South Korea*

ilya.goldin@pearson.com, roberto@it.usyd.edu.au,
erin.a.walker@asu.edu, rkumar@bbn.com, jihie.kim@gmail.com

https://sites.google.com/site/aied2015islg

## Program Committee

Jessica Andrews, *Educational Testing Service, USA*
Ari Bader-Natal, *Minerva Project, USA*
Tiffany Barnes, *North Carolina State University, USA*
Sherice Clark, *University of Pittsburgh, USA*
Girlie C. Delacruz, *University of California, USA*
Yannis Dimitriadis, *University of Valladolid, Spain*
Toby Dragon, I*thaca College, USA*
Sharon I-Han Hsiao, *Arizona State University, USA*
Seiji Isotani, *Universidade de São Paulo, Brazil*
Charalampos Karagiannidis, *University of Thessaly, Greece*
James Lester, *North Carolina State University, USA*
Alejandra Martínez Monés, *University of Valladolid, Spain*
Bruce McLaren, *Carnegie Mellon University, USA*
Jennifer Olsen, *Carnegie Mellon University, USA*

## Table of Contents

# Preface

Technological advances in the use of artificial intelligence in education (AIED) over the past two decades have enabled the development of highly effective, deployable learning environments that support learners across a wide range of domains and age groups. Alongside, mass access to and adoption of modern communication technologies have made it possible to bridge learners and educators across spatiotemporal divides. Students can now collaborate using educational technology in ways that were not previously possible.

Intelligent tutoring systems seek to individualize each student's learning experience, but this need not imply a solitary experience. Research on computer-supported collaborative learning (CSCL) has revealed the pedagogical benefits of learning in groups, as well as how to structure the activity to lead to productive interactions. A variety of recent systems have demonstrated ways in which an adaptive learning environment can benefit from the presence of multiple learners. Similarly, students using CSCL systems have been shown to benefit from the introduction of adaptive support. It is of high relevance to the AIED community to explore how AI techniques can be used to support collaborative learning, and how theories of how students learn in groups can inform the design of adaptive educational technologies.

The goal of this series of workshops is to gather the sub-community of AIED researchers interested in intelligent support for learning in groups with learning scientists to share approaches and exchange information about adaptive intelligent collaborative learning support. We invite discussion on how the combination of collaborative and intelligent aspects of a system can benefit the learner by creating a more productive environment. Over the past few years, the AIED research community has started investigating extension of the fundamental techniques (student modeling, model-based tutors, integrated assessment, tutorial dialog, automated scaffolding, data mining, pedagogical agents, and so on) to support collaborative learning. We aim to explore ways that the current state of the art in intelligent support for learning in groups can be informed by learning sciences research on collaborative learning principles.

June, 2015
Ilya Goldin, Roberto Martinez-Maldonado,
Erin Walker, Rohit Kumar, and Jihie Kim

# Negotiating Individual Learner Models in Contexts of Peer Assessment and Group Learning

Susan Bull and Lamiya Al-Shanfari

Electronic, Electrical and Systems Engineering, University of Birmingham, UK
s.bull@bham.ac.uk, LSA339@bham.ac.uk

**Abstract.** This paper introduces learner model negotiation not only as a means to increase the accuracy of the learner model and promote metacognitive activities as in past examples, but also as a way to help learners correct peer assessment entries in their learner model, that they consider inaccurate. While open learner models are not new, and negotiated learner models have been developed before, in today's learning contexts of potentially big data from many sources including other learners, some kind of approach to managing the data as well as helping learners to understand and accept it, or correct it, is needed.

## 1. Introduction

Benefits of a range of approaches to learning in groups have been argued (e.g. [9]), and there is strong interest in the field of Artificial Intelligence in Education in developing useful support for group learning [18]. Peer assessment and feedback have also been advocated as beneficial to the learning process (e.g. [27],[30]). We introduce a negotiated open learner model (OLM) approach to supporting students in the peer assessment situations that are becoming more common in today's learning contexts.

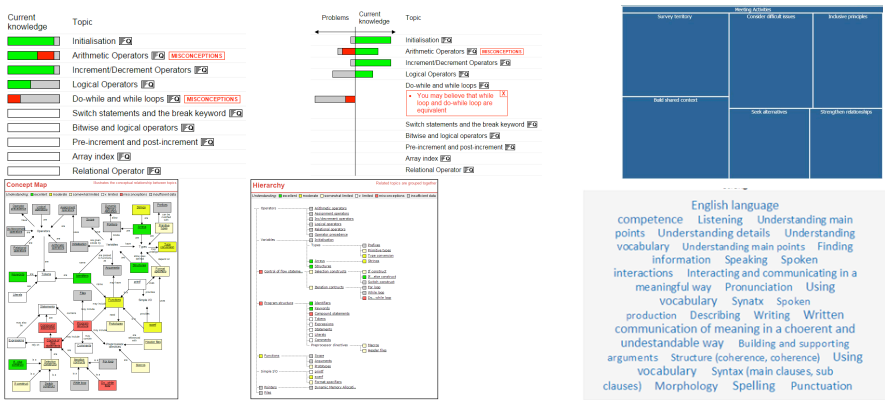



Fig. 1. Examples of open learner model visualisations

OLMs are learner models that are externalised to users in an understandable form, often to support collaboration or metacognitive behaviours [4]. Figure 1 gives OLM

visualisation examples of simple skill meters [2], structured concept map and hierarchical tree [20], and newer visualisation approaches of overview-zoom treemap and word cloud [3]. While OLMs to support group learning have been developed (e.g. [1],[2],[6],[28]), the range of activities a student may be engaged in will likely include individual activities. Thus, in this paper we reflect on individual learner models that may be used in a group context. We focus in particular on situations in which peer feedback or assessment contributes to the individual learner model, which may follow the production of an artefact for assessment, or participation in a group activity.

## 2. OLMs and Peer Assessment in Modern Learning Contexts

Learner modelling has broadened, now being found in contexts with rich collections of digital materials [14]. Recent advances in learner modelling have aimed to address the use of new technologies, e.g.: learner models holding diverse data from different sources [3],[7],[21],[22]); combining e-portfolios and viewable learner models [23]; and OLMs to help learners monitor progress and plan their learning in MOOCs [15].

Peer assessment has become more prevalent in modern learning contexts such as MOOCs [17],[25] and e-portfolios [12],[31]; as well as individual online systems that allow peer assessment and feedback to be given and received [19]. OLMs that include peer assessment and feedback have been proposed [11], and developed (see [3]) in the context of peer assessments (numerical, contributing to the learner modelling algorithm) alongside automated data from a variety of external applications, and feedback (non-interpreted text, to help explain the numerical value of a peer assessment to an assessee). However, although there are many learning benefits for both peer assessors and assessees, there can also be cases of motivation decreasing if a student considers a peer assessment to be unjust [17]; or a learner feels there to be a lack of effort/attention from a peer assessor [10]. Another issue that may cause concern is the outcome of group assessments where there has been unequal contribution from group members [24]. For example, a student who engaged minimally in a group activity or project may receive the same assessment as the other participants. Experiences such as the above can cause strong emotional responses, and a method for learners to either understand learner model representations originating from peer assessors, or to challenge them, would help to relieve this frustration. The solution should allow individuals to understand the reasons for peer assessments and the system's perspective on them, as well as justify why they believe these representations or reasons to be inappropriate. We address these problems in the context of the LEA's Box OLM, where a learner model negotiation mechanism is being developed (based on [5]).

## 3. Maintaining the Learner Model through Negotiation

Building on the Next-TELL OLM [3], the LEA's Box learner model data may originate from a range of applications. In some cases, activities may be completed away from any tracking software. To address the latter, teacher, self and peer assessments can be entered alongside automated assessments. However, these may themselves

differ in quality according to effort, experience and expertise of the assessor. While a learner may accept an automated assessment, or assessment by a teacher, they may be less happy with peer assessments and, indeed, may retain a negative attitude towards peer assessments over teacher assessments [13]. Even though a single peer assessment may ultimately contribute little to the value(s) in their learner model, this negative affective state may remain strong.

Some OLMs have allowed the learner model to be negotiated, where student and system have the same powers and negotiation moves [5],[8],[16]; or to be discussed in some other way, e.g. one partner has greater control over the discussion outcome [26],[29],[32]. Advantages of discussing or negotiating learner models include: the possibility to increase the accuracy of the learner model by allowing the learner to challenge the representations [5]; motivation may be increased by offering an alternative task [26]; significant learning gains may be achieved as a result of the negotiation process [16]. We here add a new benefit resulting from the inclusion of peer-entered data in an individual's learner model, for the increasing number of contexts in which multiple sources of data, including human contributions, are incorporated in the learner model. As well as increasing the accuracy of the learner model, individuals have the opportunity to redress any perceived injustices introduced by peer assessment.

Discussion of learner models typically involves moves such as agree/disagree; requesting information; challenging the other partner (learner or system); stating one's viewpoint; and justifying one's viewpoint by referring to evidence. For example:

- LEARNER: My value for [multiplying matrices] should be [higher].
- SYSTEM: Your last [five] attempts in [OLMlets] showed that you have [multiplied the corresponding items in each matrix]. You are adopting an approach used to [add matrices] when you are trying to [multiply matrices].
- LEARNER: I have since [participated in peer discussion].
- SYSTEM: Please [multiply] the following: ……

This example shows a student-initiated attempt to change a learner model representation (square brackets indicate variables). Having looked at their OLM, the learner notices that they disagree with the strength of the value for multiplying matrices. The learner has the misconception that the process mirrors that of adding matrices. The system has modelled this misconception, and is able to provide the evidence to back its inferred value – that the user's most recent attempts at this type of problem using the OLMlets system [2] illustrate the misconception. On receiving this evidence, the learner may accept the existing value or try to persuade the system further (in this example, stating that they have discussed this problem in their group). In this case the system offers a quick test to see whether it accepts the learner's claim as valid. Negotiation of learner models can take place using a variety of methods – menu-based selection of arguments [5]; buttons to initiate dialogue moves in dialogue games [8]; natural language discussion with a chatbot [16]. In most cases, even where negotiation takes place through a graphical interface, a text version is recorded for reference during the negotiation and afterwards. This record can also be kept as further evidence for the system, in subsequent discussions of the learner model.

As stated above, in LEA's Box the learner model data comes from a variety of activities. These may be simple quizzes, intelligent educational systems, or self or peer assessments. Thus, negotiation may not always be so clearly focussed on specific

points such as a misconception as in the above example. However, the evidence used in negotiation can still be meaningful and, indeed, beneficial for both increasing the accuracy of the learner model and facilitating learner reflection. For example:

- LEARNER: My value for [matrices] should be [higher].
- SYSTEM: Your use of [OLMlets] showed [some difficulties].
- LEARNER: In [group discussion 2] I understood [well]. The [peer assessment] is [too low].
- SYSTEM: [Group discussion 2] was [5 days] ago and the [peer assessment] was [4 days ago]. You used [OLMlets] [1 day] ago. The level of [matrices] in [OLMlets] was [easy].

In this example, the system accesses the timestamp of data: in this case data from OLMlets [2]; and a peer assessment following a group discussion. It is able to explain that the first set of data was older, and also that the more recent OLMlets data was from a quite basic task. If the learner did not wish to accept the reasoning, the system could further explain that easier exercises can lead to higher scores, and that the learner was now working on more complex tasks, so old data would be less relevant. Through negotiation, as well as determining the correct representation for the learner model, the learner should come to better recognise their skills as they are required to think about the evidence provided by the OLM as well as in any justifications that they themselves give, supporting their claim. In addition, if the learner has disputed a peer assessment, the interaction will allow them to better understand that assessment, or have the opportunity to persuade the system to correct the disputed value.

Thus, the LEA's Box approach that is currently under development draws on the benefits of OLMs as meaningful visualisations of learning, as well as the benefits of negotiated learner models that can increase the accuracy of the learner model while also promoting learner reflection and other metacognitive behaviours. This is particularly useful when learners may be using disjointed applications, and when the learner model data includes data from other users. For the latter, in addition to the potential to increase the accuracy of the learner model, the process allows learner frustrations and perceived unjust assessments to be handled.

The current method of learner modelling uses a simple weighted algorithm, applying heavier weighting to more recent data, regardless of their origin [3]. However, teachers can adjust the weightings for individual activities according to the relevance of an activity for the learner model. As well as the recency of data as indicated above, the learner model negotiation will take account of these teacher weightings, and include these in its reasoning when 'defending' a representation during negotiation.

## 4. Summary

We have explained how benefits of negotiating learner models can be applied in today's contexts of multiple applications contributing learner data, as well as other activities which may include group interaction and peer assessment. By giving self, peer and teacher assessments the same status as automated data from various sources, such assessments can offer valuable insights to the learner's current learning state. Including this data also allows a system to better gauge the learner's viewpoint on

their understanding (through self assessments), and also take into account learning outcomes from non-computer-based or non-tracked activities (from self, peer and teacher assessment). By negotiating the learner model, users can help maintain their learner model, and through this process they should also benefit from the critical thinking required to justify their viewpoints if they disagree with any representations. In addition, learner model negotiation allows a method to verify peer assessment values, and a means to allow a learner to try to update the learner model in cases of unfairness or perceived unfairness resulting from peer contributions to their model.

## Acknowledgement

This work is supported by the European Commission (EC) under Information Society Technology priority FP7 for R&D, contract 619762 LEA's Box. This document does not represent the opinion of the EC and the EC is not responsible for any use that might be made of its contents.

## References

1. Bakalov, F., Hsiao, I-H., Brusilovsky, P. & Koenig-Ries B. (2011). Visualizing Student Models for Social Learning with Parallel Introspective Views, Workshop on Visual Interfaces to the Social Semantic Web, ACM IUI 2011, Palo Alto, US.
2. Bull, S. & Britland, M. (2007). Group Interaction Prompted by a Simple Assessed Open Learner Model that can be Optionally Released to Peers, Proceedings of PING Workshop, User Modeling 2007.
3. Bull, S., Johnson, M.D., Demmans Epp, C., Masci, D., Alotaibi, M. & Girard, S. (2014). Formative Assessment and Meaningful Learning Analytics, ICALT 2014.
4. Bull, S. & Kay, J. (2013). Open Learner Models as Drivers for Metacognitive Processes, in R. Azevedo & V. Aleven (eds), International Handbook of Metacognition and Learning Technologies, Springer, New York, 349-365.
5. Bull, S. & Pain, H. 'Did I Say What I Think I Said, And Do You Agree With Me?': Inspecting and Questioning the Student Model, Proceedings of World Conference on Artificial Intelligence and Education, AACE, Charlottesville VA, 1995, 501-508.
6. Chen, Z-H., Chou, C-Y., Deng, Y-C. & Chan, T-W. (2007). Active Open Learner Models as Animal Companions: Motivating Children to Learn through Interacting with My-Pet and Our-Pet, Int. Journal of Artificial Intelligence in Education 17(2), 145-167.
7. Cruces, I., Trella, M., Conejo, R. & Galvez J. (2010). Student Modeling Services for Hybrid Web Applications, Int. Workshop on Architectures and Building Blocks of Web-Based User-Adaptive Systems, http://ceur-ws.org/Vol-609/paper1.pdf.
8. Dimitrova, V. (2003). StyLE-OLM: Interactive Open Learner Modelling, Int. Journal of Artificial Intelligence in Education 13(1), 35-78.
9. Gillies, R.M. & Ashman, A.F. (2003). An Historical Review of the Use of Groups to Promote Socialization and Learning, in R.M. Gillies & A.F. Ashman (eds), Co-operative Learning: the Social and Intellectual Outcomes of Learning in Groups, RoutledgeFalmer, London, 1-18.
10. Hanrahan, S.J. & Isaacs, G. (2001) Assessing Self- and Peer Assessment: The Students' Views, Higher Education Research & Development 20(1), 53-70.

11. Hu, Q., Huang, Y. & Liu, C. (2012). The Design and Implementation of Learner Models in Online Peer Assessment to Support Learning, Int. Conference on Consumer Electronics, Communications and Networks, IEEE, 2961-2963.
12. Hung, S-T.A. (2012). A Washback Study on E-portfolio Assessment in an English as a Foreign Language Teacher Preparation Program, CALL 25(1), 21-36.
13. Kaufmann, J.H. & Schunn, C.D. (2011). Students' Perceptions about Peer Assessment for Writing: their Origin and Impact on Revision Work, Instructional Science 39, 387-406.
14. Kay, J. (2012). AI & Education: Grand Challenges, IEEE Intelligent Systems 27(5), 66-69.
15. Kay, J., Reimann, P., Diebold, E. & Kummerfeld, B. (2013). MOOCs: So Many Learners, So Much Potential, IEEE Intelligent Systems 28(3), 70-77.
16. Kerly, A. & Bull, S. (2008). Children's Interactions with Inspectable and Negotiated Learner Models, Proc. ITS, Springer-Verlag, Berlin Heidelberg, 132-141.
17. Kulkarni, C, Wei, K.P., LE, H., CHIA, D., Papadopoulos, K., Cheng, J., Koller, D., & Klemmer, S.R. (2013). Peer and Self Assessment in Massive Online Classes, ACM Transactions on Computer-Human Interaction 9(4), Article 39.
18. Kumar, R. & Kim, J. (2014). Editorial: Special Issue on Intelligent Support for Learning in Groups, International Journal of Artificial Intelligence in Education 24, 1-7.
19. Lu, J. & Law, N. (2012). Online Peer Assessment: Effects of Cognitive and Affective Feedback, Instructional Science 40, 257-275.
20. Mabbott, A. & Bull, S. (2004). Alternative Views on Knowledge: Presentation of Open Learner Models, Proc. ITS, Springer-Verlag, Berlin Heidelberg, 689-698.
21. Mazzola, L. & Mazza, R. (2010). GVIS: A Facility for Adaptively Mashing Up and Presenting Open Learner Models, Proc. EC-TEL, Springer, Berlin Heidelberg, 554-559.
22. Morales, R., Van Labeke, N., Brna, P. & Chan, M.E. (2009). Open Learner Modelling as the Keystone of the Next generation of Adaptive Learning Environments, in C. Mourlas & P. Germanakos (eds), IUI, Information Science Reference, ICI Global, London, 288-312.
23. Raybourn, E.M. & Regan, D. (2011). Exploring E-Portfolios and Independent Open Learner Models: Toward Army Learning Concept 2015, Interservice/Industry Training, Simulation, and Education Conference Proceedings, Florida USA.
24. Subramanian, R. & Lejk, M. (2013). Enhancing Student Learning, Participation and Accountability in Undergraduate Group Projects through Peer Assessment, South African Journal of Higher Education 27(2), 368-382.
25. Suen, H.K. (2014). Peer Assessment for Massive Open Online Courses (MOOCs), The International Review of Research in Open and Distance Learning 15(3), 312-327.
26. Thomson, D. & Mitrovic, A. (2010). Preliminary Evaluation of a Negotiable Student Model in a Constraint-Based ITS, Research and Practice in Technology Enhanced Learning 5(1), 19-33.
27. Topping, K. (1998). Peer Assessment between Students in Colleges and Universities, Review of Educational Research 68(3), 249-276.
28. Upton, K., & Kay, J. (2009). Narcissus: Group and Individual Models to Support Small Group Work, Proc. User Modeling, Adaptation and Personalization, Springer, 54-65.
29. Van Labeke, N., Brna, P. & Morales, R. (2007). Opening Up the Interpretation Process in an Open Learner Model, Int. Journal of Artifial Intelligence in Education 17(3), 305-338.
30. Van Zundert, M., Sluijsmans, D. & van Merrienboer, J. (2010). Effective Peer Assessment Processes: Research Findings and Future Directions, Learning and Instruction 20(4), 270-279.
31. Welsh, M. (2012). Student Perceptions of Using the PebblePad E-portfolio System to Support Self- and Peer-Based Formative Assessment, Technology, Pedagogy and Education 21(1), 57-83.
32. Zapata-Rivera, J-D. & Greer, J.E. (2004). Interacting with Inspectable Bayesian Student Models, Int. Journal of Artificial Intelligence in Education 14(2), 127-163.

# Predicting Student Attrition in MOOCs using Sentiment Analysis and Neural Networks

Devendra Singh Chaplot, Eunhee Rhim, and Jihie Kim

Samsung Electronics Co., Ltd.
Seoul, South Korea
{dev.chaplot,eunhee.rhim,jihie.kim}@samsung.com

**Abstract.** While there is increase in popularity of massive open online courses in recent years, high rates of drop-out in these courses makes predicting student attrition an important problem to solve. In this paper, we propose an algorithm based on artificial neural network for predicting student attrition in MOOCs using sentiment analysis and show the significance of student sentiments in this task. To the best of our knowledge, use of user sentiments and neural networks for this task is novel and our algorithm beats the state-of-the-art algorithm on this task in terms of Cohen's kappa.

**Keywords:** Student Attrition, MOOC, Educational Data Mining, Sentiment Analysis, Neural Network

## 1 Introduction

Massive Open Online Courses (MOOCs) have been gaining lot of interest in academia and industry in last few years. The key reasons in growing popularity of MOOCs include accessibility to every person in the world who has internet, scalability to handle any number of students with wide diversity of needs and expectations, and flexibility they provide to learners to study according to their routine. However, issues such as lack of instructor attention and absence of social learning environment, have led to high rates of attrition in MOOCs. With various unique benefits they offer over traditional classroom setting, online courses have the potential to transform future of education system, which brings out the importance of predicting student attrition in MOOCs.

With scalability, MOOCs also offer huge amounts of data of student activity, which can be utilized to train models for predicting attrition. The absence of physical learning environment makes the forums in MOOCs only medium of interaction with the instructor and peers. In this paper, we analyze the importance of sentiment analysis on these forum posts in predicting student attrition and study the effectiveness of neural network in modeling this problem.

The rest of the paper is divided into the following sections. Section 2 covers related work regarding machine learning techniques used to predict attrition and different kind of features used in them. Our algorithm is described in detail in Section 3. The experiments and results are presented in Section 4. Conclusions and future work are covered in Section 5.

## 2 Related Work

Recently, there have been many efforts to predict student attrition in MOOCs by extracting a wide variety of features from learner activity data and applying different machine learning approaches. [11] operationalize video lecture clickstream to capture behavioral patterns in student's activity, which is used to construct students' information processing index. [4] use feature such as number of threads viewed, number of forum posts, percentage of lectures watched, etc to predict student attrition. [12] construct a graph to capture sequence of active and passive learner activity, and use graph metrics as features for predicting attrition. [2] use quiz related (attempts and submissions) and activity related (length of action sequences, counts of various activities) features while [7] and [10] extract more than 15 features indicating learner activity and engagement from clickstream log. All these methods use variety of machine learning techniques including Logistic Regression, SVMs, Hidden Markov Models and random forest method.

There has not been much work on use of student sentiments in predicting attrition. [1] conclude that sentiment of students for assignments and course material has positive effects on successful completion of course. [14] also find correlation between sentiment expressed in the course forum posts and student drop out rate while they advice prudence against inconsistencies.

## 3 Proposed Algorithm

We have used click stream log and forum posts data from Coursera MOOC, 'Introduction to Psychology', which was prepared for MOOC Workshop at EMNLP 2014. The data consists of over 3 million student click logs and over 5000 forum posts. The click stream logs contain clicks made while watching video lectures and requests for viewing forums, threads, quiz, course wiki, etc. with time stamp of each click. More details about the dataset can be found in [7]. The following input features were extracted from the dataset:

- **User ID:** Unique numerical ID of the student.
- **Course Week:** Number of weeks since course has begun.
- **User week:** Number of weeks since student has joined the course.
- **Number of clicks** by the student in the current week.
- **Number of study sessions** by the student in the current week.
- **Number of course pages viewed** by the student in current week which include all pages except the video lectures.
- **Number of forum pages viewed** by the student in current week.
- **Student sentiment** of forum posts in the current week.

All the input features except Student Sentiments were indicated to be most effective by previous works mentioned in Section 2. The output of the algorithm is 1 indicating the user will drop out of the course in next week, and 0 otherwise. Note that we are predicting the exact week when the student is going to drop-out unlike [11] who predict whether student is going to finish the course or not. Our algorithm pinpoints the time when student is predicted to drop-out, which allows the course instructor and his team to take necessary steps to prevent or reduce student attrition during the course.

### 3.1 Sentiment Analysis

We follow a lexicon-based approach to extract sentiment from forum posts using SentiWordNet 3.0 [3] as the knowledge resource. It assigns a sentiment score to each synset in the WordNet [8]. Given the forum post, we pass the stem of each content word (using MIT JWI [6]) and its POS Tag (using Stanford POS Tagger [13]) to the SentiWordNet which returns a sentiment score. The sentiment score of the forum post is calculated by summing up the sentiment scores of all the words in the post. Fig. 1 shows a block diagram of this process.

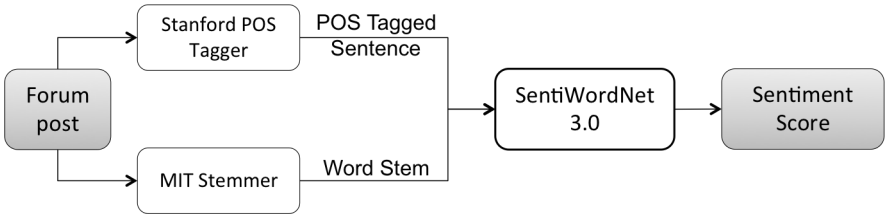



**Fig. 1.** Block Diagram of lexicon-based sentiment analysis using SentiWordNet 3.0.

### 3.2 Neural Network

Artificial neural networks are suitable to model the problem of predicting student attrition as there are a large number of inputs, and any mathematical relationship between input and output is unknown. Unlike many other machine learning techniques, neural networks are able to model the output as any arbitrary function of inputs and considered extremely robust if network structure, cost function and learning algorithm are selected appropriately through experiments. Downside of neural networks is inability to interpret the model.

We construct an artificial neural network consisting of 7 nodes in input layer: Course Week, User week, Number of clicks, Number of sessions, Number of page views, Number of forum views and Student sentiment as described above. Output layer consists of single node predicting whether student is going to drop-out in the next week. Each input feature is normalized to take values between 0 and 1. We add a hidden layer of 6 neurons in the neural network between the input and output layer. The number of neurons in the hidden layer were experimentally determined to get best possible results. Fig. 2 shows the structure of the neural network used to predict student attrition. To train the neural network, we use resilient propagation heuristic. It gave best results in our experiments among back propagation, Manhattan propagation and quick propagation.

## 4 Experiments & Results

In predicting student attrition, our focus is to capture all students who are going to drop-out and thus, minimizing false negative rate is important. False negative rate is the ratio of students who are predicted to stay in the course (predicted negative) in next week but actually drop out in the next week. While minimizing false negative rates, its also necessary to maintain overall accuracy so as to not produce too many false positives for the course instructor to handle.

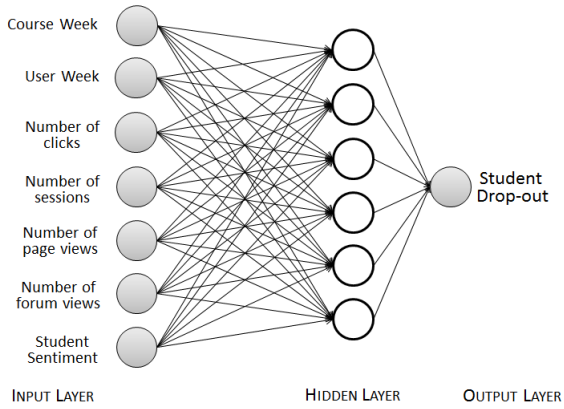



**Fig. 2.** The structure of neural network used to predict student attrition.

Since we are predicting whether student will drop-out in next week, our data set is highly imbalanced towards negative (will not drop-out) class. This is because a student who joins the course in 1st week, and drops out in 11th week, will have 9 negative class data points (week 1 to 9) and 1 positive class data point (week 10). Since the data set consists of student logs over 19 weeks, it is highly imbalanced with only 22.56% positive data points. Due to high imbalance in data set, we believe comparison of Cohen's kappa [5] is more suitable than comparing total accuracy directly. [9] show that Cohen's Kappa provides a unbiased estimate of performance of a classifier, and is thus much more meaningful than Recall, Precision, Accuracy, and their biased derivatives. It is more robust than total accuracy as it excludes proportion of correct predictions occurring by chance which is important in case of imbalanced data set, as a simple majority classifier would get 77.44% accuracy in this task.

In Table 1 we report our results with and without using student sentiments using 5-fold cross validation and compare them with some other approaches mentioned in Section 2. The proposed algorithm provides the best Cohen's Kappa values as compared to previous algorithms. Fall in accuracy and false negative rate when our algorithm doesn't use student sentiments indicates its importance in predicting attrition. Note that the algorithm which provides the best accuracy [10] also has the highest number of false negatives and the algorithm with best false negative rate has the lowest accuracy (Sinha-14 Baseline + Graph). This is due to imbalance in data which is explained in the following subsection. Note that the proposed algorithm has either better accuracy or better false negative rates than each of the previous algorithms, and this is reason behind better Kappa values. Since the dataset is from a MOOC which had free enrollment, there are many initial lurkers in the first week of the course who just want to browse the contents of the course. Thus, we believe predicting student attrition in first week is not very useful. Substantial improvement in performance of our algorithm without using first week's data is also shown in Table 1.

| Algorithm | Accuracy | False Neg. | Kappa |
|---|---|---|---|
| Balakrishnan-13 Stacking [4] | 80.5% | 0.353 | - |
| Balakrishnan-13 Cross-Product [4] | 80.1% | 0.442 | - |
| Sharkey-14 [10] | **88.0%** | 0.460 | - |
| Sinha-14 Baseline + Graph [12] | 62.4% | **0.095** | 0.277 |
| Sinha-14 Graph [12] | 69.2% | 0.157 | 0.365 |
| Neural Network (NN) | 70.7% | 0.199 | 0.365 |
| NN with Sentiment Analysis (SA) | 72.1% | 0.141 | 0.403 |
| NN with SA & without Week 1 | 74.1% | 0.132 | **0.432** |

**Table 1.** Comparison of accuracy and false negative rates with and without using student sentiments. The best results in each column is marked in **bold**.

### 4.1 Problem of data imbalance

The high data imbalance leads to biasing of the classifier towards the majority class. The problem of data imbalance in the same task is also addressed by [2] who try to solve it by oversampling the minority class, but were unsuccessful. We counter this problem by setting the boundary for classification to the ratio of drop out data points to total number of data points in the training set. This means that if the value of output neuron is greater than this ratio, then student is predicted to drop out in the next week, and vice-versa otherwise. If complete data set is used as training set, then this boundary would be 0.2256, meaning student is predicted to drop-out if value of output neuron is greater than 0.2256, rather than 0.5 by default. This adjustment to the boundary allows us to train the neural network on highly unbalanced dataset and still achieve very good recall over minority class while maintaining the overall accuracy.

The boundary is essentially a trade-off between accuracy and false negative rate. It can be adjusted to get better accuracy or false negative rates depending upon the application. This boundary can also be calculated using receiver operating characteristic (ROC) Curve.

## 5 Conclusion & Future Work

We propose an algorithm to predict student attrition using an artificial neural network. Sentiment analysis of forum posts is shown to be an important feature to predict student attrition in MOOCs. We also provide an approach to tackle the problem of data imbalance which can be extended to wide variety of applications in many other domains. This approach allows to find a good middle ground between accuracy and false negative rates and leads our algorithm to beat the previous algorithms in terms of Cohen's Kappa.

Most methods provide analysis of MOOC data which indicate factors responsible for attrition. In contrast, we provide a method to pin-point students who are likely to drop-out during in the following week. Since our algorithm has a very low false negative rate, it can be used in MOOCs to capture most students who are likely to drop-out in near future and take necessary actions specific to the student to prevent them from dropping out. Apart from MOOCs, the proposed algorithm can also used in smart schools using digital methods for learning and interaction, which are becoming increasingly popular in recent years.

## References


1. Adamopoulos, P.: What makes a great MOOC? An interdisciplinary analysis of student retention in online courses. In: Proceedings of the International Conference on Information Systems, ICIS 2013, Milano, Italy (2013)
2. Amnueypornsakul, B., Bhat, S., Chinprutthiwong, P.: Predicting Attrition Along the Way: The UIUC Model. In: Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs. pp. 55–59. Association for Computational Linguistics, Doha, Qatar (October 2014)
3. Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta (2010)
4. Balakrishnan, G.: Predicting Student Retention in Massive Open Online Courses using Hidden Markov Models. Master's thesis, EECS Department, University of California, Berkeley (May 2013)
5. Cohen, J.: A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 20(1), 37 (1960)
6. Finlayson, M.: Java libraries for accessing the princeton wordnet: Comparison and evaluation. In: Orav, H., Fellbaum, C., Vossen, P. (eds.) Proceedings of the Seventh Global Wordnet Conference. pp. 78–85. Tartu, Estonia (2014)
7. Kloft, M., Stiehler, F., Zheng, Z., Pinkwart, N.: Predicting MOOC Dropout over Weeks Using Machine Learning Methods. In: Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs. pp. 60–65. Association for Computational Linguistics, Doha, Qatar (October 2014)
8. Miller, G.A.: Wordnet: A lexical database for english. Commun. ACM 38(11), 39–41 (Nov 1995)
9. Powers, D.M.W.: Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Tech. Rep. SIE-07-001, School of Informatics and Engineering, Flinders University, Adelaide, Australia (2007)
10. Sharkey, M., Sanders, R.: A Process for Predicting MOOC Attrition. In: Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs. pp. 50–54. Association for Computational Linguistics, Doha, Qatar (October 2014)
11. Sinha, T., Jermann, P., Li, N., Dillenbourg, P.: Your click decides your fate: Inferring Information Processing and Attrition Behavior from MOOC Video Clickstream Interactions. ArXiv e-prints (Jul 2014)
12. Sinha, T., Li, N., Jermann, P., Dillenbourg, P.: Capturing "attrition intensifying" structural traits from didactic interaction sequences of MOOC learners. CoRR abs/1409.5887 (2014)
13. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1. pp. 173–180. NAACL '03, Association for Computational Linguistics, Stroudsburg, PA, USA (2003)
14. Wen, M., Yang, D., Rosé, C.P.: Sentiment analysis in mooc discussion forums: What does it tell us. In: Proceedings of Educational Data Mining (2014)

# Adapting Collaborative Chat for Massive Open Online Courses: Lessons Learned

Oliver Ferschke, Gaurav Tomar, and Carolyn P. Rosé

Carnegie Mellon University, School of Computer Science
Pittsburgh, PA 15213, USA
{ferschke,gtomar,cprose}@cs.cmu.edu

**Abstract.** In this paper we explore how to import intelligent support for group learning that has been demonstrated as effective in classroom instruction into a Massive Open Online Course (MOOC) context. The Bazaar agent architecture paired with an innovative Lobby tool to enable coordination for synchronous reflection exercises provides a technical foundation for our work. We describe lessons learned, directions for future work, and offer pointers to resources for other researchers interested in computer supported collaborative learning in MOOCs.

## 1 Introduction

The field of Computer Supported Collaborative Learning (CSCL) has a rich history extending for nearly two decades, covering a broad spectrum of research related to learning in groups, especially in computer mediated environments. In this paper we describe the initial stages of a research program designed to import findings from a history of successful classroom research in the field of CSCL to the challenging environment of MOOCs.

In order to support the growth of student discussion skills, it is necessary to design environments with affordances that encourage transactive behaviors and other valuable learning behaviors. The most popular approach to providing such affordances in the past decade has been that of script-based collaboration [2, 7, 6]. A script is a schema for offering scaffolding for collaboration. Some typical forms of scripts come in the form of instructions that structure a collaborative task into phases, or structured interfaces that reify certain types of contributions to the collaboration. Prior work on dynamic conversational agent based support built on a long history of work using tutorial dialogue agents to support individual learning with technology [11, 10, 5, 12].

The MOOC environment presents a number of challenges that must be addressed in order to introduce synchronous collaboration opportunities into MOOCs. From a research perspective, interesting challenges include exploration of group composition questions with MOOC student populations, which are far more diverse with respect to culture, age, educational level, and goals than typical classroom populations. Another interesting methodological challenge is the lack of control over the context. In a classroom context, a certain amount of time

may be set aside for an activity, and students can be expected to be present for the whole activity. In a MOOC, students may come and go as they please, and since they may be logging in from anywhere, any number of events could interfere with the task proceeding as planned. While a collaborative task may have been carefully designed with roles for each student to perform in a serious learning task, those roles may play out differently than intended in cases where the students who take on those roles are actually multiple students, students with a seriously inadequate preparation for the task, or even students with far more expertise than anticipated.

Before any of these issues may even be touched upon, a number of more practical issues must be addressed first to lay a foundation for this research. A major challenge in MOOCs is coordination. Whereas in a face-to-face course and traditional small-scale online courses, students can be expected to be amenable to stipulated meeting times, students in MOOCs typically come from different time zones around the world. The great majority of students make use of resources at their convenience, when they happen to have time to log in, rather than planning ahead and arriving at a scheduled time. The sheer numbers of students make it challenging to coordinate plans for meeting times. Furthermore, not all students adopt the same orientation towards following instructions in general or engaging in a task as presented in particular. Some students may click on an activity in an exploratory or playful fashion rather than with a serious intention of completing the activity. Thus, there is a danger of introducing students into a group in a way that engenders conflict or mismatched expectations.

In the remainder of the paper we first introduce the technical approach we adopted in an initial MOOC deployment. We then summarize our main results and lessons learned. We conclude with directions for continued work and resources to share with the community. Further discussion of the results of our deployment can be found in two separate publications [3, 4].

## 2 Technical Approach

In order to gain a deeper understanding of the problems that may arise from synchronous collaborative activities in MOOCs, we integrated a chat environment with interactive agent support in a recent 9-week long MOOC on learning analytics (DALMOOC) that was hosted on the edX platform from October to December 2014.

In order to facilitate the formation of ad-hoc study groups for the chat activity, we make use of a simple setup referred to as a Lobby. The Lobby introduces an intermediate layer between the edX platform and the synchronous chat tool. Even though the Lobby allows groups of arbitrary sizes to be formed, we decided that agent-guided discussions in groups of two students are the suited setup in the context of this MOOC. Students enter the Lobby with a simple, clearly labeled button click integrated with the edX platform. In order to increase the likelihood of a critical mass of students being assigned to pairs, we suggested a couple of two hour time slots during each week of the MOOC when students

might engage in the collaborative activities. These timeslots were advertised in weekly newsletters. However, the chat button was live at all times so that students were free to attempt the activity at their convenience. Upon entering the lobby, students are asked to enter the name that will be displayed in the chat. Once registered in the lobby, the student waits to be matched with another participant. If they are successfully matched with another learner who arrives at the Lobby within a couple of minutes to interact with, they and their partner are then presented with a link to click on to enter a chat room created for them in real time. Otherwise they are requested to come back later. A visualization is presented to them that illustrates the frequency of student clicks on the button at different times of the day on the various days of the week so that they are able to gauge when would be a convenient time for them to come back when the likelihood of a match would be higher. In the beginning of the course, the graph was based on experiences with past MOOCs while it was later updated with real data from the DALMOOC logs.

When the successfully matched students click on the provided link, they enter a private chat room. This chat setup has been used in earlier classroom research [1]. It provides opportunities for students to interact with one another through chat as well as to share images. The chat environment furthermore has built-in support for conversational agents who appear as regular users in the chat.

In contrast to our earlier work where we support collaborative chat dynamically with conversational agents triggered by real time monitoring of student interactions [1], we employ statically scripted agents in DALMOOC which guide the students with course-related discussion questions (Figure 1). Even though the scripts are linear, the agent prompts are not strictly timed but rather allow the students to interact in their own pace and take as much time as needed to discuss the given topic. Once a team wants to proceed with the discussion, they can move on with the *We're ready*-button. The agent will proceed with the next prompt as soon as both students indicated that they are ready. In case the students never signal their readiness, the agent will inquire after a predefined timeout in order to move forward with the discussion and avoid the students to lose focus.

## 3 Main Results

Though we encountered many challenges during the DALMOOC deployment, the main results suggest value added by the intervention. In order to begin to assess the added value of integrating reflective chat activities with a MOOC platform, we compared our synchronous collaborative chats with two other communication contexts, namely Twitter and the MOOC discussion forum [3]. What we found is that different subpopulations of learners within DALMOOC tended to gravitate towards different communication contexts. Furthermore, each context was associated with its own unique profile in terms of content focus and the nature of the discussion. The chat conversations showed the highest average of reflective contributions across all the platforms we observed. Furthermore,

| | |
|---|---|
| **Prompt 1** | In this collaborative activity, we will reflect on what you have learned about the field of learning analytics. First, take a couple of minutes to introduce yourselves. |
| **Prompt 2** | Now that you have viewed the videos, share what you found most interesting about learning analytics. |
| **Prompt 3** | Regarding learning analytics tools, did you find the classifications of a) proprietary/open source and b) single functionally/Integrated suites to be useful? How would you improve these classifications to make them more relevant to educators starting with analytics toolsets? |
| **Prompt 4** | Reflect on the structure of the dual-layer structure of the course. Describe your experience of coming to understand different course elements. |
| **Prompt 5** | Now this activity has come to an end. Thanks for a great chat! Why don't you exchange contact information to stay in touch? |

Fig. 1: Agent prompts for the collaborative chat activity in the first week

the one-on-one conversations in Bazaar exhibit a strong constructive character where reflective statements are not merely precompiled by each student and then exchanged, they are rather collaboratively constructed in the course of the conversation. We see ample evidence within contributions across media pertaining to social connection that these MOOC learners crave continuing social engagement with other individuals participating in their MOOC course. The analysis suggests that there is value in providing a diverse set of discussion contexts but that it creates a need for greater efforts towards effective bridging between media and channeling of students to pockets of interaction that are potentially of personal benefit.

We also used a survival analysis to evaluate the impact of participation in collaborative chats on attrition over time in the course [4]. The results suggest a substantial reduction in attrition over time, specifically a reduction by more than a factor of two, when students experience a match for a synchronous collaborative reflection exercise. Nevertheless, these results must be treated with some caution as we experienced significant difficulty in managing the logistics of matches. Even with 20,000 students enrolled in the course, some students had to make as many as 15 attempts to be matched with a partner before a match was made.

## 4 Lessons Learned

In this first deployment study, we learned valuable lessons that will help to improve our experimental setup in future cycles of our iterative design based research process. In this section, we first describe the main lessons learned and then briefly discuss future directions we are planning to take.

Integrating a synchronous collaborative activity in an inherently asynchronous learning environment used by students in different time zones was one of the greatest organizational challenges to overcome. As mentioned earlier, we attempted to alleviate the problem by introducing dedicated chat hours to increase the likelihood of students getting matched with each other. Nevertheless,

the majority of students who entered the lobby could not be matched with a chat partner within 10 minutes. This was a frequent cause of frustration which lead to students abandoning the chat activity in the course of the MOOC.

Since students are matched randomly in pairs for each chat activity, their interaction is naturally limited to a single chat session. Whenever they return to the chat, they will be connected with a different student. From the logs we have seen that especially after longer discussions, students expressed the desire to connect with each other and continue the interaction beyond the chat activity. On several occasions, they exchanged contact information in order to reconnect for further collaboration. However, the intervention did not provide any support for continued social engagement between paired learners.

We are currently developing new strategies for tackling these problems in future deployments of the intervention. First, we will employ a single-chatroom setup that allows students to directly enter at their own volition without the need for explicit matchmaking. The agent in this continuous chatroom will then adapt to the student population in the room at any given time. For instance, a single user in the room would be prompted to reflect on the course material on their own. Once a second user enters, the agent summarized the reflection of the other student and composes a discussion topic for the two users to collaboratively reflect on. The agent keeps track of the topics already discussed by the users currently present in the room to avoid redundant prompts.

Second, we will explore a scheduling system that allows students to sign up for a set of predefined timeslots. This approach has proven effective for multi-party voice chats in MOOCs [8]. Even though the necessity to schedule discussions ahead might negatively affect the engagement of users who merely interact with the MOOC on an ad-hoc basis, the approach could nevertheless help to reduce overall friction by offering an easier transition from the asynchronous nature of the MOOC to the synchronous nature of the chat.

## 5 Conclusions

This research was motivated by the goal to import best practices and technologies from the field of Computer Supported Collaborative Learning into MOOCs [9]. It is part of a broader effort drawing from two decades of research in Computer Supported Collaborative Learning, where we are working to design an extension of the edX platform to enhance instructionally beneficial discussion opportunities available to students[1]. We are partnering with edX as a satellite collaborative, seeking to involve researchers and developers from multiple universities, foundations, and industrial organizations. Our long term vision is to seek to leverage insights and methodologies from the field of Human-Computer Interaction more broadly and encompassing both synchronous and asynchronous communication very broadly. Our vision includes text, speech, and video based interactions, instrumented with all sorts of intelligent support powered by state-of-the-art

[1] `http://dance.cs.cmu.edu`

analytics and leveraging language technologies and artificial intelligence more broadly in order to offer contextually appropriate support.

**Acknowledgments** This work was funded in part by NSF grants OMA-0836012 and DATANET 1443068 and funding from Google and Bosch.

## References

1. Adamson, D., Dyke, G., Jang, H., Rosé, C.P.: Towards an agile approach to adapting dynamic collaboration support to student needs. International Journal of Artificial Intelligence in Education 24(1), 92–124 (2014)
2. Dillenbourg, P.: Over-scripting CSCL: The risks of blending collaborative learning with instructional design. In: Kirschner, P.A. (ed.) Three worlds of CSCL. Can we support CSCL?, pp. 61–91. Heerlen, Open Universiteit Nederland (2002)
3. Ferschke, O., Howley, I., Tomar, G., Yang, D., Rosé, C.P.: Fostering Discussion across Communication Media in Massive Open Online Courses. In: Proceedings of the 11th International Conference on Computer Supported Collaborative Learning. Gothenburgh, Sweden (2015)
4. Ferschke, O., Yang, D., Tomar, G., Rosé, C.P.: Positive Impact of Collaborative Chat Participation in an edX MOOC. In: Proceedings of the 17th International Conference on Artificial Intelligence in Education. Madrid, Spain (2015)
5. Graesser, A., Vanlehn, K., TRG Group, NLT Group: Why2 report: Evaluation of why/atlas, why/autotutor, and accomplished human tutors on learning gains for qualitative physics problems and explanations. Tech. rep. (2002)
6. Kobbe, L., Weinberger, A., Dillenbourg, P., Harrer, A., Hmlinen, R., Hkkinen, P., Fischer, F.: Specifying computer-supported collaboration scripts. International Journal of Computer-Supported Collaborative Learning 2(2-3), 211–224 (2007)
7. Kollar, I., Fischer, F., Hesse, F.: Collaboration scripts  a conceptual analysis. Educational Psychology Review 18(2), 159–185 (2006)
8. Kulkarni, C., Cambre, J., Kotturi, Y., Bernstein, M.S., Klemmer, S.R.: Talkabout: Making distance matter with small groups in massive classes. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW 2015, Vancouver, BC, Canada, March 14 - 18, 2015. pp. 1116–1128 (2015)
9. Rosé, C.P., Goldman, P., Sherer, J.Z., Resnick, L.: Supportive technologies for group discussion in moocs, current issues in emerging elearning. Current Issues in Emerging eLearning. Special issue on MOOCs (2015)
10. Rosé, C.P., Jordan, P., Ringenberg, M., Siler, S., V.K., Weinstein, A.: Interactive Conceptual Tutoring in Atlas-Andes pp. 256–266 (2001)
11. Wiemer-Hastings, P., Graesser, A., Harter, D.: The foundations and architecture of autotutor. In: Goettl, B.P., Halff, H.M., Redfield, C., Shute, V.J. (eds.) Intelligent Tutoring Systems, Lecture Notes in Computer Science, vol. 1452, pp. 334–343. Springer Berlin Heidelberg (1998)
12. Zinn, C., Moore, J.D., Core, M.G.: A 3-tier planning architecture for managing tutorial dialogue. In: Proceedings of the 6th International Conference on Intelligent Tutoring Systems. pp. 574–584 (2002)

# Exploring the Effects of Open Social Student Model Beyond Social Comparison

Julio Guerra[1], Yun Huang[2], Roya Hosseini[2], and Peter Brusilovsky[1]

[1] School of Information Sciences, University of Pittsburgh, Pittsburgh, PA, USA
[2] Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA
{jdg60,yuh43,roh38,peterb}@pitt.edu

**Abstract.** In our journey exploring the effects of Open Student Model (OSM) on students working with programming problems and examples, we have incorporated the idea of social visualizations to extend OSM to Open Social Student Modeling (OSSM). Although comparison features in OSSM, where a student can compare herself to the group or individual peers, have shown to increase students' work, we now shift our attention to other effects. The goal is to explore the OSSM effects beyond comparison, particularly metacognitive support, and we propose a representation of the OSSM towards these lines.

**Keywords:** Open Student Model, Open Social Student Model, Metacognition, Self-Regulated Learning, Group-Awareness

## 1 Introduction

Open Student Model (OSM, also called Open Learner Model or OLM) consists of a set of features, usually visual and sometimes interactive, that shows data of progress, mastery of knowledge, or other statistics of the activity of the student to herself [3]. This data comes from the internal user model that the computer-based educational system maintains to bring in adaptive and tutoring functionalities [2]. By showing the user model to the learner, OSM fosters metacognitive processes like self-awareness [4] and can be further used as a navigational tool. In the past we have explored different forms of guidance based on OSM. KnowledgeZoom (KZ) [1] implements a fine-grained student model based on concepts hierarchically organize in an Ontology of Java programming. KZ presents the student model using treemap that shows different levels of details as the student "enters" each of the concepts. This approach allows the student to have an overall view and a detailed view of her progress and knowledge gaps just few clicks away. We have also incorporated the idea of social visualizations and extended the OSM to an Open Social Student Modeling (OSSM) [8, 10]. OSSM seeks for sharing aggregated or individual OSM among the students allowing social comparison and social guidance dynamics. Figure 1 shows a screenshot of the MasteryGrids system, our current OSSM implementation. The first 3 rows represent the progress of the current student, the comparison between the student and the group, and the progress of the rest of the class, respectively. Cells

represent topics, ordered as they are covered in the course. Darker colors mean higher progress in the content. The student progress is colored in shades of green, and the group (the average of the class) is represented with a blue color palette. The middle row shows a differential color that turns green when the student is more advanced than the group and blue otherwise. By clicking in a cell, the student has the access to educational material included in the selected topic (in the figure, the cell corresponding to the topic *Loops While* has been clicked.) The second group of cells shows the progress of all individual students in the class, anonymized, ranked by the amount of progress (advanced students at the top) using shades of blue.

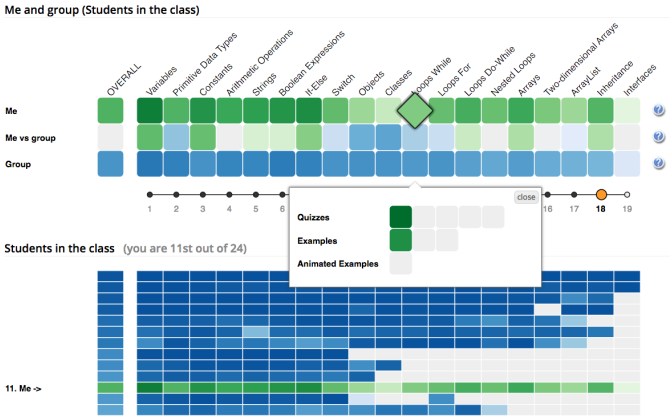



**Fig. 1.** MasteryGrids OSSM interface.

Overall, our efforts to implement OSSM have been focused on exploiting comparison effects and we have observed in classroom studies that this kind of features make students work more and follow others [8, 10]. Also, the sort of guidance produced by advanced students over non-advanced students is quite conservative, and we further proposed a guidance approach incorporating OSSM and adaptive navigation features (work presented as a poster in AIED 2015 [1]). We now shift our attention to explore the OSSM effects beyond comparison, particularly how OSSM can be applied to support metacognitive processes involved in self-regulated learning activities. The motivation for our vision comes from two areas. On the one hand, the strong ideas behind OSM are related to metacognitive support: OSM increases self-awareness and self-control of the learning process [4]. We believe that the metacognitive support of OSM reaches another level in OSSM. For example, OSSM can give students a sense of common difficulties helping them to make better self-judgments when facing failure. On the other hand, although our approach to OSSM does not incorporate direct interaction and collaborative tools, there is a sort of "indirect interaction" or "soft

[1] Poster title: *Off the Beaten Path: The Impact of Adaptive Content Sequencing on Student Navigation in an Open Social Student Modeling Interface*

collaboration" happening mediated by the cognitive aspects of the group information displayed. This social dimension can be used to enrich OSM features, for example, to guide students using the traces of others. In the next sections we explore related work about open student model, metacognition in self-regulated learning, the measures of metacognitive processes in computer-based learning environments, and social awareness in computer-based collaborative environments, and from these ideas we propose a representation of OSSM.

## 2 Open Student Model and Metacognition

Open Student Model (or Open Learner Model) discloses the user model that the system maintains to the learner. As a result, OSM is a tool for self-awareness and learning monitoring. In the review of OSM work, Bull and Kay [4] described different systems incorporating OSM features or indicators supporting metacognition, including high level indicators of performance, OSM negotiation features (the learner can negotiate her user model with the system), and fine-grained indicators at finer levels of knowledge components (for example, concepts). Fine-grained conceptual representations of OSM, where the student can discover gaps in her knowledge that are hidden in high level indicators, have been attempted in a number of works [11, 13]. A common approach to fine-grained models involves a detailed domain model that can be represented as a concept-map or concept tree where nodes are concepts in the domain linked by the ontological or semantic relations among them. The learner model is an *overlaid* status of the learner in each of the concepts and it is represented by using, for example, colors [3]. A common problem of fine-grained models is that they can become very complex and hard to understand by the student. Visual techniques has been proposed to deal with this issue, for example, our system KnowledgeZoom uses semantic-zooming [1]. Open Student Model is also acknowledged to be of benefit when shared. For example, the instructor can perform a detailed monitoring of the learners, the learner can find collaborators by inspecting other models, compare with suitable ones, or improve group awareness through open group model [3]. Our vision of OSM incorporates the idea of sharing the OSM (we call it Open Social Student Model) and a fine-grained model that serves the student to make a more precise judgement of her own learning process.

## 3 Self-Regulated Learning and Metacognition

The research in Self-Regulated Learning (SRL) puts importance to feedback mechanisms in the development of cognitive and metacognitive processes. Feedback is not only related to the learner seeing summaries of her activity traces or providing information to others or the educational system, but also the internal feedback processes the learner develops while reflecting on the activities, for example the update of beliefs about herself and beliefs about the content of study [5]. Moreover, Butler and Winne [5] noted the heterogeneous and adaptive nature of self-regulation (here *adaptive* refers to the behavior of the learner that

adapts during the learning experience) and they stressed the need of study it in a finer grain, i.e. continuously, while the learning activity is being performed. They proposed a broader view of self-regulated learning and feedback by describing four stages or elements: knowledge and beliefs, selection of goals, tactics and strategies, and monitoring. We take on this view and see ideas that can be applied in OSSM in each of these 4 stages. For example, for knowledge and beliefs, OSSM might project conflicting information to learner's self-efficacy beliefs (as the learner can compare her performance against others), and this discrepancy could be set to improve self-beliefs when possible. About goals, feedback can help the learner to set her goals and to make a good decision while navigating the content. Establishing a proper strategy to accomplish a goal might be difficult when the task is unfamiliar to the student, and here OSSM can use traces of other students to implement navigational guidance. Monitoring processes need to be supported by feedback information regarding both the current goal and about the progress on the learning activities.

Greene and Azevedo [7] saw the opportunity that Computer-Based Learning Environments (CBLEs) introduce for observing and measuring the learning process in detail, and reported a number of works using different forms of metacognitive measures and interventions in CBLEs. According to them, there are three types of techniques for measuring metacognition: 1) by self-reported instruments usually applied before or/and after the learning activity, 2) by using activity logged by the system or collected by sensors, and 3) inferred from explicit feedback given by the learner as the result of interacting with the system. They emphasized the idea that fine-grained metacognitive measure allows different levels of analyses, including semantic and statistical analyses of the activity, and analysis of sequences of actions in the time, which is in line with what Butler and Winne [5] recognized as necessary to study metacognition: continuous and on-line measurements. From the summary of Greene and Azevedo [7], we consider three broad ideas to incorporate in OSSM. First, it is important that the OSSM system collects all possible information while the learner performs the learning activities. Second, richer analyses and guidance can be achieved by incorporating some sort of dialogue or active interaction in the system that can be used to capture self-reported metacognitive state in real-time (for example asking the student what was the most difficult exercise, or asking the student to verify her model and write down her corrections). Third, the representation of the user model evolution over time (for example the progress in the last week), along with representation of the sequences of actions of the student and of the group or peers, can contribute to a better monitoring and planning tasks.

## 4 Social Awareness Tools

Janssen and Bodemer [9] summarized ideas of cognitive group awareness and social group awareness in collaborative activities supported by computer. Awareness, a process of inherent metacognitive character, can be of the type *Cognitive Group Awareness*, mainly about the knowledge and expertise of others (content

space), or of the type *Social Group Awareness*, mostly about the levels of participation or engagement of peers and the quality of the interaction (relational space). Both broad types of group awareness are not exclusive of each of the spaces. For example, Cognitive Group Awareness also interacts with the relational space. Following this framework, we situate our idea of OSSM as a Cognitive Group Awareness (OSSM shows the knowledge/progress model of peers and the group), and we see the value of incorporating a dimension of Social Group Awareness, for example, by showing indicators of visits, attempts, and other current activity made by peers. Also, as Janssen and Bodemer suggested [9], using Cognitive Group Awareness features will also produce an impact in the relational space and we should not ignore it. For example Glahn, Sprecht and Koper [6] observed that even though a group awareness indicator (showing average of the group performance) produces the longest and strongly positive effect in the amount and the quality of work, it also produces frustration in non-contributing users and in some cases, the belief of vicious competitive behaviors of others. Moreover, the question is how to grasp the benefits of the group awareness features in OSSM on both content and relational spaces. Different group-awareness tools are implemented by Papanikolaou [12] in the system INSPIREus, including indicators of effort, progress, working style, personalization features, and social construction of knowledge (summarizing the type of discussions in forums). Students reported that the indicators allowed them to better understand their weaknesses and helped them to better plan their activities. We take on these ideas to incorporate different types of indicators for reflection, self-monitoring and comparison, specially, by combining indicators of activity with pedagogical information that sets the context of the desired metacognitive processes. Similar to INSPIREus, we maintain a domain model consisting of concepts mapped to the content material and activities, and structured using different semantic relationships, which can be used to provide indicators at different levels, for example high level indicators summarizing a topic.

## 5 A Concept-Map OSSM

We propose to complement our current OSSM MasteryGrids with a network representation of the concepts as the student progress in the learning activities. Activities are mapped to a set of finer grained concepts and these concepts get connected as the student practices activities containing pairs of concepts. Thus, the network representation or concept map, gets more connected as the student practices the concepts with different other concepts. We recognize that in many domains *mastery* is reached as the student is able to connect different concepts. We hypothesize that this concept map will allow students to have a finer and detailed view of their models, thus engage them in deeper metacognitive processes. On the other hand, the representation grows naturally as the student *connects* concepts, thus giving an idea of the dynamic progress or advance in pursuing learning goals. We plan to incorporate features supporting other metacognitive processes of goal setting and learning strategy. The learner should be able to

choose concepts she wants to target, and the OSSM incorporates an indicator of the overall progress of the goal set. Recommendation and navigational clues are giving to signal concepts that should be targeted first and which activities to do to progress on those concepts. Collaborative filtering techniques can be used to grasp the *wisdom of the crowd* in order to power such recommendation mechanisms. For example, once a goal is set, the system can find the traces of other students that set similar goals in the past and use this information to recommend which activities to do. Each concept in the map can show information of the overall activity of the group related to the concept, for example to give an average of the progress on the concept. One important aspect on OSM is letting the learner correct or negotiate the model. Our implementation should allow students to change their knowledge levels through selected assessment items.

## References


1. Brusilovsky, P., Baishya, D., Hosseini, R., Guerra, J., Liang, M.: Knowledgezoom for java: A concept-based exam study tool with a zoomable open student model. In: Advanced Learning Technologies (ICALT), 2013 IEEE 13th International Conference on. pp. 275–279. IEEE (2013)
2. Brusilovsky, P., Millán, E.: User models for adaptive hypermedia and adaptive educational systems. In: The adaptive web. pp. 3–53. Springer-Verlag (2007)
3. Bull, S., Kay, J.: Open learner models. In: Advances in intelligent tutoring systems, pp. 301–322. Springer (2010)
4. Bull, S., Kay, J.: Open learner models as drivers for metacognitive processes. In: International Handbook of Metacognition and Learning Technologies, pp. 349–365. Springer (2013)
5. Butler, D.L., Winne, P.H.: Feedback and self-regulated learning: A theoretical synthesis. Review of educational research 65(3), 245–281 (1995)
6. Glahn, C., Specht, M., Koper, R.: Visualisation of interaction footprints for enagement in online communities (2010)
7. Greene, J.A., Azevedo, R.: The measurement of learners self-regulated cognitive and metacognitive processes while using computer-based learning environments. Educational psychologist 45(4), 203–209 (2010)
8. Hsiao, I.H., Bakalov, F., Brusilovsky, P., König-Ries, B.: Progressor: social navigation support through open social student modeling. New Review of Hypermedia and Multimedia 19(2), 112–131 (2013)
9. Janssen, J., Bodemer, D.: Coordinated computer-supported collaborative learning: Awareness and awareness tools. Educational Psychologist 48(1), 40–55 (2013)
10. Loboda, T.D., Guerra, J., Hosseini, R., Brusilovsky, P.: Mastery grids: An open source social educational progress visualization. In: Open Learning and Teaching in Educational Communities, pp. 235–248. Springer (2014)
11. Mabbott, A., Bull, S.: Alternative views on knowledge: Presentation of open learner models. In: Intelligent Tutoring Systems. pp. 689–698. Springer (2004)
12. Papanikolaou, K.: Constructing interpretative views of learners interaction behavior in an open learner model
13. Perez-Marin, D., Alfonseca, E., Rodríguez, P., Pascual-Nieto, I.: Automatic generation of students conceptual models from answers in plain text. In: User Modeling 2007, pp. 329–333. Springer (2007)

# Dual Eye Tracking as a Tool to Assess Collaboration

Jennifer K. Olsen[1], Michael Ringenberg[1], Vincent Aleven[1], and Nikol Rummel[12]

[1]Human Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, USA
{jkolsen, mringenb, aleven}@cs.cmu.edu
[2]Institute of Educational Research, Ruhr-Universität Bochum, Germany
nikol.rummel@rub.de

**Abstract.** In working towards unraveling the mechanisms of productive collaborative learning, dual eye tracking, a method where two people's eyes are tracked as they collaborate on a task, is a potentially helpful tool to identify moments when students are collaborating effectively. However, we are only beginning to understand how eye gaze relates to effective collaborative learning and how it fits in with other data streams. In this paper, we present three broad areas of analysis where we believe dual eye tracking will promote our understanding of collaborative learning. These areas are: (a) How eye gaze is associated with other communication measures, (b) how eye gaze is associated with task features, and (c) how eye gaze relates to learning outcomes. We present exploratory analyses in each of the three areas using a dataset of 28 4th and 5th grade dyads working on an Intelligent Tutoring System for fractions. Our analyses illustrate how dual eye tracking could be used in conjunction with other data streams to assess collaborative learning.

**Keywords:** collaborative learning, intelligent tutoring system, dual eye tracking

## 1 Introduction

Collaboration can be an effective tool for learning; however, it can be difficult to identify the mechanisms of collaboration and how students' actions may lead to learning when working in a group. The communication between partners plays a large role in the success of the group [3], and there are many different processes that happen during a collaborative session that can affect learning such as speech, joint attention, and tutor feedback. By analyzing these different processes separately and together, we may be able to develop a better understanding of the collaborative learning process. In this paper, we specifically focus on dual eye tracking, a method where two people's eyes are tracked as they collaborate on a task, with an Intelligent Tutoring System (ITS) and explore how it could be used with other data streams to analyze students' collaborative interactions. By using multiple data streams that include eye gaze, we may be able to have insights into collaboration that were not otherwise possible.

Research shows eye gaze is tied to communication, making eye tracking a promising method to use for the analysis of collaborative learning [9]. Previous research has shown that there is a link between eye gaze and speech [4], [9]. When people hear a

reference through speech, their eye gaze will follow that object [9], and when people are describing a picture, their eye gaze will look at the relevant part of the picture before it is described [4]. These studies show a link between speech and eye gaze that goes in both directions. This same pattern follows when people are working on a task together. There is a coupling of the collaborators' eye gaze around a reference [12]. The eye gaze has a closer coupling when each of the collaborators has the same initial information and when there is a shared selection [7], [12], suggesting that task features influence eye gaze. The coupling of eye gaze between collaborating partners may be an indicator of quality interaction and better comprehension [6], [11]. It also may be associated with better learning because there is more comprehension and understanding from the interactions with a closer coupling of eye gaze. Much of the previous work has focused on the correlation of eye gaze with speech, but it is still an open question of how dual eye tracking can be used to assess the effectiveness of collaboration in terms of learning and how it is associated with other process data, especially within an ITS.

In this paper, we will explore three types of broad questions that can be answered by using dual eye tracking: (a) How is eye gaze associated with other communication measures, (b) how is eye gaze associated with task features, and (c) how is eye gaze associated with learning outcomes. By answering these questions we may have a better understanding of how the interface of an ITS relates to both speech and the learning process while students are collaborating. There are multiple measures that can be gathered through dual eye tracking to understand eye gaze. In this paper, we will focus on one such measure, joint attention, which measures the relative amount of time two students are looking at the same area at the same time and corresponds to a very close coupling of eye gaze. Using a dataset of $4^{th}$ and $5^{th}$ grade students working on a fractions ITS, we explore a specific question in each of these three broad areas. These exploratory analyses demonstrate the potential of questions involving dual eye tracking and other data streams to be used to analyze collaborative learning.

## 2 Methods

### 2.1 Experimental Design and Procedure

Our data set involves 14 $4^{th}$ and 14 $5^{th}$ grade dyads from a larger study [10]. The dyads were engaged in a problem-solving activity in a collaborative ITS for fractions learning while communicating through audio only using Skype. Each dyad worked with the tutor for 45 minutes in a lab setting at their school. The morning before working with the tutor and the morning after working with the tutor, students were given 25 minutes to complete a pretest or posttest individually on the computer to assess their learning. Through the lab set-up in the school, we were able to collect dual eye tracking data, transcript data, and tutor log data in addition to the pretest and posttest measures for multiple stream of data.

### 2.2 Tutor Design

The ITS was developed using Cognitive Tutoring Authoring Tools and consisted of two problem sets, targeting procedural and conceptual knowledge. The tutor provides standard ITS support, such as hints and feedback [14], combined with embedded collaboration scripts. Each student had their own view of the collaborative tutor that allowed the students to have a shared problem space and synchronously work while being able to see slightly different information and to take different actions. Three different features supported the student collaboration. On some tutor steps, the students were *assigned roles* where they were either responsible for entering the answer or for asking questions of their partner and providing help with the answer. We supported other problem steps through *individual information* [13]. Here the students were each provided with a different piece of information that they needed to share with their partner. The final feature that was used to support collaboration was *cognitive group awareness* [5]. This feature was implemented in the tutor by providing each student an opportunity to answer a question individually before seeing each other's answers and being asked to provide a consensus answer.

### 2.3 Data and Dependent Measures

A computer-based test was developed to closely match the target knowledge covered in the tutors. The test comprised of 5 procedural and 6 conceptual test items, based on pilot studies with similar materials. Two isomorphic sets of questions were developed, and there were no differences in performance on the test forms, $t(79) = 0.96$, $p = 0.34$. The presentation of these forms as pretests and posttests was counterbalanced.

In addition, to pretest and posttest measures, we also collected process data including tutor log data, transcript data, and dual eye tracking data. The log data consisted of the transactions that the students took with the ITS. These include attempts at solving each step together with the request of hints and errors.

We coded the dialogue transcript data using a rating scheme with four categories: interactive dialogue, constructive dialogue, constructive monologue, and other. For our analysis, we focused on the interactive dialogue, in which students engage in actions such as co-construction and sequential construction. Interactive dialogue aligns with ICAP's joint dialogue pattern [2]. Our rating scheme was developed to look at groups of utterances associated with subgoals (i.e., a group of steps that all are for the same goal) to account for the interactions between the students. An inter-rater reliability analysis was performed to determine consistency among raters (Kappa= 0.72).

In addition to collecting log data and transcript data, we also collected dual eye tracking data using two SMI Red 250 Hz infrared eye tracking cameras. We calculated a measure of joint attention through gaze recurrence [1], [8]. Gaze recurrence is the proportion of times where the fixations are at the same location for each student. To calculate the joint attention from the gaze data, we used gaze recurrence with a distance threshold of 100 pixels to approximate the percentage of time that students were looking at the same thing at the same time. This distance threshold was chosen to align with prior research [6] and is close to the size of the interface elements.

## 3 Research Questions and Analysis

The first broad area of analysis is how eye gaze is associated with other communication measures. Within this area, we investigated how joint attention differs between subgoals without talk and subgoals with talk. We also explored whether or not there is an interaction with the subgoals that have errors. Based on previous work, we hypothesize that subgoals with talk will have a higher level of joint attention than subgoals with no talk since talk has been found to be coupled with eye gaze and speech might guide the visual attention [9]. In addition, we hypothesize that subgoals where an error occurred will have a higher level of joint attention compared to subgoals where no error occurred because there will be a visual red mark on the screen for the students to discuss [12]. To investigate the association between talk and joint attention, we used a hierarchical linear model with two nested levels to analyze how the talk during subgoals related to the joint attention. At level 1, we modeled if talk occurred and if one or more errors occurred for the subgoals. At level 2, we accounted for random dyad differences. We found no effect of errors on joint attention, so it was removed from the model. We found greater joint attention for subgoals that had talk ($M$ = 0.25, $SD$ = 0.13) versus those that did not ($M$ = 0.22, $SD$ = 0.14), $t(1705)= 12.66$, $p < .001$, showing a coupling between talk and joint attention that extends previous results to younger learners working in an ITS environment.

The second broad area of analysis is how eye gaze is associated with task features. For this area, we investigated how eye gaze is associated with the tutor's three types of collaboration support. Based on previous work, we hypothesize that subgoals supported through individual information would have the lowest joint attention since there is no joint reference for the students on the screen [7]. To investigate the association between collaboration features and joint attention, a hierarchical linear model with two nested levels was used to analyze how collaboration features relate to the joint attention. At level 1, we modeled the type of collaboration support of the subgoals along with the talk type to control for this covariate. At level 2, we accounted for random dyad differences. We found that the joint attention for subgoals that were supported through cognitive group awareness ($M$ = 0.19, $SD$ = 0.11) was lower than that for subgoals supported through roles ($M$ = 0.25, $SD$ = 0.14), $t(1705)= -4.19$, $p < .001$, indicating that task type has an impact on joint attention.

The third broad area of analysis is how eye gaze is associated with learning. Within this area, we investigated how joint attention correlates with learning gains for conceptual and procedural knowledge. Based on previous work where we analyzed the first four questions (opposed to the entire session) [1], we hypothesize that joint attention will be correlated with conceptual learning gains, but not procedural learning gains, because a deeper understanding is needed to acquire the conceptual information that can be supported through joint attention [11]. To investigate this question, we computed a linear regression between posttest score and joint attention while controlling for pretest scores. Individual pretest and posttest scores were averaged for each member of the dyad for a single score for each dyad, and the joint attention was calculated for each dyad for the entire 45-minute session. Our results replicate previous findings, where for the conceptual condition, there were no significant results for

conceptual or procedural posttest scores. For the procedural condition, there was no significance for procedural posttest scores, but joint attention significantly predicts conceptual posttest scores when controlling for conceptual pretest score, $t(10) = 2.6$, $p = 0.03$, showing joint attention may be more important for gaining conceptual knowledge on procedural problems, whereas students working on the conceptual problems were able to learn the same information with less joint attention.

## 4 Discussion

Although the correspondence of eye gaze with speech has been studied before, it is still an open question of how dual eye tracking can be used to assess the effectiveness of collaboration in terms of learning and how it is associated with other process data. In this paper, we explore the importance of eye gaze for collaborative learning analysis by presenting three different areas of analysis for using dual eye tracking data. Although the results are preliminary, these questions provide a broad structure and illustrate the potential of dual eye tracking to be used with other data streams.

Can dual eye tracking be used to understand the collaborative learning process? Through our analysis, we found that subgoals where talk occurs have a higher level of joint attention, extending previous results to younger learners and an ITS environment [12]. This result indicates that in an environment where there is step-by-step guidance and steps are revealed one at a time, which may guide eye gaze, there is still a benefit of speech for referencing items on the screen. Although we did not find any impact of errors on joint attention, analyzing the joint attention immediately after an error may provide a better indication of the effect of errors on joint attention. In addition, we found that subgoals supported through cognitive group awareness had a lower level of joint attention compared to those supported through roles showing the importance of the task features on collaboration. The difference between collaborative features on joint attention may be because the students would be looking at different points while answering individually and would then be looking at their partner's answer after it is revealed on cognitive group awareness, which may split the attention of the partners. We also used dual eye tracking to identify moments where collaboration may successfully support learning. We found joint attention as a significant predictor of conceptual posttest scores in the procedural condition, showing collaboration and joint attention may be important for conceptual knowledge specifically when it is not being directly supported. Although the results are preliminary, they show the potential of using dual eye tracking along with other data streams to better understand collaboration. For collaborative learning, dual eye tracking can provide insights into tasks that elicit collaboration as well as providing insights into how joint eye gaze interacts with other communication measures to impact learning.

For future work, we would like to expand the three areas of analysis around dual eye tracking beyond joint attention. There are other measures such as AOIs (areas of interest) analyses and gaze patterns that would be of interest in each of the three areas and can be measured through dual eye tracking. These different measures of eye gaze would not only provide additional ways of comparing collaboration within groups by

looking at AOIs and gaze patterns that occur for partners at the same time, but would also allow the comparison to students working individually to see how collaboration affects the learning process. In addition, in our analyses so far, we have analyzed joint attention at the subgoal level and the dyad level, but analysis at additional grain sizes, such as a few seconds around errors and the problem level, would allow us to ask a wider range of questions. This future work will build upon the analysis presented in this paper to further explore the three broad areas of analysis for dual eye tracking to shed light on the mechanisms of collaborative learning.

**Acknowledgments.** We thank the CTAT team, Daniel Belenky, and Amos Glenn for their help. This work was supported by Graduate Training Grant # R305B090023 and by Award # R305A120734 both from the US Department of Education (IES).

## 5 References

1. Belenky, D.M., Ringenberg, M., Olsen, J. K., Aleven, V., & Rummel, N.: Using dual eye-tracking to evaluate students' collaboration with an intelligent tutoring system for elementary-level fractions. Paper in the *36th Annual Meeting of the Cog. Sci. Society.* (2014)
2. Chi, M.T.H.: Active-constructive-interactive: a conceptual framework for differentiating learning activities. Topics in Cognitive Science, 1, 73-105 (2009)
3. Chi, M. T., & Wylie, R.: The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist*, *49*(4), 219-243 (2014)
4. Griffin, Z. M., & Bock, K.: What the eyes say about speaking. Psychological science, 11(4). (2000)
5. Janssen, J., & Bodemer, D.: Coordinated computer-supported collaborative learning: Awareness and awareness tools. Educational Psychologist, 48(1), 40-55 (2013)
6. Jermann, P., Mullins, D., Nüssli, M.-A., and Dillenbourg, P.: Collaborative Gaze Footprints: Correlates of Interaction Quality. In Proceedings of CSCL Conference 2011, 184-191 (2011)
7. Jermann, P. and Nüssli, M. A.: Effects of sharing text selections on gaze recurrence and interaction quality in a pair-programming task. In Proceedings of the 2012 ACM CSCW conference. (2012)
8. Marwan, N., Romano, M. C., Thiel, M., Kurths, J.: Recurrence Plots for the Analysis of Complex Systems, Physics Reports, 438(5-6), 237-329 (2007)
9. Meyer, A. S., Sleiderink, A. M., & Levelt, W. J. M.: Viewing and naming objects: Eye movements during noun phrase production. Cognition, 66(2), (1998)
10. Olsen, J. K., Belenky, D. M., Aleven, V., & Rummel, N.: Using an intelligent tutoring system to support collaborative as well as individual learning. In 12th International Conference on Intelligent Tutoring Systems, 134-143 (2014)
11. Richardson, D. C., & Dale, R.: Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. Cognitive Science, 29, 1045-1060 (2005)
12. Richardson, Daniel C., Dale, R., & Kirkham, N. Z.: The Art of Conversation Is Coordination. Psychological Science, 18(5), (2007). doi:10.1111/j.1467-9280.2007.01914.x
13. Slavin, R. E.: Research on cooperative learning and achievement: What we know, what we need to know. Contemporary educational psychology, 21(1), 43-69 (1996)
14. VanLehn, K.: The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Pyschologist. 46(4) 197-221 (2011)

# Workshop on Les Contes du Mariage: Should AI stay married to Ed?

*held in conjunction with*

Seventeenth International Conference on Artificial Intelligence in Education (AIED 2015)

Monday, June 22, 2015
Madrid, Spain

**Workshop Co-Chairs:**

Kaska Porayska-Pomsta,[1] Gord McCalla[2] and Benedict du Boulay[3]

[1]*UCL Institute of Education, London, WC1H 0AL, UK*
[2]*University of Saskatchewan, Saskatoon, S7N 5C9, Canada,*
[3]*University of Sussex, Brighton, BN1 9QJ, UK*

http://www.sussex.ac.uk/Users/bend/aied2015

# Table of Contents

# Preface

At its origin, the field of Artificial Intelligence in Education (AIEd) aimed to employ Artificial Intelligence techniques in the design of computer systems for learning. The 25th anniversary of the IJAIEd is a good opportunity to interrogate the aims and aspirations of the field, its past and current achievements, while the AIED conference constitutes a timely forum for such an interrogation. This workshop explores questions such as:

- What is and what should be the role of AI in Education and conversely of Education in AI? Specifically, in the early days of AIED there seemed to be lots of AI in AIED, but now AI is more often a placeholder for any kind of advanced technology.
- What is and what should be the motivation of AIEd as a field? Supporting learning has been considered a great "challenge domain" for AI in that many of the big AI questions must be answered, at least to some extent, to build a sophisticated learning environment. But, it seems that the ideas generated in AIED are neither influencing AI nor Education in any serious way. Why not?
- What is and what should be the balance of respective contributions to AIED from AI and Education as distinct fields of research and practice? Both fields have well-established methodologies and practices, but the extent to which these are cross-fertilising under AIED is not clear.
- A related question relates to the extent to which the results of AIEd research are meaningful to real educational practices? Does the community even care?
- What are the future directions for the field that could justify and maintain its unique identity? How does AIED differ from related disciplines such as Learning Sciences, ITS, and CSCL? Or are these just labels for essentially the same research discipline?

We thank the many people who contributed toward making this workshop a reality. In particular, we are grateful to the authors and to the Programme Committee, listed below:

- Vincent Aleven, Carnegie Mellon University, USA
- Nicolas Balacheff, Centre National de la Recherche Scientifique, Grenoble, France
- Sara Bernardini, King's College, London, UK
- Bill Clancey, Florida Institute for Human and Machine Cognition, USA
- Tak-Wai Chan, National Central University, Taiwan
- Pierre Dillenbourg, Swiss Federal Institute of Technology, Lausanne, Switzerland
- Vania Dimitrova, University of Leeds, UK
- Claude Frasson, University of Montreal, Canada

- Jim Greer, University of Saskatchewan, Canada
- Lewis Johnson, President and CEO, Alelo Inc., USA
- Judy Kay, University of Sydney, Australia
- Chad Lane, University of Illinois, USA
- James Lester, NC State University, USA
- Chee-Kit Looi, National Institute of Education, Singapore
- Rose Luckin, UCL Institute of Education, London, UK
- Manolis Mavrikis, UCL Institute of Education, London, UK
- Richard Noss, UCL Institute of Education, London, UK
- Helen Pain, University of Edinburgh, UK
- Elliot Soloway, University of Michigan, USA
- Julita Vassileva, University of Saskatchewan, Canada
- Beverly Woolf, University of Massachusetts, USA

Kaska Porayska-Pomsta, Gord McCalla and Benedict du Boulay
May 2015

# Whither or wither the AI of AIED?

Judy Kay

School of Information Technologies, University of Sydney, Australia
judy.kay@sydney.edu.au

**Abstract.** This position paper explores the relationship between the historic roots of AIED and the challenges of restricting our vision to EdTech that has AI. It argues that the founders of AIED had a broad vision of the field, primarily driven by the *goals* of creating advanced technology for personalised learning. They were not wedded to a techno-centric view, demanding use of particular techniques that are now thought of as "AI". The paper argues that we have accepted work with no AI, notably in Open Learner Modelling. We discourage, either directly or just because of our name, work that is true to the AIED founders' vision. In doing so, we miss many exciting and promising ways to create better technology for education**.**

## 1 What was the AI in the initial vision of AIED?

So how did we come to be called AIED in the first place? In the early days of computing research, AI had a very broad brief. It was driven by the vision that computers would one day be able to emulate the actions we describe as intelligent when people do them. What a bold vision this was --- at a time when computers were very slow, expensive and available only in research labs, military and business contexts. AI research stood in stark contrast to other the major areas of computing, such as hardware, operating systems, programming languages and numerical analysis. It was AI that looked to real world applications and creating the visions of science fiction.

AIED was born in the 1970s, with its first conferences in the 1980s (Self, 2015). It aspired to create applications that could help people learn. This was long before it was possible for most learners to even see, let alone use, a computer. A widely cited driver for our AIED research was the vision that computers could help achieve Bloom's famous 2-sigma learning benefits from *personalized teaching* by an expert teacher (Bloom, 1984). Our community is still committed to this goal. But it is useful to consider what it meant.

The classic early work in AIED identified four key elements:

- domain expertise;
- teaching expertise;
- student model; and
- user interface.

And so, the goal of researchers was to explore any or all of these architectural elements, towards building what was called an Intelligent Tutoring Systems (ITS) or AIED system. Overall, for both AIED and ITS, one key goal was to create computer systems that could provide *personalised teaching*, just as a knowledgeable teacher with expert teaching skills could do. We still aim to do this. And another goal was to support excellent *user interfaces* --- with what we may now call natural user interfaces (such as natural language and speech) and rich forms of interaction (such as graphical user interfaces that are now the norm). The spirit of their vision included creating systems and interfaces that both mimic human expert teaching and to use other techniques that are better suited to machines.

Since our early days, when the AIED community chose its name, a great deal has changed for AI, computing broadly, even for the behemoth of formal education and the commercial interests associated with those institutions and broader education. In parallel, AIED research has evolved in important ways. The next part of this paper explores these differences as a foundation for arguing that AI still has a place in AIED, but that it is not necessary for the still worthy and, as yet, unreached core vision of our founders.

## 2 How has AI changed since the birth and naming of AIED?

AI has become mainstream in the sense that it is part of the technology that each of us uses each day. This is well illustrated in the following descriptions from the EdX Introduction to AI[1].

Artificial intelligence is already all around you, from web search to video games. AI methods ***plan*** your driving directions, ***filter*** your spam, and ***focus your cameras on***

[1] https://www.edx.org/course/artificial-intelligence-uc-berkeleyx-cs188-1x-0#.VQzBWUaI0k4

***faces***. AI lets you guide your phone with your ***voice*** and ***read foreign newspapers in English***. Beyond today's applications, AI is at the core of many new technologies that will shape our future. From self-driving cars to household robots, advancements in AI help transform science fiction into real systems.

I have added the bold font to highlight the sampler of technical areas alluded to: planning, filtering, vision, natural language translation. AI has been so successful that it has resulted in many off-the-shelf tools for these tasks, and for many other core AI tasks. AI has also changed from its focus on deep reasoning to large-scale statistical methods. This partly reflects the huge drop in the cost of memory and processing, along with the availability of networking. So, for example, an area like natural language translation has shifted from an early focus on user modeling and deep reasoning to statistical techniques for machine learning that makes use of large corpus data, particularly text which occurs naturally in online materials such as books, newspapers, social media sites, Wikipedia…. Where early work often involved complex reasoning, now it is possible, and sensible, to explore far simpler methods that harness huge amounts of data to achieve more robust and practical systems.

AI has earned a place as part of a standard computing undergraduate degree. Similarly, some other core areas of the computing syllabus include databases, HCI, software engineering, graphics. Such areas have now established a substantial collection of techniques that belong in the computing professional's toolkit. All of these, not just AI methods, should be used to achieve the core goals of AIED.

AI has achieved much in its long history, often resulting in new communities that are more problem-, rather than technique-focused. For example, robotics researchers have their own publication venues; while they may also publish in AI venues when they create a new contribution to the body of knowledge in AI, their core goals are to create effective robots. High impact research may be based on new ways to make effective use of *existing* software tools for AI, database, graphical, language, vision systems…. Similarly, separate communities have emerged in areas that are central to the AIED vision of effective interfaces, notably natural language generation and understanding and systems based on vision and depth sensing to provide NUI, natural user interaction. This offers support for learning away from the desktop. It

opens possibilities for just-in-time learning, teachable moments and kinesthetic interaction that can be valuable for learning.

*In summary, AI is pervasive and it is just one, of many, software tools that AIED researchers should draw upon to create the future of technology to enhance learning and education.*

## 3 How has education changed since the birth and naming of AIED?

Over the history of AIED, computing has changed radically. Every potential learner in the developed world now has easy access to many forms of computers in their daily lives. And they will have many more, including personal devices, wearables, mobiles, portables and desktops and well as embedded systems such as interactive tables and walls and smart environments. The interface will have input modalities that include natural language, speech, gaze and gestures as well as keyboard and mouse. Diverse sensors will provide indirect input, such as eye-tracking, mood detectors and activity trackers. Even in the developing world, there is increasing availability of personal technology, particularly mobile phones.

This explosion of computing devices has finally begun to have a deep impact on education, both formal and informal. Our educational institutions make extensive use of computers. Those uses range from core productivity tools, through to tools for particular disciplines as well as personalized and collaborative learning tools. They link the formal and informal, for life-wide learning support.

This has seen the emergence of communities that follow the AIED founder vision for using technology to enhance education. One recent example has seen the emergence of the *Learning Analytics* community. They represent the mainstream of education exploring ways to harness data from even administrative tools (such as those used to capture details of student demographics) and certainly for widespread learning tools, such as Learner Management Systems.

Another emerging example, this time for lifelong, life-wide learning is due to *sensor technology.* For example, wearable activity trackers can be viewed as a valuable data source to an AIED system. They are a form of the interface element, just as surely as a keyboard, drawing tablet or spoken input is. Such sensors can play a key role for personalized teaching, such as interfaces to help people set effective goals and plans, self-monitor progress on these, discover which personal strategies are effective for achieving goals and to learn about new strategies.

Yet another recent EdTech innovation is the *MOOC*. This is exciting on several levels. MOOCs offer the possibility for a very broad population of learners to have access to high quality personalised learning opportunities. MOOC platforms emerged from the elite computer science research world. This is striking as computer scientists, with outstanding expertise in diverse areas of computing, have so clearly committed to creating innovative teaching systems. MOOCs provide exciting green fields for EDM and for translating our years of AIED research into widely used software systems.

These illustrate just three of many trends that matter for AIED. They are pervasive and have high impact. All are currently outside the core of what some members of our community see as AIED. There is a real risk that a paper reporting any of these would be rejected for lacking AI. And authors may assume this, and submit such work elsewhere. Yet all three do offer personalized learning, as the term is described in the broader community. All have data about learners and it is widely recognized that this data is important for informing the learning. Should we call that data a learner model? Why not? Do those communities consider it a learner model? Probably not. Should we object to calling such data a learner model representation just because it is simple by AI standards, rather than complex. Surely these classes of EdTech are within the scope of the vision of the AIED founders.

## 4 How has AIED changed? And not? Personal case studies.

The last section suggested that AIED has not changed enough to keep up with the dramatic shifts in the real world of education. This section explores some of the ways that the AIED community has already made steps towards accepting research that has little or no AI. There have

been AIED papers dealing with essentially the software engineering aspects of sophisticated AI systems. For example, these include the creation of interfaces to make it easier for non-technical users to design and modify the teaching in a complex AIED system; such work tackles the problem that an AIED system needs a better user-friendly interface.

But there has also been work that has no element of AI at all. Lest I risk offending others, I illustrate this in terms of my own work that has been published in AIED and ITS venues but does not have AI. As a young researcher, I was excited at the AIED vision of creating personalized teaching system. I concluded that a key is the learner model because it drives the personalisation, based on its data about the learner. But I was also committed to treating the learner model as the personal data of the learner and to respect the asymmetry in the relationship that should exits between a person and a machine, where the person should be able to maintain a sense of control.

This focus led me to work on creating learner models that respected the learner's right to *control* their own data, to help the learner to be *responsible* for their own learning. As a foundation for learner control, I concluded that it was important to create learner modeling middleware that was designed, from its foundations, to enable the learner to *scrutinize* the *learner model* and the associated *personalization processes*. Issues of personal data privacy are not mainstream AI concerns. But they are important for real world deployments. This is reflected in the 2012 workshop by leaders of the MOOC community, resulting in the Asilomar Convention for Learning Research in Higher Education[2]. While the philosophical standpoint of learner control was a key driver for my research, there are also more pragmatic aspects. One relates to the deeply fallible process of learner modeling. Since the data about learners is generally noisy, unreliable and incomplete, I wanted to create interfaces to the learner model, Open Learner Models (OLMs), that enabled the learner to see their model and how teaching applications interpret and use it. This could enable them to correct it. They could also alter it in other ways if they wished to introduce incorrect data. (The underlying representation avoids this from corrupting the model, and supports multiple views and interpretations of the model). That work was accepted by the AIED and ITS communities, as evidenced by

[2] http://asilomar-highered.info/

publications, such as Kay (2000; 2000a), Kay & Lum (2005) and Czarkowski et al (2005). The learner model representations in that work did not require, or make use of, sophisticated AI.

Concerns for *systems* aspects led to my work on user and learner model servers. This is important for practical systems, but it is not AI (Kay et al 2002; Brusilovsky, 2003; Brusilovsky et al, 2005; Assad et al, 2007; Kay and Kummerfeld, 2012). Designing OLM interfaces is essentially HCI, with a strong focus on user-centred design, rather than AI. The challenge of building systems that work effectively also makes it desirable to create the simplest technical solution that is effective, in that it achieves the intended task. This is good software engineering, good sense and also an excellent foundation for creating OLM interfaces that are simple enough make the model understandable and scrutable. In line with the view of learning data as belonging to the user and under their control, even my earliest implementations of the learner model placed it outside any single application (Kay 1994). The move to learner model servers (Kay et al, 2002) continued the move towards a cloud-based independent learner model as a first class citizen (Kay 2008; Bull and Kay 2010). None of these concerns are AI.

Learner models are clearly core to AIED; they are one of the four elements of personalized teaching. Papers on OLMs have been published in our journal and conferences, as reviewed by Desmarais and Baker (2012). Some have used sophisticated AIED representations, such as cognitive and constraint-based models and Bayesian nets. However, my own work, and key work by other prominent OLM researchers has typically had rather simple learner models. There was no need for complex AI techniques. The defining characteristic of an OLM is that it provides an interface onto a data structure where both were explicitly designed to provide a view of the learner model that would be useful to the learner.

A foundation for designing a learner model is the definition of the domain ontology and the processes to transform learning data into inferences about that learning ontology. In my work, it could more accurately be described as defining the curriculum in terms of the learning objectives. Then the inference is essentially a mapping from learning data onto that curriculum, using the simplest effective interpretation. While

some reviewers have criticized some of this work for the lack of AI, they have never explained why a more complex AI approach would be useful or how such modest and simple approaches are inadequate to the task. Nor have they argued the work is not useful. I believe that OLM research is true to the aspirations of the founders of the AIED community, even if it has no element of what is currently AI.

While OLM research is accepted in AIED, my other current research involves creating interfaces for surface computing, with large screen interactive tabletops and walls. This is exciting stuff. Some of it has made it into AIED venues (Martinez-Maldonado et al, 2011, 2012, 2013, 2014). This work used the data from small group interaction at a tabletop to model the effectiveness of collaboration. This used EDM methods to interpret the raw data, to distinguish more, and less, effective collaboration in groups of students. We trialled that work in a lab setting. However, when we moved into the wild, with real classrooms and real teachers, the actual demands of the classroom called for far simpler learner models. For this real world context, we took the same digital footprints of the learners, but this time presented them in very simple OLMs (Martinez-Maldonado et al, 2012, 2014). That was what met the teacher's needs; it did not have or need AI for the core of the research. Some of it seemed to have enough AI or OLM content to make to our conferences, much did not.

In summary, the publications of the AIED community already include some research that provides innovative teaching systems but does not need AI and reports none. But we still exclude other interesting and innovative work, or authors self-exclude it.

## 5 Summary

This position paper has argued that the foundation vision for AIED was to create personalised learning systems, with highly effective interfaces, and that this vision is still relevant to the AIED community. There is much that remains to be done if we are to create the four core components of AIED architectures. But over the last 25 years, AI has changed, as has education and EdTech. We run the real risk of being left behind some of the most exciting and novel directions if we insist on restrict-

ing our research to systems that create or use AI, as it is understood today.

**References**


1. Assad, M., Carmichael, D. J., Kay, J., & Kummerfeld, B. (2007). PersonisAD: Distributed, active, scrutable model framework for context-aware services. In *Pervasive Computing* (pp. 55-72). Springer Berlin Heidelberg.
2. Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher*, 4-16.
3. Brusilovsky, P. (2003). A component-based distributed architecture for adaptive webbased education. In *Artificial intelligence in education* (pp. 386-388).
4. Brusilovsky, P., Sosnovsky, S., & Shcherbinina, O. (2005). User modeling in a distributed e-learning architecture. In *User Modeling 2005* (pp. 387-391). Springer Berlin Heidelberg.
5. Bull, S., & Kay, J. (2007). Student models that invite the learner in: The SMILI:() Open learner modelling framework. *International Journal of Artificial Intelligence in Education*, *17*(2), 89-120.
6. Bull, S., & Kay, J. (2010). Open learner models. In *Advances in intelligent tutoring systems* (pp. 301-322). Springer Berlin Heidelberg.
7. Czarkowski, M., Kay, J., & Potts, S. (2005). Scrutability as a core interface element. In *Proceedings of the 2005 conference on Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology* (pp. 783-785). IOS Press.
8. Desmarais, M. C., & Baker, R. S. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, *22*(1-2), 9-38.
9. Kay, J. (1994). The um toolkit for cooperative user modelling. *User Modeling and User-Adapted Interaction*, *4*(3), 149-196.
10. Kay, J. (2000). Stereotypes, student models and scrutability. In *Intelligent Tutoring Systems* (pp. 19-30). Springer Berlin Heidelberg.
11. Kay, J. (2000a). Accretion representation for scrutable student modelling. In *Intelligent Tutoring Systems* (pp. 514-523). Springer Berlin Heidelberg.
12. Kay, J. (2008). Lifelong learner modeling for lifelong personalized pervasive learning. *Learning Technologies, IEEE Transactions on*, *1*(4), 215-228.
13. Kay, J., Kummerfeld, B., & Lauder, P. (2002). Personis: a server for user models. In *Adaptive Hypermedia and Adaptive Web-Based Systems* (pp. 203-212). Springer Berlin Heidelberg.
14. Kay, J., & Kummerfeld, B. (2012). Creating personalized systems that people can scrutinize and control: Drivers, principles and experience. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *2*(4), 24.
15. Kay, J., & Lum, A. (2005). Exploiting Readily Available Web Data for Scrutable Student Models. In *AIED* (pp. 338-345).

16. Martinez, R., Wallace, J. R., Kay, J., & Yacef, K. (2011). Modelling and identifying collaborative situations in a collocated multi-display groupware setting. In *Artificial Intelligence in Education* (pp. 196-204). Springer Berlin Heidelberg.
17. Maldonado, R. M., Kay, J., Yacef, K., & Schwendimann, B. (2012). An interactive teacher's dashboard for monitoring groups in a multi-tabletop learning environment. In *Intelligent Tutoring Systems* (pp. 482-492). Springer Berlin Heidelberg.
18. Martinez-Maldonado, R., Kay, J., & Yacef, K. (2013). An automatic approach for mining patterns of collaboration around an interactive tabletop. In *Artificial Intelligence in Education* (pp. 101-110). Springer Berlin Heidelberg.
19. Martinez-Maldonado, R., Clayphan, A., Yacef, K., & Kay, J. (2014). Towards Providing Notifications to Enhance Teacher's Awareness in the Classroom. In *Intelligent Tutoring Systems* (pp. 510-515). Springer International Publishing.
20. Self, J. (2015) The birth of IJAIED, International Journal of Artificial Intelligence in Education. To appear.

# A is for Adaptivity, but What is Adaptivity? Re-Defining the Field of AIED

Vincent Aleven

Human-Computer Interaction Institute
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213

aleven@cs.cmu.edu

**Abstract** This paper proposes to define the field currently known as AIED not in terms of the technology used, but in terms of system behavior. Specifically, it is proposed that AIED is the science and engineering of systems that adapt to learners, so as to help bring about effective, efficient, and enjoyable learning experiences. But what, in general, is adaptivity? Intuitively, being adaptive means that the system adjusts the course of instruction in nuanced and effective ways based on learner differences, for example the goals and needs of individual learners and group of learners. It is difficult to state necessary and sufficient conditions for the concept of adaptivity. Instead, I stipulate that a system is more adaptive to the degree that: (a) its design is grounded in a thorough (empirical) understanding of learners in the given task domain, (b) it is appropriately interactive, and (c) it takes into account, in its pedagogical decision making, how individual learners measure up along different psychological dimensions. These factors help in comparing systems in terms of their degree of adaptivity. They imply that the presence of Artificial Intelligence technology is not a defining factor, even if it can be (and often is) instrumental in bringing about adaptivity.

## Introduction

How we define our field (currently called AIED) influences how we position it vis-à-vis other efforts to create learning technologies. This positioning is not merely academic. It may influence public perception and acceptance of our technologies. For example, it may influence how MOOC developers see the need for AIED technology in their courses, and may influence how the technology is accepted and spreads. Experts do not agree about how to define AIED or (relatedly) intelligent tutoring systems [25, p. 21], so the issue is not straightforward. How can we define our field in a way that is inclusive and honors its interdisciplinary nature, while also honoring the range of technologies that are typically being applied, whether AI technologies or not?

As with all educational technology, the goal of our field is to develop a science and practice for the design and implementation of technologies that can support effective, efficient, and pleasurable learning experiences for learners, groups of learners, instructors, and other stakeholders in the educational process. What sets our field apart is that we strive to make our systems "intelligent" or "adaptive," so as to be highly effective with a very wide range of learners. But what do these terms mean? Although the notion of intelligent and adaptive educational technologies is ill-defined, a shared intuition among researchers and practitioners may be that in order to be considered adaptive, a system must be sensitive to important learner differences; a system must have a nuanced way of deciding what, for a given learner or team of learners in a given situation, might be the best way of supporting them, given their learning history and learning goals. Such systems "understand learners" or, more broadly, "care," as John Self famously argued [18].

Artificial Intelligence (AI) can often contribute to creating such systems. It has brought to our field a focus on representation and reasoning, and has highlighted modeling and investigations into the nature of knowledge as a key emphasis in the early days of AIED and intelligent tutoring systems (e.g., [19, 24]). Nonetheless, in my opinion, our field cannot and should not be defined in terms of whether the system has AI or not. One problem is that AI is an ill-defined concept – so it would merely be replacing one ill-defined concept ("adaptive learning technologies") with another. More importantly, AI is neither necessary nor sufficient in order for learning technologies to be adaptive. The use of AI does not in and of itself make a system adaptive in a manner that supports learners effectively. Conversely, not all systems that are adaptive use AI. Also, defining our technologies in terms of the underlying technology seems fundamentally to be barking up the wrong tree. What matters is how learning is supported and whether learning is supported effectively. This viewpoint implies a focus on the behavior of systems [22] much more so than the underlying technology. The question whether AI to stay married to Ed is an interesting one. Perhaps this marriage, which started out so interestingly, needs to now become an open marriage. Better yet, perhaps it needs to be reconceptualized, replaced with a broader, more productive vision, with a renewal of the vows! Definitely, AI should and will remain a central aspect of what we do but it should not be the defining characteristic.

## Intuitively, What is Adaptivity?

Proposing that adaptivity should be the defining characteristic of AIED system begs the question, what is adaptivity? Intuitively, we assume that learners differ along (possibly) many dimensions (e.g., prior knowledge, affect, self-regulated learning skills) and that, all else being equal, instruction that takes these differences into account tends to be more effective than instruction that treats all learners as the same. Adaptivity is not binary, something a learning environment either has or does not have. Adaptivity is a matter of degree. Below I offer a more formal definition of adaptivity, first presented in Aleven, Beal, and Graesser [4]. The discussion in the current paper discussion in a paper currently under review [6], although it also broadens and

elaborates that discussion. Before I do so, perhaps it helps to get some obvious examples and non-examples on the table. We can then look at more borderline cases and offer a general definition for what it means for a system to be adaptive.

Obvious (i.e., non-controversial) non-examples of adaptive learning technologies are for example textbook problems with final answers to each problem in the back of the book, especially when every student in the same class is assigned the same problems. Other examples that are probably not controversial are online text, lectures with Powerpoint slides, video lectures of famous professors, and documentaries. I am not claiming that these types of instructional material have no place in the educational process [14, 17]. They very well may but they seem to lack adaptivity.

An obvious example of an adaptive learning technology may be an intelligent tutoring system, but what is it that makes it adaptive? A typical answer from our field may be, a rich student model with many student-related variables (knowledge, affect, metacognition, motivation, social factors), updated in real-time, in a sophisticated manner, inferring the unobservable from the observable, and used in sophisticated pedagogical decision making at multiple levels. Each learner or team of learners gets the instruction that is most effective, efficient, or pleasurable for them. Instructional decisions are always based on nuanced, fully up-to-date information.

It may be relevant also to point out that in many discussions about MOOCs and e-learning, a very low bar is used when talking about personalization or adaptivity. For example, Daphne Koller, one of the Coursera co-founders, in her Ted Talk (https://www.youtube.com/watch?v=U6FvJ6jMGHU), hails the ability to provide an error-specific feedback message (on an error discovered through data mining) as an important aspect of personalization of instruction in MOOCs. Further, in a widely-used learning management system such as Moodle (https://moodle.org/) [16], even simple branching structures are considered to be adaptive forms of instruction, in contrast to the intuitions of many ITS researchers.

## A Somewhat Unsatisfactory Way to Define Adaptivity?

Let me now examine a prior proposed definition of our concept of interest. The argument has been put forward that a key criterion for adaptivity in learning technologies is that the system has an inner loop [22], meaning that it provides step-level guidance during complex, multi-step problem solving or dialogues. This form of guidance is to be contrasted with answer-level guidance, in which feedback is provided only at the end of each problem. In his 2006 paper, VanLehn views the presence of an inner loop as a defining criterion for intelligent tutoring systems: "Systems that lack an inner loop are generally called Computer-Aided Instruction (CAI), Computer-Based Training (CBT) or Web-Based Homework (WBH). Systems that do have an inner loop are called Intelligent Tutoring Systems (ITS)" [22, p. 233]. In a later article [21], however, he seemed to back off: "Most intelligent tutoring systems have step-based or substep-based granularities of interaction, whereas *most other tutoring systems* [emphasis added] (often called CAI, CBT, or CAL systems) have answer-based user interfaces." Importantly, he points out that systems that provide step-based tutor-

ing tend to have a stronger positive effect on student learning outcomes, compared to no tutoring conditions (i.e., a greater effect size) than systems that provide answer-based tutoring (i.e., do not have an inner loop). VanLehn's definition is attractive in many ways: It emphasizes adaptive behavior as a hallmark of intelligence, which seems right to us. It avoids debates about system architectures or about the thorny question, what is AI? It aligns with key empirical evidence. On the other hand, it is not without its shortcomings, reason perhaps that VanLehn seems to have backed off. Step-based guidance may not be very adaptive if the tutor can only recognize one particular set of steps through each problem. Also, certain desirable forms of adaptivity may not easily be viewed as step-level support (e.g., reacting to student affect or adaptive selection of problems in the system's outer loop). Also, some systems that are commonly considered intelligent or adaptive have rather minimal inner loops such as ASSISTments [12], Wayang Outpost/Mathsprings [9], and Hint Factory tutors [20]. These systems all have a legitimate claim to being adaptive and intelligent. ASSISTments and Wayang Outpost/Mathsprings may not have an elaborate inner loop, but they have other features, such as being designed with a fundamental and sound understanding of student learning. Also, Wayang Outpost in its outer loop adapts to student metacognition and affect in certain ways. Similarly, Hint Factory tutors do not have on-board intelligence, yet behave like an intelligent tutor because of the next-step hint capability.

In this discussion, it is interesting to consider the degree to which specific forms of adaptivity are supported by empirical investigations (e.g., task analysis) and/or rigorous research. For example, step-level feedback and cognitive mastery are strongly supported in the empirical ITS literature, as enhancing student learning [7, 8, 11, 15]. Although the ability to support multiple student strategies within a given problem is widely viewed as desirable, the only study I know that tested this assumption did not find evidence to support it [23].

## Adaptivity: A Proposed Definition

Given these considerations, let me now highlight an alternative definition of adaptivity, first presented in a recent article by Aleven, Beal, and Graesser [4], who listed three key elements of advanced learning technologies. For purposes of the current discussion, we can take this term to be synonymous with AIED; the key elements can therefore be viewed of key elements of the kind of adaptivity or intelligence we would like to see in our smart systems for education.

> "Although defining ALTs (advanced learning technologies) is difficult, ALTs have 3 key elements to varying degrees:
>
> - First, these technologies are created by designers who have a substantial theoretical and empirical understanding of learners, learning, and the targeted subject matter.
> - Second, these systems provide a high degree of interactivity, reflecting a view of learning as a complex, constructive

activity on the part of learners that can be enhanced with detailed, adaptive guidance.

- Third, the system is capable of assessing learners, while they use the system, along different psychological dimensions, such as mastery of the targeted domain knowledge, application of learning strategies, and experiences of affective states. On the basis of these assessments, the systems make pedagogical decisions that attempt to adapt to the needs of individual learners."

This definition lists factors, rather than necessary and sufficient conditions, thus acknowledging that adaptivity is an open-textured concept, that is, a concept whose meaning needs to be interpreted as we go, perhaps on a case-by-case basis, and perhaps with a shift in meaning over time, as our field evolves and develops new and innovative forms of instructional support. Listing factors helps with defining the concept flexibly in a way that enables us to talk about degrees of adaptivity, rather than view it as binary. It is interesting to point out, further, that these elements are technology-agnostic; no specific technologies are mentioned or assumed. It is reasonable to think that the second and third key elements (interactivity with detailed guidance based on learner variables assessed by the system) will often involve AI technology. AI might be a particularly good match, given its emphasis on knowledge representation, reasoning, and problem solving, its concerns with diagnostic processes needed to infer and update learner models, and its concern with the nature of knowledge to be learned (e.g., [24]). Nonetheless, AI cannot be the one defining ingredient of what makes our systems adaptive.

On a personal note, this definition marks an expected return to a central theme of my dissertation, which dealt with a tutoring system, CATO, for case-based legal argumentation, a quintessential ill-defined task domain [1, 2, 3]. CATO was designed to help beginning law students learn skills of argument by analogy, a common form of argument in the legal domain. That is, this work addresses debates about whether a given new case (a problem situation about which a legal claim has arisen) properly belongs to an open-textured category, which, as in our current discussion, was defined by factors, rather than necessary and sufficient conditions. A key mode of analyzing, exploring, and arguing is to compare the new case to carefully selected past cases with favorable and unfavorable decisions [10], with the factors functioning as key dimensions of comparison. In the legal domain, comparisons with past cases that have been authoritatively classified often bring substantial clarity, although not often provably correct answers. And so it is with our question of what it means for a learning environment to be adaptive, although with an interesting twist: Our own domain lacks authoritative classifications; we do not have a supreme arbiter of whether systems are officially AIED systems or not (nor, of course, should we strive to have such an arbiter). We do have paradigm cases, however, landmark intelligent tutoring systems and even the hypothetical intelligent tutoring system sketched above. These systems can play an important role as anchors in enlightened discussions about the foundations of our field.

## Element 1: Design Based on an Empirically-Grounded Understanding of Learners

Perhaps it helps to elaborate on each of the three key elements (or factors). Interestingly, the first element (i.e., the requirement that the designers have "a substantial theoretical and empirical understanding of learners, learning, and the targeted subject matter") relates to the *design* of the system, not to system features or techniques/methods/algorithms under the hood. (The discussion of this factor is informed by debates I have had with my colleague Ken Koedinger.) This requirement could be met in many different ways. Specific to the concerns of the field of AIED, the first part of this definition emphasizes (implicitly) the use of cognitive task analysis and educational data mining to guide system design or redesign, development, and cyclical improvement. A particularly attractive scenario is that the designers carry out cognitive task analysis activities up front to study learners' ways of thinking in the given domain including their strategies and informal shortcuts, but also including the specific conceptual and procedural difficulties they experience. This scenario continues with the data-driven refinement of the system, preferably in ways in which the overall effectiveness of the system, in terms of out-of-system transfer of learning outcomes, preparation for future learning, learner (and instructor) satisfaction, and so forth, is continuously assessed, so that improvement from cycle to cycle is clearly visible. It may be clear that this vision fits particularly well with the current emphasis of big data in education. The fields of EDM and AIED can be at the forefront of this movement (see, e.g., [5]).

A somewhat different way of thinking about this requirement may be that the project team has specialists in a variety of fields, not just technology experts but also researchers in relevant branches of psychology, in education in the given subject area (e.g., math education, science education, legal reasoning, and so forth), as well practitioners.

This first factor implies a substantial broadening of how we think about adaptivity, compared for example to the intuitive notion discussed above and more generally, compared to how we, as a field, have construed the notion of adaptivity up until now. It raises the possibility of considering the design of systems, including even the choice of problems sets and detailed learning objectives, as part of what makes a system adaptive. It may even make it possible to see a modicum of adaptivity in some of our prime examples of instructional materials previously considered as non-controversial non-examples, such video lectures. When designed to target known challenges in learning, they meet the first factor, the more so when based on extensive empirical investigations of what is hard for learners to learn. They would however not be strong examples, as they would not meet the second and third factors.

## Element 2: Interactivity

The second requirement for adaptivity is that a system supports a high degree of interactivity, to provide guidance in complex and constructive learning activities. I do

not mean to say that more interactivity is always better; rather, in emphasizing the adaptive nature of the guidance that the system gives, the system is capable of providing an appropriate amount of guidance for the given learner(s) at the specific junction in their learning process. How much guidance is appropriate at what stages of learning is an interesting question [13].

The second factor was included to help capture the emphasis that our field places on constructive learning activities and on learning by doing, rather than learning by (merely) reading, watching or listening. An interesting data mining study of data from a psychology online course suggests that learning by doing yields six times greater learning than reading online text in the course or watching the video lectures [14]. A clear cut case of the second factor would be an intelligent tutoring systems with detailed guidance in their inner loop, even if we do not consider the presence of an inner loop as a defining characteristic. I do not mean to rule out systems or projects that focus on enhancing reading, watching, or listening by means of interactive support for comprehension or metacognitive strategies, for example. The second factor was included partly to help rule out (or at least, help view as low on the adaptivity scale) the non-controversial non-examples listed above, such as fixed problem sets with only answer-level feedback in the back of the book, or long video lectures without interactive activities

Frankly, this second factor is the factor that I am the least sanguine about; it may be somewhat redundant with the third factor, and it is difficult to view interactivity per se as a good thing, contributing to learning. Then again, discussions around the notion of interactivity are interesting, as long as the discussants are mindful that it is not interactivity per se that matters, but how it supports learning or other desirable educational outcomes. Further, this factor highlights an important connection, namely, that of our field with the broader field of human-computer interaction.

## Element 3 –Change Instruction Depending on Learner Differences

The third requirement, as mentioned, is that the system in its pedagogical decisions takes into account that learners differ and that the same learner is not the same for very long; learners change as they learn. For example, different learners have different prior knowledge, may experience different affect during a given learning task, tend to have different goals for learning the material, and may differ in how they regulate their own learning. In collaborative learning situations, learners may have different collaboration skills and social skills; they may be a good or a poor match regarding prior knowledge or personality, and so forth. A system should be considered as more adaptive to the extent that it adjusts its instruction, both in the inner and outer loop, based on these learner variables and perhaps others.

This requirement is consistent with Woolf's emphasis on a system having a student model and using it to adapt instruction [25], traditionally viewed as a hallmark of "intelligent" tutoring. The system builds up and maintains a student model by continuously assessing learners along various psychological dimensions (cognitive, metacognitive, motivational, and so forth). This student model is then used as the basis for

individualization. Perhaps the requirement that it is the system doing the assessing is too stringent. Perhaps the viewpoint that what is being assessed is the learner is too stringent as well. Alternative viewpoints would be that a group of learners is being assessed or perhaps that the system interprets the situation more than the learner(s) or group of learners, if that distinction makes sense (it may not). I do not mean to argue, however, that we define our field in terms of whether or not systems have a student model. That is, I do not mean to equate AIED with the field of UMAP. For example, it is conceivable that a system could be strong with respect to the first two factors but not the third and be generally accepted as belonging to the field of AIED.

There are many interesting open questions regarding how systems (as well learning environments not strongly supported by technologies) *should* adapt to learners and which learner variables (or learner group variables) are most important in this regard. In my opinion, our field is uniquely positioned to extend the science of how instruction should adapt to individual differences. Of the three factors, the third reflects most clearly how we have traditionally viewed our field.

## Final Remarks

In closing, it may be worth re-iterating that the proposed definition of adaptivity does not place emphasis on particular technologies; rather, it emphasizes the behavior of systems, much in line with VanLehn's seminal 2006 article [22] and also in line with the Turing test as a behavioral test of intelligence. Another attractive property of this definition is that also honors the interdisciplinary foundations of our field. In my view, AIED was never only about technology (CS/AI, computational linguistics, and so forth); its strength has always been that it included people and methodologies from different fields, such as human-computer interaction, psychology (cognitive, educational, developmental, social), education, design, statistics, and so forth. The field and its methodologies are interdisciplinary. Empirical evaluation of systems building has always been highly valued in our field, sometimes even to a fault (e.g., when interesting new technology developments were not given air time at conferences before there are proven results). The emphasis on high-quality empirical work is enormously important toward the goal of creating a science for the design and implementation of technologies that can support effective, efficient, and pleasurable learning experiences for a wide range of learners.

An implication of the proposed definition is that reviewer comments that "there is no AI in the system" or "the work does not push the envelop in terms of AI algorithms applied to education" should be a thing of the past. Instead, reviewer feedback should refer to the factors listed above: systems not being designed with deep insight into learning and learners' difficulties, not being interactive, and not being able to react in nuanced ways that make learning better.

The way for AI to stay married to Ed is perhaps not to declare it an open marriage, but rather, to re-define the marriage so it is appropriately broad and open-ended, a way of renewing the vows. We hope that the thoughts offered in this paper can be helpful.

Finally, what's in a name? A lot, I would argue. Our name reflects how we view ourselves, and in turn, how the rest of the world views us. Our current name honors AI as a central component as we do. I would much prefer that the disciplinary diversity and focus on behavior of systems be central. How about:

AIED = Adaptive Instruction: Evaluation and Design?

Or, if we are willing to tolerate AIEDD, how about:

AIEDD = Adaptive Instruction: Evaluation, Development, and Design

## Acknowledgments

This paper benefited from discussions with my colleague Ken Koedinger as well as from the helpful encouragement and pushback from the anonymous reviewers.

## References

1. Aleven V.: Using background knowledge in case-based legal reasoning: A computational model and an intelligent learning environment. Artificial Intelligence 150, 183-237 (2003)
2. Aleven V.: An intelligent learning environment for case-based argumentation. Technology, Instruction, Cognition, and Learning 4, 191-241 (2006)
3. Aleven V., Ashley K.D.: Teaching case-based argumentation through a model and examples empirical evaluation of an intelligent learning environment. In: du Boulay B., Mizoguchi R. (eds.) Artificial Intelligence in Education, Proceedings of AI-ED 97, pp. 87-94. IOS Press, Amsterdam (1997)
4. Aleven V., Beal C.R., Graesser A.C.: Introduction to the special issue on advanced learning technologies. Journal of Educational Psychology 105, 929-931 (2013, Nov)
5. Aleven V., Koedinger K.R.: Knowledge component approaches to learner modeling. In: Sottilare R., Graesser A., Hu X., Holden H. (eds.) Design recommendations for adaptive intelligent tutoring systems, pp. 165-182. US Army Research Laboratory, Orlando, FL (2013)
6. Aleven V., McLaren B.M., Sewall J., Popescu O., et al: Towards tutoring at scale: Reflections on "A new paradigm for intelligent tutoring systems: Example-Tracing tutors". International Journal of Artificial Intelligence in Education (under review)
7. Anderson J.R., Conrad F.G., Corbett A.T.: Skill acquisition and the LISP tutor. Cognitive Science 13, 467 - 505 (1989)
8. Arroyo I., Beck J., Woolf B.P., Beal C.R., Schultz K.: Macroadapting AnimalWatch to gender and cognitive differnces with respect to hint interactivity and symbolism. In: Gauthier G., Frasson C., VanLehn K. (eds.) Proceedings of the 5th International Conference on Intelligent Tutoring Systems (ITS 2000), pp. 574-583. Springer Verlag, Berlin (2000)
9. Arroyo I., Woolf B.P., Burleson W., Muldner K., et al: A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. International Journal of Artificial Intelligence in Education 24, 387-426 (2014, Dec)

10. Ashley KD. Modeling legal argument: Reasoning with cases and hypotheticals. Cambridge, MA: MIT press; 1991.
11. Corbett A., McLaughlin M., Scarpinatto K.C.: Modeling student knowledge: Cognitive tutors in high school and college. User Modeling and User-Adapted Interaction 10, 81-108 (2000)
12. Heffernan N.T., Heffernan C.L.: The assistments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. International Journal of Artificial Intelligence in Education 24, 470-497 (2014, Dec)
13. Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. Educational Psychology Review, 19(3), 239-264.
14. Koedinger K.R., Kim J., Jia J.Z., McLaughlin E.A., Bier N.L.: Learning is not a spectator sport: Doing is better than watching for learning from a MOOC. In: Kiczales G., Russell D.M., Woolf B. (eds.) Proceedings of the Second (2015) ACM Conference on Learning @ Scale, pp. 111-120. ACM, New York (2015)
15. McKendree J.: Effective feedback content for tutoring complex skills. Human-Computer. Interaction 5, 381-413 (1990, Dec)
16. Rice W. Moodle 2.0 e-learning course development a complete guide to successful learning using moodle 2011:
17. Schwartz D.L., Bransford J.D.: A time for telling. Cognition and Instruction 16, 475-5223 (1998)
18. Self J.: The defining characteristics of intelligent tutoring systems research: ITSs care, precisely. International Journal of Artificial Intelligence in Education 10, 350-364 (1998)
19. Sleeman D, Brown JS. London: Academic Press; 1982.
20. Stamper J., Eagle M., Barnes T., Croy M.: Experimental evaluation of automatic hint generation for a logic tutor. International Journal of Artificial Intelligence in Education 22, 3-17 (2013)
21. VanLehn K.: The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist 46, 197-221 (2011)
22. VanLehn K.: The behavior of tutoring systems. International Journal of Artificial Intelligence in Education 16, 227-265 (2006)
23. Waalkens M., Aleven V., Taatgen N.: Does supporting multiple student strategies lead to greater learning and motivation? Investigating a source of complexity in the architecture of intelligent tutoring systems. Computers & Education 60, 159 - 171 (2013)
24. Wenger E. Artificial intelligence and tutoring systems: Computational and cognitive approaches to the communication of knowledge. Los Altos, CA: Morgan Kaufmann; 1987.
25. 24. Woolf BP. Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning. Burlington, MA: Morgan Kaufmann; 2009.

# All that glitters (in the lab) may not be gold (in the field)

Amruth N. Kumar

Ramapo College of New Jersey, Mahwah, USA
amruth@ramapo.edu

**Abstract.** AI-ED community has hewed to rigorous evaluation of software tutors and their features. Most of these evaluations were done *in-ovo* or *in-vivo*. Can the results of these evaluations be replicated in *in-natura* evaluations? In our experience, the evidence for such replication has been mixed. We propose that the features of tutors that are found to be effective *in-ovo*/*in-vivo* might need motivational supports to also be effective *in-natura*. We speculate that some features may not transfer to *in-natura* use even with supports. Recognition of these issues might bridge the gap between AI-ED community and educational community at large.

**Keywords:** In-ovo, in-vivo, in-natura, replication of results.

## 1 Introduction

Evaluation of software tutors may be carried out in one of three settings:

- *In-ovo*: Research subjects hand-picked for the evaluation use the software tutor in a laboratory setting, typically under tightly controlled conditions, and under the supervision of the researcher.
- *In-vivo*: Students enrolled in a course use the software tutor in the class room, typically under tightly controlled conditions and under the supervision of the researcher or course instructor.
- *In-natura*: Students enrolled in a course use the software tutors, typically after class, on their own time, and unsupervised.

These three types of evaluation are summarized in Table 1.

| Type | Location | Subjects | Conditions | Supervised |
|---|---|---|---|---|
| *In-ovo* | Laboratory | Recruited | Controlled | Yes |
| *In-vivo* | Classroom | Students enrolled in a course | Controlled | Yes |
| *In-natura* | After-class | | Not controlled | No |

Table 1: Types of evaluation of software tutors

AI-ED community has reported frequently using *in-ovo* and *in-vivo* evaluations in its studies of the effectiveness of software tutors and their features. Researchers have strictly controlled the conditions of these studies – what a subject can do or not do during the study, whether the subject is exposed to any distractions during the study, etc. – so as to minimize the influence of extraneous factors.

However, in real-life, especially at baccalaureate level, software tutors are less used as in-class exercises than as after-class assignments or study aides. The reasons for such use are many, including: course instructors may not want to spend valuable class time using software tutors; and students may not have access to (sufficient numbers of) computers during class.

When software tutors are used for after-class assignments, mandatory or otherwise, issues of intrinsic and extrinsic motivation play a much larger role in their use and utility. For starters, the popular aphorism *If you build it, they will come* does not apply to software tutors – unless students are required to use a software tutor, they will not use it (in any significant numbers). This significantly drives down participation and may skew evaluation results because of the self-selected nature of subjects. When they do use it, extrinsic motivation often plays a larger role than intrinsic motivation – if they are awarded course grade proportional to how well they do on the software tutor, they are more likely to engage seriously with the tutor. On the other hand, if they are given credit simply for using the software tutor, they are likely to do the least amount of work possible to qualify for such credit.

Given these considerations, do the research results elicited under carefully controlled conditions *in-ovo* or *in*-vivo extend to *in-natura* use of software tutors? In other words, can results obtained *in-ovo* or *in-vivo* be replicated *in-natura*? Our experience has been mixed. We will present results from evaluations of two features – reflection and self-explanation - vouched for by the AI-ED community that did not pan out in our *in-natura* evaluations.

For our evaluations, we used software tutors for programming concepts, called problets (problets.org). These tutors are being used every semester by 50-60 schools, both undergraduate and high-school. Since problets are deployed over the web, students have access to the software tutors anytime, anywhere. Problets are set up to automatically administer pre-test-practice-post-test protocol every time they are used [5]. They have been continually used and evaluated *in-natura* since fall 2004.

## 2 Reflection

The benefits of post-practice reflection have been studied by several researchers (e.g., [3]). In problets, we introduced reflection in the form of a multiple-choice question presented after each problem. The question states "This problem illustrates a concept that I picked based on your learning needs. Identify the concept." The learner is provided five choices, each of which is a different concept in the domain. The learner must select the most appropriate concept on which the problem might be based, and cannot go on to the next problem until (s)/he correctly selects it. The problet records the number of unique concepts selected by the learner up to and including the most appropriate concept. See Figure 1 for a snapshot of the reflection question presented after the student has solved a problem on selection statements.

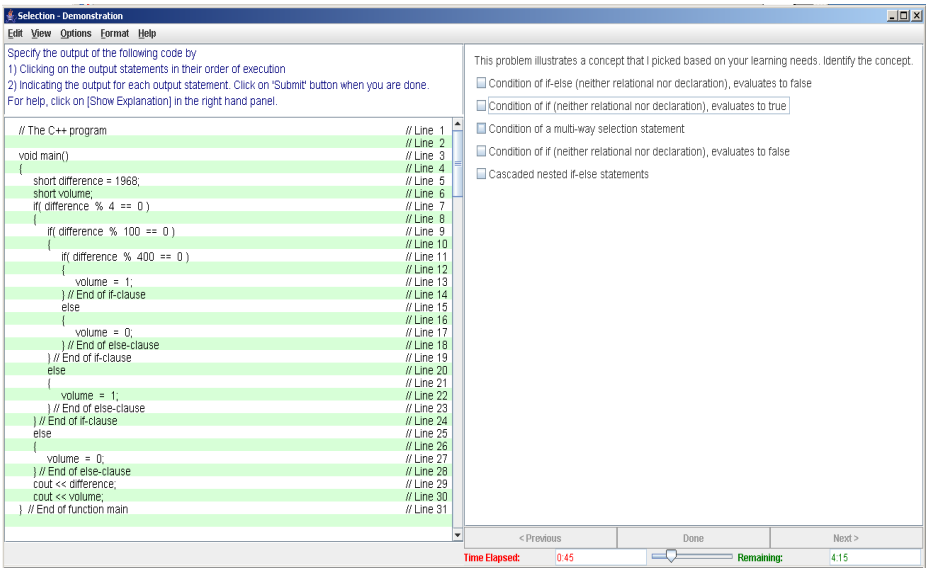



Figure 1: Selection tutor: Problem in the left panel; Reflection question in the right panel

We conducted several controlled evaluations of reflection [8] using selection and while loop tutors in 2006-07. Control group was never presented any reflection questions. Test group was presented a reflection question after each problem during pre-test, practice and post-test. If a student solved a problem incorrectly, the student was required to answer the subsequent reflection question correctly before going on to the next problem.

Practice was adaptive, and based on the student's performance on the pre-test. The entire protocol was limited to 30 minutes for control group and 33 minutes for test group. For analysis purposes, we considered only *practiced* concepts [5], i.e., concepts on which the student solved a problem incorrectly during pre-test, solved one or more problems during adaptive practice and also solved the post-test problem before running out of time.

Table 1 lists the score per problem on pre-test and post-test of all *practiced* concepts. No significant difference was found between control and test groups, indicating that the two groups were comparable. However, no significant difference was found in their pre-post improvement either, suggesting no differential effect of reflection on their learning. Please see [8] for additional details of the evaluation.

| Score per problem | Pre-Test | Post-Test | Pre-post $p$ |
|---|---|---|---|
| Control Group (Without Reflection) ($N$ =89) | | | |
| Mean | 0.118 | 0.736 | < 0.001 |
| Standard-Deviation | 0.177 | 0.353 | |
| Test Group (With Reflection) ($N$ =152) | | | |
| Mean | 0.144 | 0.787 | < 0.001 |
| Standard-Deviation | 0.183 | 0.319 | |
| Between groups $p$ | 0.283 | 0.266 | |

Table 1: Both the groups improved significantly from pre-test to post-test; the difference between the two groups was not significant on either the pre-test or the post-test

## 3 Self-Explanation

The effectiveness of providing self-explanation questions in worked examples has been well documented by AI-ED community (e.g., [1]).

Selection tutor was used for this study. When the student solves a problem incorrectly, the tutor presents feedback including step-by-step explanation of the correct execution of the program in the fashion of a fully worked-out example. Self-explanation questions were presented embedded in this step-by-step explanation, as shown in Figure 2. Each self-explanation question is a drop-down menu that deals with the semantics of the program, e.g., the value of a variable, the line to which control is transferred during execution, etc. The questions were independent of each other, but answering them required the student to closely read the step-by-step explanation/worked out example and understand the behavior of the program in question.

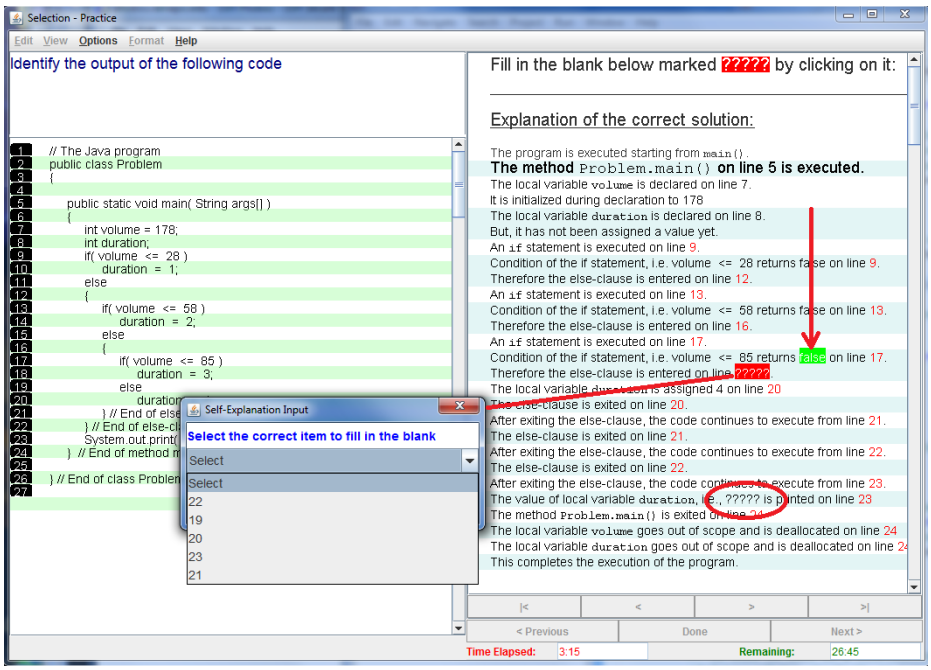



**Figure 2.** Snapshot of selection tutor with self-explanation questions displayed in the right panel

So as not to overwhelm the student, the tutor limited the number of self-explanation questions per problem to three. The student was allowed as many attempts as needed, but had to answer each self-explanation question correctly before

proceeding to the next question, and had to answer all the self-explanation questions correctly before proceeding to the next problem. A version of the tutor was used for the control group that did not present any self-explanation questions. This version of the tutor allowed the learner to advance to the next problem as soon as it displayed step-by-step explanation of the current problem.

Controlled evaluation of selection tutor was conducted *in-natura* over three semesters: fall 2012-fall 2013 [4]. No significant difference was found in the average score per pre-test problem between control (N = 395) and test (N = 335) groups [$F(1,729) = 1.018$, $p = 0.313$]. So, the two groups were equivalent. The mean number of concepts practiced by control group was 1.62, and by test group was 1.78. However, since control group was allowed 30 minutes to practice with the tutor and test group was allowed 40 minutes, univariate analysis of the number of concepts practiced was conducted with self-explanation as the fixed factor and total time spent as the covariate. The difference between the two groups was found to be significant [$F(2,597) = 62.207$, $p < 0.001$]: accounting for the extra time allowed, control group practiced $1.72 \pm 0.11$ concepts whereas test group practiced $1.662 \pm 0.12$ concepts. Therefore, test group practiced significantly fewer concepts than control group. No significant difference was found between the two groups on the pre-post change in score on practiced concepts, suggesting no differential effect of self-explanation on learning. Please see [4] for additional details of the evaluation.

## 4 Discussion

In both the studies – on reflection and self-explanation – we have verified that our implementation is behaviorally similar to, if not the same as described in at least some of the literature on the topic published in the AI-ED community. Even if our interpretation of both reflection and self-explanation behaviors differs enough from those reported in literature to render our treatments ineffective, we would expect that the increased time-on-task due to these faux treatments would have still yielded some learning benefits.

Our evaluations cannot be faulted for inadequate participation – our evaluations have typically involved 200-300 students, which is an order of magnitude larger than the number of subjects reported in typical *in-ovo* and *in-vivo* evaluations.

We have used standard protocols for evaluation – controlled studies, pre-test-practice-post-test protocol and partial crossover design. We have used ANOVA for data analysis. In our studies, we have considered only *practiced* concepts – concepts on which students solved problems during all three stages of the protocol: pre-test, practice and post-test, so noise is not an issue in the analyzed data.

These practices have been effective - not all our evaluations have come up empty, e.g., we have found significant effect of providing error-flagging feedback on test performance (e.g., [6]), and significant stereotype threat (e.g., [7]).

An explanation for the lack of results might be the difference in student motivation in *in-ovo*/*in-vivo* versus *in-natura* evaluation. Apart from issues of extrinsic motivation mentioned earlier, it may also be argued that given the lack of supervision in *in-*

*natura* evaluation, students are less likely to experience *Hawthorne effect* [2]. So, the features of tutors that are found to be effective in *in-ovo*/*in-vivo* evaluations might need motivational support to also be effective in *in-natura* evaluations.

Then again, even with motivational support, students may resent having to perform tasks (such as answering questions on reflection) that they do not perceive as directly contributing to their assignment at hand, and may not participate in, or may not be amenable to benefiting from what they view as a chore. In other words, some features may not be transferable from the laboratory to the field regardless of the supports provided.

While we have focused on the transferability of evaluation results from lab/classroom to after-class setting, researchers have reported similar issues transferring results from the lab to the classroom, e.g., in a study of politeness in intelligent tutors [9], researchers reported finding weaker results when the study was conducted in a classroom rather than a laboratory. They speculated that grades, an extrinsic motivational factor, may be to blame. Furthermore, they wrote [9], "In the rough-and-tumble of the classroom, with its noise, question-asking, and social environment, students *may simply not concentrate as much on the feedback provided by the computer tutor*. The lab setting, on the other hand, is a quiet environment where subjects work on their own with few distractions, and certainly none from classmates and a teacher" (italics not in the original). The noise, distractions and lack of structure used to describe a classroom as compared to laboratory setting are the very same terms, magnified, that could be used to describe an after-class setting as compared to a classroom. In other words, when it comes to noise, distractions and lack of structure, laboratory and after-class setting are at opposite ends of a spectrum, with the classroom situated in between. That *students may not concentrate as much on the feedback provided by the tutor* may explain why reflection and self-explanation, both provided as part of feedback, failed to live up to expectation in our *in-natura* evaluations.

It appears that *in-natura* use of software tutors entails more than just large-scale/unsupervised deployment of *in-vivo* results and *in-vivo* use entails more than just live-classroom deployment of *in-ovo* results. Motivational supports may be needed to transition results from the laboratory to the field and some results found in the laboratory may fail to transfer to the field even with motivational supports. Treating *in-natura* use of software tutors as being distinct from *in-ovo/in-vivo* uses is reminiscent of the outgrowth of Chemical Engineering as a discipline of the field from Chemistry as a discipline of the laboratory. While Chemistry is the study of properties of materials, Chemical Engineering is the study of the production of materials on an industrial scale, albeit with its basics firmly rooted in Chemistry. In the early years, chemists refused to accept Chemical Engineering as anything more than Chemistry, and engineers refused to recognize Chemical Engineering as an engineering discipline [10], but not so any more. May be AI-ED community should treat *in-natura, in-vivo* and *in-ovo* as three independent, necessary and valuable stages in the evaluation of any treatment. May be, *in-natura* evaluation is what is needed for educational community at large (especially higher-education community) to recognize and incorporate the important pedagogical insights being offered by AI-ED community.

**Acknowledgments.** Partial support for this work was provided by the National Science Foundation under grants DUE-0817187 and DUE-1432190**.**

## 5 References

1. Conati, C. and VanLehn, K. (1999) Teaching meta-cognitive skills: implementation and evaluation of a tutoring system to guide self-explanation while learning from examples. Proc. AI-ED 99, 297-304.
2. Franke, R.H. and Kaul, J.D. The Hawthorne experiments: First statistical interpretation. *American Sociological Review.* Vol 43. 1978. 623-643.
3. Katz, S., O'Donnell, G., Kay, H. (2000) An Approach to Analyzing the Role and Structure of Reflective Dialogue. International Journal of Artificial Intelligence in Education, 11, 320-343.
4. Kumar, A.N. An Evaluation of Self-Explanation in a Programming Tutor. In Proc. Of ITS 2014, Hawaii, June 2014. 248-253.
5. Kumar, A.N. A Model for Deploying Software Tutors. IEEE 6th International Conference on Technology for Education (T4E). Amritapuri, India, 12/18-21/2014, 3-9.
6. Kumar, A.N. Limiting the Number of Revisions While Providing Error-Flagging Support During Tests. Proc. Intelligent Tutoring Systems (ITS 2012), LNCS 7315, Chania, Crete, 6/14-18/2012, 524-530.
7. Kumar, A.N. A Study of Stereotype Threat in Computer Science. Proceedings of Innovation and Technology in Computer Science Education (ITiCSE 2012). Haifa, Israel, 7/3-5/2012, 273-278.
8. Kumar, A.N. Promoting Reflection and its Effect on Learning in a Programming Tutor. Proceedings of 22nd International FLAIRS conference on Artificial Intelligence (FLAIRS 2009) Special Track on Intelligent Tutoring Systems, Sanibel Island, FL, May 19-21, 2009, 454-459.
9. McLaren, B.M., DeLeeuw, K.E., and Mayer, R.E. Polite web-based intelligent tutors: Can they improve learning in classrooms? *Computers and Education.* 56(3): 574-584, 2011.
10. Reynolds, Terry S. Engineering, Chemical, in Rothenberg, Marc, *History of Science in United States: An Encyclopedia*, New York City: Garland Publishing, 2001. ISBN 0-8153-0762-4

# Why AIED Needs Marriage Counselling by Cognitive Science (to Live Happily Ever After)

Björn Sjödén

Lund University Cognitive Science, Sweden
Bjorn.Sjoden@lucs.lu.se

**Abstract.** In this position paper, I reflect upon the question "Should AI stay married to Ed?", specifically referring to how research in AI and Education should cross-fertilize to define AIED as an independent practice, beyond its composite fields. In my view, a mix of approaches, inspired by cognitive science, should serve to formulate characteristic research questions for the AIED community. Such questions may be derived from considering the social context of learning and how it is applied in artificial systems, as exemplified by educational games and ITS with Teachable Agents. I conclude by suggesting two discussion points of emergent interest to AIED research: (1) How can we formulate scientifically based guidelines for the use and evaluation of educational software? (2) Is there anything such as "unique AIED competence" and, if so, what does this imply for the AIED identity?

**Keywords:** AIED, marriage, multidisciplinary research, educational games, ITS, Teachable Agents.

## 1 Introduction

What *is* and what *should be* the role of AI in Education and conversely of Education in AI?

For all its successes, the very need to reassert AIED's position as a research field after 25 years may reflect two critical shortcomings: a failure to appreciate its relative independence from both AI and education, on the one hand, and an underused and conservative application of AI for educational purposes that has not been fully embraced by educators, on the other. One might compare to fields like HCI or interaction design, which have successfully defined and built research communities around cross-disciplinary domains that focus on people's use of technology.

A stumbling-block to AIED practitioners might be that the field has no obvious "core", that is, it has no clearly defined subject of investigation, such as "computers" or "interactive systems" or even an abstract topic like "instructional strategies". The most concise, official description of the field appears in operational terms, with respect to the scope of the AIED journal, as "the application of artificial intelligence techniques and concepts to the design of systems that support learning" (from ijaied.org). This leaves room for a great variety of research and different approaches – which is good – but the field seems to lack a common conceptual framework for relating advances in AI to advances in education research that would inform characteristic AIED research questions. Is it at all clear for the field's different practitioners what the common denominator of AIED research is?

I posit that it means something to be knowledgeable in AIED and being skilled in AIED research as such, beyond having expertise in AI and Education as distinct fields. The identity of the AIED field is then formed by the content of this "AIED competence", its unique contributions and necessary limitations to other areas. In effect, other considerations become important for AIED than in traditional AI research that does not necessarily apply to education. For example, the AIED researcher with an AI background might be more concerned with "weak AI" as a means to make students learn better or pay more effort, while having to take into account what can be realistically implemented and evaluated in a school or classroom setting, on different technical platforms (tablets, smart phones, laptops etc.) and for different groups of students. Likewise, an AIED researcher with an education or pedagogy background would look to how the use of technology can add to present pedagogical strategies and teaching methods as a means to achieve the same goals. Eventually, as research from both ends cross-fertilize, they may transform educational practices by setting new learning goals defined by the use of technology (e.g. "21st century skills") [1].

In this text, I present a view of AIED that develops from practical considerations for a functioning relationship between AI and Ed, but also forms a new area of research for educational purposes. The educational context both constrains and opens up a largely unexplored scene for novel applications of AI techniques that further motivates the growth of AIED as an independent field. As such, I argue that AIED should aspire to achieve two overarching goals: (1) Improve human learning, and (2) Inform and expand the scientific basis of education. (Notably, the AIED Society has set as its aims to promote knowledge and research in AIED but does not explicate the aims of the field itself.)

As a field of empirical, scientific inquiry (and not just the pragmatic "application of AI techniques"), AIED may be fruitfully compared to Cognitive science. The success of cognitive science as an academic discipline shows how intrinsically different fields – among those psychology, biology, computer science, anthropology and philosophy – have found a common identity in the pursuit of certain well-recognized research questions under the multidisciplinary banner of "cognition". Notably, one does not have to be an expert in all these fields to become an expert cognitive scientist, and it is possible to work within any of these fields without doing research of intrinsic interest to cognitive science. Thus, cognitive science found an identity of its own from combining perspectives and methods from various disciplines, in principle not different from how AIED can develop from merging aspects of AI and education research.

The first question then becomes how the multidisciplinary AIED field should be conceptualized in relation to its history and previous accounts. Second, we need to know what the content of the practice is – the research outcomes and applications – that motivates AI and Ed's relationship. By setting the example, cognitive science might be just the marriage counsellor that AI and Ed need to develop their common interests and secure the future well-being of AIED.

## 2 Reconceptualizing AIED as a multidisciplinary field

Looking back, Cumming and McDougall [2] already in 2000 speculated how AIED might be "mainstreaming into education" in the (then) future of 2010. They argued

both for relabeling the field ("especially the 'AI'", p. 204, which they considered does not communicate the field well) and its crucial need for "AI expertise of the highest order" in order to "keep at the forefront of all of the contributing disciplines" (p. 205). This appears, to express it mildly, as a tall order for AIED to take. Above all, it seems to indicate that the field has long had an unclear identity, particularly when it comes to defining the kind of AI expertise needed for being an AIED, rather than an AI or educational, researcher.

I will not propose a new label for AIED research, but perhaps its identity should not be formed on basis of its historical, composite fields, but rather from what motivates AIED research as a multidisciplinary practice in the present and for the future. As to the topics of research, there is a vast array of educational technologies available today that did not exist at the field's inception 25 years ago. In short, things have changed, and besides emerging new technologies, there is an emerging new generation of AIED researchers.

Looking forward, I approach this question from the perspective of a beginning researcher in the field, who needs to define his future area of expertise. While being actively involved in the AIED community [e.g. 3, 4, 5, 6], I do not see myself as belonging either to the "AI field" or the "Education field" or at least not exclusively so. Rather, and for reasons outlined in the introduction, I would attest to Cumming and McDougall's [2] observation that "Many AIED researchers would be happy to be described as cognitive scientists." (p. 198) and their suggestion that AIED "should overlap with cognitive science" (p. 205).

Like cognitive science, albeit on a smaller scale, AIED may play a crucial role for bringing computer science-oriented (AI) and psychology/pedagogy-oriented (Education) research together. Figure 1 illustrates how this view of an emerging AIED field differs from previous conceptions of bringing AI and Ed together.

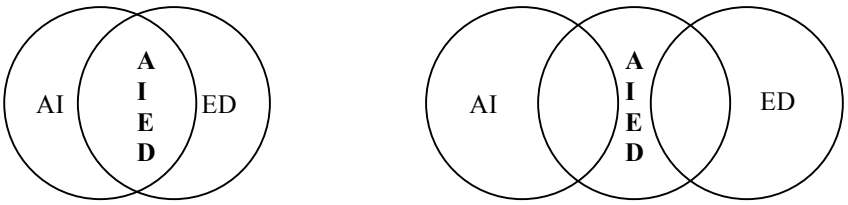



**Fig. 1.** Two alternative conceptions of AI + Ed: (left) AIED as the combinatory interests of AI and Education research; (right) AIED as an independent, multidisciplinary field, defining its own aims and scope in between the respective fields of AI and Education.

The emancipation of AIED research from its surrounding disciplinary boundaries (Fig. 1, left) would make room for its own defining research questions (Fig. 1, right). Whereas AI puts machine learning and human-like intelligence in focus, Education focuses on fostering human learning and intelligence. AIED knowledge should then serve to bridge this gap by informing techniques to promote more efficient and intelligent interactions with humans that improve educational outcomes (rather than, say, aiming to reproduce human abilities or solving computational problems that do not feed back to students). A combination of methodological approaches is likely needed,

as every method carries with it implicit assumptions about what knowledge it can produce (from an AI as well as an education perspective).

Often multidisciplinary knowledge is needed to appreciate the educational applications of different technologies. A classic example would be how computers in school are most often used as word processing and communication tools, whereas the sophistication of the underlying technology would make the computer the obvious arena for students to learn from and by AI technologies in any school subject. However, just like book not read, the computer becomes meaningless as an educational tool unless it is engaging enough for students to actually use it. Student engagement puts the interactive qualities of the system in the forefront, not in terms of superficial "usability", but rather as to "learnability" and "teachability". This poses an array of non-trivial AIED research questions as to how the technology functions in the educational context and how AI may serve to improve the scientific basis of education. Next, I consider some of these challenges in greater detail and how they are dealt with in two types of AIED applications that exemplify the present and future potential of the field.

## 3 What AIED Brings to Artificial intelligence for Authentic Learning

Education is a social process characterized by learning in interaction: between teachers and students, among students, and, if AIED has a say, between "intelligent" artefacts and both teachers and students. As extensively demonstrated by Reeves and Nass' "media equation" [7], adding interactivity to a system naturally invites social behavior. Accordingly, as computerized learning environments become increasingly interactive and adaptive, they can be said to expand upon the social dimension that affects the learning process. This has important implications for AIED, first for distinguishing AIED applications from other, "static" learning material such as text books; second, for which methods should be used to study learning outcomes (e.g. some would take this as an argument for a situated perspective on learning, or against "media comparison studies" that undervalue the instructional process as such [e.g. 8, 9]). Stressing the interactivity aspect also brings forth the "social intelligence" of the system as an important consideration for new AI techniques.

With the more advanced interactivity that comes with technical development, it makes sense for a field like AIED to take social motivations for learning with artificial systems to the core of its interest. Considering that students may vary as much in what keeps them motivated and engaged as they do in cognitive abilities, AI techniques devoted to exercise social influence (e.g. by virtual agents and feedback) that adapt to the individual student would allow for unique educational arrangements. More specifically, AIED may serve to dissolve previous conceptions of "education", noted also by Cumming and McDougall, as something that takes place in groups (e.g. in schools and classes) versus "learning" as something that happens in an individual (sometimes with a book or a computer).

Schwartz and colleagues [10] argue for a specific form of computer interactivity that generates a "learning sweet spot" of both high motivation and high learning from students' social motivations. This "sweet spot" is achieved through designing an environment that encourages shared initiative and engagement in interaction, in their case

with a teachable pedagogical agent. The researchers make the important point that the technology is not used with the goal to "perfectly" model human traits, conversation or intelligence, but only to be *sufficient* to elicit the social schemas (e.g. that of teacher/student) that engage students in productive interactions for learning.

Notably, the research question and goal of making a human-like functioning, artificial system (e.g. "How can we model human intelligence and learning?") are essentially different from those of making a system designed to help students organize and reason with their own concepts (e.g. "How can we visualize students' knowledge?", "What level of prompting from the system is optimal for triggering students to contribute with their own knowledge?"). One can also say that the goal of the former is to produce an *autonomously* intelligent system, whereas the goal of the latter is to produce a *jointly* (with the student and to the advantage of the student) intelligent system.

The perspective on "joint intelligence" takes more of the social context into account, which brings AIED closer to traditional educational research, drawing from established pedagogical strategies such as peer tutoring and learning-by-teaching. However, what makes AIED unique as a research field is that it deals with variables known to have an impact on learning but that have never before been possible to manipulate independently and systematically, such as social roles (including altering avatar representations of gender or ethnicity), others' skills or knowledge level (using virtual peers), and parameters for individually adapting material for teaching in groups. The educational impact of such manipulations makes a prominent subject for AIED research.

In sum, the social nature of interaction points to a range of issues crucial to understanding the impact of AI techniques when employed in real-world educational settings, which AIED should serve to make explicit for both the AI and education community:

First, it is essential to understand which social factors drive students' learning, since technical functions, even when used, may turn out to be used in unintended ways that diverge from the original pedagogical principles [6].

Second, it is important to realize that employing AI techniques may bring "added value" to traditional teaching but possibly also "reduced value", if it takes time and resources from educational needs that are better met by human teachers or other means [11]. It should be a concern of educators to determine when and for what purpose to use AI-based systems for students' learning, and it should be a primary concern of AIED research to distinguish between the "added" and the "reduced" values for different knowledge needs.

Third, and arguably the most crucial point for positioning the AIED research field, the shortcomings of technology, as well as the shortcomings of educational practices for predicting successful learning, leave space for new and original research on what forms students' learning experience in their interaction with technology. I take two examples to show how our expectations of how AI techniques "should" work can be as important for the outcome of the interaction as the underlying technology itself. This also shows that even if the social context cannot be fully controlled or predicted, a careful design can devote AI techniques to create certain "illusions" of intelligent behavior that promote students' learning.

### 3.1 Example 1: "Educational" games

Many computer games make use of some AI; however, AI techniques appear strikingly underused in the subcategory of so-called "educational games". However, there is an ambiguity in the term "educational games" that both confuses researchers and confounds educational practices. This confusion is mirrored in the debate of whether or not "computer games" as such are effective for learning (for contrasting accounts, see [12, 13]).

In short, the "educational" in games might refer to the educational *subject content* in terms of topics relevant to the curriculum (such as "a math game training fractions"[1] or "a political strategy game that models the global economy"[2]), or it might refer to the (intended) educational *use* of the game in a school context (such as playing a commercial game in the *Halo* or *Assassin's Creed* series that employ AI techniques and that some teachers may use for training "strategy thinking" or "problem-solving" skills, though these are not explicit aims of the game). Games with subject-relevant content typically include explicit exercises or "game tasks" for the intended skill (e.g. counting, spelling tasks) whereas the educational use of other games typically assumes that relevant skills are learned implicitly, through the practice of other kinds of overarching "game goals".

As to the vast offer of games that claim educational content, there is rarely any advanced AI to direct or scaffold the learning process. For example, several AIED-relevant review- and development articles have remarked that the vast majority of math games in the open market (e.g. in AppStore) do not adhere to even basic, cognitive design principles and they seem to contain little more than simple 'drilling' exercises with limited feedback [14, 15, 16].

Using other, commercial computer games for educational purposes, may have great effects on student engagement but little or no effect on learning [13]. This is because game-players may utilize the affordances in a game in a relatively superficial way, learning only "what to press when" to achieve certain results, such that good game performance and progression do not necessarily require the deeper cognitive processing wished for good education. Linderoth reaches the thought-provoking conclusion that the educational appeal of computer games may come from maintaining an "*illusion of learning"* (ibid, p. 59).

The task for the educator is further complicated by the fact that some commercial games might indeed require great skill (e.g. for solving puzzles) but it is hard to determine how much of these abilities are trained by the game itself and, if so, to what extent they transfer to school-relevant tasks (e.g. solving physics problems or mathematical equations).

Nevertheless, it seems safe to say that the education industry has failed to take on board the creative application of the relatively sophisticated "game AI" techniques used in the commercial gaming industry. Rather than the "illusion of learning" on

---

[1] http://www.mathsgames.com/fraction-games.html

[2] http://www.positech.co.uk/democracy3/index.php

behalf of the student, an alternative and productive use of "game AI" (or "weak AI") would be to maintain an "illusion of intelligence" [17] on behalf of the software that keeps the student engaged in intellectual activities for actual learning, just like the game-player is kept engaged in playing for entertainment. In other words, a game might not need cognition-like computational power to have educational value in terms of meaningful learning, but the resources it does use should be dedicated to educationally relevant goals. It is to the latter point commercial games often cut short.

Whereas the above may be bad news to educational games, it is good news to the AIED research field, because it shows that there are both *potentially* effective AI techniques already in place and an enormous interest from the education community to employ them (e.g. Education apps being the second largest category in AppStore after Games, both in numbers of 100.000+). Games appear as a domain where AI and Educational interests merge but where AI techniques have been underemployed *for learning*. This makes "educational games" a primary topic for AIED research, a brain child still in its infancy, which calls for more attention and better interdisciplinary upbringing by AI and Ed.

### 3.2 Example 2: ITS including Teachable Agents

Intelligent Tutoring Systems (ITS) might represent the most dedicated and successful use of AI for student-centered learning, since it actively employs AI techniques not only to structure and present material but also for communication purposes that (more or less) model that of a human teacher. I chose this example (and to include Teachable Agents, or TA, as a "reversed" model of tutoring) because it clearly illuminates how AI-based system can make use of familiar social schemas [e.g. 18, 19].

ITS and TA exploit and benefit from social learning mechanisms (most visibly so when represented by a visual character on screen although the system could be entirely text-based) derived from the student-teacher relationship. As these systems become increasingly advanced, the knowledge needs about people's social motivations and social psychology in general become of greater importance to the AIED field.

But is it a realistic, or even wanted aspiration, for AIED purposes (i.e. for use in teaching and learning) to develop virtual agents that are as life-like or sociable as a real person? This is an important question for the future of AIED because it poses where resources are better spent; for instance, how should the overwhelming task of producing human-like AI be balanced against working out effective instructional strategies that can be formalized and computed?

Importantly, artificial systems can invoke social responses to improve learning without having to employ AI. For example, Okita et al [20] showed that the mere belief in "real" social interaction when interacting with a computer agent had positive effects on learning, again an effect exploited in the TA metaphor [18]. Some of my own AIED research [3, 4] suggests that social effects of interacting with a Teachable Agents might also transfer from the learning situation to being tested on one's knowledge; students took on harder problems and performed better on those problems if tested "in company" with the TA they had previously worked with in a learning game. In short, remarkably simple social stimuli may trigger complex and beneficial learning behaviors.

As an interesting contrasting example, some researchers have employed AI techniques to create an “illusion of teachability”, by making an agent appear more socially sensitive to the student’s input than it actually is [21]. In this case, the system construes a mental model of the human student that informs the agent’s responses so it appears as “teachable”, though it actually only reflects the kind of knowledge gaps and mistakes that the student has displayed. In effect, the student has to “teach” exactly the things needed to improve his/her own shortcomings (and not necessarily those of a third-party agent). This adds to the power of the social schema by showing that not only the intentional belief of teaching drives the effect, but also the belief in how the tutee (the agent) responds.

My point here is that an important topic of AIED research is to disentangle the social and cognitive mechanisms underlying the effects of ITS and TA, both for the general understanding of such systems and for developing resource-effective systems. For example, the “teachability” features of a TA may be theoretically divided into the underlying (AI-governed) mechanisms that direct the information processing, and its social appearance, as constituted by its visual looks, the things it says, and the types of choices it offers. A key contribution of technology is to offer means to control and regulate these factors through digitalized and personalized "social" responses that can avoid the pitfalls of human socializing (such as distraction from the task and negative stereotyping) while maintaining and even adding to the benefits (such as constructive feedback and active engagement).

In sum, ITS and TA represent a case of true cross-fertilization of the AI and Education domains that produce some unique results, never before seen in human history: semi-independent, virtual beings whose sociable qualities place them somewhere in between artificial and human agents, more like active “educational peers” than passive information systems. In this sense, AIED breaks up the traditional teacher/student dichotomy and includes a third party in the educational design. Students’ social motivations to engage in interaction with this party might be more a matter of the effective representation of social features as learned and recognized from the outside world, than how its knowledge is represented inside the system. For use in the social context of a classroom, this makes a strong argument for bringing in more of the educator’s experience of “what works” into the design of AI systems.

## 4 Moving On

Taking the example of cognitive science, I aimed to illustrate AIED as a multidisciplinary practice that forms its identity in relation to technical development as well as pedagogical methods and social learning theories. To the extent that “AI” and “Education” hold separate identities as distinct fields that do not seamlessly combine or “marry”, it might be more productive to focus on what they can form together, as a common theme for their future. Educational games and ITS with TA provide example domains that cannot be said to be either “AI” or “Ed” but very much AIED. Relating to those examples, I conclude by suggesting two further discussion points that AIED should take into consideration when moving on together:

1. As to educational software (including games, ITS, simulations and other digital learning environments), AIED still seems predominantly concerned with development and design aspects, whereas little has been done to serve educators' need for sound evaluation and scientifically based, qualitative assessment of existing applications. How do design criteria for learning-effective software translate to evaluation criteria? Considering the vast selection of educational apps to date, perhaps the best way to guide teachers is to formulate meta-criteria that help inform their own selection and recognize well-designed content? How can AIED assist in making this judgment scientifically informed?

2. Considering the range of issues an AIED researcher may have to confront, as exemplified in this text, what is the essence of the "AIED competence" – what does an AIED researcher (need to) know that others don't? Is there anything such as "interdisciplinary expertise" in its own right and then, how does this show, and how is it applied, within AIED research? Is the explication of specific AIED knowledge areas required (or just helpful) for forming a unique identity of the field?

## References


1. West, D.: Digital schools: How technology can transform education. Brookings, Washington D.C. (2012)
2. Cumming, G., & McDougall, A.: Mainstreaming AIED into education?. International Journal of Artificial Intelligence in Education (IJAIED), 11, 197-207. (2000)
3. Sjödén, B., Tärning, B., Pareto. L., Gulz, A.: Transferring teaching to testing, An unexplored aspect of teachable agents. In: Proc. of the 15th Int. Conf. on Artificial Intelligence in Education (AIED 2011), LNAI, vol. 6738 (pp. 337-344). Springer, Heidelberg (2011)
4. Sjödén, B., & Gulz, A.: From Learning Companions to Teaching Companions: Experience with a Teachable Agent motivates students' performance on summative tests. In: Proc. of the 17th Int. Conf. on Artificial Intelligence in Education (AIED 2015), *in press*
5. Pareto, L., Haake, M., Lindström, P., Sjödén, B., & Gulz, A.: A teachable agent based game affording collaboration and competition – evaluating math comprehension and motivation. Educational Technology Research and Development, 60(5), 723-751 (2012)
6. Lindström, P., Gulz, A., Haake, M., Sjödén, B.: Matching and mismatching between the pedagogical design principles of a math game and the actual practices of play. Journal of Computer Assisted Learning, 27(1), 90-102 (2011)
7. Reeves, B., & Nass, C.: The Media Equation. How people treat computers, television, and new media like real people and places. CSLI Publications and Cambridge university press. (1996)
8. Clark, R. E.: Reconsidering the research on learning from media. Review of Educational Research, 53(4), 445-459 (1983)
9. Ross, S. M., Morrison, G. R., & Lowther, D. L.: Educational technology research past and present: Balancing rigor and relevance to impact school learning. Contemporary Educational Technology, 1(1), 17-35 (2010)
10. Schwartz, D. L., Blair, K. P., Biswas, G., Leelawong, K., & Davis, J.: Animations of thought: Interactivity in the teachable agent paradigm. In R. Lowe & W. Schnotz (Eds.), Learning with Animation: Research and Implications for Design (pp. 114-140). UK: Cambrige University Press. (2007)

11. Chin, D. B., Dohmen, I. M., Cheng, B. H., Oppezzo, M. A., Chase, C. C., & Schwartz, D. L.: Preparing students for future learning with teachable agents. Educational Technology Research and Development, 58(6), 649-669 (2010)
12. Gee, J. P.: What Video Games Have to Teach Us About Learning and Literacy. Palgrave Macmillan, New York (2003)
13. Linderoth, J.: Why gamers don't learn more: An ecological approach to games as learning environments, Journal of Gaming and Virtual Worlds, 4: 1, pp. 45–62, doi: 10.1386/jgvw.4.1.45_1 (2012)
14. Larkin, K.: Mathematics Education. Is there an App for that?. In: Mathematics education: Yesterday, today, and tomorrow, pp. 426-433. Mathematics Education Research Group of Australasia (MERGA) (2013)
15. Veenstra, B., Van Geert, P. L. C., Van der Meulen, B. F.: Is edutainment software really educational? A feature analysis of Dutch edutainment software for young children. Netherlands Journal of Psychology, 66(2), 50-67 (2011)
16. Ginsburg, H. P., Jamalian, A., Creighan, S.: Cognitive guidelines for the design and evaluation of early mathematics software: The example of MathemAntics. In: Reconceptualizing early mathematics learning (pp. 83-120). Springer Netherlands (2013)
17. Buckland, M.: Programming game AI by example. Jones & Bartlett Learning, Burlington (2005)
18. Chase, C., Chin, D., Oppezzo, M., Schwartz, D.: Teachable agents and the protégé effect: Increasing the effort towards learning. J. of Sci. Edu. and Tech., 18, 334-352 (2009)
19. Ogan, A., Finkelstein, S., Mayfield, E., D'Adamo, C., Matsuda, N., Cassell, J.: Oh dear stacy!: social interaction, elaboration, and learning with teachable agents. In: Proc. of the SIGCHI Conference on Human Factors in Computing Systems, pp. 39-48. ACM (2012)
20. Okita S, Bailenson J, Schwartz D.: The mere belief of social interaction improves learning. In: McNamara DS, Trafton JG (eds) The proceedings of the 29th meeting of the cognitive science society. August, Nashville, pp 1355–1360 (2007)
21. Lenat, D. B., & Durlach, P. J.: Reinforcing Math Knowledge by Immersing Students in a Simulated Learning-by-teaching Experience. International Journal of Artificial Intelligence in Education, 24(3), 216-250. (2014)

# AI and Education: Celebrating 30 years of Marriage

Beverly Park Woolf[1]

[1]College of Information and Computer Sciences
University of Massachusetts Amherst, USA

**Abstract.** This article describes contributions that artificial intelligence (AI) has made and needs to continue to make towards long-term educational goals. The article articulates two challenges in education that require the use of AI: personalizing teaching and learning 21st century skills. This article first describes AI and some of its history and then suggests why AI is invaluable to development of instructional systems. Instructional systems that use AI technology are described, e.g., computational tools that personalize instruction, enhance student experience and supply data for development of novel education theory development. Additionally, some intelligent tutors supply researchers with new opportunities to analyze vast data sets of instructional behavior and learn how students behave.

## 1 A Brief History of Artificial Intelligence in Education

The field of Artificial Intelligence in Education is focused on research into, development of and evaluation of computer software that improves teaching and learning. Several long term goals have been espoused, such as to interpret complex student responses and learn as they operate; to discern where and why a student's understanding has gone astray, to offer hints to help students understand the material at hand and ultimately to simulate a human tutor's behavior and guidance. Personalized tutors have been envisioned that adapt to an individual student's needs or to teach to groups of students, e.g., classified by gender, achievement level, amount of time for lesson, etc. Another goal is to use Artificial Intelligence (AI) techniques learn about teaching and learning and to contribute to the theory of learning.

AI techniques are needed for almost every phrase in the definition of intelligent tutors above, including *interpret* complex student responses, *learn* as they operate, *discern* where and why a student's understanding has gone astray and *offer* hints. The central problems (or goals) of AI research include reasoning, knowledge, planning, learning, natural language processing (communication), perception and the ability to move and manipulate objects [1]. AIED has been applied to complex domains, e.g. physics, programming, writing essays, and reading. These tutors learn about the strengths and weaknesses of students in these domains and also about students' skills, and emotion. How effective are intelligent tutors? Several tutors have been shown to be very effective in the classroom. Researchers looking at student skills at end of experiments and also at the end of course and large scale standardized testing evaluations found dramatic improvement understanding and learning [2]. Intelligent online tutors are an AI success story [3], though researchers seek to move beyond domain dependence and to support learning of multiple tasks and domains.

To mentor effectively and support individuals or groups, intelligent tutors will assess learning activities and model changes that occur in learners. Estimates of a learner's competence or emotional state, stored in user models, represent what learners know, feel, and can do. When and how was knowledge learned? What pedagogy worked best for this individual student? Machine learning and data mining methods, both derived from the field of AI, are needed to explore the unique types of data that derive from educational settings and use those methods to better understand students and the settings in which they learn (see [2, 4]).

Technology cannot impact education in isolation, rather it operates as one element in a complex adaptive system that considers domain knowledge, pedagogy and environments that students, instructors and technology co-create [5]. AI and Education researchers need to be driven by the problems of education practice as they exist in school settings. The emerging forms of technology described here will challenge, if not threaten, existing educational practices by suggesting new ways to learn [6]. Policy issues that involve social and political considerations, need to be addressed, but are beyond the scope of this document.

## 2 AI called by a different name: AI behind the scenes

Many components of intelligent instructional systems have their roots in artificial instructional research, e.g., adaptive curriculum, modeling (student, teacher, domain), educational data mining, speech recognition and dialogue systems. All began by using artificial intelligence (AI) techniques. Yet once these algorithms and techniques begin to appear as parts of larger tutors, the tutors are no longer considered AI and AI receives little or no credit for their successes. Many of AI's greatest innovations have been reduced to the status of just another item in the tool chest of instructional designers or computer science. Nick Bostrom explains "A lot of cutting edge AI has filtered into general applications, often without being called AI because once something becomes useful enough and common enough it's not labeled AI anymore." [7] "After all, all smart technologies currently in use (in the classrooms or homes), from tablet computers to smart phones, from Internet search engines to social networking sites, have a growing reliance on techniques derived from AI." [7] The AI effect began in the larger AI field and "occurs when onlookers discount the behavior of an artificial intelligence program by arguing that it is not real intelligence." [7] Pamela McCorduck writes: "It's part of the history of the field of artificial intelligence that every time somebody figured out how to make a computer do something—play good checkers, solve simple but relatively informal problems—there was chorus of critics to say, 'that's not thinking'." [8] AI researcher Rodney Brooks complains "Every time we figure out a piece of it, it stops being magical; we say, 'Oh, that's just a computation.'" [9].

Intelligent personal assistants in classrooms or in smartphones use algorithms that emerged from lengthy AI research. IBM's question answering system, Watson, which defeated the two great Jeopardy champions by a significant margin, was derived from basic AI research in natural language processing, information retrieval, knowledge representation, automated reasoning, and machine learning technologies to the field of

open domain question answering [10]. In addition, the Kinect, which provides a 3D body–motion interface for the Xbox 360 and the Xbox One was derived from basic AI research [7].

**AI is whatever hasn't been done yet.** Software and algorithms developed by AI researchers are now integrated into many applications, without really being called AI, e.g., speech understanding as part of online travel reservations, expert systems that save companies millions of dollars (US). Michael Swaine reports "AI advances are not trumpeted as artificial intelligence so much these days, but are often seen as advances in some other field."[11] "AI has become more important as it has become less conspicuous," Patrick Winston says. "These days, it is hard to find a big system that does not work, in part, because of ideas developed or matured in the AI world." [12]**.**

## 3 Impact on Education

A related question about AIED relates to the impact of AI on education and focuses on the extent to which the results of AIED research are meaningful to real educational practice [13]. Does the education community even care? Similar to many fields aspiring to scientific rigor, the AIED community can showcase dozens of studies demonstrating the statistical significance of this or that approach or system or their individual components through rigorously designed studies, but it is not always clear how the results of many of those studies actually translate into real educational teaching and learning practices raising a question as to whether all this rigor may not be happening in a vacuum.

For example, schools in the USA are not thriving. Too many schools *teach in traditional ways* and aren't preparing the next generation to meet new challenges. When today's students graduate, they'll be asked to fill the jobs of tomorrow—ones we can't even imagine [14]. And they'll be asked to tackle global problems like climate change, endemic hunger, and refugee problems. Additionally, the current use of digital resources in K12 and higher education can be described as dysfunctional: many school stakeholders can't find sufficient effective digital resources, while large collections of resources exist and sit online, waiting to be discovered. Some solutions have been proposed to migrate successful evidence-based digital resources into classrooms. One solution is to define a roadmap that moves well-tested resources towards publishers and software companies and ultimately into classrooms.

More than 4 million USA students at the K12 level took an online course in 2011, up significantly from just 1 million three years earlier. During the coming decade education should shift from print to digital and from batch processing to personalized learning [15]. In addition to virtual schools, online learning is increasingly being incorporated into traditional settings that blend the best of online and face- to-face learning. A shift to online learning is happening in K12 in the USA due in part to the need to implement college- and career- ready standards, the shift to next-generation assessments, and the prevalence of affordable devices. Online learning may move

standardized teaching towards more personalized instruction without increasing the number of teachers.

The field of AIED, now nearly thirty years old, has finally achieved some of its oldest goals. Thirty years is calculated from the first Intelligent Tutoring Systems Conference, 1988, organized by Claude Frasson in Montreal, Canada. Some long-term goals are currently being worked on, including understanding and responding to student knowledge, meta-knowledge (thinking about learning), and affect [16-19]. Educational games and new forms of digital learning are being investigated. In many cases evaluation of student progress shows improvement in learning. Some of the success is due to increasing computer power and some due to researchers focusing on specific isolated problems and pursuing them with the highest standards of scientific accountability. The reputation of AIED, in the education world at least, is still not very positive, because few tutors are robust enough to work consistently in a classroom environment.

# 4 Future directions for AIED to justify and maintain its unique identity

AI techniques are essential to develop new representations and reasoning about cognitive insights, to provide a rich appreciation of how people learn and to measure collaborative activity. Communities of researchers offer distinct clues to further refine individual instruction in online environments and also require far deeper knowledge about human cognition, including dramatically more effective constructivist and active instructional strategies [20].

## 4.1 Personalize teaching

One-to-one attention is very important for learning at any age. Research has also shown that students' emotions influence achievement outcomes: confidence, boredom, confusion, stress, and anxiety are all strong predictors of achievement [21, 22]. However, teachers are unable to provide attention based on intimate knowledge of each student. Providing personalized teaching for every learner begins by providing timely and appropriate guidance for student cognition, meta-cognition and emotion [20]. In other words, online tutors should determine in real-time what to say, when to say it, and how to say it. This process grows increasingly complex as the topics become more difficult and the required detectors becomes more complex, e.g., detectors for students' knowledge, skills, or emotion. The field of Learning Science has provided a wealth of knowledge about how to deliver effective feedback and how to teach with new methods (e.g., problem-based learning [23]. Rich, multi-faceted models of instruction go beyond providing simple statements about correctness and provide feedback appropriate to each student's learning needs.

Mentoring systems should support learners with decision-making and reasoning, especially in volatile and rapidly changing environments. Learners often need to make informed decisions and justify them with evidence, gathered through collaboration and communication (see [24, 25]). Students need to learn science practices, scientific reasoning and how to apply facts and skills they have acquired. In collaborative

learning, students share their experiences and perhaps persuade others to see their point of view, and articulate what they need to learn more about. They "mess about" and generate their own questions about the targeted science. Groups of students need to be supported as they discuss their methods and results, ask questions and make suggestions.

**Respond to student affect.** Student emotion while learning is critical to understanding student behavior. Researchers are developing intelligent tutoring systems that interpret and adapt to the different student emotional states [26, 27]. Humans do not just use cognitive processes to learn; they also use affective processes. For example, learners learn better when they have a certain level of disequilibrium (frustration), but not enough to make the learner feel completely overwhelmed [28]. This has motivated researchers in affective computing to produce and creating intelligent tutoring systems that can interpret the affective process of students. An intelligent tutor can be developed to read an individual's expressions and other signs of affect in an attempt to find and guide the student to the optimal affective state for learning. There are many complications in doing this since affect is not expressed in just one way but in multiple ways so that for a tutor to be effective in interpreting affective states it may require a multimodal approach (tone, facial expression, etc.). One example of a tutor that addresses affect is Gaze Tutor that was developed to track students' eye movements and determine whether they are bored or distracted and then the system attempts to reengage the student [29].

AI might be a game changer in education. It provides tools to build computational models of students' skills and to scaffold learning. AI methods can act as catalysts in learning environments to provide knowledge about the domain, student and teaching strategies through the integration of cognitive and emotional modeling, knowledge representation, reasoning, natural language question-answering and machine learning methods [30]. When such tutors work smoothly they provide flexible and adaptive feedback to students, enabling content to be customized to fit personal needs and abilities and to augment a teacher's ability to respond. AI techniques appear to be essential ingredients for achieving mentors for every learner.

User models are being developed that leverage advanced reasoning and inference-making tools from AI, represent inferences about users, including their level of knowledge, misconceptions, goals, plans, preferences, beliefs, and relevant characteristics (stereotypes) along with records of their past interactions with the system. They might also include information on the cultural preferences of learners [31] and their personal interests and learning goals. When modeling groups of learners, the model should make inferences to identify the group skills and behavior.

Finally, providing a mentor for every learning group means improving the ability of intelligent tutors to provide timely and appropriate guidance. In other words, tutors need to determine in real-time what to say, when to say it, and how to say it. This grows more complicated as the skills demanded by society increase in complexity. The learning sciences have provided a wealth of knowledge about how to deliver

effective feedback, but the challenge is to incorporate 21st century skills, such as creativity and teamwork.

### 4.2 Teach 21st Century Skills

Citizens of the 21st century require different skills than did citizens from earlier centuries [20]. 21st century skills include cognitive skills (non-routine problem solving, systems thinking and critical thinking), interpersonal skills (ranging from active listening, to presentation skills, to conflict resolution) and intrapersonal skills (broadly clustered under adaptability and self-management /self-development personal qualities) [32]. We describe two AI techniques that can improve teaching for 21st Century skills: dialogue systems and inquiry learning.

**Dialogue Systems.** One key development for teaching 21st century skills is implementation of strong dialogue and communication systems. Human tutors can understand a student's tone and inflection within a dialogue and interpret this to provide continual feedback through ongoing dialogue. Intelligent tutoring systems are still limited in dialogue and feedback. Systems that begin to simulate natural conversations have been developed [33, 34]. However, more research is needed to understand student tone, inflection, body language, and facial expression and then to respond to these. Dialogue modules in tutors should ask specific questions to guide students and elicit information while supporting them to construct their own knowledge [33, 34]. The development of more sophisticated dialogues between computers and students partially addresses the current limitations in human-computer communication and creates more constructivist teaching approaches.

The 21st century worker needs both 'hard' skills (traditional domains, such as, history, mathematics, science) as well as 'soft' skills (teamwork, reasoning, disciplined thinking, creativity, social skills, meta-cognitive skills, computer literacy, ability to evaluate and analyze information). Further, working in today's knowledge economy requires a high comfort with uncertainty, a willingness to take calculated risks, and an ability to generate novel solutions to problems that evade rigorous description. Unfortunately, many of today's classrooms look exactly like 19th century classrooms; teachers lecture and students remain passive and work alone on homework problems that do not require deep understanding or the application of concepts to realistic problems. Our system of education is behind and the gap grows wider each day.

As we know, changes in educational policy, practice and administration tend to happen slowly. For example, in the U.S. about 25 years are required for an individual to receive a sufficiently well-rounded education to become a proficient educator [30, 35]. The impact of that individual's teaching cannot be seen in subsequent learners for another 20 years. Thus the total cycle time for learning improvement is on the order of 45 to 50 years. Very few challenges in research or social policy cover such a long time scale [36].

**Inquiry and Collaborative Learning.** What type of technology is needed to mentor students as they learn complex, ill-structured problems? How can technology support exploratory behavior and creativity? Open-ended and exploratory inquiry-based

systems support learners to question and enhance their understanding about new areas of knowledge [37, 38]. Innovative instructional approaches, such as preparation for future learning, have uncovered ways to increase comfort with uncertainty and promote development of adaptive expertise [39].

Engagement in the information society often requires people to collaborate and exchange real-time responses over lengthy time periods [20]. A single individual working alone over time often cannot provide enough expertise to solve modern problems (e.g., environmental issues, sustainability, security). Technology is needed to support small groups, class discussions, 'white boarding,' and the generation of questions. To support learners in groups, networking tools are needed to facilitate individuals to learn within communities, communities to construct knowledge, and communities to learn from one another [40-43]. AI software is needed to support students in collaboration, researchers to examine learning communities and learning communities to morph into global communities. For example, how do learning communities sustain, build on, and share knowledge? Students clearly do not construct original knowledge in the same way as do research communities, but they can learn from community-based project work [44].

Support for inquiry and collaboration is needed as students become exposed to diverse cultures and viewpoints. What is the process by which teams generate, evaluate, and revise knowledge? How can we enhance learners' communication skills and creative abilities? Which tools match learners with other learners and/or mentors taking into account learner interests? Finally research is needed to support exploratory, social, and ubiquitous learning. How can software both support collaboration and coach about content? Can technology support continuous learning by groups of learners in ways that enable students to communicate what they are working on and receive help as needed? Learning communities, networking, collaboration software and mobile and ubiquitous computing are being used to create seamless social learning [41]. Socially embedded and social driven learning is pervasive.

In a society built on knowledge, citizens need to acquire new knowledge quickly, to explore alternative problem solving approaches regularly and to form new learning communities effectively [20]. People need to tackle knowledge challenges and opportunities. For educators, this requires rapid revision of what is taught and how it is presented to take advantage of evolving knowledge in a field where technology changes every few years. As an example of rapid change and unpredictability, consider the Internet itself. It first appeared in the mid-1990s. By 2015, 37.3% of the Earth's population uses it. Internet services and applications apply to virtually every aspect of modern human life (e.g., research, banking, shopping, meeting people, health, travel, job seeking). How can education prepare students for a society that changes so dramatically and rapidly? In just 25 years the Internet has become a major factor in nearly every civilized activity and applies to virtually every aspect of human life. At the minimum, students need to be taught how to search it, learn from it, evaluate its information, use it wisely, and contribute to it with well-vetted information. One answer lies in improved and expanded learner competencies. Learners must be more creative, more agile, and more able to learn in groups; they

must know how to learn. Key features include skills in critical thinking, creativity, collaboration, meta-cognition and motivation.

## 5 Discussion

This article described why AI is vital in Education and identified two challenges: personalized teaching and learning 21st century skills. Specifically, personalized learning should be supported by tools that enhance student and group experience, reflection, analysis, and theory development. Learning 21st century skills should be facilitated by resources that improve human-computer interfaces (dialogue systems) and inquiry-based and collaborative learning. We also expect AI technology to contribute to richer experiences for learners who will then be able to reflect on their own learning. Learning scientists with AI tools will have new opportunities to analyze vast data sets of instructional behavior collected from rich databases, containing elements of learning, affect, motivation, and social interaction.

Research shows that skilled workers have more job opportunities than do less skilled workers [45]. As technology advances, educated workers tend to benefit more, and workers with less education tend to have their jobs automated.

Over the next few years we expect intelligent online instruction to increasingly be a part of the online learning landscape [46]. Maybe in five years, children will increasingly be online with educational games and simulation environments; behind the scene will be intelligent tutoring capabilities adapting the environment. Similar to working with Google, people may not know what the adaptation algorithm is doing, but it is changing the individual search ranking in the background [46]. Algorithms are there and making search more effective. Similarly, students will see action like this in the educational material they use, with intelligence in the background. Intelligent tutors may provide many of the benefits of a human tutor and also provide real-time data to instructors and developers looking to refine teaching methods.

## References


1. Russell, S.J. and P. Norvig, *Artificial Intelligence: A Modern Approach*. 3rd ed. 2010, Upper Saddle River: Prentice Hall.
2. Koedinger, K., et al., *New potentials for data-driven intelligent tutoring system development and optimization*, in *AI Magazine, Special Issue on Intelligent Learning Technologies*. 2013.
3. Anderson, J.R., et al., *Cognitive tutors: Lessons learned.* The Journal of the Learning Sciences, 1995. **4**(2): p. 167-207.
4. Conati, C. and S. Kardan, *Student modeling: Supporting personalized instruction, from problem solving to exploratory open ended activities.* AI Magazine, 2013. **34**(3): p. 13-26.
5. Oblinger, D.G., *Game Changers: Education and Information Technology*. 2012, Educause: Washington, D.C.
6. McArthur, D., Lewis, M. and M. Bishay, *The Roles of Artificial Intelligence in Education: Current Progress and Future Prospects*. 1994: RAND DRU-472-NSF.
7. AIE. *AI Effect*. 2015; Available from: Retrieved from http://en.wikipedia.org/wiki/AI_effect.

8. McCorduck, P., *Machines Who Think*. 2nd Edition ed. 2004, Natick, MA: A. K. Peters, Ltd.
9. Kahn, J., *It's Alive*, in *Wired*. 2002.
10. Watson. 2015; Available from: retrieved from http://en.wikipedia.org/wiki/Watson_(computer).
11. Swaine, M. *AI: It's OK Again!, Dr. Dobbs, The World of Software Development*. 2007 May 20, 2015]; Available from: Retrieved from http://www.drdobbs.com/architecture-and-design/ai-its-ok-again/201804174.
12. Hofstadter, D., *Gödel, Escher, Bach: an Eternal Golden Braid*. 1980: Basic Books.
13. AIED. *Call for workshop proposals AIED*. 2015; Available from: Retrieved from http://WWW.sussex.ac.uk/Users/bend/aied2015/.
14. Hewlett, W., *Programs in deeper learning*. 2013, William and Flora Hewlett Foundation.
15. VanderArk, T. and C. Schneider, *How Digital Learning Contributes to Deeper Learning*. 2012, Getting Smart: http://gettingsmart.com/wp-content/uploads/2012/12/Digital-Learning-Deeper-Learning-Full-White-Paper.pdf. p. 26.
16. Arroyo, I., et al., *Emotion sensors go to school*, in *Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED'09)*, V. Dimitrova, et al., Editors. 2009, IOS Press: Brighton, UK. p. 17-24.
17. Chaouachi, M., et al., *Affect and mental engagement: Towards adaptability to intelligent systems*, in *The 23th International FLAIRS Conference*. 2010, AAAI Press: Daytona Beach, FL, USA.
18. D'Mello, S., A. Graesser, and R.W. Picard, *Toward an affect-sensitive AutoTutor.* Intelligent Systems, IEEE, 2007. **22**(4): p. 53-61.
19. Frasson, C. and A. Heraz, *Emotional Learning. Encyclopedia of the Sciences of Learning*. 2011: Springer Verlag.
20. Woolf, B.P., et al., *AI Grand Challenges for Education.* AI Magazine, 2013. **34**(4): p. 66-83.
21. Goleman, D., *Emotional Intelligence: why it can matter more than IQ*. Bloomsbury. 1996, London.
22. Pekrun, R., et al., *Boredom in achievement settings: Exploring control–value antecedents and performance outcomes of a neglected emotion.* Journal of Educational Psychology, 2010. **102**(3): p. 531-549.
23. Hmelo, C.E., *Problem-based learning: Effects on the early acquisition of cognitive skill in medicine.* The Journal of the Learning Sciences, 1998. **7**(2): p. 173-208.
24. Rus, V., et al., *Recent Advances in Conversational Intelligent Tutoring Systems*, in *AI Magazine, Special Issue on Intelligent Learning Technologies*. 2013.
25. Swartout, W., et al., *Virtual humans for learning*, in *AI Magazine, Special Issue on Intelligent Learning Technologies*. 2013.
26. D'Mello, S.K. and A. Graesser, *Language and discourse are powerful signals of student emotions during tutoring.* IEEE Transactions on Learning Technologies, 2012. **5**(4): p. 304-317.
27. Woolf, B., et al., *The effect of motivational learning companions on low achieving students and students with disabilities.*, in *10th International Conference on Intelligent Tutoring Systems (ITS'10)*, V. Aleven, J. Kay, and J. Mostow, Editors. 2010, Springer Berlin/Heidelberg. p. 327-337.
28. Graesser, A.C., et al., *Detection of emotions during learning with AutoTutor*, in *Proceedings of the 28th Annual Meeting of the Cognitive Science Society*. 2006, Cognitive Science Society. p. 285-290.
29. D'Mello, S., et al., *Gaze tutor: A gaze-reactive intelligent tutoring system.* International Journal of human-computer studies, 2012. **70**(5): p. 377-398.
30. Woolf, B.P., *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning*. 2009, San Francisco, CA: Morgan Kauffman.

31. Blanchard, E.G. and D. Allard, *Handbook of Research on Culturally-Aware Information Technology: Perspectives and Models*. 2010, Hershey, PA: Information Science Publishing.
32. Pellegrino, J. and M. Hilton, Editors, *Committee on Defining Deeper Learning and 21st Century Skills*. 2012, National Research Council: http://www.leg.state.vt.us/WorkGroups/EdOp/Education for Life and Work- National Academy of Sciences.pdf.
33. Graesser, A.C., VanLehn, K. , Rosé, C.P., Jordan, P.W., and Harter, D., *Intelligent tutoring systems with conversational dialogue*, in *AI Magazine*. 2001, American Association for Artificial Intelligence. p. 39-51.
34. Rosé, C., et al., *Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning.* International journal of computer-supported collaborative learning, 2008. **3**(3): p. 237-271.
35. King, J., N. Sabelli, and H. Kelly, *Preamble on Policy Issues*, in *Global Resources for Online Education (GROE) Workshop*. 2009: Tempe, AZ.
36. Roschelle, J., et al., *Eight issues for learning scientists about education and the economy.* Journal of the Learning Sciences, 2011. **20**(1): p. 3-49.
37. Dragon, T., B.P. Woolf, and T. Murray, *Intelligent coaching for collaboration in ill-defined domains*, in *Proceedings of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling* 2009, IOS Press. p. 740-742.
38. Floryan, M. and B.P. Woolf, *Improving the efficiency of automatic knowledge generation through games and simulations*, in *Proceedings of 16th International Conference on Artificial Intelligence in Education (AIED-2013). LNAI 7926*, K. Yacef, Editor. 2013, Springer: Heidelberg.
39. Schwartz, D.L. and T. Martin, *Inventing to Prepare for Future Learning: The Hidden Efficiency of Encouraging Original Student Production in Statistics Instruction.* Cognition and Instruction, 2004. **22**(2): p. 129-184.
40. Suthers, D., *Representational support for collaborative inquiry*, in *System Sciences, 1999. HICSS-32. Proceedings of the 32nd Annual Hawaii InternationalConference on System Sciences* 1999, IEEE. p. 14-.
41. Suthers, D.D., *Representational guidance for collaborative inquiry*, in *Arguing to learn: Confronting cognitions in computer-supported collaborative learning environments.*, J. Andriessen, M.J. Baker, and D.D. Suthers, Editors. 2003, Kluwer Academic: Dordrecht. p. 27-46.
42. Suthers, D.D. and C. Hundhausen, *An experimental study of the effects of representational guidance on collaborative learning processes.* Journal of the Learning Sciences, 2003. **12**(2): p. 183-219.
43. Woolf , B.P., *A Roadmap for Education Technology.* 2010, A Report to the Computing Community Consortium: http://telearn.archives-ouvertes.fr/docs/00/58/82/91/PDF/groe_roadmap_for_education_technology_final_report_003036v1_.pdf. p. 80 pp.
44. Johnson, D.W., Johnson, R., *An Overview of Cooperative Learning*, in *Creativity and Collaborative Learning*, J.Thousand, A. Villa, and A. Nevin, Editors. 1994, Brookes Press: Baltimore, MD.
45. Brynjolfsson, E. and A. McAfee, *Race Against The Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy*. 2013, Lexington, MA: Digital Frontier Press.
46. Koedinger, K., *7 Things You Should Know About Intelligent Tutoring Systems*, in *EDUCAUSE Learning Initiative (ELI)*. 2013.

# AI and Ed: a Happy Open Marriage

Julita Vassileva[1] James Lester[2], Judith Masthoff[3]

[1]University of Saskatchewan, Canada, jiv@cs.usask.ca
[2]North Carolina State University, USA, lester@ncsu.edu
[3]University of Aberdeen, Scotland, j.masthoff@abdn.ac.uk

**Abstract.** We claim that this marriage has never been closed and exclusive. It started because both AI and Education share the goal of understanding the human process of knowing, and getting to know, i.e. learning. The difference is how the two areas exploit the understanding they aim to develop. AI is more focused on making machines that know and learn like people or better than them. AIED is more interested in supporting people to learn better.

**Keywords:** AI in Education, AIED, AI

## 1 Introduction

AI originated from the curiosity of understanding how the human mind works and creating models of reasoning and machines that mimic and improve on human reasoning (using the capacities of computers). The early research in AI started with theoretical studies in reasoning and knowledge representation, metacognition, and learning of a single human (single agent). This research, married the area of Cognitive Psychology and lead to the creation of the area of Cognitive Modeling. The need for practical applications drove the formation of many "children" areas of applied AI: Expert Systems, Probabilistic Reasoning, User Modeling, Ontologies (and more recently, Semantic Web and Linked Data), and Advanced Learning Algorithms (which branched more recently into Data Mining, Data Analytics, Data-warehousing etc.).

Around the mid 1990ies, the theoretical interest shifted towards situated action and social reasoning, and multi-agent architectures, leading to the creation of the area of Multi-Agent Systems. Theoretical studies in Argumentation and Negotiation followed with the creation of their own research areas. The area of Interactive Virtual Agents (IVA) emerged around the end of the 1990ies. Another "child" area of applied AI is Recommender Systems (RecSys), which deploys user modeling and advanced learning algorithms to emerging CS application areas, such as e-commerce. Around the same time, some AI researchers turned their sight to modeling other human psychological phenomena such as emotion and affect, which lead to the establishment of the Affective Computing area.

## 2 AI and CogPsy Meet Education

AI in Education has been "married" to all of these children of AI. Early ITS work in the 1980s and early 1990ies on pedagogical planning, domain knowledge modeling,

student modeling and ITS shells applied techniques from the areas of planning expert systems, and knowledge representation. The second half of the 1990s saw attention shift to agent-based tutoring systems, tutorial dialogues, animated characters, and the first works on modeling learner affect and adapting the interaction with the tutor. In the beginning of the new century the application of ontologies and semantic web technologies for learning material annotation and concept maps for domain knowledge representation took a center stage and the first applications of recommender systems for learning materials considering both content based and social recommendations, and visualizations appeared to explain both the recommendations and the student model (social navigation, open learner modeling). We have seen many research topics in AIED evolve into its own children areas, such as CSCL (a child of AIED, the Learning Sciences and CSCW) and EDM (a child of Data Mining and AIED).

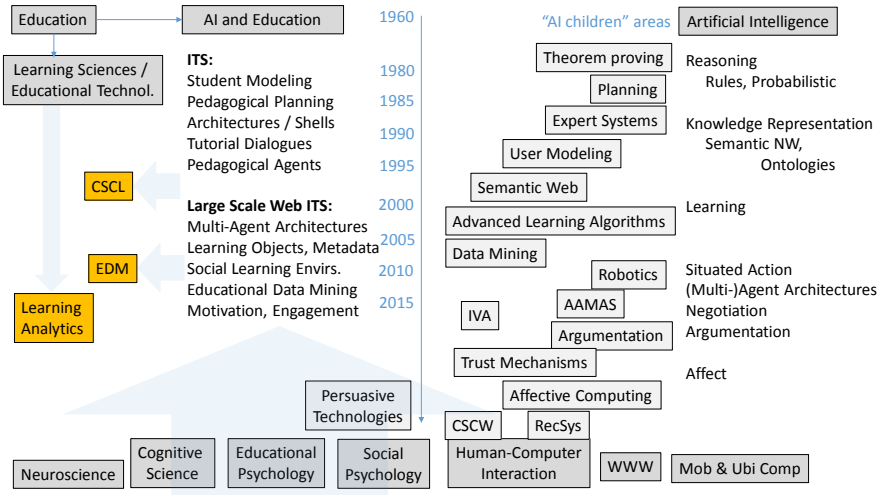



**Figure**: Approximate evolution of AIED research topics along with AI research topics and the emerging applied AI - children areas, and the influences of other areas of CS and other disciplines AI research topics and

Motivation is an important factor in learning, and the first attempts to model computationally motivational pedagogical strategies started around 1995. The OCC model of emotions triggered interest in incorporating affective factors in HCI around 2000, and it was very soon followed by work in the AIED area, on modeling affect in learning scenarios.

The realization that students engage in off-task behaviours or "game the system" around 2005 lead to increased interest in learner motivation and engagement, as well as educational games (and gamification) which started 10 years earlier. Yet the inspiration for this work is found often in other disciplines (Social Psychology, Persuasive Technology, Behaviour change and even Neuroscience), rather than in Affective Computing.

## 3. Exploration vs Rigour

The influence of the above AI-children areas, broader computer science areas and other disciplines has been not just in the choice of research topics of AIED researchers, but also in the methods used to carry out, design and evaluate the research. In the early years the focus was on constructing working ITS and a typical research paper included a detailed design justification, description and perhaps a couple of screen-shots as "proof of existence", with not much evaluation. Later on it became necessary to present evaluation data – even if it consisted only of the number of students who liked the system. With the increasing influence of Educational Psychology, evaluation methods from the behavioural sciences were introduced in the area. This coincided with the rapid development of web technologies and tools that allowed an easy design of systems and easier experimentation with more subjects as the ITS prototypes were now accessible on the web. The CMU cognitive tutors were successfully applied with thousands of children in the US, and they started producing a lot of data allowing to evaluate the learning effects on a large scale and for long term use. After 2005, statistical methods for evaluation became a standard, and a typical research paper in the area became much more like a psychology paper or a natural science paper than an engineering paper. The main point became studying the phenomenon of a human interacting with an "experimental tool" designed based on a particular theoretical foundation, and in a way, a significant part of the research in AIED became a branch of applied Cognitive Science. Researchers who were more interested in building systems than in studying human cognition wandered off to other areas that focused more on the technologies, for example ICALT, EC-Tel, Web-Based Learning.

Yet, there are still researchers interested in developing further the "tools", not only from the point of view of the underlying cognitive theories, but also, from the available new technologies developed in the meantime by the Mobile & Ubiquitous Computing community, new data-mining techniques that can allow to automatically learn and improve pedagogical decisions (not necessarily based on theory). The AIED community needs the researchers interested in technology so that the field doesn't become stagnant, overly constrained by methodology, making miniscule improvements based on the same mature AI technologies. So the marriage between AIED and the younger AI children (such as Recommender Systems, or Affective Computing, and even "embryo" areas such as Mechanism Design, Trust, and Negotiation in AAMAS) is important.

When we look at the complex map of how the research topics and children areas emerged, we can notice that in several areas the connection is bi-directional. For example, the area of User Modeling and Personalization, which emerged as a child of AI, has been strongly influenced by AIED. Similarly, IVA, and Affective Computing, have moved ahead to a large extent due to insights and case studies in the context of educational applications. Newly emerged areas, such as Persuasive Technologies also have a lot to learn from the area of AIED, and AIED has a lot to learn from them.

## 3 Conclusion

So, in conclusion, the marriage between AI and Education and AI is in name only, as much as the name AI describes the inspiration of understanding how the human mind works and creating models (with practical use) of human mind. In fact it has been more of an "open marriage" with quite a few partners – the children areas of applied AI and some other areas and disciplines (as shown in The Figure). Yet, AI is a good family to be in – a large and productive family of smart people. In many ways, it is a perfect marriage.

# AI *in* Education as a methodology for enabling educational evidence-based practice

Kaśka Porayska-Pomsta*

University College London, Institute of Education, London Knowledge Lab,
23-29 Emerald Street, London WC1N 3QS K.Porayska-Pomsta@ioe.ac.uk

**Abstract.** Evidence based practice (EBP) is of critical importance in Education where, increasingly, emphasis is placed on the need to equip teachers with an ability to independently generate evidence of their best practices *in situ*. Such contextualised evidence is seen as the key to informing educational practices more generally. One of the key challenges related to EBP lies in the paucity of methods that would allow educational practitioners to generate evidence of their practices at a low-level of detail in a way that is inspectable and reproducible by others. This position paper focuses on the utility and relevance of AI methods of knowledge elicitation and knowledge representation as a means for supporting educational evidence-based practices through *action research*. AI offers methods whose service extends beyond building of ILEs and into real-world teaching practices, whereby teachers can acquire and apply computational design thinking needed to generate the evidence of interest. This opens a new dimension for AIEd as a field, i.e. one that demonstrates explicitly the continuing pertinence and a maturing reciprocity of the relationship between AI and Education.

## 1 Introduction

AI methods of knowledge representation and knowledge elicitation can make an important contribution to supporting educational evidence-based practices (EBP) through Action Research (AR). EBP is of critical importance in education where, increasingly, emphasis is placed on the need to equip teachers with an ability to independently generate evidence of their best practices *in situ* [8]. Such evidence is seen as the key to informing educational practices more generally. One of the key challenges related to EBP lies in the lack of readily available methods that would support the generation of evidence by practitioners at a fine-grained level of detail and in a way that is reproducible by other practitioners. There is also a notable lack of consensus as to what constitutes good evidence

* My colleagues Manolis Mavrikis, Karen Guldberg, Sarah Parsons, Helen Pain and Mina Vasalou have all contributed over the years in different ways to the development of the position presented in this paper, as have all of my students taking the *Learning and Teaching with Technologies* module at the UCL Institute of Education. LeAM and TARDIS were both funded by the European Commission (FP6-IST-2003-507826 and FP7-ICT2011-7-288578 respectively).

in education, with randomised controlled studies being typically favoured due to being seen as leading to measurable results similar to those in the biological and medical sciences – currently the gold standard of scientific rigour. Unfortunately, given the inextricable dependency of educational outcomes on the context within which learning and teaching takes place, e.g. [1], the results of such studies tend to have limited generalisability. Education requires a more nuanced and transparent approach than a pill-like medical intervention approaches can offer; they need to serve as tools for teacher reflection and experimentation in order to provide an informed basis for effecting positive change on the learners.

## 2 In pursuit of a broader definition of AI in Education

AI methods used to elicit knowledge of teaching and learning processes and to represent such knowledge computationally, offer the tools needed by teachers to gather evidence in a systematic, detailed and incremental manner that can be also shared with and inspected by others. Viewing the contribution of AI to Education as a methodological one opens up an important perspective on the possible role of AI in Education than has been adopted to date. Some important fundaments for the adoption of such a perspective have been laid some thirty years ago by Alan Bundy who categorised Artificial Intelligence (AI) field in terms of three kinds of AI: (i) basic AI, aiming to explore computational techniques to simulate intelligent behaviour, (ii) applied AI, concerned with using existing AI techniques to build products for real-world use and (iii) cognitive science, or computational psychology, focusing on the study of human or animal intelligence through computational means [2]. In doing so, Bundy highlighted the diversity of motivations for *doing* AI and, consequently, of the methodologies to both inform and evaluate systems that are underpinned with AI. This diversity of motivations was also noted by Mark and Greer [10] in their exploration of the AIEd evaluations methodologies, where they highlighted the distinction between formative and summative evaluations. Retrospectively, this distinction remains crucial insofar as it allows for a more precise definition of AIEd within the wider fields of AI and Education, by bringing to the fore the dependency between the technologies engineered within AIEd and the purpose, context and design of their use. Over the years, the role of formative evaluation has been elaborated by AIEd researchers based on the growing aspirations of the community not only to establish some ground truths to inform the design and implementation of AIEd technologies, but also to connect AIEd research with educational practices.

Conlon and Pain [5], who relied on Bundy's 3-kind definition of AI to provide their own vision of AIEd, proposed a Persistent Collaboration Methodology (PCM) as a means of ensuring the real-world relevance and effectiveness of the AIEd technologies and to enhance rigour of the design, implementation and evaluation process. PCM draws equally from the key educational methodology of Action Research (AR) [4], applied AI approaches to knowledge elicitation and representation, and human-computer interaction (HCI) design. In contrast with the prevalent practices at the time, PCM advocated that early and continuous

involvement of practitioners specifically as *action researchers* in the design and evaluation of AIEd technologies is essential to securing the educational validity of such technologies, to enabling a contribution to both AI and educational theories and practices, and to achieving a balance in the emerging technologies and research between the 'technological push' and 'educational pull'. While inspirational in its effort to acknowledge and marry educational and AI methods PCM remains firmly within the boundaries of AIEd practices offering insights as to the best educational systems designs, but not necessarily as to the best educational practices more generally. In the next two sections I discuss the affordances of knowledge representation as a conceptual tool of relevance to educational practices and, using two examples, I illustrate the role of knowledge elicitation as a means for utilising and for developing this conceptual tool further.

## 3 Knowledge Representation

Knowledge representation (KR) is fundamental to AI and, arguably, to any scientific endeavour, because at its very basic (and most general), it is a *conceptual* tool for describing and reasoning about the world we inhabit. Scientific theories are in essence forms of knowledge representation about the world, albeit delivered at different levels of specificity. In AI, knowledge representation is inevitably and by definition a theory of intelligence, or more precisely – of intelligent reasoning.

Davis et al. [6] define knowledge representation in terms of five distinct roles that it plays in AI. The first and overarching role of KR, is to serve as a *surrogate* of the thing itself, i.e. the world being represented. As a surrogate, KR offers us (or a computer system) a means for reasoning about the world without having to take action in it, i.e. it allows us to determine consequences within the world we describe by thinking about them rather than by enacting them. Thus, KR provides tools for thinking about and for refining our perceptions of the world, which are, at least conceptual and, at their most usable, computational in nature.

The second role of KR is in forcing us to make *ontological commitments* that tell us how to see the world, i.e. what kind of concepts, entities, etc. and relationships between them describe the world. Since it is impractical (and impossible) to represent all of the characteristics of the world, Davis et al. refer to these ontological commitments as a "strong pair of glasses that determine what we can see, bringing some parts of the world into sharp focus, at the expense of blurring other parts.". They highlight that such focusing/blurring is the greatest affordance of KR in that it enables decisions about what to attend to and what to ignore in our world (Davis et al., [6], p.5). Although ontologies are language agnostic, the choice of representation technologies[1] will impact on what specific commitments we make; logic, rules, frames, semantic nets, etc., constitute different representation technologies, each encapsulating a specific viewpoint on what kinds of things are important in the world. For example, frames use a prototypes viewpoint, whereas logic focuses on individual entities and the relations between

[1] This is the term is used by Davis et al. to refer to "the familiar set of basic representation tools like logic, rules, frames, semantic nets, etc." (p.3)

them. These are by no means the only representation technologies available in AI and neither are they the only technologies that are possible or needed for some domains. In Education and AIEd, ontologies are relatively well understood and accepted as forms of representations of specific subject domains and of knowledge about the learner. However, while they inform us about a possible view of the world, in terms of its component parts, they do not tell us how we can reason about the world using those parts.

The third role of KR is therefore as *a theory of intelligent reasoning*, which tells us what inferences we can and should draw (*sanctions* vs. *recommendations*, respectively), given our ontological commitments. Recommendations define what inferences are appropriate to make and hence which ones are intelligent. A theory of intelligent reasoning lies at the core of AI and, arguably, of educational *practice*, because it is critically concerned with understanding intelligent action and its relationship to the external world [7];[1]. It is this relationship that resides at the heart of teachers' adaptive capabilities and it is in capturing it that one of the greatest challenges for AIEd (*and* Education) lies. This challenge is all the more, because KR related to reasoning involves making the fundamental choice of a theory of intelligent reasoning that must underpin a given representation. Given many different conceptions of intelligent reasoning (e.g. logic, psychology, biology, statistics and economics, etc.) such choice will yield very different conclusions and hence, yet again, different views of the world. For example, logic views reasoning as a form of calculation such as deduction, whereas a theory derived from psychology views intelligent reasoning as a variety of human behaviour, plausibly involving structures such as goals, plans or expectations. Education too offers a variety of different theories of learning, each engendering inferences that are possible and needed. The contrast between approaches which view learning as an outcome of a pre-designed intervention or as an outcome of a transactional experience offers one example.

The fourth role of KR is as a *medium for pragmatically efficient computation.* As such KR provides an environment in which thinking can be accomplished (and conclusions drawn). Ontological and inferential representations jointly provide a contribution to defining such an environment and although they do not in themselves guarantee full computational efficiency, the choice of the specific representation technologies and of intelligent reasoning theory must act in support of achieving such efficiency. While educational theories of learning as transactional and situated experiences are abound they tend to lack specificity as to how exactly such experiences can be captured, described and reasoned about. And while AIEd research provides numerous accounts of such mechanisms and explicitly considers computational efficiency (both as relate to problem solving and affect, e.g. [11]), those accounts tend to be limited in scope and in their power to convince educational community of their applicability to wider education.

The fifth (and final) role of KR is as a *medium of human expression*, i.e. a language through which we convey and ground our view of the world. As such KR allows us to share the different representations with other people. It is precisely the affordance of being sharable and inspectable that makes KR such a

compelling candidate as a conceptual tool for supporting evidence-based practices in education. This affordance is also of crucial relevance to AIEd practices: at least in principle, the representations created by educational practitioners can provide rich source of authentic data that can then be used to inform the AIEd systems. However, how successfully the affordances of KR as a medium for expression can *actually* be exploited at the intersection of AIEd and Education, hangs on an understanding that although it does not matter what language we employ to express our world view, the language that we do employ has to be easy to use. As Davis et al. put it "If the representation makes things possible but not easy, then as real users we may never know whether we have misunderstood the representation and just do not know how to use it, or it truly cannot express things we would like to say". Thus, a representation has to provide a language in which we can communicate without having to make a *heroic effort* (p.15).

Davis et al.'s definition of KR in AI is very useful in highlighting its role as a tool for thinking with and as a method for understanding the complexities of our internal and external experiences. There are at least four different ways in which KR as a methodology can serve education. First, it forces us to make explicit our tacit knowledge about the world and the relationships therein. Representing such tacit knowledge enables us not only to reflect on the world that we represent, but also to gain a better understanding of what it is that we actually know. Such reflection is key to educational practice because it brings into focus the strengths and weaknesses in the particular approaches to supporting learning and the kinds of priorities that may characterise such support. Second, KR allows us to create different knowledge representations of the same phenomenon without having to fundamentally change the way we act in the real world. This is important in education where any efforts to effectuate a change involve real and potentially life long impact on real people (the learners) and wherefore such efforts must always be based on informed choices. Third, KR allows us to observe the possible consequences of the different representations on the world, thus enhancing our predictive powers, without involving the actual experience of such consequences. As with the second point, this is important to our being granted access to different viewpoints on the same phenomenon, but this time we also have access to various possible consequences of adopting the different viewpoints. Fourth, KR allows us to share the different representations with other people to generate rich critiques of the different viewpoints and to enrich, update or change our existing viewpoints based on the perspectives of the others' unique experiences and understandings. As well as being shareable with others, KR can also provide a trace of our own views of the world over time and a basis for reflection and introspection on how our ideas evolved and what influenced them.

## 4 Knowledge elicitation

Knowledge elicitation (KE) is an inseparable companion of knowledge representation in that it is through KE that we engage in reflection about the world.

KE is a *process* in which we can engage alone (through self questioning) or with others, either collaboratively or as respondents to someone else's queries and the process can be either formal or informal, and structured or unstructured.

There are various forms of KE instruments that have been adopted, developed and tested in the context of AIEd. For example, questionnaires or interviews, have been borrowed directly from the social sciences, whereas methods such as post-hoc cognitive walkthroughs, gained in power and applicability with the advent of audio and video technologies, and further through logs of man-machine interactions. Other methods, e.g. Wizard of Oz (WoZ), have been devised as placeholders for yet-to-be-developed fully functional learning environments or components thereof, with the specific purpose of informing the design of technologies in a situated fine-grained level of detail way (e.g. see [12]).

Although KE is standardly employed in AIEd to inform the design of its technologies, its role as a means of explicitly informing educational practice is less well understood and it may be even regarded as somewhat out of AIEd's focus. Yet, it is precisely in examining both how KE informs the design of our technologies and how real educational practices may be affected by KE, that the idea of AI as a methodology, comes to life. It is through this two-way lens that we can start to appreciate the real value of creating a more transitive relationship between AI and Educational practices. Two research projects – LeActiveMath (in short *LeAM*[13]) and TARDIS [14] – serve to illustrate these points.

LeAM is a system in which learners at different stages in their education can engage with mathematical problems through natural language dialogue. It consists of a learner model, a tutorial component, an exercise repository, a domain reasoner and natural language dialogue capabilities. LeAM's design is based on the premise that the specific context of a situation along with the learner-teacher interaction are integral to both regulating learners emotions and to being able to recognise and act on them in pedagogically viable ways.

To inform the learner and the natural language dialogue models, studies were conducted using WoZ design and a bespoke chat interface. Specifically, the student-teacher communication channel was restricted to a typed interface with no visual or audio inputs to resemble the interface of the final learning environment. Five experienced tutors participated in the studies where they had to tutor individual learners in real time, delivering natural language feedback. They were asked to talk aloud about their feedback decisions as they engaged in tutoring and to further qualify those decisions by selecting situational factors, e.g. student confidence or difficulty of material, that they considered important in those decisions. The tutors were asked to make their factor selections through a purpose-built tool every time they provided feedback. To aid them in this task some factors were predefined (based on previous research), but these were not mandatory as the tutors could add their own factors to the existing set.

Following each completed interaction, the tutors were invited to participate in post-task walkthroughs, which synchronised a replay of (1) the recording of the student screen (2) the verbal protocol of the tutor and (3) the selected situational factors for the given interaction. Walkthroughs allowed the tutors

and the researchers to review specific interactions, to discuss them in detail, to explain their in-the-moment choices of factors, and to indicate any change in their assessment of the situations.

The data elicited provided a concrete basis for the implementation of LeAM's user and dialogue models and the corresponding knowledge representations. However, the studies also provided important insights into the potential impact that the KE process had on the participating tutors. Specifically, the demand on teachers' to report on the situational factors of importance to their feedback decisions brought to their attention that such factors may indeed play a role and forced them to think explicitly about them while making those decisions. Verbal protocols facilitated verbalisation of those decisions *while* they were made and later on provided an important tool for facilitating situated recall. Although initially, all tutors had a clear understanding of and an ability to identify the factors related to subject domain taught, e.g. the difficulty of the material or correctness of student answer, they were much less willing or fluent at diagnosing and talking about factors related to student's affective states. However, after an initial familiarisation period, involving up to two sessions, their willingness to engage in situational analysis and the fluency of their reports increased, while the tentativeness in identifying student behaviours at fine level of details decreased. This was evidenced primarily in the increased speed at which they engaged in the task, the fluency and quality of their verbal protocols and in the post-hoc interviews. Another interesting outcome was the tutors' increased attention to giving praise in their feedback, as well as a more targeted attention to possible relationship between the form of students' responses and their mental states.

The use of verbal protocols during the interactions, each of which was followed by semi-structured interviews, allowed the tutors to formulate hypotheses about the possible meanings of the students' different behaviours in terms of cognitive and affective states and to evaluate those first against the appropriateness of their feedback and then during subsequent tutoring sessions with further students. Finally, post-task walkthroughs were used with the tutors, during which situated recall was facilitated through replay of the video-recorded screens and verbal protocols. The fact that the tutors were given the opportunity to inspect their selection of situational factors and to correct them gave them an opportunity to assess the consistency of their interpretations and further, to analyse those situations where they did not agree with themselves, leading, in some tutors' own words, to deep reflection and grounding of their understanding of (a) what matters to them the most in tutoring situations and (b) the kinds of tutoring they want to be able to deliver *ideally*. The appreciation of the tutors' involvement in the LeAM's KE process was reflected in their request for a tutoring system for tutors, through which they could rehearse and perfect their understanding of the different nuances of educational interactions along with their pedagogical feedback and which they could also use to train novice tutors.

Although the realisation of the potential value of KE methods used to inform an intelligent tutoring system such as LeAM was very inspirational, the methods used, specifically, the way in which they were used, was fundamentally

research-centric. The studies were aimed specifically and exclusively to establish some ground truths about very particular kinds of educational interactions for the purpose of creating knowledge representations to underpin the system's learner modelling and natural language dialogue capabilities. As such the tutors participating in the LeAM studies were in essence merely willing informants for and testers of the technological design ideas. Because of the complexity of the studies' set up the tools and the methods used in the study did not lend themselves readily for independent use by the tutors.

The importance of practitioner independence in generating evidence of their practices is emphasised throughout the EBP literature, where it is often accompanied by the rhetoric of *action research* [4] and the call for practitioners as researchers of their own practices. This rhetoric was used to underpin the design of the TARDIS system – a serious game for coaching young people in job interview skills through interactions with intelligent conversational agents able to react to social cues and complex mental states as detected and modelled by TARDIS' user modelling tools [14]. The TARDIS project took LeAM's insights forward, by employing KE methods throughout. Apart from the goal of informing the design of the game, the goal was also to inform the *design of use* of such a game in real contexts of youth employment associations across Europe. Independence of use by practitioners as facilitators of this game was key. In TARDIS, KE was used as the basis for developing practitioners' self-observation and self-reporting skills, which were then built on in the formative evaluation studies, in which the practitioners increasingly participated as researchers, with the support by researchers being gradually removed. The whole process was divided into three stages, roughly corresponding to the three years of the project. The first stage (*familiarisation*) involved gradual preparation and training of practitioners in the application of knowledge elicitation for the purpose of knowledge representation in the domain of job interview training.

Post-hoc walkthroughs, using video replays of practice of job interview sessions between youngsters and practitioners were used to (a) access practitioners' expert knowledge to be represented in TARDIS; (b) allow the practitioners to make overt to themselves, and to the researchers, the types of knowledge and interpretations that are of particular interest in the context of job interview skills coaching and (c) allow the practitioners to reflect on their and the youngsters' needs, leading up to the specification of the necessary and sufficient elements of a technology-enhanced learning environment able to support those needs. This specification was captured in the form of requirements and recommendations, while the reflections were recorded as practitioners' videos annotations in an off-the-shelf tool called *Elan* (https://tla.mpi.nl/tools/tla-tools/elan/).

The second stage (*testing, critique and design of use*) involved a period of continuous cycles of reflection, observation, design and action scaffolded by researchers and guided by the Persistent Collaboration Methodology [5]. This stage was crucial not only to the TARDIS researchers who were able to implement ever more sophisticated prototypes, but it was also fundamental to the practitioners' growing confidence in providing targeted critique of those prototypes, to their

increased independence in using TARDIS and in experimenting with its different set-ups. Crucially, the knowledge self-elicitation skills, developed in the first year, along with their rehearsed focus on the type and form of information needed by the researchers to create the various computational models, provided the practitioners with a structure against which to report their observations and reflections to the researchers and a common language for both. One of the key outcomes of this was a growing sense of co-ownership of the tools and knowledge developed which was reflected in the independent curation of TARDIS tools by the practitioners who participated in the project to other practitioners. As such the participating practitioners became *lead-practitioners* in co-designing with their colleagues the use of TARDIS in their everyday practices. This independence was put to the test and further deepened in the third and final stage of the project, where the practitioners engaged in summative evaluation of the system with minimal support from the researchers (*independent use and research*). As well as being able to use the system independently and to explore new ways in which to utilise it within their existing practices, a key outcome was the practitioners' confidently vocal involvement in the development and testing of a schema for annotating data of youngsters engaging in job interviews. This schema was used directly in the analysis of the TARDIS evaluation data, offering the first such tool for examining job interview skills at the low level of detail needed to build user models and artificial agents in this domain [3].

The practitioners' roles and competencies have evidently changed from those of willing informants (the beginning of the project), through advisors and co-designers of the TARDIS system (middle of the project), to lead-practitioners who initiate projects independently (end of the project). At the core of this change was a gradual shift in the practitioners' way of thinking and viewing the world of their practice. Through engaging in KE and its eventual KR in terms of design recommendations and fine-grained specification of the domain and inferences therein (annotation schema), the practitioners' role in applying technology in their practices changed from that of mere consumers to its co-creators and owners. They demonstrated an ability to think about their domain and practices in terms that are by nature both computational (low level knowledge specification) and design (design of the technology's look-and-feel, functionality, as well as pedagogical design[2]). In other words the practitioners have demonstrated an emergent ability to engage in *computational design thinking*.

## 5 Conclusions

This paper argued a position that the relationship between AIEd and Education can be strengthened through the application of AI as a methodology for supporting educational evidence-based practices. AI offers to educational practitioners specific instruments for generating evidence of their practices that are inspectable and reproducible by the wider educational community. AI methods of knowledge elicitation and representation can enable practitioners to engage

[2] Note that some researchers in Education view teacher as a design science, e.g. [9]

in computational design thinking and this can engender practitioners independence in defining, creating and inspecting their real-world practices at a low-level of representational detail. Investing in educational practitioners using AI as a methodology is not entirely altruistic insofar as the specificity of the evidence thus generated creates an important opportunity for AIEd to tap into situated knowledge of educational practices in a way that supports the implementation of AIEd systems sustainably and over long-term. Such investment carries a promise of creating a dynamically generated knowledge infrastructure thereby reducing the often prohibitive cost of developing AIEd systems and by lending itself more readily to targeted mining and interpretation by the AIEd researchers and developers. Making the AI methods available to practitioners opens the AIEd research to critical, but informed inspection by some of its end-users and it offers a much needed opportunity to re-interrogate its approaches to connecting with existing educational practice, along with its future goals and aspirations more generally.

## References


1. Biesta, G.: Why 'what works' won't work: Evidence-based practice and the democratic deficit in educational research. Educational Theory 57(1) (2007)
2. Bundy, A.: What kind of field is artificial intelligence? In: DAI Research Paper No.305. Department of Artificial Intelligence, University of Edinburgh. (1986)
3. Chryssafidou, E., Porayska-Pomsta: Situated job interview coaching: Evaluating the role and efficacy of TARDIS serious game. Journal of Computers in Human Behavior (in prep)
4. Cohen, L., Manion, L.: Research Methods in Education. 2nd edn. Croom-Helm, Dover, NH (1990)
5. Conlon, T., Pain, H.: Persistent collaboration: methodology for applied AIED. Journal of Artificial Intelligence in Education 7, 219–252 (1996)
6. Davis, R., Shrobe, H., Szolovits, P.: What is knowledge representation? AI Magazine 14(1), 17–33 (1993)
7. Dewey, J.: Experience and Nature. Dover Publications Inc. (1998)
8. Hargreaves, A.: Revitalising educational research: lessons from the past and proposals for the future. Cambridge Journal of Education 29(2), 239–249 (1999)
9. Laurillard, D.: Teaching as a Design Science: Building Pedagogical Patterns for Learning and Technology. Routledge (2012)
10. Mark, M.A., Greer, J.: Evaluation methodologies for intelligent tutoring systems. Journal of Artificial Intelligence in Education 4, 129–153 (1993)
11. Porayska-Pomsta, K., Bernardini, S.: Learner modelled environments. In: Handbook of Digital Technology Research (2013)
12. Porayska-Pomsta, K., Mavrikis, M., D'Mello, S., Conati, C.and Baker, R.: Knowledge elicitation methods for affect modelling in education. International Journal of Artificial Intelligence in Education 22, 107–140 (2013)
13. Porayska-Pomsta, K., Mavrikis, M., Pain, H.: Diagnosing and acting on student affect: the tutors perspective. International Journal of Artificial Intelligence in Education 18, 125–173 (2008)
14. Porayska-Pomsta, K., Rizzo, P., Damian, I., Baur, T., Andre, E., Sabouret, N., Jones, H., Anderson, K., Chryssafidou, E.: Who's afraid of job interviews? definitely a question for user modelling. In: Proc. Conference on User Modeling, Adaptation and Personalization (2014)

# AIED Is Splitting Up (Into Services) and the Next Generation Will Be All Right

Benjamin D. Nye

Institute for Intelligent Systems, University of Memphis
365 Innovation Dr. Memphis, TN 38152
benjamin.nye@gmail.com

**Abstract.** Advanced learning technologies are reaching a new phase of their evolution where they are finally entering mainstream educational contexts, with persistent user bases. However, as AIED scales, it will need to follow recent trends in service-oriented and ubiquitous computing: breaking AIED platforms into distinct services that can be composed for different platforms (web, mobile, etc.) and distributed across multiple systems. This will represent a move from learning platforms to an ecosystem of interacting learning tools. Such tools will enable new opportunities for both user-adaptation and experimentation. Traditional macro-adaptation (problem selection) and step-based adaptation (hints and feedback) will be extended by meta-adaptation (adaptive system selection) and micro-adaptation (event-level optimization). The existence of persistent and widely-used systems will also support new paradigms for experimentation in education, allowing researchers to understand interactions and boundary conditions for learning principles. New central research questions for the field will also need to be answered due to these changes in the AIED landscape.

## 1 Introduction

Initial efforts to bring learning technology into schools faced hardware hurdles, such as insufficient computing resources. Later efforts encountered serious barriers related to matching technology to teachers' beliefs, pedagogy, and resource constraints. While all of these barriers are still relevant, learning technology is endemic in higher education and has made significant footholds in K-12 schools, with estimates of 25-30% of science classes using technology as early as 2012 (Banilower, Smith, Weiss, Malzahn, Campbell, & Weis, 2013). Correspondingly, an influx of investment into educational technology has occurred, with online learning doubling from a $50b industry to a $107b industry in only three years (Monsalve, 2014).

Future barriers will not be about getting learning technology into schools: they will be about competing, integrating, and collaborating with technologies already in schools. This is not an idle speculation, as it is already occurring. In a recent multi-year efficacy study to evaluate a major adaptive learning system, some teachers started using grant-purchased computers to use other math software as well (Craig, Hu, Graesser, Bargagliotti, Sterbinsky, Cheney, & Okwumabua, 2013). After working for

many years to get teachers to use technology, the point may come where they are using so many technologies that it is difficult to evaluate an intervention in isolation.

Some research-based artificial intelligence in education (AIED) technologies have already grown significant user bases, with notable examples that include the Cognitive Tutor (Ritter, Anderson, Koedinger, & Corbett, 2007), ALEKS (Falmagne, Albert, Doble, Eppstein, & Hu, 2013), and ASSISTments (Heffernan, Turner, Lourenco, Macasek, Nuzzo-Jones, & Koedinger, 2006). Traditionally non-adaptive systems with large user bases, such as Khan Academy and EdX, have also started to add basic adaptive learning and other intelligent features (Khan Academy, 2015; Siemens, 2013).

Large-scale online platforms are not just the future of learning, but they are also the future of research. Traditional AIED studies have been limited to dozens to hundreds of participants, sometimes just for a single session. While such studies will remain important for isolating new learning principles and collecting rich subject data (e.g., biometrics), large-scale platforms could be used to run continuously-randomized trials across thousands of participants that vary dozens or even hundreds of parameters (Beck and Mostow 2006; Liu, Mandel, Brunskill, & Popovic, 2014). Even for AIED work not based on such platforms, it is increasingly feasible to "plug in" to another system, with certain systems serving as active testbeds for 3rd-party experiments (e.g., ASSISTments and EdX).

The difference is qualitative: rather than being limited to exploring a handful of factors independently, it will be possible to explore the relative importance of different learning principles in different contexts and combinations. In many respects, this means not just a change to the systems, but to the kinds of scientific questions that can and will be studied. These opportunities raise new research problems for the field of AIED. A few areas related areas will reshape educational research: Distributed and Ubiquitous Intelligent Tutoring Systems (ITS), Four-Loop User Adaptation, AI-Controlled Experimental Sampling, and Semantic Messaging. Some new frontiers in each of these areas will be discussed.

## 2 Distributed and Ubiquitous AIED

As implied by the title, AIED technologies are approaching a juncture where many systems will be splitting up into an ecosystem of reusable infrastructure and platforms. The next generation of services will be composed of these services, which may be hosted across many different servers or institutions. More specifically, we may be reaching the end of the traditional four-component ITS architecture with four modules: Domain, Pedagogy, Student, and Communication (Woolf, 2010). While the functions of all these modules will still be necessary, there is no reason to think that any given ITS must *contain* all these components, in the sense of building them, controlling them, or owning them. The future for ITS may be to blow them up so that each piece can be used as a web-service for many different learning systems.

With respect to other online technologies, learning technology is already behind. On even a basic blog site, a user can often log in using one of five services (e.g.,

Google, Facebook), view adaptively-selected ads delivered by cloud-based web services that track users across multiple sites, embed media from anywhere on the internet, and meaningfully interact with the site on almost any device (mobiles, tablets, PC). In short, most web applications integrate and interact with many other web services, allowing them to be rapidly designed with robust functionality and data that no single application would be able to develop and maintain.

From the standpoint of AIED, moving in this direction is an existential necessity. Without pooling capabilities or sharing components, serious academic research into educational technologies may be boxed out or surpassed by the capabilities of off-the-shelf systems, many of which will have closed architectures. Unfortunately, while industry research can offer powerful results, competing pressures can lead to under-reporting: publishing research is costly, time-consuming, and can risk disclosing trade secrets or unfavorable empirical findings. While some companies make the investment to generalize their research, many others do not. By comparison, academic institutions and research-active commercial systems should be motivated to share and combine technologies to build more effective and widely-used learning technology. This model of collaborative component design stands alone in making platforms that co-exist with major commercial endeavors, such as web-browsers (FireFox), operating systems (Linux), and statistical packages (R; R Core Team, 2013). Moreover, service-oriented computing allows for a mixture of free research development and commercial licensing of the same underlying technologies.

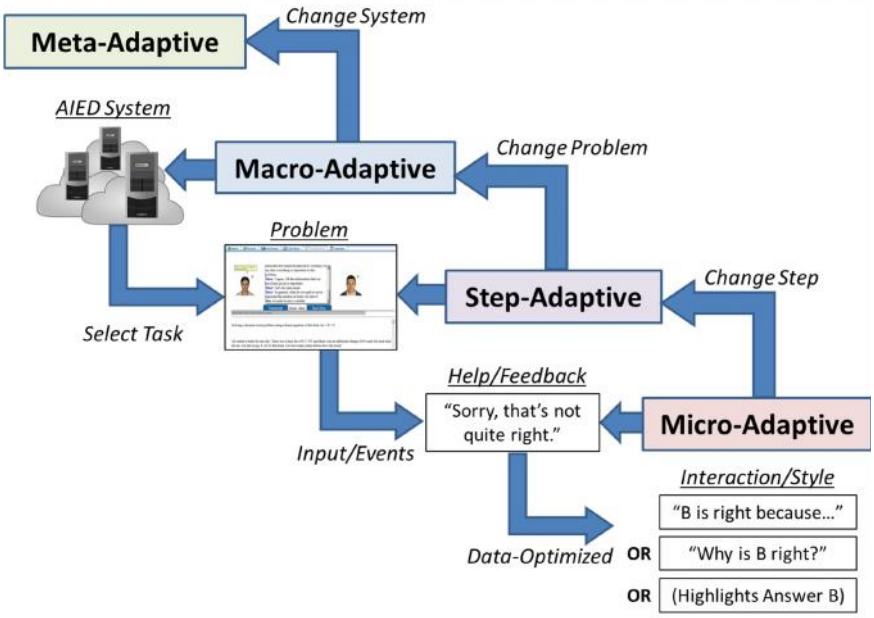
The benefits of moving toward service-oriented AIED will be substantial. First, they should enable AIED research to deeply specialize, while remaining widely applicable due to the ability to plug in to other platforms with large and sustained user bases. In such an ecosystem, user adaptation will be free to expand beyond the canonical inner loop and outer loop model (VanLehn, 2006). Composing and coordinating specialized AIED services will also demand greater standardization and focus on data sharing between systems. While this process may be painful initially, standards for integrating data across multiple systems would enable the development of powerful adaptation, analytics, and reporting functionality that would greatly reduce barriers for developing AIED technology and studying its effects on learners.

## 3 Four-Loops: Above Outer Loops and Under Inner Loops

One implication of scaling up AIED and moving beyond the standard four-component ITS model is that adaptation to users may become prevalent at grain sizes larger and smaller than traditional ITS. VanLehn (2006) framed the adaption from tutoring systems as consisting of an outer loop (selecting problems) and an inner loop (providing help and feedback on specific problem steps). These are often referred to as “macro-adaptivity” and “step-based adaptivity.” However, recent developments have shown the first steps toward “meta-adaptivity,” where the system adapts to the user by shifting the learner to an entirely different ITS system (which may then adapt to the user differently). Likewise, research on “micro-adaptivity” has looked at the benefits for using data to fine-tune interactions below the problem step level (e.g., keystroke-level

inputs, emotion detection, presentation modes or timing of feedback). This implies a four-loop model for user adaptation, as shown in Figure 1.

**Fig. 1.** Four-Loop User Adaptation




### 3.1 Meta-Adaptation: Handoffs Between Systems

Meta-adaptation has only become possible recently, due to increasing use and maturity of AIED technology. In the past, learning technologies such as ITS were trapped in sandboxes with no interaction. Due to service-oriented approaches, systems have taken the first steps toward real-time handoffs of users between systems. For example, in the recent Office of Naval Research STEM Grand Challenge, two out of four teams integrated multiple established adaptive learning systems: Wayang Outpost with ASSISTments (Arroyo, Woolf, & Beal, 2006; Heffernan et al., 2006) and AutoTutor with ALEKS (Nye, Windsor, Pavlik, Olney, Hajeer, Graesser, & Hu, in press). Other integration efforts are also underway as part of the Army Research Lab (ARL) Generalized Intelligent Framework for Tutoring (GIFT) architecture, which is built to integrate external systems (Sottilare, Goldberg, Brawner, & Holden, 2012) and version of AutoTutor has also been integrated with GIFT.

These initial integrations represent the first steps toward meta-adaptation: transferring the learner between different systems based on their needs and performance. This type of adaptation would allow learners to benefit from the complementary strengths of multiple systems. For example, learners that benefit most from animated agents might be sent to systems such agents (i.e., trait-based adaptation). Alternatively, different types of learning impasses or knowledge deficiencies may respond best to

learning activities in different systems (i.e., state-based adaptation). One problem that this approach might mitigate is the issue of wheel spinning, where an adaptive system detects that it cannot serve the learner's current needs (Beck & Gong, 2013). Meta-adaptation might also mean referring the learner to a human instructor, tutor, or peer. In general, meta-adaptation would focus on passing students and knowledge between different adaptive learning contexts (both AI-based and human).

Meta-adaptation is the maximum possible grain size, which makes it somewhat different from standard adaptation because users are transferred to an entirely different system. This type of adaptation likely requires either distributed adaptation or brokered adaptation. Distributed adaptation would involve individual systems deciding when to refer a learner to a different system and possibly trusting the other system to transfer the student back when appropriate. This would be analogous to doctors in a hospital, who rely on networks of specialists who share charts and know enough to make an appropriate referral, but may use their own judgment about when and how they make referrals. On the converse, brokered adaptation would require a new type of service whose purpose is to monitor student learning across all systems (i.e., a student model integrator) and make suggestions for appropriate handoffs. This service would be consulted by each participating AIED system, probably as part of their outer loop. In the long term, such a broker may be an important service, because it could help optimize handoffs and ensure that students are transferred appropriately. Such brokers might also play a role for learners to manage their data and privacy settings. Other models for coordinating handoffs might also emerge over time.

### 3.2 Micro-Adaptation: Data-Optimization and Event Streams

In addition to adaptivity at the largest grain size (selecting systems), research on the smallest grain sizes (micro-adaptation) is also an important future area. Micro-adaptation involves optimizing for and responding to the smallest level of interactions, even those that are not associated with a traditional user input on a problem step. For anything but simple experiments, this type of optimization and adaptation is too fine-grained and labor-intensive to perform by hand at scale, meaning that it will need to rely on data-driven optimizations such as reinforcement learning. Chi, Jordan and VanLehn (2014) used reinforcement learning to optimize dialog-based ITS interactions in the Cordillera system for Physics, which showed potential gains of up to 1σ over poorly-optimized dialog or no dialog. Dragon Box has taken a related approach by optimizing for low-level user interface and click-level data, by applying trace-based models to find efficient paths for learning behavior and associated system responses (Andersen, Gulwani, & Popovic, 2013).

These lines of research represent the tip of the iceberg for opportunities for micro-adaptation. A variety of low-level data streams have not yet been leveraged. Continuous sensor data, such as emotion sensors or speech input waveforms, may present rich opportunities for exploring fine-grained user-adaptation based on algorithmic exploration of possible response patterns. Low-level user interface optimization may also help improve learning, such as human-computer interaction design or keystroke-level events or mouse-over actions (i.e., self-optimizing interfaces).

Both the strength and the drawback of micro-optimization is that it is will tightly fit the specific user interface or content (even down to specific words in text descriptions). Optimizing for a particular presentation of a problem can lead to learning efficiency gains by emphasizing parts that are salient to learning from that specific case, while skipping or downplaying other features. However, micro-level optimization will likely suffer from versioning issues (e.g., changes to small problem elements potentially invalidating prior data and policies) and also transferability issues (e.g., an optimized case not transferring well from a desktop to a mobile context). Solutions to weight the relevance of prior data will be required to address issues related to altered problems or new contexts (e.g., mobile devices, classroom vs. home, different cultural contexts).

## 4 AI-Controlled Experimental Sampling

Techniques for micro-adaptation may also reshape experimental methods. Artificial intelligence can play a major role in the experimental process itself, which is a type of efficient search problem. Educational data mining research has already started looking at dynamically assigning subjects to different learning conditions based on multi-armed bandit models (Liu, Mandel, Brunskill, & Popovic, 2014). Multi-armed bandit models assume that each treatment condition is like a slot machine with different payout distributions (e.g., student learning gains). These models are common in medical research, where it is important to stop treatments that show harms or a consistent lack of benefit. They allow building intelligent systems that explore new strategies, while pruning ineffective ones.

The field is only taking its first baby steps for these types of experimental designs. Fundamental research is needed to frame and solve efficient-search problems present in AIED experiments. Based on varying different parameters and interactions in the learning experience, learning environments can search for interpretable models that predict learning gains. In the long term, models for automated experimentation may even allow comparing the effectiveness of different services or content modules, by randomly selecting them from open repositories of content.

The most difficult aspect of this problem is likely to be the interpretability. While multi-arm bandit models can be calibrated to offer clear statistical significance levels between conditions, models that traverse the pedagogical strategy space are often too granular to allow for much generalization. For example, some popular models for large learning environment focus on efficient paths or traces of learning behavior and associated system responses (Andersen, Gulwani, & Popovic, 2013). Unfortunately, these models are often not easily generalizable: they may capture issues tied to the specific system or may tailor instruction to specific problems so tightly that it is difficult to infer theoretical implications (Chi, Jordan & VanLehn, 2014).

New techniques are needed that can automatically explore the space of pedagogical designs, but that can also output interpretable statistics that are grounded in theories and concepts that can be compared across systems. This is a serious challenge that probably lacks a general algorithmic solution. Instead, such mappings will probably

be determined by the constraints of learning and educational processes. A second major challenge is the issue of integrating expert knowledge with statistically-sampled information. Commonly, expert knowledge is used to initially design a system (e.g., human-defined knowledge prerequisites), which is later replaced by a statistically-inferred model after enough data is collected. However, in an ideal world, these types of heterogeneous data would be gracefully integrated (e.g., treating expert knowledge as Bayesian prior weights). Future research in AIED will need to identify where this sort of expert/statistical hybrid modeling is needed, and match these problems with techniques from fields of AI and data modeling that specialize in these issues. Ultimately, a goal of this work should be to blur the lines between theory and practice by building systems that can both report and consume theoretically-relevant findings.

## 5 Semantic Messaging: Sharing Components and Data

To share technology effectively, AIED must move toward open standards for sharing data both after-the-fact (i.e., repositories) and also in real-time (i.e., plug-in architectures). The first steps in these directions have already been taken. Two notable data repository projects with strong AIED roots exist: the Pittsburg Science for Learning Center (PSLC) DataShop (Koedinger, Baker, Cunningham, Skogsholm, Leber, & Stamper, 2010) and the Advanced Distributed Learning (ADL) xAPI standards for messaging and learning record stores (Murray & Silvers, 2013). The IMS Global Specifications are also a move in this direction (IMS Global, 2015).

Due to solid protocols in messaging technologies, the technical process of exchanging data between systems at runtime is not onerous. The larger issue is for a receiving system to actually apply that data usefully (e.g., understand what it means). Hidden beneath this issue is a complex ontology alignment problem. In short, each learning technology frames its experiences differently. When these experiences and events are sent off to some other system, the designers of each system need to agree about what different semantics mean. For example, one system may say a student has "Completed" an exercise if they viewed it. Another might only mark it as "Completed" if the learner achieved a passing grade on it. These have very different practical implications. Likewise, the subparts of a complex activity may be segmented differently (e.g., different theories about the number of academically-relevant emotions). While efforts have been made to work toward standards, this seldom solves the problem: the issue with standards is that there tends to be so many of them.

So then, ontology development must play a key role for the future of ITS interoperability. There are multiple ways that this might occur. Assuming the number of standards is countable, it would be sufficient to have an occasional up-front investment to develop and update explicit mappings between ontologies by hand. While this is low-tech, it works when the number of terms is fairly small. For larger ontologies of AIED behavior and events, it may be possible to align ontologies by applying both coding systems to a shared task (e.g., build benchmark tasks that are then marked up with messages derived from that ontology).

By collecting data on messages from benchmark tasks, it may be possible to automate much of the alignment between ontologies, particularly for key aspects such as assessment. Research on Semantic Web technologies is also very active, and may offer other effective solutions to issues of ontology matching and alignment (Shvaiko & Euzenat, 2013). The final approach is to simply live without standards and allow the growth of a folksonomy: common terms that are frequently used. These terms can then become suggested labels, with tools that make their use more convenient and prevalent. The one approach that should *not* be taken is to try to develop a super-ontology or new top-down standard for the types of information that learning systems communicate. While there are roles for such ontologies, top-down ontologies have never achieved much support within research or software development communities.

## 6 Closing Remarks

The future for AIED should be a bright one: expansion of learning software into schools will ultimately result in unprecedented diversity and size of user bases. The areas noted in this paper are only the first wave for new AIED opportunities. In time, it will be possible to explore entirely new classes of questions, such as mapping out continuous, multivariate functional relationships between student factors and pedagogical effectiveness of certain behaviors. Systems such as personal learning lockers for data would allow for longitudinal study of learning over time, either in real-time or retrospectively. A major game-changer for future learning research will probably be data ownership and privacy issues: data will exist, but researchers will need to foster best-practices for data sharing, protection, and archiving.

With this wealth of data, researchers will be able to connect learning to other relationships and patterns from less traditional data sources. In 20 years, the range of commonly-available sensor data will be dizzying: geolocation, haptic/acceleration, camera, microphone, thermal imaging, social ties, and even Internet-of-Things devices such as smart thermostats or refrigerators. Moreover, the ecosystem of applications leveraging this data will likewise be more mature: your phone might be able to tell a student not only that their parents left them a voicemail, but that they sounded angry. This event might then be correlated with a recent report card, and the consequences of the interaction might be analyzed. Learning is a central facet of the human experience, cutting across nearly every part of life. To that end, as life-long learning becomes the norm, the relationship between life and learning will become increasingly important. By consuming and being consumed in a distributed and service-oriented world, AIED will be able to play a major role in shaping both education and society.

## References

1. Andersen, E., Gulwani, S., & Popovic, Z. (2013). A trace-based framework for analyzing and synthesizing educational progressions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 773-782). ACM.

2. Arroyo, I., Woolf, B. P., & Beal, C. R. (2006). Addressing cognitive differences and gender during problem solving. *International Journal of Technology, Instruction, Cognition and Learning*, *4*, 31-63.
3. Banilower, E . R., Smith, P. S., Weiss, I. R., Malzahn, K. A., Campbell, K. M., & Weis, A. M. (2013). *Report of the 2012 national survey of science and mathematics education*. Chapel Hill, NC: Horizon Research, Inc.
4. Beck, J. E., & Gong, Y. (2013, January). Wheel-spinning: Students who fail to master a skill. In *Artificial Intelligence in Education* (pp. 431-440). Springer Berlin.
5. Chi, M., Jordan, P., & VanLehn, K. (2014). When Is Tutorial Dialogue More Effective Than Step-Based Tutoring? In *Intelligent Tutoring Systems* (pp. 210-219). Springer: Berlin.
6. Craig, S. D., Hu, X., Graesser, A. C., Bargagliotti, A. E., Sterbinsky, A., Cheney, K. R., & Okwumabua, T. (2013). The impact of a technology-based mathematics after-school program using ALEKS on student's knowledge and behaviors. *Computers & Education*, *68*, 495-504.
7. Falmagne, J. C., Albert, D., Doble, C., Eppstein, D., & Hu, X. (2013). *Knowledge Spaces: Applications in Education*. Springer Science & Business Media.
8. Graesser, A. C. (2011). AutoTutor. In P. M. McCarthy, & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Identification, investigation and resolution*. Hershey, PA: IGI Global.
9. Heffernan, N. T., Turner, T. E., Lourenco, A. L., Macasek, M. A., Nuzzo-Jones, G., & Koedinger, K. R. (2006). The ASSISTment Builder: Towards an Analysis of Cost Effectiveness of ITS Creation. In *FLAIRS Conference* (pp. 515-520).
10. IMS Global (2015). Learning Tools Interoperability. Retrieved from: http://www.imsglobal.org/lti/index.html on Feb 22, 2015.
11. Khan Academy (2015). About Khan Academy. Retrieved www.khanacademy.org/about on Feb. 22, 2015.
12. Koedinger, K. R., Baker, R. S., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. In C. Romero, S. Ventura, M. Pechenizkiy, & R. S.J.d. Baker (Eds.) *Handbook of educational data mining*, 43-53.
13. Liu, Y. E., Mandel, T., Brunskill, E., & Popovic, Z. Trading Off Scientific Knowledge and User Learning with Multi-Armed Bandits. *Educational Data Mining (EDM) 2014*. 161-168. Springer
14. Monsalve, S. (2014). A venture capitalist's top 5 predictions for 2015. *Fortune*. Retrieved http://fortune.com/2014/12/29/a-venture-capitalists-5-predictions-for-2015/ on Feb. 22, 2015.
15. Mostow, J., & Beck, J. (2006). Some useful tactics to modify, map and mine data from intelligent tutors. *Natural Language Engineering*, *12*(02), 195-208.
16. Murray, K., & Silvers, A. (2013). A learning experience. *Journal of Advanced Distributed Learning Technology*, *1*(3-4), 1-7.
17. Murray, T., Blessing, S., & Ainsworth, S. (2003). *Authoring tools for advanced technology learning environments: Toward cost-effective adaptive, interactive and intelligent educational software*. Springer.
18. Nye, B. D., Windsor, A., Pavlik, P. I., Olney, A., Hajeer, M., Graesser, A. C., & Hu, X. (In Press). Evaluating the Effectiveness of Integrating Natural Language Tutoring into an Existing Adaptive Learning System. In *Artificial Intelligence in Education (AIED) 2015*.
19. R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0

20. Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive Tutor: Applied research in mathematics education. *Psychonomic bulletin & review*, *14*(2), 249-255.
21. Shvaiko, P., & Euzenat, J. (2013). Ontology matching: state of the art and future challenges. *Knowledge and Data Engineering, IEEE Transactions on*, *25*(1), 158-176.
22. Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, 0002764213498851.
23. Sottilare, R. A., Goldberg, B. S., Brawner, K. W., & Holden, H. K. (2012). A modular framework to support the authoring and assessment of adaptive computer-based tutoring systems (CBTS). In *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) 2012*.
24. VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, *16*(3), 227-265.
25. Woolf, B. P. (2010). Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning. Morgan Kaufmann.

# Education still needs Artificial Intelligence to support Personalized Motor Skill Learning: Aikido as a case study

Olga C. Santos

aDeNu Research Group. Artificial Intelligence Dept. Computer Science School, UNED. Calle Juan del Rosal, 16. Madrid 28040. Spain
http://adenu.ia.uned.es
ocsantos@dia.uned.es

**Abstract.** Motor skill learning is hardly considered in current AIED literature. However, there are many learning tasks that require consolidating motor tasks into memory through repetition towards accurate movements, such as learning to write, to draw, to play a musical instrument, to practice a sport technique, to dance, to use sign language or to train for surgery. The field of Artificial Intelligence (AI) needs new sap to cope with the challenges in the Educational (ED) domain aimed to support psychomotor learning. This new sap can be provided by novel interactive technologies around the Internet of the Things that deal with Quantified-self wearable devices, 3D modelling, Big Data processing, etc. The paper aims to identify opportunities and challenges for AI + ED that can be discussed during the workshop. Some of the issues raised are illustrated within a case study instantiated in the Aikido practice, a defensive martial art that involves learning skilled movements by training both the body and the mind, and which is not only part of extra-curricular activity in many schools, but has also been reported of value for teaching in STEM (Science, Technology, Engineering and Mathematics) education, in particular, some laws of mechanics.

**Keywords:** motor skill learning, psychomotor domain, artificial intelligence, education, Internet of the Things, personalization, Aikido, STEM.

## 1 Introduction

Motor skill learning can be defined as achieving the ability to perform a function acquired with practice that requires body and/or limb movement to accomplish the goal of an action or task [1]. Although it is not a new concept [2], up to my knowledge (grounded by a review of the papers published in the International Journal of Artificial Intelligence in Education (IJAIED) and which is reported elsewhere [3]), the physical aspects of learning have been hardly considered in the AIED research. Nevertheless, consolidating specific motor tasks into memory through repetition (thus, creating long-term muscle memory for a given task) is very relevant in diverse educational scenarios that support learning processes involving not only brain activi-

ty, but also physical activity, such as learning to write, to draw, to play a musical instrument, to practice a sport technique, to dance, to use sign language or to train for surgery that require long-term physical training, as reported in [3]. In these situations, learners have to train by repeating basic and very specific movements till they learn the best way to carry them out effectively without conscious effort. It has to be remarked here that learning physical skills (i.e., the proficiency of individual movements, also called sensomotor habits [4]) goes beyond mere muscle memory, but involve blending motor skills and cognitive, meta-cognitive and affective skills. In fact, psychomotor skills cannot be acquired by multiple repetitions of given motor pattern without considering the importance of feedback between cognitive processes and motor actions [5]. However, the focus of the discussion that this paper aims to bring to the workshop is mainly on how the physical part related to the psychomotor learning domain (which deals with physical movement, coordination and the use of the motor skill areas [6]) can be supported from an AIED perspective, both in 1) the modelling of the learner physical interaction, and 2) the provision of the required personalized support during the learning. In my view, this is a new dimension that is worth to be explored by combining AI + ED research. The cognitive, meta-cognitive and affective dimensions are already being widely addressed in AIED literature.

In addition, at this point in time, technology has evolved in such a way that it can monitor the movements carried out by the learners through diverse types of sensors (e.g., inertial, optical, position, electromyography, etc.) and timely feedback can be provided through diverse actuators (such as resistance, force, vibration, etc. as well as servo motors) to help the learner improve the performance of the corresponding movement. Quantified-self approaches (based on data gathered from wearable devices such as electronic bracelets and intelligent t-shirts) allow personal awareness and reflection for behavioral monitoring in many situations, such as physical exercise or affective support. Big Data allows processing real time data streams gathered from heterogeneous information sources. 3D models of real objects can be produced with low-cost scanners and printers. These technologies (among others) support the so called Internet of the Things (IoT), that is, the connection of *physical things* to the Internet, which makes possible to access remote sensor data and to control the physical world from a distance [7]. In this context, the do-it-yourself movement supports non-experts in getting familiar with these novel interaction technologies and in being able to build ad-hoc electronic components for their own needs. Thus, AIED researchers can take advantage of this supportive context so the learning curve of integrating above technologies with AI techniques can be feasible for the field.

As a result, this paper proposes to explicitly open a new research line for the AIED field where ED can benefit from AI techniques enriched with emerging novel interactive technologies around the Internet of the Things. This new research direction, framed within the psychomotor learning domain, requires a shift towards supporting physical practice (i.e., training) rather than supporting instructional teaching. This implies that the physical actions carried out while practicing need to be monitored, modelled and, when needed, corrected, to achieve successful motor skill learning (i.e., skills learning at a physical level).

## 2 Opportunities and Challenges

As discussed in [3], the synergy of Artificial Intelligence techniques with novel interactive technologies opens new opportunities for researching the physical (i.e., corporal) aspects of learning. For instance, it seems to be possible to provide intelligent real time feedback to scaffold physical skill learning by using sensors, actuators, 3D scanning and modelling, data streams processing, etc. And in order to improve performance, tangible scaffolding could be provided to guide motor skill learning in a personalized way through embodiment technology. A case study that illustrates some issues involved is outlined in Section 3.

In any case, by integrating novel interactive technology, the foreseen goal is that AIED researchers can produce systems that sense the learner's corporal behavior as she learns specific skilled movements, and then guide the learner on how to react in an optimal way (taking into account the learner's current performance, corporal features and the particularities of the specific movement to perform) by providing personalized feedback during the learning process (rather than just giving directions of what to do and how to do, as in traditional AIED intervention approaches). Procedural learning in terms of motor skill is usually difficult to explain by the instructor and to understand by the learner. In fact, this procedural tutoring support is of major relevance in the case of novice learners, as they might get into a wrong habit if no timely feedback is provided to them while practicing by their own and, thus, they cannot understand why the movement is not correct.

In order to build procedural learning systems that can personalize motor skill learning, both AI and ED research need to revise the application of their theoretical and methodological approaches to the particularities of the psychomotor learning domain. From the AI point of view, there is a need for modelling the individual functional and corporal features, her interaction and the accurate movement, by processing the simultaneously and continuously data streams produced by diverse and heterogeneous sensors, and then controlling the robotics to physically deliver the intervention to the learner. From the ED point of view, the focus has to be put on identifying what is the most appropriate intervention in each case (considering cognitive, meta-cognitive, affective and behavioral dimensions) and when and how it should be delivered in order to make a positive impact in the learning process.

Therefore, as discussed in [3], there exist challenges regarding 1) modelling and representing the movements of the learner by building the learner physical interaction model as well as the accurate movement model, and 2) providing the appropriate personalized physical support in the most efficient way for each learner in each training context. More specifically, regarding the modelling of movements, there seem to be challenges related to: i) detecting the physical interaction, ii) modelling the movements to be trained, iii) error diagnosing and intervention modelling, and iv) modelling the learner. In turn, regarding the provision of the appropriate personalized physical support, challenges might exist in order to: i) deciding upon adaptation, ii) evaluating the user activity, iii) visualization of movement performance, and iv) sharing progress and social learning.

## 3 A case study for AI + ED: supporting personalized psychomotor learning in Aikido

In order to facilitate the discussion on existing challenges for AI + ED to support personalized motor learning skill learning, a case study is presented. This case study focuses on Aikido martial art. Since it might surprise the reader the selection of this domain from an ED perspective, first some of the reasons for its selection are discussed. Then, some technological advances that can help AI to provide personalized motor skill training within the Aikido psychomotor learning domain are presented. They intend to include in the AIED research agenda ideas that can be explored.

### 3.1 Aikido & ED: more than just a psychomotor learning domain

Aikido is a non-aggressive Japanese martial art that consists of entering and turning movements that redirect the momentum of an opponent's attack, and a throw or joint lock that terminates the technique [8]. The word is formed by Ai (coordination, accord, harmony, blending), Ki (psychological energy, spirit, universal force) and Do (way of life, philosophy of living) [9]. It is guided by defending oneself while also protecting the attacker from injury. In fact, it is based on the principle that in order to control an attacker, the defender must meet the attack in a state of perfect balance [10]. Properly carrying out the technique requires years of training by repeating over and over the sequence of movements that makes up each Aikido technique.

Martial arts do not only involve complex manipulations of human anatomy and physiology [9], but they aim to train both the body and the mind, since training consist of improving mental disposition and motor skills (i.e., fitness and coordination) [4]. According to these authors [4], the technique of self-defense can be defined as a specific sequence of movements constituting a partial or total resolving of various dynamic situations. These movements imply eccentric and concentric muscle work, rotation of the trunk and hips, translocation of the body mass center and adequate leg work. Interplay of muscle tension and relaxation combined with accurate decisions is needed. This requires the development of skills in body movement control that combine mental balance and appropriate motor actions, where the general motor fitness is adjusted to the individual level of motor abilities (i.e., quality is more important than strength). Automation of movements occurs when mental processes are free of controlling individual movements. An ability of psycho-physical self-controls is also required to allow for efficient performance under stressful situations.

Since Aikido practice involves the execution of paired movements between the attacker (*uke*: receiver of the technique) and the defendant (*tory*: doer of the technique), it helps understanding cooperation and timing in movement [11]. Recent studies using electroencephalography and electromyography techniques have shown that the postural control training using Aikido improves psychomotor performance [10].

Nonetheless, the benefits of Aikido go beyond physical fitness and motor abilities. For instance, some studies suggest that Aikido training increases mindfulness [11]. In particular, since practitioners are taught to be mindful of the technique, breathing,

balance, center of gravity and their connection to the other person, it may facilitate increasing one's awareness of body position, of others around, practitioner's emotional states and how other people's emotions may affect the Aikido practitioner's emotional states. As compiled by these authors, benefits of increased mindfulness may include better concentration, stronger awareness, improved immune system functioning and decreases in stress related physical symptoms [12, 13]. In this way, Aikido training may enhance awareness and resolution of problematical situations, as during training sessions, the practitioner learns to deal with multiple stressors concurrently, and this is learnt to do in an effective manner while remaining calm, which suggests that Aikido seem to teach practical problem solving and acceptance of circumstances [11]. In this sense, Aikido is one of the more spiritual martial arts as it studies the energy within oneself, her partner and the world through the physical principles of entering, turning and securing, and thus, focuses directly on the energy involved in dealing with one's emotions, perceptions of trust and fear, and conceptions of reality as well as the energy and demands in relating with another human being [14]. In this authors' viewpoint, Aikido can contribute to relationship encounters, conflict resolution, motivation and personal energy by an effective management of energy, improving interpersonal relationships and facilitating stress reduction. Following these ideas, studies have shown that including martial arts such as Aikido in school programs can enhance student's awareness of violence prevention and allow them to react calmly and without panic, reducing violence in schools [15].

In addition to above benefits, Aikido has also potential to be used in education, not only for physical education (i.e., development of motor abilities, mental and physical health benefits, violence reduction…) but also in STEM education (i.e., Science, Technology, Engineering and Mathematics). In this sense, there are studies where some laws of Physics are taught with Aikido practice (see [15]) that show statistically significant improvements in the scores on biomechanics (i.e., mechanics principles of human movement) tests as well as statistically significant correlations between the results in those tests and the performance of the Aikido techniques. From these works, it seems that solid-state mechanics concepts such as the law of momentum conservation, second law of motion for angular motion, centrifugal force and composition of resultant forces and moments of force, can be explained more effectively with the practice of Aikido, facilitating the understanding of how forces act on a person while in translator or rotary motion.

Since the practice of Aikido seems to improve not only motor skills, but also some cognitive abilities (i.e., acquiring the knowledge of mechanics required by the scholar curriculum), this martial art has been chosen to discuss how a psychomotor learning domain like this could benefit from an AIED procedural learning environment. In this sense, some ideas on how to provide some tangible scaffolding when needed to guide motor skill learning in a personalized way using novel interactive technology from the IoT are discussed next. The research question behind is: *How to design and implement a personalized procedural learning environment that can physically train and guide the particular way each learners' body and limbs should move in order to achieve a specific learning goal that is related to improving learners' motor skills acquisition, such as the needs identified in the Aikido practice?*

### 3.2 Improving AI based personalized motor skill learning in Aikido with novel interactive technologies

The goal of Aikido is to hold the *uke* (attacker) in a compromised and secured position with a minimal amount of effort [17]. To achieve this, Aikido practice involves the manipulation of various joints of the body and is based on effective anatomical principles to subdue a training partner by twisting the limbs or locking up the skeletal system. In order to better understand the body's responses and improve the proficiency of applying specific techniques, anatomical studies on cadavers that investigated the nerves, bones, muscles, tendons and tissues manipulated by each technique have been carried out in the past [9]. However, novel interactive technologies, such as those provided by quantified-self wearable devices, can be used to gather dynamic indicators while making the movement. This can help to understand how the movements are performed and improve training. For instance, the movements carried out by a person can be monitored using diverse types of sensors (inertial, optical, position, physiological, etc.) [18] for real time motion study outside the laboratory [19]. This technology is becoming less and less intrusive, to the point that sensors that allow complex movement patterns tracking are getting embedded directly into clothes [20]. The interaction data streams continuously collected by these sensors in real time need to be processed. Due to its volume, variability and speed, Big Data mining techniques need probably to be applied [21].

In addition, as introduced above, Aikido requires long-term physical training to learn how to carry out the movements in the most efficient way. Very often, the execution of the corresponding techniques involves practitioners moving along a curve and lowering one's center of gravity in order to employ the centrifugal force acting on the opponent and one's own gravity [16]. Forces applied are notably subtle and intricate, and thus, difficult to learn without the direct tutelage by an experienced *sensei* (teacher) [17]. This is not easy to put into practice without being repeatedly told what is done wrong and what should be done right. In order to be able to compare how the movement is performed, a model of the accurate movement needs to be built. In the field of virtual reality, there are works that build virtual skeletal models for videogames from the information collected using wearable technology (e.g., biomechanical or inertial sensors), which both map the movement as well as recognize gestures with AI techniques [22]. The movement controlled by sensors can also be represented in 3D models of the human body [23].

The next step is to provide some guided feedback. Since the situations where the applied techniques are never the same (e.g., the degree and direction of force is different, the position of the *tory* is not always the same, body shape and muscular structure differ from *uke* to *uke*, perception and timing change) the application of the technique must change accordingly [24]. This means that the provided feedback should be personalized to the current situation, including *uke* and *tory* body built. With respect to defining the appropriate feedback to give, an initial proposal can be to provide some tangible scaffolding through embodiment technology that corrects the learner's movements by physically controlling and guiding the movement of the learner till her ideal movements (considering the learner's own body built) are achieved. Feedback

with different levels of complexity (simple verification, try again and elaborated) provided through different channels (visual, audio and haptic) [17] should be considered. For instance, in order to provide motor intervention, some works use electromyography sensors (i.e., the measure of the electrical activity produced by the skeletal muscles) to detect movement intentions and help to carry them out through exoskeletons (i.e., physical shells) moved with servo motors [25]. Resistive sensors have also been used to move body parts through vibrations [26]. Inertial sensors and vibro-tactil feedback is also used to replicate referred postures and correct those that are not alike [27]. A forced feedback system to guide fingers movement to improve motor skills when playing the piano has been implemented with a simple exoskeletal robotics [28]. The technology for 3D modelling can be used to build physical prototypes of tangible objects. As an example, combining available technologies, a 3D printed hand has been controlled with Arduino using servomotors [29].

However, guiding the learner by delivering forced haptic feedback when the movement performed does not reflect the reference movement might not be the most appropriate psycho-educational approach to achieve long-term learning, although it might help to increase motivation by contributing to short-term performance [30]. Therefore, there is a need to research the appropriate personalized support to provide. Here, the application of TORMES methodology [31] (or an extended version of it that addresses the particularities required by the psychomotor learning domain and the requirements to sense the environment and provide tangible support) can be of value to model the personalized dynamic psychomotor support to be provided in specific situations. In particular, TORMES extends the design cycle of interactive systems as defined by ISO 9241-210 with the life cycle of e-learning and the layered evaluation of adaptive systems, and combines user centered design methods (which can be applied to gather tacit knowledge from psychomotor experts as well as experienced Aikido teachers and practitioners) with (big) data mining techniques (that can be used to analysis performance indicators regarding the movements carried out gathered from Aikido training sessions, for instance, using wearable devices).

There is a commercial software (i.e., Aikido 3D[1]) that recreates with animated characters the movements of a high degree Aikido black belt using motion capture technology. The goal of this tool is to facilitate visualizing how the techniques are to be carried out, so the learner can see it from different perspectives, in slow motion, zoomed, etc. It provides a technological improvement on top of what takes place in Aikido *dojos* (i.e., training places) around the world, but the approach behind is similar: learner watches how an expert (in this case, an animated character whose behavior has been modelled with the movements of an expert) carries out the technique and then tries to reproduce (imitate) the same movements with a partner. However, an AIED support through a procedural learning environment could improve the learning experience by physically controlling and guiding the movements of the learner when appropriate, so she can correct them till she masters the movements for the technique (considering the learner's own body built and skills, as well as the context where the movement is carried out, including the opponent features). This requires the follow-

[1] https://www.aikido3d.com

ing: 1) sensing the learner's movement and the context in which this movement takes place (e.g., the physical features and abilities of the opponent), 2) comparing it against the accurate movement (e.g., how an expert in the technique would carry the movement out considering the same physical features and abilities of the learner and the opponent), 3) deciding whether it is appropriate to provide the tangible support at this moment (dealing with focusing on short term performance vs. long-term learning), and 4) if appropriate, then provide the tangible support in an effective non-intrusive way, for instance with vibro-tactil feedback through actuators sew on the *Aikidogi* (i.e., the Aikido training uniform).

## 4 Conclusion

There is a challenge and opportunity to take advantage of AI and ED research to develop personalized procedural systems that can support learners while acquiring psychomotor abilities. Learning and improving motor skills is of relevance in many domains, such as learning to write, to draw, to play a musical instrument, to practice a sport technique, to dance, to use sign language or to train for surgery.

In this paper, the relevance of Aikido practice and the support it can obtain from AIED based procedural learning environments has been discussed for the first time in the literature. In addition, the application of novel interaction technologies that are being used by the Internet of the Things (such as quantified-self wearable devices, big data processing and 3D modelling) to build an AIED procedural learning environment has been proposed by reporting works that partially address some of the technological issues discussed. Although the assimilation of new technologies is always costly, the do-it-yourself movement, which encourages people in creating Internet of the Things applications by their own [32], can simply their learning curve and thus, their usage should be feasible for the AIED research community, provided that many people around the world are taking advantage of them without a wide specialized technological background. In turn, non-specialized users benefit from the feeling of belonging to a community that characterizes this kind of developing culture (as well as the open source and open hardware philosophy underneath it) and receive on-line peer support both on search (i.e., looking for information with the help of web search engines or within specialized repositories) and on demand (i.e., asking in specialized forums).

In addition, it can also be noted here that most of the approaches referenced in the previous section can be controlled by an Arduino based infrastructure. Arduino is an open source electronics prototyping platform, which is based on easy to use hardware and software [33]. As reported in previous work, Arduino can be used to gather contextual information from sensors [34] and deliver ambient intelligent feedback [35].

In summary, the motivation of this paper is to propose a new research direction to the AIED field, where novel interactive technologies enrich Artificial Intelligence techniques to deal with some challenges within the Educational domain. This proposal will be discussed further during the workshop "Les Contes du Mariage: Should AI stay married to ED? A workshop examining the current and future identity of the AIED field" taking place during the 17th International Conference on Artificial Intel-

ligence in Education (AIED 2015). Outcomes from the discussion in the workshop will be included in a paper for the IJAED Special Issue "The next 25 Years: How advanced, interactive educational technologies will change the world" [3].

## Acknowledgements

This work has been carried out in the context of the project *"Towards tangible recommender systems: personalizing kinematic recommendations modelling with low cost technology"* (***tangibREC***). Funding has been requested to the Spanish Ministry of Economy and Competitiveness (MINECO) in the Call 2014 for Explore Science and Explore Technology projects.

## References

1. Christensson, J. Individualising Surgical Training in a Virtual Haptic Environment. MSc Thesis in Interaction Design. IT University of Göteborg (Sweden), 2005.
2. Fitts, P.M. Perceptual-Motor skills learning. In: Melton AW, eds. Categories of Human Learning. New York, NY: Academic Press Inc., 243-286, 1964.
3. Santos, O.C. Training the Body: the Potential of AIED to support Personalized Motor Skill Learning. Special Issue "The next 25 Years: How advanced, interactive educational technologies will change the world". International Journal of Artificial Intelligence in Education, 2015 (in preparation).
4. Harasymowicz, J., and Kalina, R.M. Training of psychomotor adaptation – a key factor in teaching self-defence. Archives of Budo, 1, 19-26, 2005Schmidt, R.A. Motor learning and performance. Human Kinetics, Champaign, IL. 1991.
5. Harrow, A. A Taxonomy of Psychomotor Domain: A Guide for Developing Behavioral Objectives; David McKay: New York, NY, USA, 1972.
6. Kopetz, H. Internet of Things, Real-Time Systems Series, 307-323, 2011.
7. Ueshiba, K. The Art of Aikido: Principles and Essential Techniques. Kodansha International, 2004.
8. Seitz, F.C., Olson, G.D., Stenzel, T.E. A martial arts exploration of elbow anatomy: Ikkyo (Aikido's first teaching). Perceptual and Motor Skills, 73, 1227-34 (1991).
9. Bazanova, O.M., Kholodina, N, Kurose-Payet, Y., Payet, J., Nikolenko, E.D., Podoinikov, A. Postural control training using Aikido improves psychomotor performance. International Journal of Psychophysiology, 94(2), 165-165, 2014.Lothtes, J., Hakan, R. and Kassab, K. Aikido experience and its relation to mindfulness: a two-part study. Perceptual and Motor Skills, 116(1), 30-9, 2013.
10. Shapiro, S., Carlson, L., Astin, J., and Freedman, B. Mechanisms of Mindfulness. Journal of Clinical Psychology. 62 (3), 373-386, 2006.
11. Kabat-Zinn, J. Wherever you go there you are: mindfulness meditation in everyday life. New York: Hyperion, 1994.
12. Seitz, F.C., Olson, G.D., Locke, B. and Quam, R. The martial arts and mental health: the challenge of managing energy. Perceptual and Motor Skills, 70(2), 459-464, 1990.
13. Lu, C. Martial Arts, Violence, and Public Schools, Brock Education, Vol 18, 2008, 1-12.
14. Mroczkowski A. Using the Knowledge of Biomechanics in Teaching Aikido. In Injury and Skeletal Biomechanics. Goswami, T. (Ed.). InTech, 2012.

15. Olson, G.D., Cook IV, M., Brooks, L. Aikido's Arm-Lock (Ude-Gatame) Technique: What Tissues are Affected? Journal of Asian Martial Arts, 8(2), 42-49, 1999.
16. Schneider, J., Börner, D., van Rosmalen, P., Specht, M. Augmenting the Senses: A Review on Sensor-Based Learning Support. Sensors, 15, 4097-4133, 2015.
17. Fong, D.T.-P.; Chan, Y.-Y. The Use of Wearable Inertial Motion Sensors in Human Lower Limb Biomechanics Studies: A Systematic Review. Sensors, 10, 11556-11565, 2010.
18. Fleury, A.; Sugar, M.; Chau, T. E-textiles in Clinical Rehabilitation: A Scoping Review. Electronics, 4, 173-203, 2015.
19. Fan, W., Bifet, A. Mining big data: current status, and forecast to the future. SIGKDD Explor. Newsl. 14, 2 (April 2013), 1-5, 2013,
20. Arsenault, D. A Quaternion-Based Motion Tracking and Gesture Recognition System Using Wireless Inertial Sensors. Master of Applied Science in Human-Computer Interaction. Carleton University, 2014.
21. Bae, J., Haninger, K., Wai, D., Garcia, X., Tomizuka, M. A Network-Based Monitoring System for Rehabilitation. The 2012 IEEE/ASME International Conference on Advanced Intelligent Mechatronics, 232-237, 2012.
22. Saotome, M. Aikido and the harmony of nature. Boulogne, France: Serirep. 1986.
23. Gopura, R. A. R. C., Kiguchi, K. EMG-Based Control of an Exoskeleton Robot for Human Forearm and Wrist Motion Assist. IEEE International Conference on Robotics and Automation, 731-736, 2008.
24. Rush, R.P Sensation Augmentation to Relieve Pressure Sore Formation in Wheelchair Users. Eleventh International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'09), 275-276, 2009.
25. Ding, Z.Q., Luo, Z.Q., Causo, A., Chen, I.M., Yue, K.X. , Yeo, S.H., Ling, K.V. Inertia sensor-based guidance system for upperlimb posture correction. Medical Engineering & Physics, vol. 35 (2), 269-276, 2013.
26. Datta, S. Forced fingers, Available from Github: https://github.com/dattasaurabh82/Forced-Fingers
27. Huluta, E., da Silva, R.F., de Oliveira, T.E.A. Neural network-Based hand posture control of a humanoid Robot Hand. IEEE Int. Conf. on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), 124-128, 2014.
28. Soderstrom, N.C. and Bjork, R.A. Learning Versus Performance: An Integrative Review. Perspectives on Psychological Science, vol. 10(2) 176–199, 2015.
29. Santos, O.C., Boticario, J.G. Practical guidelines for designing and evaluating educationally oriented recommendations. Computers & Education, 81: 354-374, 2015.
30. De Roeck, D., Slegers, K., Criel, J., Godon, M., Claeys, L., Kilpi, K., and Jacobs, A. I would DiYSE for it!: a manifesto for do-it-yourself internet-of-things creation. In Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design (NordiCHI '12). ACM, New York, NY, USA, 170-179, 2012.
31. Banzi, M. Getting started with Arduino. O'Reilly Media, 2009.
32. Santos, O.C., Boticario, J.G. Exploring Arduino for Building Educational Context-Aware Recommender Systems that Deliver Affective Recommendations in Social Ubiquitous Networking Environments. Web-Age Information Management. LNCS, vol. 8597, 272-286, 2014.
33. Santos, O.C., Saneiro, M., Boticario, J.G., Rodriguez-Sanchez, C. Toward Interactive Context-Aware Affective Educational Recommendations in Computer Assisted Language Learning. In New Review of Hypermedia and Multimedia, 2015 (in press).

# Realizing the Potential of AIED

Lewis Johnson

*Alelo Inc.*

**Abstract.** This is a time of opportunity and promise for AIED as a field. The field has had some major successes, and is having an impact with significant numbers of learners. Now that Big Data has arrived in education, opportunities are opening up to generate analytics from that data and use it to personalize learning. There is however potential to have an even greater impact on education, and make greater use of AI technologies. The field should focus on realizing this potential, and not divorce itself from either AI or Ed. Achieving impact will require more effective dialog and collaboration with educators, learners, and people in industry.

**Keywords:** Educational impact, partnering with education, partnering with industry, participatory design, technology transfer

## 1 The Time for AIED has arrived

These are exciting times for learning technologies. Technology is becoming integrated into education at all levels, as online learning, blended learning, and smart classrooms are becoming the norm. The global market in technology-enabled learning is projected to grow at an annualized rate of 20.3% to $220bn in 2017 (MarketsandMarkets, 2014). Large as this is it is still a small fraction of the $5.89tr that is projected to be spent on education in 2015 (Next Up Research, 2010); this suggests there will be even greater opportunities in the future. Technology-enabled education is enabling and fuelling demand for personalized and adaptive learning and assessment (Borden, 2011; Getting Smart, 2012), capabilities which AIED systems are well positioned to provide.

AIED-based systems are contributing to this innovation in learning. Alelo's language and culture training systems (Johnson, 2010; Camacho et al., 2009; Johnson et al., 2012) are in widespread use throughout the world, with well over 100,000 learners to date. They have had a significant effect on the cultural and linguistic competence of the learners who use them. For example the 3rd Battalion, 7th Marines, the first American Marine unit in the Iraq war to complete their tour of duty without any combat fatalities, learned Iraqi Arabic language and culture using Alelo's Tactical Iraqi learning environment (Marine Corps Center for Lessons Learned, 2008). Another AIED success story is the ASSISTments system, which is being used throughout the United States by nearly 20,000 or more students per year (Gelfand, 2011). And perhaps the biggest success so far has been the Carnegie Learning curriculum and software, which

as of 2010 had been used by over 500,000 students (Institute of Education Sciences, 2010).

The workshop call for papers questions whether the ideas of AIEd are influencing AI or Education in any major way. The above examples illustrate that it is AIED is in fact having an impact. One could perhaps argue as to whether they are having a major impact, but they certainly intend to do so.

Yet these examples are just the beginning, and AIED has the potential to have an even greater impact on education in the future. The challenge for the AIED community is to realize that potential. It needs more success stories – examples of AIED research that is having an impact. The more instances there are of research that is having an impact, the more impact the field as a whole will have.

I regret that other obligations do not permit me to participate in person in the workshop in Madrid. However remote participation is becoming commonplace in technology-enabled learning, so I hope it is also possible for a major international conference on technology-enabled learning such as AIED. In any case I feel compelled to contribute this position paper and hopefully offer some constructive suggestions.

## 2 Connect AIED to Educational Problems

I have a number comments on the questions posed in the call for papers, but I will focus here on just one: the extent to which the results of AIED research are meaningful to real educational practices. Or to put it another way: What steps can people in the AIED community take to ensure that their research has meaningful educational impact? Here are some recommendations.

**Talk with educational leaders.** More than individual teachers, educational leaders and managers have a broad view of how where the unmet educational needs are, and may be open to innovative approaches that can meet those needs. Many of these are needs that AIED technologies can address. If you have a promising AIED technology, show it educational leaders and listen to what they have to say. They might help you make the connection to education needs, or if not you will come away with a better understanding of what the critical educational needs really are. They may be able to put you in touch with schools and teachers that are receptive to innovative solutions.

**Talk with people in the edtech industry.** There is not enough dialogue between AIED researchers and people in the edtech industry, which leads me to suspect that that there may be an insufficient appreciation of what researchers can learn from such dialogue. People in edtech have an understanding of what it takes to make a real impact on real educational problems with technology. They may be aware of educational problems that they themselves are not in a position to address, but they wish someone else would.

**Engage in effective iterative, participatory design.** The workshop call for papers suggests that participatory research is often a matter of rhetoric rather than practice. The question as I see it is how to make such participatory research achieve more effective results. Dialogue with educational leaders prior to the start of the design process can help, to make sure that the design is focusing on the right problems. So can iterative participatory design, in which researchers show teachers and learners

partial prototypes and ask for input on how to improve it. Participatory design can be very effective when people have something concrete to respond to.

**Learn from research programs that value educational contributions.** The US National Science Foundation's Cyberlearning program is an example of research program whose projects address learning research questions as well as learning technology questions. The program requires research teams to carefully evaluate the educational impact of the designs that they develop, instead of simply focusing on technology development. Other AIED researchers can draw useful lessons from this and similar programs.

The RALL-E project (Alelo, 2015) is an example of an exploratory AIED research project that has undertaken each of these steps. With funding from the National Science Foundation's Cyberlearning program, we have developed a lifelike robot that can converse in Chinese, using the Robokind's Zeno-R25 robot as a platform. We developed the concept with advice from the Virginia Department of Education, which made us aware of critical needs in their state such as the lack of availability of qualified language teachers in many schools and the lack of access to high-quality interactive learning materials in many of those schools. We designed RALL-E as an interactive language-learning tool that students can use to develop their conversational skills, with or without the presence of a teacher. The Virginia Department of Education introduced us to the principal of a receptive test site, the Thomas Jefferson High School for Science and Technology (TJ) in Alexandria, Virginia. We have developed the robot iteratively, and have conducted a series of focus group tests with students and teachers at TJ. This has helped us refine the technical concept, as well as develop a better understanding of how it might be used in an educational context. This gives us confidence that students and teachers will respond positively to the completed solution. And finally, we talk with other people in the edtech industry, to determine how this technology might be relevant to educational needs that they see.

As more AIED projects draw lessons from projects that have had good impact, it will help the field overall to realize its potential of improve education. The rapid increase in availability of computing resources is multiplying the opportunities for the field to make a difference. If we seize these opportunities the prospects for the future of AIED are bright indeed.

## Acknowledgments

This research was supported in part by the National Science Foundation under Grant IIS-1321056. Any opinions, findings, and conclusions expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

## References

1. Alelo (2015). Alelo Develops an Interactive Robot for Learning Chinese. Retrieved March 16, 2015 from http://www.prweb.com/releases/chinese/robot/prweb12538156.htm.

2. Borden, J. (2011). The Future of Online Learning. *eLearn Magazine*, August 2011. Retrieved March 16, 2015 from http://elearnmag.acm.org/featured.cfm?aid=2024704.
3. Camacho, J., Johnson, W.L., Valente, A., & Bushika, M. (2009). Cultural training: The Web's newest gaming frontier. In Proceedings of I/ITSEC 2009.
4. Gelfand, A. (2011). Unleashing the Potential. *Worcester Polytechnic University Annual Research Magazine*. Retrieved March 16, 2015 from http://www.wpi.edu/research/magazine/unleashing.html.
5. Getting Smart (2012). New Survey Shows Demand for Personalized & Adaptive Learning Assessments. Retrieved March 16, 2015 from http://gettingsmart.com/2012/02/new-survey-shows-demand-for-personalized-adaptive-learning-assessments/.
6. Institute of Education Sciences (2010). Intervention: Carnegie Learning Curricula and Cognitive Tutor® Software. Retrieved March 16, 2015 from http://ies.ed.gov/ncee/wwc/reports/hs_math/cog_tutor/info.asphttp://ies.ed.gov/ncee/wwc/reports/hs_math/cog_tutor/info.asp.
7. Johnson, W.L. (2010). Serious use of a serious game for language learning. Int. J. of Artificial Intelligence in Ed. 20(2), 175-195.
8. Johnson, W.L., Friedland, L., Watson, A.M., & Surface, E.A. (2012). The art and science of developing intercultural competence. In P.J. Durlach & A.M. Lesgold (Eds.), Adaptive Technologies for Training and Education, 261-285. New York: Cambridge University Press.
9. Marine Corps Center for Lessons Learned (MCCLL) (2008). "Tactical Iraqi Language and Culture Training System" *Marine Corps Center for Lessons Learned Newsletter* 4 (8), 4.
10. MarketsandMarkets.com (2014). Smart Education & Learning Market by Hardware (IWB & SBL), Software (LMS/LCMS, Open Source & Mobile Education Apps), Educational Content (Digital Content, Test And Assessment & Digital Text Book) - Global Advancements, Market Forecast and Analysis (2014 - 2019). Retrieved March 16, 2015 from http://www.marketsandmarkets.com/Market-Reports/smart-digital-education-market-571.html.
11. Next Up Research (2010). NeXt Knowledge Factbook 2010. Retrieved March 16, 2015 from www.kellogg.northwestern.edu/faculty/jones-ben/htm/NextKnowledgeFactbook2010.pdf.

# Second Workshop on Simulated Learners

*held in conjunction with*

Seventeenth International Conference on Artificial Intelligence in Education (AIED 2015)

Friday, June 26, 2015
Madrid, Spain

**Workshop Co-Chairs:**

John Champaign[1] and Gord McCalla[2]

[1] *University of Illinois at Springfield, Springfield, IL, 62703, USA*
[2] *University of Saskatchewan, Saskatoon, S7N 5C9, Canada*

https://sites.google.com/site/simulatedlearners/

# Table of Contents

# Preface

This workshop, a follow-up to the successful first Simulated Learners workshop held at AIED 2013, is intended to bring together researchers who are interested in simulated learners, whatever their role in the design, development, deployment, or evaluation of learning systems. Its novel aspect is that it isn't simply a workshop about pedagogical agents, but instead focuses on the other roles for simulated learners in helping system designers, teachers, instructional designers, etc.

As learning environments become increasingly complex and are used by growing numbers of learners (sometimes in the hundreds of thousands) and apply to a larger range of domains, the need for simulated learners (and simulation more generally) is compelling, not only to enhance these environments with artificial agents, but also to explore issues using simulation that would be otherwise be too expensive, too time consuming, or even impossible using human subjects. While some may feel that MOOCs provide ample data for experimental purposes, it is hard to test specific hypotheses about particular technological features with data gathered for another purpose. Moreover, privacy concerns, ethics approval, attrition rates and platform constraints can all be barriers to this approach. Finally, with thousands of learners at stake, it is wise to test a learning environment as thoroughly as possible before deployment.

Since this is a follow-up to the 2013 workshop, we build on some of the ideas that emerged there (see proceedings at: http://goo.gl/12ODji).

The workshop explores these and other issues with the goal of further understanding the roles that simulated learners may play in advanced learning technology research and development, and in deployed learning systems.

John Champaign and Gord McCalla
Workshop Co-Chairs

# An Approach to Developing Instructional Planners for Dynamic Open-Ended Learning Environments

Stephanie Frost and Gord McCalla

ARIES Lab, Department of Computer Science, University of Saskatchewan, Canada
stephanie.frost@usask.ca, mccalla@cs.usask.ca

**Abstract.** *Instructional planning* (IP) technology has begun to reach large online environments. However, many approaches rely on having centralized metadata structures about the learning objects (LOs). For *dynamic open-ended* learning environments (DOELEs), an approach is needed that does not rely on centralized structures such as prerequisite graphs that would need to be continually rewired as the LOs change. A promising approach is collaborative filtering based on learning sequences (CFLS) using the ecological approach (EA) architecture. We developed a CFLS planner that compares a given learner's most recent path of LOs (of length $b$) to other learners to create a neighbourhood of similar learners. The future paths (of length $f$) of these neighbours are checked and the most successful path ahead is recommended to the target learner, who then follows that path for a certain length (called $s$). We were interested in how well a CFLS planner, with access only to pure behavioural information, compared to a traditional instructional planner that used explicit metadata about LO prerequisites. We explored this question through simulation. The results showed that the CFLS planner in many cases exceeded the performance of the simple prerequisite planner (SPP) in leading to better learning outcomes for the simulated learners. This suggests that IP can still be useful in DOELEs that often won't have explicit metadata about learners or LOs.

**Keywords:** instructional planning, collaborative filtering, dynamic open-ended learning environments, simulated learning environments, simulated learners, ecological approach

## 1 Introduction

Online courses need to be able to personalize their interactions with their many learners not only to help each learner overcome particular impasses but also to provide a path through the learning objects (LOs) that is appropriate to that particular individual. This is the role of *instructional planning* (IP), one of the core AIED sub-disciplines. IP is particularly needed in open-ended learning environments (OELEs), where learners choose their own goals, because it has been shown that sometimes learners require an outside push to move forward

[11]. An added challenge is what we call a *dynamic open-ended* learning environment (DOELE), where both the learners and LOs are constantly changing. Some learners might leave before finishing the course, while others may join long after other learners have already begun. New material (LOs) may need to be added in response to changes in the course or the material, or to learner demand. Sometimes new material will be provided by the course developers, but the big potential is for this to be crowd sourced to anybody, including learners themselves. Other material may fade away over time.

Note that a DOELE is similar to, but not the same as, a "traditional" open-eded learning environment [8, 11]. A traditional open-ended environment also gives students choice, but mostly in the problems they solve and how they solve them, with the course itself fixed in its content, order and goals. In a DOELE everything is open-ended and dynamic, including even what is to be learned, how deeply, when it needs to be learned, and in what order.

An impediment to IP in a DOELE is that there is no centralized representation of knowledge about the content or the learners. Work has been done to make IP possible in online environments, such as [7], where authors showed that by extending the LO metadata, instructional plans could be improved to adapt based on individual learning styles as well as a resource's scheduling availability. But for IP to work in DOELEs, an approach to IP is needed where centralized course structures would not need to be continually revamped (by instructional designers, say) as learners and LOs change.

We wish to explore how IP can be done in a DOELE. We model a DOELE in the ecological approach (EA) architecture [14]. In the EA there is no overall course design. Instead, courses are conceived as collections of learning objects each of which captures usage data as learners interact with it. Over time this usage data accumulates and can be used for many pedagogical purposes, including IP [2]. Drawing inspiration from work like [1, 5], we propose a new IP algorithm based on collaborative filtering of learning sequences (CFLS). For a given learner our planner finds other learners who have traversed a similar sequence of learning objects with similar outcomes (i.e. similar paths). Then it suggests paths to the learner that were successful for these similar learners (peers) going forward.

To evaluate IP techniques in such an environment, one could implement a real course with thousands of learners using the EA to capture learner interactions with the various LOs in the course. However, after doing this it would take several years for enough learners to build up enough interactions with each LO to provide useful data to be used by an instructional planner. Also, in a course with thousands of learners, there is risk of causing confusion or inconvenience to a vast multitude if there are problems while the planner is under development. Finally, there are unanswered design questions such as the criteria to use for identifying an appropriate peer, how many LOs should be recommended for a learner before re-planning occurs, and appropriate values for many other parameters that would be used by the planner. In order to overcome these challenges and gain insight into these questions immediately, we have thus turned to simulation.

## 2 Simulation Environment

Before describing the CFLS planner and experiment in detail, we describe the simulation environment. The simulation is low-fidelity, using very simple abstractions of learners and LOs, as in our earlier work [6]. Each of the 40 LOs has a difficulty level and possible prerequisite relationships with other LOs. Each simulated learner has an attribute, *aptitude-of-learner*, a number between (0,1) representing a learner's basic capability for the subject and allows learners to be divided into groups: low ($\leq .3$), medium ($.4 - .7$) and high aptitude ($\geq .8$).

A number called *P[learned]* is used to represent the learning that occurred when a learner visits a LO, or the probability that the learner learned the LO. P[learned] is generated by an *evaluation function*, a weighted sum: 20% of the learner's score on a LO is attributed to *aptitude-of-learner*, 50% attributed to whether the learner has mastered all of the prerequisite LOs, 20% attributed to whether the learner had seen that LO previously, and 10% attributed to the difficulty level of the LO. We feel this roughly captures the actual influences on how likely it is that real learners would master a learning object.

The simulated learners move through the course by interacting with the LOs, one after another. After each LO is encountered by a simulated learner, the above evaluation function is applied to determine the learner's performance on the LO, the P[learned] for that learner on that LO. In the EA architecture, everything that is known about a learner at the time of an interaction with a LO (in this case, including P[learned]) is captured and associated with that LO. The order of the LOs visited can be set to random, or it can be determined by a planner such as the CFLS planner. To allow for the comparison of different planning approaches without advantaging one approach, each simulated learner halts after its 140th LO regardless of the type of planner being used.

## 3 Experiment

By default, the simulation starts with an empty history - no simulated learners have yet viewed any LOs. However, because the CFLS planner relies on having previous interaction data, it is necessary to initialize the environment. Thus, a simple prerequisite planner (SPP) was used to initialize the case base with a population of simulated learners. The SPP is privy to the underlying prerequisite structure and simply delivers LOs to learners in prerequisite order. As Table 1 shows, the SPP works much better than a random planner. The data from the 65 simulated learners who used the SPP thus was used to initialize the environment before the CFLS planner took over. This interaction data generated by the SPP also provides a baseline for comparison with the CFLS planner. Our simulation experiment was aimed at seeing if, with appropriate choices of $b$ and $f$ (described below) the CFLS planner could work as well or better than the SPP.

We emphasize that the CFLS planner has no knowledge about the underlying prerequisite structure of the learning objects. This is critical for CFLS planning to work in a DOELE. However, there are two places where clarification

**Table 1.** Baseline results for each group of simulated learners (high, medium and low aptitude) when visiting LOs randomly and following a simple prerequisite planner.

| Planning Type / Aptitude | low | medium | high |
|---|---|---|---|
| **Random** | N=21 | N=26 | N=18 |
| Average Score on Final Exam (P[learned]) | 0.107 | 0.160 | 0.235 |
| **Simple Prerequisite Planner (SPP)** | N=21 | N=26 | N=18 |
| Average Score on Final Exam (P[learned]) | 0.619 | 0.639 | 0.714 |

is required. First, while the SPP is running, the evaluation function will be used by the simulation to calculate P[learned] values for each LO visited. This usage data will contain implicit evidence of the prerequisite relationships. So, at a later time when the CFLS planner is given access to the same usage data, the CFLS planner could implicitly discover prerequisite relationships from the interaction data. Second, during the CFLS planner execution, the underlying prerequisite structure is still being consulted by the evaluation function. However, the CFLS planner knows nothing about such prerequisites, only the P[learned] outcome provided by the evaluation function. When simulated learners are replaced with real learners, the evaluation function would disappear and be replaced with a real world alternative, such as quizzes or other evidence to provide a value for P[learned]. Similarly, the CFLS planner does not require knowledge of the difficulty level of each LO, nor does it require knowledge of the aptitude of each learner; these are just stand-in values for real world attributes used by the simulation and would disappear when the planner is applied in a real world setting.

Different studies can use simulated student data in varying ways. In some cases, low fidelity modelling is not adequate. For example, in [4] it was found that the low fidelity method of generating simulated student data failed to adequately capture the characteristics of real data. As a result, when the simulated student dataset was used for training the cognitive diagnosis model, its predictive power was worse than when the cognitive diagnosis model was trained with a simulated student dataset that had been generated with a higher fidelity method. In our study, using a low fidelity model is still informative. We are less concerned with the exactness of P[learned] and are more interested in observing possible relative changes of P[learned] for certain groups of students, as different variations of the planner are tried on identical populations of simulated students.

The CFLS planner works as follows. For a given target learner the CFLS planner looks backward at the $b$ most recent learning objects traversed. Then, it finds other learners who have traversed the same $b$ learning objects with similar P[learned] values. These $b$ LOs can be in any order, a simplification necessary to create a critical mass of similar learners. These are learners in the target learner's "neighbourhood". The planner then looks forward at the $f$ next LOs traversed by each neighbour and picks the highest value path, where value is defined as the average P[learned] achieved on those $f$ LOs ahead. This path is then recommended to the learner, who must follow it for at least $s$ (for "sticky") LOs before replanning occurs. Of course, $s$ is always less than $f$. In our research

we explored various values of $b$ and $f$ to find which leads to the best results (we set $f = s$ for this experiment). "Best results" can be defined many ways, but we focused on two measurements that were taken for each learner at the end of each simulation: the percentage of LOs mastered, and the score on a final exam. A LO is considered to be mastered when a score of P[learned] = 0.6 or greater is achieved. The score on the final exam is taken as the average P[learned] on the LOs that are the leafs of the prerequisite graph (interpreted as the ultimate target concept, which in the real world might well be final exams).

There is still a cold start problem even after the simulation has been initialized with the interaction data from the SPP. This is because the simulated learners who are to follow the CFLS planner have not yet viewed any LOs themselves as they begin the course, so there is no history to match the $b$ LOs to create the plan. In this situation, the CFLS planner matches the learner with another arbitrary learner (from the interaction data from the SPP), and recommends whatever initial path that the other learner took when they first arrived in the course. While another solution to the cold start problem could be to start the new learner with the SPP, we did this to avoid any reliance whatsoever on knowing the underlying prerequisite structure.

The most computationally expensive part of the CFLS planner is finding the learners in the neighbourhood, which is at worst linear on the number of learners and linear on the amount of LO interaction history created by each learner. Each learner's LO interaction history must be searched to check for a match with $b$, with most learners being removed from the list during this process. The forward searching of $f$ is then executed using only the small resulting dataset.

## 4 Results

We ran the CFLS planner 25 different times with all pairings of the values of $b$ and $s$ ranging from 1 to 5, using a population of 65 simulated learners. This population had the same distribution of aptitudes as the population used to generate the baseline interaction data described above. The heat maps in Figs. 1 and 2 show the measurements for each of the 25 simulations, for each aptitude group, with the highest relative scores coloured red, mid-range scores coloured white, and the lowest scores coloured blue. In general, simulated learners achieved higher scores when following the CFLS planner than when given LOs randomly. The CFLS planner even exceeded the SPP in many cases.

A success triangle is visible in the lower left of each aptitude group. The success triangles can be interpreted to mean that if a path is going to be recommended, never send the learner any further ahead ($s$) than you have matched them in the past ($b$). For example if a learner's neighbourhood was created using their $b = 2$ most recent LOs, then never make the learner follow in a neighbour's steps further than $s = 2$ LOs. One reason for the eventual drop at high values of $b$ is that no neighbour could be found and a random match is used instead. However, the abrupt drop at $b > s$ was unexpected. To be sure the pattern was real, an extended series of simulations was run. We ran $b = 6$ and $s = 5$ to see

if there would be a drastic drop in performance, and indeed this was the case. We also ran another row varying $b$ with a fixed $s = 6$, and again found a drop at $b = 7$.

| LOW | | | | |
|---|---|---|---|---|
| b=1 s=1 | b=2 s=1 | b=3 s=1 | b=4 s=1 | b=5 s=1 |
| 100 | 21.9 | 32.5 | 31 | 37.7 |
| b=1 s=2 | b=2 s=2 | b=3 s=2 | b=4 s=2 | b=5 s=2 |
| 89.6 | 86 | 36.9 | 36.9 | 40.4 |
| b=1 s=3 | b=2 s=3 | b=3 s=3 | b=4 s=3 | b=5 s=3 |
| 72.1 | 68.6 | 62 | 43.21 | 40.1 |
| b=1 s=4 | b=2 s=4 | b=3 s=4 | b=4 s=4 | b=5 s=4 |
| 77.3 | 74.4 | 72.1 | 66.1 | 49.3 |
| b=1 s=5 | b=2 s=5 | b=3 s=5 | b=4 s=5 | b=5 s=5 |
| 68.1 | 70.7 | 67.5 | 67.4 | 63.5 |

| MEDIUM | | | | |
|---|---|---|---|---|
| b=1 s=1 | b=2 s=1 | b=3 s=1 | b=4 s=1 | b=5 s=1 |
| 100 | 50 | 45.8 | 42.2 | 44.1 |
| b=1 s=2 | b=2 s=2 | b=3 s=2 | b=4 s=2 | b=5 s=2 |
| 100 | 100 | 42.9 | 40.3 | 38 |
| b=1 s=3 | b=2 s=3 | b=3 s=3 | b=4 s=3 | b=5 s=3 |
| 100 | 99.4 | 98.6 | 42.4 | 43 |
| b=1 s=4 | b=2 s=4 | b=3 s=4 | b=4 s=4 | b=5 s=4 |
| 100 | 99.3 | 99.5 | 99.4 | 50.8 |
| b=1 s=5 | b=2 s=5 | b=3 s=5 | b=4 s=5 | b=5 s=5 |
| 100 | 100 | 100 | 100 | 100 |

| HIGH | | | | |
|---|---|---|---|---|
| b=1 s=1 | b=2 s=1 | b=3 s=1 | b=4 s=1 | b=5 s=1 |
| 100 | 75 | 39.3 | 39.7 | 41.1 |
| b=1 s=2 | b=2 s=2 | b=3 s=2 | b=4 s=2 | b=5 s=2 |
| 100 | 100 | 41.7 | 34.9 | 40 |
| b=1 s=3 | b=2 s=3 | b=3 s=3 | b=4 s=3 | b=5 s=3 |
| 100 | 100 | 100 | 42.5 | 50 |
| b=1 s=4 | b=2 s=4 | b=3 s=4 | b=4 s=4 | b=5 s=4 |
| 100 | 100 | 100 | 100 | 61 |
| b=1 s=5 | b=2 s=5 | b=3 s=5 | b=4 s=5 | b=5 s=5 |
| 100 | 100 | 100 | 100 | 100 |

**Fig. 1.** Average % Learning Objects Mastered by aptitude group

| LOW | | | | |
|---|---|---|---|---|
| b=1 s=1 | b=2 s=1 | b=3 s=1 | b=4 s=1 | b=5 s=1 |
| 0.6587 | 0.1036 | 0.1314 | 0.1283 | 0.146 |
| b=1 s=2 | b=2 s=2 | b=3 s=2 | b=4 s=2 | b=5 s=2 |
| 0.5178 | 0.4387 | 0.1398 | 0.1248 | 0.1363 |
| b=1 s=3 | b=2 s=3 | b=3 s=3 | b=4 s=3 | b=5 s=3 |
| 0.4051 | 0.266 | 0.2256 | 0.1586 | 0.132 |
| b=1 s=4 | b=2 s=4 | b=3 s=4 | b=4 s=4 | b=5 s=4 |
| 0.4138 | 0.2984 | 0.3016 | 0.2755 | 0.176 |
| b=1 s=5 | b=2 s=5 | b=3 s=5 | b=4 s=5 | b=5 s=5 |
| 0.357 | 0.2884 | 0.2859 | 0.2679 | 0.2249 |

| MEDIUM | | | | |
|---|---|---|---|---|
| b=1 s=1 | b=2 s=1 | b=3 s=1 | b=4 s=1 | b=5 s=1 |
| 0.6894 | 0.1851 | 0.2105 | 0.2099 | 0.2425 |
| b=1 s=2 | b=2 s=2 | b=3 s=2 | b=4 s=2 | b=5 s=2 |
| 0.7004 | 0.698 | 0.2058 | 0.22 | 0.1972 |
| b=1 s=3 | b=2 s=3 | b=3 s=3 | b=4 s=3 | b=5 s=3 |
| 0.6942 | 0.6761 | 0.6715 | 0.1944 | 0.2152 |
| b=1 s=4 | b=2 s=4 | b=3 s=4 | b=4 s=4 | b=5 s=4 |
| 0.6931 | 0.6867 | 0.6874 | 0.6856 | 0.2292 |
| b=1 s=5 | b=2 s=5 | b=3 s=5 | b=4 s=5 | b=5 s=5 |
| 0.6912 | 0.6884 | 0.6924 | 0.6965 | 0.6899 |

| HIGH | | | | |
|---|---|---|---|---|
| b=1 s=1 | b=2 s=1 | b=3 s=1 | b=4 s=1 | b=5 s=1 |
| 0.7641 | 0.2514 | 0.2805 | 0.2866 | 0.2702 |
| b=1 s=2 | b=2 s=2 | b=3 s=2 | b=4 s=2 | b=5 s=2 |
| 0.77 | 0.7694 | 0.2673 | 0.2738 | 0.2748 |
| b=1 s=3 | b=2 s=3 | b=3 s=3 | b=4 s=3 | b=5 s=3 |
| 0.7653 | 0.7638 | 0.7727 | 0.3019 | 0.3097 |
| b=1 s=4 | b=2 s=4 | b=3 s=4 | b=4 s=4 | b=5 s=4 |
| 0.768 | 0.7697 | 0.7633 | 0.7697 | 0.3431 |
| b=1 s=5 | b=2 s=5 | b=3 s=5 | b=4 s=5 | b=5 s=5 |
| 0.7601 | 0.7612 | 0.7591 | 0.7644 | 0.7636 |

**Fig. 2.** Average Score on Final Exam (P[learned]) by aptitude group

A hot spot of successful combinations of $b$ and $s$ appeared for each aptitude group. For low aptitude learners, it was best to only match on the $b = 1$ most recent learning objects, and to follow the selected neighbour for only $s = 1$ LOs ahead before replanning. This combination of $b$ and $s$ is the only time when the CFLS planner outperformed the SPP for the low aptitude group. However, for the medium and high aptitude groups, the CFLS planner outperformed the SPP in all cases within the success triangle. Looking at final exam scores (Fig. 2), medium aptitude learners responded well to being matched with neighbours using $b = 1$ or 2 and sticking with the chosen neighbour for the same distance ahead. The high aptitude group responded very well to using neighbourhoods created with $b = 3$ and recommending paths of $s = 3$.

Within the success triangles, the rows and columns of Fig. 2 were checked to see if there existed an ideal $b$ for a given $s$, and vice versa. Wherever there appeared to be a large difference, Student's t-test was used to check for statistical significance. We are able to use paired t-tests because the simulated learners have exactly the same characteristics in all the simulation runs, the only difference being the order in which LOs were interacted with. For example, learner #3 always has *aptitude-of-learner* = .4, so, there is no difference in that learner

between simulation runs. We used a two-tailed t-test because it was not certain whether one distribution was going to be higher or lower than the other.

Looking along the rows, when $s$ is held the same, there are some cases where one value of $b$ is better than another. For the low aptitude group, for the most part the lower the $b$, the better. For the medium aptitude group, there were no significant advantages to changing $b$. For the high aptitude group, when $s = 3$, the t-test was used to check if $b = 3$ was significantly more advantageous than using $b = 2$. The measurements for Score on the Final Exam for the high aptitude learners were compared between both simulation results, ($b = 2$ and $s = 3$) and ($b = 3$ and $s = 3$). With N=19 learners in this group, the calculated p-value was 0.009, indeed a statistically significant difference.

Looking along the columns, when $b$ is held the same there was a case where increasing $s$, i.e. sticking to a longer plan ahead, was statistically advantageous. In the medium aptitude group, when $b = 1$ it was statistically better to use $s = 2$ than to use $s = 1$ with a p-value of 0.011. None of the increases of $s$ with the same $b$ were significant for the high aptitude group, and there were no increases for the low aptitude group.

## 5 Analysis and Future Work

Through simulation, we have shown that a CFLS planner can be "launched" from an environment that has been conditioned with interaction data from another planner, such as an SPP, and operate successfully using only learner usage data kept by the EA and not needing centralized metadata such as a prerequisite graph. This is one of the key requirements for DOELEs. Like biological evolution, the EA is harsh in that it observes how learners succeed or fail as various paths are tried. Successful paths for particular types of learners, regardless of whether they follow standard prerequisites, is the only criterion of success. New learners or new learning objects will find their niche - some paths will work for some learners but not for others, and this is discovered automatically through usage.

More experiments are needed to explore the many possibilities of the simulation environment. While this experiment was not a true test of a DOELE because new learners and LOs were not inserted, this can be readily explored in future work. New additions could be matched randomly a few times in order to build enough data in the EA, and then automatically incorporated into neighbourhood matches or into future plans.

Given the evaluation function that was selected, we found that planning ahead and sticking to the plan worked best for high aptitude learners and a reactive approach (planning ahead but sticking to the plan for only a short time) worked best for the low aptitude learners. Would a different pattern emerge if a different evaluation function were chosen? Would a different threshold for mastery than P[learned] $> 0.6$ make any difference? In future work, would it be worthwhile to break down the aptitude groups into six: very-high, high, medium-high, medium-low, low, and very-low? This may assist with more easily tuning the weights of the evaluation function, as there was not much difference in our

results between the high and medium aptitude groups. In addition, more experiments where $s < f$ are needed to answer the question of whether the drop along the edge of each success triangle was because of $s$ or $f$. Also, in this work we did not look at the many different types of pedagogical interactions (ex. asking the student a question, giving a hint etc.) and focused on very abstract representations. More work is needed to explore this approach on systems later in the design process, when more detail about the content and the desired interactions with learners is known.

Future work could also investigate the usage of a differential planner, where different settings are tuned for different situations. For example, when creating a neighbourhood for a low aptitude learner, medium aptitude learners could be allowed into the neighbourhood if they have a matching $b$. Results could reveal situations where for example a low aptitude learner is helped by following in the steps of a medium aptitude learner. A differential planner could also dynamically choose the values of $b$ and $s$ for a given individual instead of using the same values for everyone at all times. For example, in a real world setting a CFLS planner may try to create a plan using a neighbourhood of $b = 3$, knowing it is optimal, but if for the specific case there is not enough data, it could change to $b = 2$ on the fly. Other aspects that could be changed are the criteria for creating the neighbourhood: rather than filtering by aptitude, another attribute could be chosen such as click behaviour or learning goals.

## 6 Conclusion

In this paper, we have described the need for instructional planning in DOELEs with many LOs aimed at large numbers of learners. Instructional planners such as [13] use AI planning technology that is based on states, actions and events, which are difficult to infer from an unstructured online environment. In recent years, instructional planning has been replaced by instructional design approaches such as [3]. Advanced instructional planners from the 1990s, such as PEPE and TOBIE [16] can blend different teaching strategies to appropriate situations. We have shown that instructional planning can still be done in the less rigid courses envisioned by the EA architecture and likely to be commonplace in the future, using only learner usage data kept by the EA and not needing centralized metadata about the course.

We have shown a specific planning technique, the CFLS planner, that is appropriate for DOELEs, and how to experiment in this domain. The simulation experiment revealed the number of LOs from a target learner's recent browsing history should be used for creating a neighbourhood ($b$), a question that has also been investigated by other researchers, such as in [18]. We have also found recommendations for settings for how far ahead to plan ($s$ and $f$) for different groups of learners, and identified questions for future work. As is the case with collaborative filtering and case-based approaches, the quality of the plans created is limited to the quality of LOs within the repository and the quality

of interactions that have previously occurred between learners and sequences of LOs.

The bottom-up discovery of prerequisite relationships has been investigated by others, such as [17]. When the need for centralized metadata about a course is discarded, and when the further step is taken that different paths can be found to work better for different learners, then a shift in thinking occurs. Each individual learner could effectively have a unique ideal (implicit) prerequisite graph. Whether or not a prerequisite relationship even exists between two LOs could vary from learner to learner. The notion of prerequisite can thus be viewed not only as a function of the content relationships, but also as a function of the individual learner.

Making recommendations of sequences has also been identified as a task in the recommender systems domain [9]. An approach such as a CFLS planner is a step in the direction of building recommender systems that can use sequence information to recommend sequences. This has also been accomplished with standards approaches such as [15]. Simulation with the EA provides another method for developing and testing such approaches.

Overall, the research we have done to date and the questions it raises, shows the value of exploring these complex issues using simulation. We were able to essentially generate some 25 different experiments exploring some issues in instructional planning, in a very short time when compared to what it would have taken to explore these same issues with real learners. Others have also used simulation for developing an educational planner, such as [10] for social assessment games. To be sure our simulation model was of low fidelity, but we suspect that there are some properties of the CFLS planner that we have uncovered that apply in the real world (the lower triangles seem to be very strong and consistent patterns). And, there are some very real issues that we can explore fairly quickly going forward that might reveal other strong patterns, as discussed. We believe that it isn't always necessary to have simulations with high cognitive fidelity (as in SimStudent [12]) to find out interesting things. Low fidelity simulations such as the ones we have used in this and our earlier work [6] (and those of [2]) have a role to play in AIED. Especially as we move into the huge questions of dynamic open-ended learning environments with thousands of learners and big privacy issues, the sharp minimalist modelling possible with low fidelity simulation should allow quick and safe experimentation without putting too many real learners at risk and without taking years to gain insights.

**Acknowledgements**
We would like to thank the Natural Sciences and Engineering Research Council of Canada for funding some aspects of this research.

## References

[1] Cazella, S., Reategui, E., and Behar, P.: Recommendation of Learning Objects Applying Collaborative Filtering and Competencies. IFIP Advances in Information and Communication Technology, 324, pp 35-43 (2010)

[2] Champaign, J.: Peer-Based Intelligent Tutoring Systems: A Corpus-Oriented Approach. Ph.D. Thesis, University of Waterloo, Waterloo, Canada (2012)
[3] Drachsler, H., Hummel, H. and Koper, R.: Using Simulations to Evaluate the Effects of Recommender Systems for Learners in Informal Learning Networks. SIRTEL Workshop (Social Information Retrieval for Technology Enhanced Learning) at the 3rd EC-TEL (European Conf. on Technology Enhanced Learning) Maastricht, The Netherlands: CEUR-WS.org, online CEUR-WS.org/Vol-382/paper2.pdf (2008)
[4] Desmarais, M., and Pelczer, I.: On the Faithfulness of Simulated Student Performance Data. In de Baker, R.S.J. et al. (Eds.), Proc. of the 3rd Int. Conf. on Educ. Data Mining, pp 21-30. Pittsburg USA (2010)
[5] Elorriaga, J. and Fernández-Castro, I.: Using Case-Based Reasoning in Instructional Planning: Towards a Hybrid Self-improving Instructional Planner. Int. Journal of Artificial Intelligence in Educ., 11(4), pp 416-449 (2000)
[6] Erickson, G., Frost, S., Bateman, S., and McCalla, G.: Using the Ecological Approach to Create Simulations of Learning Environments. In Lane, H.C. et al. (Eds), Proc. of the 16th Int. Con. on AIED, pp 411-420. Memphis USA: Springer (2013)
[7] Garrido, A. and Onaindia, E.: Assembling Learning Objects for Personalized Learning: An AI Planning Perspective. Intelligent Systems, IEEE, 28(2), pp 64-73 March/April (2013)
[8] Hannafin, M.J.: Learning in Open-Ended Environments: Assumptions, Methods and Implications. Educational Technology, 34(8), pp 48-55 (1994)
[9] Herlocker, J., Konstan, J., Terveen, L., and Riedl, J.: Evaluating Collaborative Filtering Recommender Systems. ACM Transactions on Information Systems (TOIS) 22(1), pp 5-53 (2004)
[10] Laberge, S., Lenihan, T., Shabani, S., and Lin, F.: Multiagent Coordination for Planning and Enacting an Assessment Game. Workshop on MultiAgent System Based Learning Environments of Int. Tutoring Systems (ITS) Honolulu, USA (2014)
[11] Land, S.: Cognitive Requirements for Learning with Open-Ended Learning Environments. Deuce. Technology Research and Development, 48(3), pp 61-78 (2000)
[12] Matsuda, N., Cohen, W. and Koedinger, K.: Teaching the Teacher: Tutoring SimStudent Leads to More Effective Cognitive Tutor Authoring. Int. Journal of Artificial Intelligence in Educ., 25(1), pp 1-34 (2014)
[13] Matsuda, N., and VanLehn, K.: Decision Theoretic Instructional Planner for Intelligent Tutoring Systems. In B. du Boulay (Ed.), Workshop Proc. on Modelling Human Teaching Tactics and Strategies, ITS 2000 pp 72-83. (2000)
[14] McCalla, G.: The Ecological Approach to the Design of e-Learning Environments: Purpose-based Capture and Use of Information about Learners. Journal of Interactive Media in Educ., http://jime.open.ac.uk/jime/article/view/2004-7-mccalla (2004)
[15] Shen, L. and Shen, R.: Learning Content Recommendation Service Based on Simple Sequencing Specification. In Liu W et al. (Eds.) Advances in Web-Based Learning - ICWL 2004 3rd Int. Conf. Web-based Learning, LNCS 3143, pp 363-370. Beijing, China:Springer (2004)
[16] Vassileva, J. and Wasson, B.: Instructional Planning Approaches: from Tutoring Towards Free Learning. Proceedings of Euro-AIED'96, Lisbon, Portugal (1996)
[17] Vuong, A., Nixon, T., and Towle, B.: A Method for Finding Prerequisites Within a Curriculum. In Pechenizkiy, M. et al. (Eds.) , Proc. of the 4th Int. Con. on Educ. Data Mining, pp 211-216. Eindhoven, the Netherlands (2011)
[18] Zhang, Y., and Cao, J.: Personalized Recommendation Based on Behavior Sequence Similarity Measures. In Cao, L. et al. (Eds.) Int. Workshop on Behaviour and Social Informatics / Behaviour and Social Informatics and Computing (BSI/BSIC 2013), Gold Coast QLD Australia / Beijing China, LNCS 8178, pp 165-177 (2013)

# Exploring the Issues in Simulating a Semi-Structured Learning Environment: the SimGrad Doctoral Program Design

David Edgar K. Lelei and Gordon McCalla

ARIES Laboratory, Department of Computer Science, University of Saskatchewan
davidedgar.lelei@usask.ca and mccalla@cs.usask.ca

**Abstract.** The help seeking and social integration needs of learners in a semi-structured learning environment require specific support. The design and use of educational technology has the potential to meet these needs. One difficulty in the development of such support systems is in their validation because of the length of time required for adequate testing. This paper explores the use of a simulated learning environment and simulated learners as a way of studying design validation issues of such support systems. The semi-structured learning environment we are investigating is a graduate school, with a focus on the doctoral program. We present a description of the steps we have taken in developing a simulation of a doctoral program. In the process, we illustrate some of the challenges in the design and development of simulated learning environments. Lastly, the expected contributions and our research plans going forward are described.

**Keywords:** Simulated learners, Simulated learning environment, Agent-based simulation, Help seeking, Doctoral learners, Multi-agent system.

## 1 Introduction

Artificial Intelligence in Education (AIED) is one of the research fields whose focus is the use of technology to support learners of all ages and across all domains[1]. Although, one shortcoming of AIED research is the limited research attention that very dynamic and semi-structured domains, such as a graduate school, have received. There is little research that investigates how technology can be used to help connect learners (help seeker and potential help givers) in the graduate school domain. Consequently, there is a gap in our understanding of how such technology may mitigate graduate learners' attrition rates and time-to-degree. We have suggested the use of reciprocal recommender technology to assist in the identification of a suitable helper [1]. However, the nature of graduate school means that validation of any education system designed to be used in a semi-structured environment would take a long time (measured in years). This paper aims to address this challenge by exploring the use of

[1] http://iaied.org/about/

simulated learning environment and simulated learners as a potential way of validating educational technologies designed to support doctoral learners.

In this paper, we first describe the nature and the metrics used by interested stakeholders to measure the success or lack therefore of a doctoral program. Following this, we briefly discuss the uses of simulation as it relates to learning environment. We then introduce the research questions we are interested in answering using simulation. We go on to describe the architectural design of our simulation model. Further, we show how data about the 'real world' target domain is used to inform the parameters and initial conditions for the simulation model. This provides the model with a degree of fidelity. Throughout this model development process, we illustrate some of the challenges in the design and development of simulated learning environments. We conclude the paper with a discussion of the expected contributions and our research plans going forward.

## 2 Understanding Doctoral Program

Graduate school is a very dynamic and complex social learning environment. A doctoral program in particular is a dynamic, semi-structured, and complex learning environment. Most doctoral programs have some structure in the sense that there are three distinct stages that doctoral learners must go through: admission stage, coursework stage, and dissertation stage. While coursework stage is fairly structured, the dissertation stage is not. Further, the dissertation stage have various milestones that include: comprehensive exam, thesis proposal, research, writing, and dissertation defense. As time passes, learners move from one stage to the next and their academic and social goals change. There is need for self-directed learning and individual doctoral learners are responsible for their own learning pace and choice of what to learn especially in the dissertation stage.

The dynamic nature of the program ensures that there is constant change; there are new learners joining the program, other learners leaving the program either through graduation or deciding to drop out, and still other learners proceeding from one stage to the next. There are two key aspects that influences learners to decide whether to persist or drop out of a learning institution: academic and social integration [2], [3] which are impacted by learner's initial characteristics and experiences during their duration in the program. The various stages of the doctoral program (e.g., coursework) and learning resources can be seen as factors that directly influence the academic integration of a doctoral learner. Peers and instructors/supervisors can be viewed as supporting the social aspects of the doctoral program and hence, directly impact the social integration of doctoral learners. As time passes, doctoral learners continually interact with both the academic and social facets of the doctoral program. As a result, there is constant change in learners' commitment to their academic goal and the social sides of the learning institution

Time-to-degree, completion rates, and attrition rates are important factors influencing the perception and experience of graduate education by interested stakeholders [4], [5]. Research on doctoral attrition and time-to-completion indicates that on aver-

age, the attrition rate is between 30% and 60% [5]–[8]. Long times to completion and a high attrition rate are costly in terms of money to the funding institution and the learning institution; and in terms of time and effort to the graduate student(s) and supervisor(s) [8]. Lack of both academic and social integration (isolation) have been shown to affect graduate learners decision to persist [2], [3], [9]. Learners facing academic and social integration challenges should be enabled to engage in a community of peers to foster interaction and hence, encourage peer help and personalized collaboration [10]. Understanding the nature of learner-institution interactions that foster doctoral learners' persistence to degree is important to both the learning institution and its learners. We use simulation to achieve this feat.

Simulation is an established third way of exploring research questions in addition to qualitative and quantitative methods [11], [12]. VanLehn [13] has identified three main uses of simulation in learning environments: 1) to provide an environment for human teachers to practise their teaching approaches; 2) to provide an environment for testing different pedagogical instructional design efforts; 3) to provide simulated learners who can act as companions for human learners. Our research is mainly focused on the first and the second uses – to enable deep insight into the complex interaction of the factors affecting doctoral learners' attrition and time-to-degree leading to a better design of an educational system. Therefore, our research questions are formulated around investigations of how various factors influence time-to-degree, completion rates, and dropout rates of doctoral students. We are interested in answering the following research questions:

1. How does the number of classes (as a platform for social integration with peers – potential helpers) offered by a program(s) or taken by a learner, influence learners' time-to-degree and their propensity to persist or drop out?
2. How does the average class size (as basis of learners' social integration) attended by learners, impact learners' time-to-degree and their inclination to persist or drop out? What is the optimum class size?
3. How does the overall population size of the learners (a few learners vs many learners) influence learners' time-to-degree and their likelihood to persist or drop out?
4. Does timely help affects doctoral learners' time-to-degree and their decision to persist or drop out? If so, how?
5. How does the level of reciprocation influence the formation of a 'helpful community' of learners and adaptive help seeking behavior of the learners?

Use of simulation enables us to explore the aforementioned issues in a fine-grained controlled environment. For example, it would be almost impossible in the 'real world' setting to examine the impact of different number of course to take or class size to attend. Two cohorts of learners will have different attributes. Simulation allows us to tweak the number of courses or class size without touching the other characteristics of learners. Hence, we are able to see the real impact of one variable at a time. Before any exploration and insight can be gained on these issues, there is need to design and implement the simulation model.

## 3 Building an Initial Prototype of *SimGrad*

In this section we demonstrate the steps we have taken in the development of our initial prototype of our simulated doctoral learning environment: *SimGrad*. We show how a designer of an educational technology can develop a model of their target learning environment and inform its initial condition with available 'real world' data.

### 3.1 *SimGrad* Design

We need to design a simulation model by addressing two key challenges. First, we need to consider issues related to the modeling of the learning environment: how do we design conceptual and computational models of a doctoral program and what stakeholders should be included in these models? The second concern is about modeling of simulated learners: what doctoral learners' features affect persistence and time-to-degree, what factors do we model, and can we inform these features with available 'real world' data?

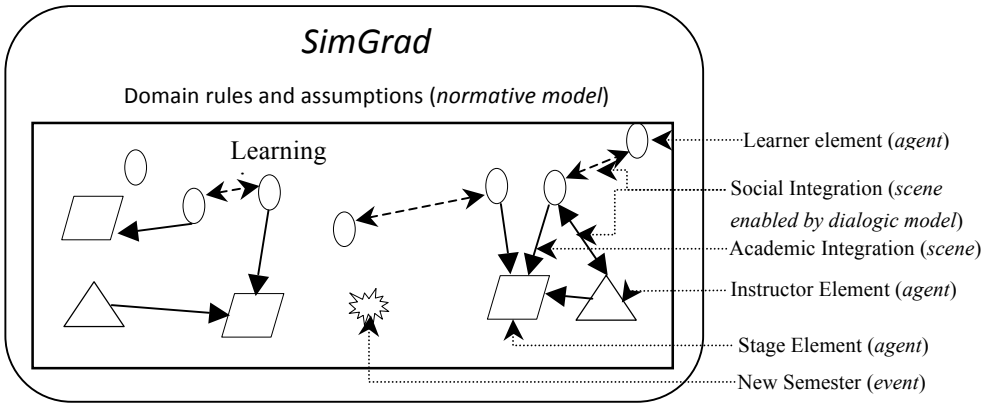



**Fig. 1.** SimGrad conceptual framework, its three elements, and the possible interaction between the elements

We have designed our conceptual model of the different aspects of simulated doctoral learners and doctoral learning environment based on the simulated learning environment specifications suggested by Koper et al. in [14], and features for building an electronic institution proposed by Esteva et al. [15]. We name our conceptual framework, *SimGrad*. Its core elements include: *normative model* - specifies requirements and constraints to guide agent actions and behavior; *dialogic model* – deals with interaction strategies and communication mechanism; *events* – refers to happenings in the model that trigger (re)action by agents; *scene* – description of an interaction between elements; *elements (agents)* – represent key stakeholders of the target domain that are modeled. Elements are modeled as agents. Each of the agents has attributes and behavior which are informed by our assumptions guided by our research questions and factors that influence learners' decision to persist. Every element of interest is to be

modeled within the learning environment and all possible interactions and operations within the learning setting is guided by domain rules represented by the normative model. See **Fig. 1**.

In our simulation model, we have chosen to model three types of elements: class, instructor, and learner. In this paper, in keeping with model simplicity, both the class and the instructor agents are passive while the learner agent is modeled to be active and reactive to its environment. Also, the only instructor's attributes we are interested in are related to classes (see **Table 1**). We modeled only one type of instructor agent. Another instructor type agent that can be modeled is the supervisor.

Each learner agent has the following properties: autonomy, social ability, reactivity, proactivity, and a degree of intentionality. We have also identified the following key attributes for our agent learner model: state – (busy, available), program, stages, course taken, peer interactions (pertaining challenges), academic integration, social integration, and motivation (see **Table 2**). In our model, peer interaction and state contribute to a learners' social integration, while research area, stage, course taken impact to their academic integration. Motivation combines both the social and academic integration and hence, is the main factor that determines whether an agent continues to persist or chooses to drop out of the program.

**Table 1.** Comparison of computed attributes of the three agent types

| *Attribute – data (value range)* | *Agent learner* | *Agent instructor* | *Agent class* |
|---|---|---|---|
| Total number of classes take, taught, or frequency of offering within 10 years – *numeric (0-20)* | X | X | X |
| Grade obtained, average awarded, or average obtained by learners – *numeric (0,12)* | X | X | X |
| Take classes from, teach classes in, or class offered in various programs – *textual (program id)* | X | X | X |
| Instructors teaching a class – *array list (instructor id)* | X | - | X |
| What is the class size – *numeric (1-5)* | X | | X |
| Number of classes taken or taught per year - *numeric (0,4)* | X | X | - |
| Which classes are taken or taught – *textual (class id)* | X | X | - |

The main intentions of each agent is to persist through doctoral requirements to graduation and to do so in a timely manner. However, each of these agents reacts to the different challenges at various stages of graduate school in divergent and autonomous ways. At the coursework stage, agents have the goal of taking courses that are relevant to their field and that they will perform well. When facing a course choice challenge or any other particular challenge, we have modeled our agents to proactively associate with peers to seek help. Each peer makes individual choice on whether to or not to respond to a request for help from others. The dialogic model

handles the agent to agent interaction and communication through a message passing mechanism [16].

**Table 2.** Attributes and parameters considered for an agent learner model for learners, their description and how each of them changes.

| *Attribute* | *Value - description* | *How it changes* |
|---|---|---|
| Enrolment | Date (MM/YYYY)<br>Indicate the month a year an agent enrolled in the program | Does not change |
| Graduation | Date (MM/YYYY)<br>Target graduation date | Evaluated whenever an agent completes a milestone |
| State | Textual (busy, available)<br>Indicates an agent availability to help others, assigned based on the smallest time unit model | Changes whenever an agent experiences a challenge |
| Program | Textual (program id)<br>Identify an agent's closer community within the larger community of learners | Does not change during a simulation run |
| Stage | Textual (admission, coursework, dissertation, timeout, dropout) | Admission stage is like an event. Learner move to the coursework immediately after admission. They more to dissertation after completing their course load. |
| Courses taken | Array [course, mark, instructor id](0-6)<br>Record courses taken by an agent and the marks obtain in each course | Every end of semester that the student took classes, this array is updated |
| Peer interaction | Array [learner id, challenge, result],<br>Keep track of an agent interactions with others and the outcome of the interaction | Changes whenever two agents interact |
| Academic integration | Numeric (-1,1)<br>Measures the academic satisfaction | Changes whenever an agent learner interacts with agent stage (i.e., completes a milestone or experience a challenge) |
| Social integration | Numeric (-1,1)<br>Measures a learners sense of belonging to the learning environment | Changes whenever an agent learner interacts with its peers or agent instructors |
| Motivation | Numeric (-1,1)<br>Measures the propensity of an agent to still want to persist. A motivation value above 0.3 indicates persistence. A value between -0.3 and 0.3 indicate help seeking needed. A value below -0.3 means the agent drops out | Whenever there is a change in the social and academic integration values. Its value is the average of the integration values. |

### 3.2 Informing *SimGrad* behavior and evaluation functions

Having identified the important agents and their key attributes, there are two sets of important functions for each element that need to be modelled: behaviour functions and evaluation functions [17]. Behaviour functions inform the decision making of the active elements and dictates the interaction patterns between them and the other modeled elements (e.g., how many classes a given agent takes). Evaluation functions indicate whether or not various interactions between the different agents in a simulation were successful (e.g., determine what grade a given agent attains in a class it took). Informing such functions with 'real world' data allows the simulation to behave in a way consistent with reality.

Simulation model fidelity is an issues that might arise when using simulation to study a target real world phenomenon. However, the most important issue to consider is the research question to be answered. While Champaign [18] used a very low fidelity model, Matsuda et al. [19] used a model with high cognitive fidelity to reach compelling conclusion. Further yet, Erickson et al. [17] also demonstrated that is possible to use a medium fidelity model and uncover interesting results. In some situations it might not be possible to have a high fidelity model because of lack of data. A case in point is our simulation scenario. Where possible, we inform our simulation functions with data received from the U of S on their doctoral program. An investigation into the U of S data showed that we will not be able to inform every aspect of our simulation model. It would be desirable to inform every initial aspects of our simulation model with 'real world' data but, we do not have data on the dissertation stage.

We are provided information on student id, years a student is registered, year of graduation (if graduated), student's program, classes taken and marks obtained, class instructor, and students instructional responsibilities. From this dataset we are able to inform the admission and coursework stages of our model (academic integration). However, there is no information concerning the dissertation stage and the social integration aspects. While it possible to inform various behaviour and evaluation functions for our simulation model, in this paper we focus on describing the steps we took to inform two functions of our simulation: learning environment admission behaviour function, and learners' class interactions behaviour function.

As already mentioned, admission is an important part of a doctoral program that contributes to it dynamic nature. The admission process is complex and involves a lot of stakeholders and processes, but we are concerned only with determining the year to year patterns in how many students are admitted. To provide some fidelity to our simulated learning environment admission, we analyzed data provided to us by the U of S University Data Warehouse[2]. The provided dataset contained information on doctoral learners registered in the 10 years 2005-2014. In this time there were 2291 doctoral learners with a total of 52850 data points on class registration. The 2005 registration included learners who had joined the program earlier than 2005. In order to get a clean admission pattern, we only considered learners who were registered from the year 2006 onwards. This reduced the population size to 1962.

[2] http://www.usask.ca/ict/services/ent-business-intelligence/university-data-warehouse.php

We were able to identify three admission periods, September, January, and May. We then obtained values for each of the admissions months for the years 2006-2014. This provided a distribution for each month that we used to generate a scatter plot of admission numbers. A sigmoidal pattern emerged. Next, we performed a non-linear curve fitting to the scatter plot so that the admission function can be represented in the form $Y = St*(c + x)$, where c is a constant, $St$ is a variable dependent on the admission period, and $x$ is the admission period. We then ran a regression to find values of each of these variables. This allowed us to model the admission patterns observed in the U of S dataset.

Next we derived the number of classes taken. To introduce some realism to the number classes taken behaviour, we had to further prune the data. We only considered data for students whose cohorts would have been registered for at least 3 years by the end of the year 2014 and hence, we considered class taking behaviour of 1466 U of S doctoral learners.

We obtained the number of classes each of the remaining learners we registered in and created a histogram. This histogram showed us the distribution of the number of students registered for a certain number of classes. Next, we transformed this distribution graph into a cumulative distribution function. We then took an inverse of the cumulative distribution function to achieve a quantile function. The quantile function, when run over many learners, assigns learners a class count that mimics the initial histogram. We use this quantile function to inform the number of classes a learner can take.

In this section we have described the importance of informing a simulation model with 'real world' data. We have described two functions that are informed with U of S dataset. Other examples of functions that can be informed using the U of S dataset include: class performance evaluation function, dropout behaviour function, time to degree behaviour function, and flow through behavior function (main as pertains to coursework stage). We have identified that missing data values is a major hindrance in this endeavor. There are possible ways of informing simulation attributes where there are no 'real world' data to derive from. A designer can either assign common sense values, generate and assign random values, or refer to the research literature to identify patterns that have been found by other researchers. Since we have the enrolment dates and the graduate dates for learners who graduate, we choose to derive common sense values with these two dates guiding the process and the value range.

## 4 Discussion, Expected Contributions, and Future Research Plans

Despite the growth in the use of simulation as a method for exploration and learning in many areas such as: engineering, nursing, medicine [20], and building design [21], research in the used of simulation within AIED is still at an early stage. There is need for more research to demonstrate that the outputs of simulation runs are desirable and informative to the AIED community. In this paper, we aim at contributing to this notion and by promoting the use of simulation in educational research and presenting

an agent based simulation conceptual framework for building simulated learning environment, with a focus on the semi-structured ones. Simulated learning environment and simulated learners are important in exploring and understanding a given learning domain. Further, it helps with the generation of system validation data.

The expected contributions to AIED include: providing a conceptual framework for simulated graduate school learning environment – an architecture that enables investigations into factors affecting doctoral learners progress through their program; shedding light on learner modeling issues in dynamic learning environments; and demonstrating the importance of simulation in exploring various AIED research domains, particularly semi-structured domains.

Current research work is focused on the implementation of the simulation model and the refinement of the various behaviour and evaluation functions. Once the implementation is done, we will validate our model against the dataset we have from the U of S before proceeding to explore the impact of various environmental factors. Since we are informing the simulation with both common sense assumptions and U of S dataset, the goal is to tweak the common sense assumptions such that when the model is run we get similar results as the U of S data in terms of class performance, dropout rate, and time-to-degree. Achieving this, would give us confidence that we have captured reality in some measurable way. We can then start exploring the various impact of measures we are interested in examining. As earlier indicated, we are interested in exploring the interactions of a number of variables: number of classes taken which will impact the availability of potential peer helpers, the effect of reciprocity on help seeking and help giving, and the effect of help seeking and other factors on doctoral learners' time-to-degree and attrition rates.

## Acknowledgement

We would like to thank University of Saskatchewan University Data Warehouse team for giving us access to 10 year dataset. Specifically, we would like to thank Mathew Zip for processing and explaining to the first author the nature of the dataset. We also wish to acknowledge and thank the Natural Science and Engineering Research Council of Canada (NSERC) for funding our research.

## References

[1] D. E. K. Lelei, "Supporting Lifelong Learning: Recommending Personalized Sources of Assistance to Graduate Students," in *Artificial Intelligence in Education*, 2013, pp. 912–915.

[2] V. Tinto, "Taking student success seriously: Rethinking the first year of college.," *Ninth Annu. Intersession Acad. Aff. Forum,* vol. 19, no. 2, pp. 1–8, 2005.

[3] V. Tinto, "Dropout from higher education: A theoretical synthesis of recent research," *Rev. Educ. Res.*, vol. 45, no. 1, pp. 89–125, 1975.

[4] H. Groenvynck, K. Vandevelde, and R. Van Rossem, "The Ph.D. Track: Who Succeeds, Who Drops Out?," *Res. Eval.*, vol. 22, no. 4, pp. 199–209, 2013.

[5] F. J. Elgar, “Phd Degree Completion in Canadian Universities,” Nova Scotia, Canada, 2003.
[6] M. Walpole, N. W. Burton, K. Kanyi, and A. Jackenthal, “Selecting Successful Graduate Students: In-Depth Interviews With GRE Users,” Princeton, NJ, 2002.
[7] L. Declou, “Linking Levels to Understand Graduate Student Attrition in Canada,” McMaster University, Hamilton, Ontario, Canada, 2014.
[8] B. E. Lovitts, *Leaving the Ivory Tower: The Causes and Consequences of Departure from Doctoral Study*, Illustrate., vol. 32. Rowman & Littlefield Publishers, 2001, p. 307.
[9] A. Ali and F. Kohun, “Dealing with isolation feelings in IS doctoral programs,” *Int. J. Dr. Stud.*, vol. 1, no. 1, pp. 21–33, 2006.
[10] Computing Research Association, “Grand research challenges in information systems,” Washington, D.C, 2003.
[11] R. Axelrod, “Advancing the Art of Simulation in the Social Sciences,” in *Proceedings of the 18th European Meeting on Cybernetics and Systems Research*, 2006, pp. 1–13.
[12] N. Gilbert and K. G. Troitzsch, “Simulation and Social Science,” in *Simulation for the Social Scientist*, McGraw-Hill International, 2005, pp. 1–14.
[13] K. VanLehn, S. Ohlsson, and R. Nason, “Applications of Simulated Students: An Exploration,” *J. Artif. Intell. Educ.*, vol. 5, no. 2, pp. 1–42, 1994.
[14] R. Koper and B. Olivier, “Representing the Learning Design of Units of Learning,” *Educ. Technol. Soc.*, vol. 7, no. 3, pp. 97–111, 2003.
[15] M. Esteva, J.-A. Rodriquez-Aguilar, C. Sierra, P. Garcia, and J. L. Arcos, “On the Formal Specification of Electronic Institutions,” in *Agent Mediated Eectronic Commerce*, 2001, pp. 126–147.
[16] H. J. C. Berendsen, D. Van Der Spoel, and R. Van Drunen, “GROMACS: A message-passing parallel molecular dynamics implementation,” *Comput. Phys. Commun.*, vol. 91, no. 1, pp. 43–56, 1995.
[17] G. Erickson, S. Frost, S. Bateman, and G. Mccalla, “Using the Ecological Approach to Create Simulations of Learning Environments,” in *In Artificial Intelligence in Education*, 2013, pp. 411–420.
[18] J. Champaign, “Peer-Based Intelligent Tutoring Systems: A Corpus-Oriented Approach,” University of Waterloo, 2012.
[19] N. Matsuda, W. W. Cohen, K. R. Koedinger, V. Keiser, R. Raizada, E. Yarzebinski, S. P. Watson, and G. Stylianides, “Studying the Effect of Tutor Learning Using a Teachable Agent that Asks the Student Tutor for Explanations,” in *2012 IEEE Fourth International Conference On Digital Game And Intelligent Toy Enhanced Learning*, 2012, pp. 25–32.
[20] A. L. Baylor and Y. Kim, “Simulating Instructional Roles Through Pedagogical Agents,” *Int. J. Artif. Intell. Educ.*, vol. 15, no. 2, pp. 95–115, 2005.
[21] G. Augenbroe, “Trends in Building Simulation,” *Build. Environ.*, vol. 37, no. 8, pp. 891–902, 2002.

# Exploring the Role of Small Differences in Predictive Accuracy using Simulated Data

Juraj Nižnan, Jan Papoušek, and Radek Pelánek

Faculty of Informatics, Masaryk University Brno
{niznan,jan.papousek,xpelanek}@mail.muni.cz

**Abstract.** Research in student modeling often leads to only small improvements in predictive accuracy of models. The importance of such improvements is often hard to assess and has been a frequent subject of discussions in student modeling community. In this work we use simulated students to study the role of small differences in predictive accuracy. We study the impact of such differences on behavior of adaptive educational systems and relation to interpretation of model parameters. We also point out a feedback loop between student models and data used for their evaluation and show how this feedback loop may mask important differences between models.

## 1 Introduction

In student modeling we mostly evaluate models based on the quality of their predictions of student answers as expressed by some performance metric. Results of evaluation often lead to small differences in predictive accuracy, which leads some researchers to question the importance of model improvements and meaningfulness of such results [1]. Aim of this paper is to explore the impact and meaning of small differences in predictive accuracy with the use simulated data. For our discussion and experiments in this work we use a single performance metric – Root Mean Square Error (RMSE), which is a common choice (for rationale and overview of other possible metrics see [15]). The studied questions and overall approach are not specific to this metric.

Simulated students provide a good way to study methodological issues in student modeling. When we work with real data, we can use only proxy methods (e.g., metrics like RMSE) to evaluate quality of models. With simulated data we know the "ground truth" so we can study the link between metrics and the true quality of models. This enables us to obtain interesting insight which may be useful for interpretation of results over real data and for devising experiments. Similar issues are studied and explored using simulation in the field of recommender systems [7, 17].

We use a simple setting for simulated experiments, which is based on an abstraction of a real system for learning geography [12]. We simulate an adaptive question answering system, where we assume items with normally distributed difficulties, students with normally distributed skills, and probability of correct

answer given by a logistic function of the difference between skill and difficulty (variant of a Rasch model). We use this setting to study several interrelated question.

### 1.1 Impact on Student Practice

*What is the impact of prediction accuracy (as measured by RMSE) on the behavior of an adaptive educational system and students' learning experience?*

Impact of small differences in predictive performance on student under-practice and over-practice (7-20%) has been demonstrated using real student data [18], but insight from a single study is limited. The relation of RMSE to practical system behavior has been analyzed also in the field of recommender systems [2] (using offline analysis of real data). This issue has been studied before using simulated data is several studies [5, 6, 10, 13]. All of these studies use very similar setting – they use Bayesian Knowledge Tracing (BKT) or its extensions and their focus is on mastery learning and student under-practice and over-practice. They differ only in specific aspects, e.g., focus on setting thresholds for mastery learning [5] or relation of moment of learning to performance metrics [13]. In our previous work [16] have performed similar kind of simulated experiments (analysis of under-practice and over-practice) both with BKT and with student models using logistic function and continuous skill.

In this work we complement these studies by performing simulated experiments in slightly different setting. Instead of using BKT and mastery learning, we use (variants of) the Rasch model and adaptive question answering setting. We study different models and the relation between their prediction accuracy and the set of items used by the system.

### 1.2 Prediction Accuracy and Model Parameters

*Can RMSE be used to identify good model parameters? What is the relation of RMSE to the quality of model parameters?*

In student modeling we often want to use interpretable models since we are interested not only in predictions of future answers, but also in reconstructing properties of students and educational domains. Such outputs can be used to improve educational systems as was done for example by Koedinger at al. [9]. When model evaluation shows that model A achieves better prediction accuracy (RMSE) then model B, results are often interpreted as evidence that model A better reflects "reality". Is RMSE a suitable way to find robust parameters? What differences in metric value are meaningful, i.e., when we can be reasonably sure that the better model really models reality in better way? Is statistical significance of differences enough? In case of real data it is hard to answer these question since we have no direct way to evaluate the relation of a model to reality. However, we can study these questions with simulated data, where we have access to the ground truth parameters. Specifically, in our experiments we study the relation of metric values with the accuracy of reconstructing the mapping between items and knowledge components.

### 1.3 Feedback between Data Collection and Evaluation

*Can the feedback loop between student models and adaptive choice of items influence evaluation of student models?*

We also propose novel use of simulated students to study a feedback loop between student models and data collection. The data that are used for model evaluation are often collected by a system which uses some student model for adaptive choice of items. The same model is often used for data collection and during model evaluation. Such evaluation may be biased – it can happen that the used model does not collect data that would show its deficiencies. Note that the presence of this feedback loop is an important difference compared to other forecasting domains. For example in weather forecasting models do not directly influence the system and cannot distort collected data. In student modeling they can.

So far this feedback has not been thoroughly studied in student modeling. Some issues related to this feedback have been discussed in previous work on learning curves [6, 11, 8]. When a tutoring system uses mastery learning, students with high skill drop out earlier from the system (and thus from the collected data), thus a straightforward interpretation of aggregated learning curves may be misleading. In this work we report experiment with simulated data which illustrate possible impact of this feedback loop on model evaluation.

## 2 Methodology

For our experiments we use a simulation of a simplified version of an adaptive question answering systems, inspired by our widely used application for learning geography [12]. Fig. 1 presents the overall setting of our experiments. System asks students about items, answers are dichotomous (correct/incorrect), each student answers each item at most once. System tries to present items of suitable difficulty. In evaluation we study both the prediction accuracy of models and also sets of used items. This setting is closely related to item response theory and computerized adaptive testing, specifically to simulated experiments with Elo-type algorithm reported by Doebler et al. [3].

**Simulated Students and Items** We consider a set of simulated students and simulated items. To generate student answers we use logistic function (basically the Rasch model, respectively one parameter model from item response theory): $P(correct|\theta_s, d_i) = 1/(1 + e^{-(\theta_s - d_i)})$, where $\theta_s$ is the skill of a student $s$ and $d_i$ is difficulty of an item $i$.

To make the simulated scenarios more interesting we also consider multiple knowledge components. Items are divided into disjoint knowledge components and students have different skill for each knowledge component. Student skills and item difficulties are sampled from a normal distribution. Skills for individual knowledge components are independent from one another.

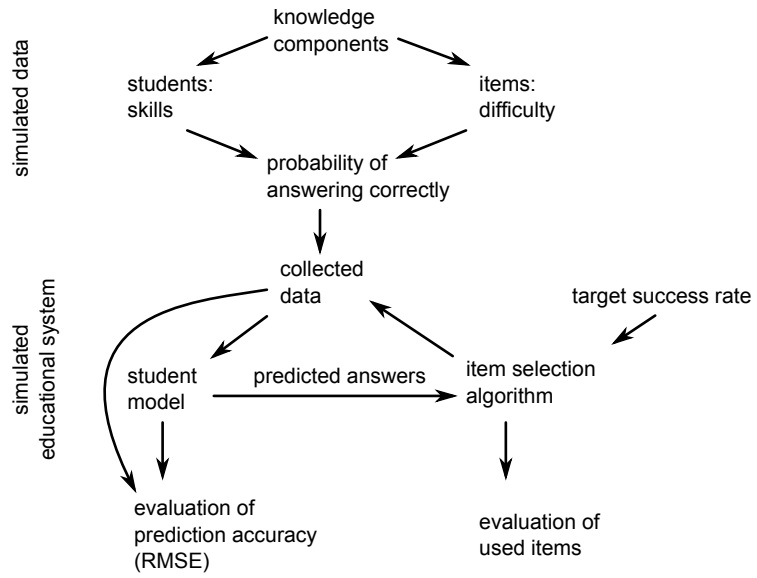



**Fig. 1.** Setting of our experiments

**Item Selection Algorithm** The item selection algorithm has as a parameter a target success rate $t$. It repeatedly presents items to a (simulated) student, in each step it selects an item which has the best score with respect to the distance of the predicted probability of correct answer $p$ and the target rate $t$ (illustrated by gray dashed line in Fig. 3). If there are multiple items with the same score, the algorithm randomly selects one of them.

**Student Models** Predictions used by the item selection algorithm are provided by a student model. For comparison we consider several simple student models:

- Optimal model – Predicts the exact probability that is used to generate the answer (i.e., a "cheating" model that has access to the ground truth student skill and item difficulty).
- Optimal with noise – Optimal model with added (Gaussian) noise to the difference $\theta_s - d_i$ (before we apply logistic function).
- Constant model – For all students and items it provides the same prediction (i.e., with this model the item selection algorithm selects items randomly).
- Naive model – Predicts the average accuracy for each item.
- Elo model – The Elo rating system [4, 14] with single skill. The used model corresponds to the version of the system as described in [12] (with slightly modified uncertainty function).
- Elo concepts – The Elo system with multiple skills with correct mapping of items to knowledge components.

– Elo wrong concepts – The Elo system with multiple skills with wrong mapping of items to knowledge components. The wrong mapping is the same as the correct one, but 50 (randomly choosen) items are classified incorrectly.

**Data** We generated 5,000 students and 200 items. Items are divided into 2 knowledge components, each user has 2 skills corresponding to the knowledge components and each item has a difficulty. Both skills and difficulties were sampled from standard normal distribution (the data collected from the geography application suggests that these parameters are approximately normally distributed). The number of items in a practice session is set to 50 unless otherwise noted.

## 3 Experiments

We report three types of experiments, which correspond to the three types of questions mentioned in the introduction.

### 3.1 Impact on Student Practice

Our first set of experiments studies differences in the behavior of the simulated system for different models. For the evaluation of model impact we compare the sets of items selected by the item selection algorithm. We make the assumption that the algorithm for item selection using the optimal model generates also the optimal practice for students. For each user we simulate practice of 50 items (each item is practiced at most once by each student). To compare the set of practiced items between those generated by the optimal model and other models we look at the size of the intersection. We assume that bigger intersection with the set of practiced items using the optimal model indicates better practice. Since the intersection is computed per user, we take the mean.

This is, of course, only a simplified measure of item quality. It is possible that an alternative model selects completely different set of items (i.e., the intersection with the optimal set is empty) and yet the items are very similar and their pedagogical contribution is nearly the same. However, for the current work this is not probable since we are choosing 50 items from a pool of only 200 items. For future work it would be interesting to try to formalize and study the "utility" of items.

**Noise Experiment** The optimal model with noise allows us to easily manipulate differences in predictive accuracy and study their impact on system behavior. Experiment reported in the left side of Fig. 2 shows both the predictive accuracy (measured by RMSE) and the impact on system behavior (measured by the size of the intersection with the optimal practiced set as described above) depending on the size of noise (we use Gaussian noise with a specified standard deviation). The impact of noise on RMSE is approximately quadratic and has a slow rise –

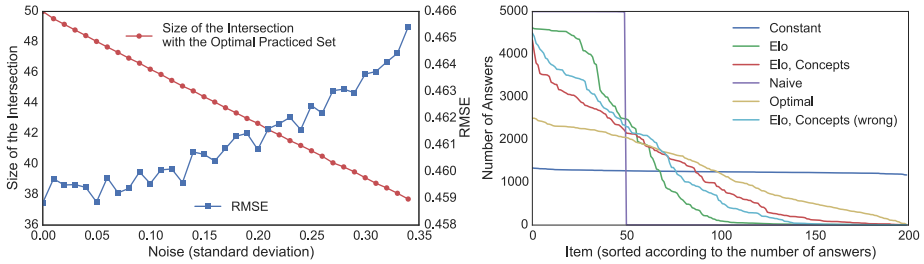



**Fig. 2.** Size of the intersection with the optimal practiced set of items and RMSE depending on Guassian noise in optimal model (left side). Distribution of answers over the items based on the given model (right side).

this is a direct consequence of the quadratic nature of the metric. The impact on used items is, however, approximately linear and rather steep. The most interesting part is for noise values in the interval $[0, 0.1]$. In this interval the rise in RMSE values is very small and unstable, but the impact on used items is already high.

**Model Comparison** Right side of the Fig. 2 shows the distribution of the number of answers per item for different models. The used models have similar predictive accuracy (specific values depend on what data we use for their evaluation, as discussed below in Section 3.3), yet the used model can dramatically change the form of the collected data.

When we use the optimal model, the collected data set covers almost fairly most items from the item pool. In the case of worse models the use of items is skewed (some items are used much more frequently than others). Obvious exception is the constant model for which the practice is completely random. The size of the intersection with the optimal practiced set for these models is – Constant: 12.5; Elo: 24.2; Elo, Concepts: 30.4; Elo, Concepts (wrong): 28.5; Naive: 12.0. Fig. 3 presents a distribution of answers according to the true probability of their correctness (given by the optimal model). Again there is a huge difference among the given models, especially between simple models and those based on Elo.

### 3.2 Prediction Accuracy and Model Parameters

Metrics of prediction accuracy (e.g., RMSE) are often used for model selection. Model that achieves lower RMSE is assumed to have better parameters (or more generally better "correspondence to reality"). Parameters of a selected model are often interpreted or taken into account in improvement of educational systems. We checked validity of this approach using experiments with knowledge components.

We take several models with different (random) mappings of items to knowledge components and evaluate their predictive accuracy. We also measure the

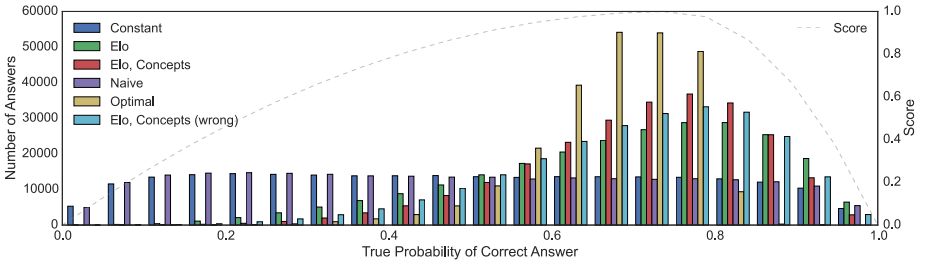



**Fig. 3.** Distribution of answers according to the true probability of correct answer. The gray dashed line stands for the score function used by the algorithm for item selection.

quality of the used mappings – since we use simulated data, we know the ground truth mapping and thus can directly measure the quality of each mapping. Quality is expressed as the portion of items for which the mapping agrees with the ground truth mapping. The names of the knowledge components are irrelevant in this setting. Therefore, we compute quality for each one-to-one mapping from the names of the components in the model to the names of the components in the ground truth. We select the highest quality as the quality of the model's item-to-component mapping. To focus only on quality of knowledge components, we simplify other aspects of evaluation, specifically each student answers all items and their order is selected randomly.

These experiments do not show any specific surprising result, so we provide only general summary. Experiments show that RMSE values correlate well with the quality of mappings. In case of small RMSE differences there may be "swaps", i.e., a model with slightly higher RMSE reflects reality slightly better. But such results occur only with insufficiently large data and are unstable. Whenever the differences in RMSE are statistically significant (as determined by t-test over different test sets), even very small differences in RMSE correspond to improvement in the quality of the used mappings. These results thus confirm that it is valid (at least in the studied setting) to argue that a model A better corresponds to reality than a model B based on the fact that the model A achieves better RMSE than the model B (as long as the difference is statistically significant). It may be useful to perform this kind of analysis for different settings and different performance metrics.

### 3.3 Feedback between Data Collection and Evaluation

To study feedback between the used student model and collected data (as is described in subsection 1.3) we performed the following experiment: We choose one student model and use it as an input for adaptive choice of items. At the same time we let all other models do predictions as well and log answers together with all predictions.

Fig. 4 shows the resulting RMSE for each model in individual runs (data collected using specific model). The figure shows several interesting results. When

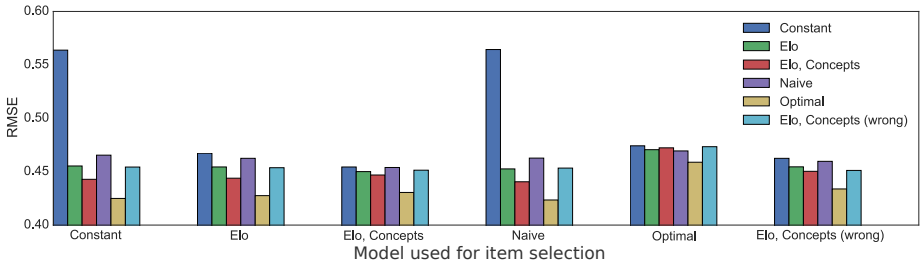



**Fig. 4.** RMSE comparison over data collected using different models.

the data are collected using the optimal model, the RMSE values are largest and closest together; even the ordering of models is different from other cases. In this case even the constant model provides comparable performance to other models – but it would be very wrong to conclude that "predictive accuracy of models is so similar that the choice of model does not matter". As the above presented analysis shows, different models lead to very different choice of items and consequently to different student experience. The reason for small differences in RMSE is not similarity between models, but characteristics of data ("good choice of suitable items"), which make predictions difficult and even a naive predictor comparatively good.

Another observation concerns comparison between the "Elo concepts" and "Elo concepts (wrong)" models. When data are collected by the "Elo concepts (wrong)" model, these two models achieve nearly the same performance, i.e., models seem to be of the same quality. But the other cases show that the "Elo concepts" model is better (and in fact it is by construction a better student model).

## 4 Conclusions

We have used simulated data to show that even small differences in predictive accuracy of student models (as measured by RMSE) may have important impact on behavior of adaptive educational systems and for interpretation of results of evaluation. Experiments with simulated data, of course, cannot demonstrate the practical impact of such small differences. We also do not claim that small differences in predictive accuracy are always important. However, experiments with simulated data are definitely useful, because they clearly illustrate mechanisms that could play role in interpretation of results of experiments with real student data. Simulated data also provide setting for formulation of hypotheses that could be later evaluated in experiments with real educational systems.

Simulated data also enable us to perform experiments that are not practical for realization with actual educational systems. For example in our experiment with the "feedback loop" we have used different student models as a basis for item selection. Our set of models includes even a very simple "constant model",

which leads to random selection of practiced item. In real setting we would be reluctant to apply such a model, as it is in contrary with the advertised intelligent behavior of our educational systems. However, experiments with this model in simulated setting provide interesting results – they clearly demonstrate that differences in predictive accuracy of models do not depend only on the intrinsic quality of used student models, but also on the way the data were collected.

Our analysis shows one particularly interesting aspect of student modeling. As we improve student models applied in educational systems, we should expect that evaluations of predictive accuracy performed over these data will show worse absolute values of performance metrics and smaller and smaller differences between models (even if models are significantly different), just because virtues of our models enable us to collect less predictable data.

## References


1. Joseph E Beck and Xiaolu Xiong. Limits to accuracy: How well can we do at student modeling. In *Proc. of Educational Data Mining*, 2013.
2. Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 39–46. ACM, 2010.
3. Philipp Doebler, Mohsen Alavash, and Carsten Giessing. Adaptive experiments with a multivariate elo-type algorithm. *Behavior research methods*, pages 1–11, 2014.
4. Arpad E Elo. *The rating of chessplayers, past and present*, volume 3. Batsford London, 1978.
5. Stephen E Fancsali, Tristan Nixon, and Steven Ritter. Optimal and worst-case performance of mastery learning assessment with bayesian knowledge tracing. In *Proc. of Educational Data Mining*, 2013.
6. Stephen E Fancsali, Tristan Nixon, Annalies Vuong, and Steven Ritter. Simulated students, mastery learning, and improved learning curves for real-world cognitive tutors. In *AIED Workshops*. Citeseer, 2013.
7. Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.
8. Tanja Käser, Kenneth R Koedinger, and Markus Gross. Different parameters-same prediction: An analysis of learning curves. In *Proceedings of 7th International Conference on Educational Data Mining. London, UK*, 2014.
9. Kenneth R Koedinger, John C Stamper, Elizabeth A McLaughlin, and Tristan Nixon. Using data-driven discovery of better student models to improve student learning. In *Artificial intelligence in education*, pages 421–430. Springer, 2013.
10. Jung In Lee and Emma Brunskill. The impact on individualizing student models on necessary practice opportunities. *International Educational Data Mining Society*, 2012.
11. R Charles Murray, Steven Ritter, Tristan Nixon, Ryan Schwiebert, Robert GM Hausmann, Brendon Towle, Stephen E Fancsali, and Annalies Vuong. Revealing the learning in learning curves. In *Artificial Intelligence in Education*, pages 473–482. Springer, 2013.

12. Jan Papoušek, Radek Pelánek, and Vít Stanislav. Adaptive practice of facts in domains with varied prior knowledge. In *Proc. of Educational Data Mining*, pages 6–13, 2014.
13. Zachary A Pardos and Michael V Yudelson. Towards moment of learning accuracy. In *AIED 2013 Workshops Proceedings Volume 4*, page 3, 2013.
14. Radek Pelánek. Application of time decay functions and Elo system in student modeling. In *Proc. of Educational Data Mining*, pages 21–27, 2014.
15. Radek Pelánek. Metrics for evaluation of student models. *Journal of Educational Data Mining*, 2015. To appear.
16. Radek Pelánek. Modeling student learning: Binary or continuous skill? In *Proc. of Educational Data Mining*, 2015.
17. Al Mamunur Rashid, Istvan Albert, Dan Cosley, Shyong K Lam, Sean M McNee, Joseph A Konstan, and John Riedl. Getting to know you: learning new user preferences in recommender systems. In *Proceedings of the 7th international conference on Intelligent user interfaces*, pages 127–134. ACM, 2002.
18. Michael V Yudelson and Kenneth R Koedinger. Estimating the benefits of student model improvements on a substantive scale. In *EDM 2013 Workshops Proceedings*, 2013.

# Using Data from Real and Simulated Learners to Evaluate Adaptive Tutoring Systems

José P. González-Brenes[1], Yun Huang[2]

[1] Pearson School Research & Innovation Network, Philadelphia, PA, USA
jose.gonzalez-brenes@pearson.com
[2] Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA
yuh43@pitt.edu

**Abstract.** Classification evaluation metrics are often used to evaluate adaptive tutoring systems— programs that teach and adapt to humans. Unfortunately, evidence suggests that existing convention for evaluating tutoring systems may lead to suboptimal decisions. In a companion paper, we propose Teal, a new framework to evaluate adaptive tutoring. In this paper we propose an alternative formulation of Teal using simulated learners. The main contribution of this novel formulation is that it enables approximate inference of Teal, which may useful on the cases that Teal becomes computationally intractable. We believe that this alternative formulation is simpler, and we hope it helps as a bridge between the student modeling and simulated learners community.

## 1 Introduction

Adaptive systems teach and adapt to humans and improve education by optimizing the subset of *items* presented to students, according to their historical performance [3], and on features extracted from their activities [6]. In this context, items are questions, or tasks that can be graded individually. Adaptive tutoring may be evaluated with randomized control trials. For example, in a seminal study [3] that focused on earlier adaptive tutors, a controlled trial measured the time students spent on tutoring, and their performance on post-tests. The study reported that the adaptive tutoring system enabled significantly faster teaching, while students maintained the same or better performance on post-tests

Unfortunately, controlled trials can become extremely expensive and time consuming to conduct: they require institutional review board approvals, experimental design by an expert, recruiting and often payment of enough participants to achieve statistical power, and data analysis. Automatic evaluation metrics improve the engineering process because they enable less expensive and faster comparisons between alternative systems.

The adaptive tutoring community has tacitly adopted conventions for evaluating tutoring systems [4]. Researchers often evaluate their models with classification evaluation metrics that assess the *student model* component of the tutoring system— student models are the subsystems that forecast whether a learner will answer the next item correctly. However, automatic evaluation metrics are

intended to measure an outcome of the end user. For example, the PARADISE [9] metric used in spoken dialogue systems correlates to user satisfaction scores. We are not aware of evidence that supports that classification metrics correlate with learning outcomes; yet there is a growing body of evidence [2, 5] that suggests serious problems with them. For example, classification metrics ignore that an adaptive system may not help learners— which could happen with a student model with a flat or decreasing learning curve [1, 8]. A decreasing learning curve implies that student performance decreases with practice; this curve is usually interpreted as a modeling problem, because it operationalizes that learners are better off with no teaching.

We study a novel formulation of the Theoretical Evaluation of Adaptive Learning Systems (Teal) [5] evaluation metric. The importance of evaluation metrics is that they help practitioners and researchers quantify the extent that a system helps learners.

## 2 Theoretical Evaluation of Adaptive Learning Systems

In this section, we just briefly summarize Teal and do not compare it with a related method called ExpOppNeed [7]. Teal assumes the adaptive tutoring system is built using a single-skill Knowledge Tracing Family model [3, 6]. Knowledge Tracing uses a Hidden Markov Model (HMM) per skill to model the student's knowledge as latent variables. It models whether a student applies a practice opportunity of a skill correctly. The latent variables are used to model the latent student proficiency, which is often modeled with a binary variable to indicated mastery of the skill.

To use Teal on data collected from students, we first train a model using an algorithm from the Knowledge Tracing family, then we use the learned parameters to calculate the effort and outcome for each skill.

- Effort: Quantifies how much practice the adaptive tutor gives to students. In this paper we focus on counting the number of items assigned to students but, alternatively, amount of time could be considered.
- Outcome: Quantifies the performance of students after adaptive tutoring. For simplicity, we operationalize performance as the percentage of items that students are able to solve after tutoring. We assume that the performance on solving items is aligned to the long-term interest of learners.

Algorithm 1 describes our novel formulation. Teal calculates the expected number of practice that an adaptive tutor gives to students. We assume that the tutor stops teaching a skill once the student is very likely to answer the next item correctly according to a model from the Knowledge Tracing Family [6]. The adaptive tutor teaches an additional item if two conditions hold: (i) it is likely that the student will get the next item wrong— in other words, the probability of answering correctly the next item is below a threshold $\tau$; and (ii) the tutor has not decided to stop instruction already.

The inputs of Teal are:

– Real student performance data from $m$ students practicing a skill. Data from each student is encoded into a sequence of binary observations of whether the student was able to apply correctly the skill at different points in time.
– A threshold $\tau \in \{0 \dots 1\}$ that indicates when to stop tutoring. We operationalize this threshold as the target probability that the student will apply the skill correctly.
– A parameter $T$ that indicates the number of practice opportunities each of the simulated students will practice the skill.

**Algorithm 1** Teal algorithm for models with one skill per item

**Require:** real student data $\mathbf{y}^{(1)} \dots \mathbf{y}^{(m)}$, threshold $\tau$, # of simulated time steps $T$
**function** TEAL
  $\theta \leftarrow$ Knowledge_Tracing($\mathbf{y}^{(1)} \dots \mathbf{y}^{(m)}$)
  $e \leftarrow \{\ \}$
  $s \leftarrow \{\ \}$
  **for** $\hat{\mathbf{y}} \in$ get_simulated_student($\theta, T$) **do**:
    $e \leftarrow$ calculate_effort($\hat{\mathbf{y}}, \theta, \tau$)
    **if** $e < T$ **then**
      $s \leftarrow$ calculate_score($\hat{\mathbf{y}}$, e)
    **else**
      $s \leftarrow$ imputed_value
  **return** mean($e$), mean($s$)

Teal learns a Knowledge Tracing model from the data collected from real students interacting with a tutor. Our new formulation uses simulated learners sampled from the Knowledge Tracing parameters. This enables us to decide how many simulated students to generate. Our original formulation required $2^m$ sequences to be generated, which can quickly become computationally intractable. If an approximate solution is acceptable, our novel formulation allows more efficient calculations of Teal. Teal quantifies the effort and outcomes of students in adaptive tutoring. Even though measuring effort and outcomes is not novel by itself, Teal's contribution is measuring both without a randomized trial. Teal quantifies effort as how much practice the tutor gives. For this, we count the number of items assigned to students. For a single simulated student, this is:

$$\text{calculate\_effort}(y_1, \dots, y_T, \theta, \tau) \equiv \arg\min_t p(y_t | y_1 \dots y_{t-1}, \theta) > \tau \qquad (1)$$

The threshold $\tau$ implies a trade-off between student effort and scores and responds to external expectations from the social context. Teal operationalizes the outcome as the performance of students after adaptive tutoring as the percentage of items that students are able to solve after tutoring:

$$\text{calculate\_score}(y_1, \dots, y_T, e) \equiv \sum_{t=e} \frac{\delta(\mathbf{y}_t, \text{correct})}{T - e} \qquad (2)$$

Here, $\delta(\cdot, \cdot)$ is the Kronecker function that returns 1 iff its arguments are equal.

## 3 Discussion

Simulation enables us to measure effort and outcome for a large population of students. Previously, we required Teal to be computed exhaustively on all student outcomes possibilities. We relax the prohibitively expensive requirement of calculating all student outcome combinations. Our contribution is that Teal can be calculated with a simulated dataset size that is large yet tractable.

## References


1. R. Baker, A. Corbett, and V. Aleven. More Accurate Student Modeling through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. In B. Woolf, E. Aïmeur, R. Nkambou, and S. Lajoie, editors, *Intelligent Tutoring Systems*, volume 5091 of *Lecture Notes in Computer Science*, pages 406–415. Springer Berlin / Heidelberg, 2008.
2. J. Beck and X. Xiong. Limits to accuracy: how well can we do at student modeling? In S. K. D'Mello, R. A. Calvo, and A. Olney, editors, *Proceedings of the 6th International Conference on Educational Data Mining, Memphis, Tennessee, USA, July 6-9, 2013*, pages 4–11. International Educational Data Mining Society, 2013.
3. A. Corbett and J. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278, 1995.
4. A. Dhanani, S. Y. Lee, P. Phothilimthana, and Z. Pardos. A comparison of error metrics for learning model parameters in bayesian knowledge tracing. Technical Report UCB/EECS-2014-131, EECS Department, University of California, Berkeley, May 2014.
5. González-Brenes and Y. José P., Huang. Your model is predictive— but is it useful? theoretical and empirical considerations of a new paradigm for adaptive tutoring evaluation. In J. G. Boticario, O. C. Santos, C. Romero, and M. Pechenizkiy, editors, *Proceedings of the 8th International Conference on Educational Data Mining*, Madrid, Spain, 2015.
6. J. P. González-Brenes, Y. Huang, and P. Brusilovsky. General Features in Knowledge Tracing: Applications to Multiple Subskills, Temporal Item Response Theory, and Expert Knowledge. In M. Mavrikis and B. M. McLaren, editors, *Proceedings of the 7th International Conference on Educational Data Mining*, London, UK, 2014.
7. J. I. Lee and E. Brunskill. The impact on individualizing student models on necessary practice opportunities. In K. Yacef, O. R. Zaïane, A. Hershkovitz, M. Yudelson, and J. C. Stamper, editors, *Proceedings of the 5th International Conference on Educational Data Mining*, pages 118–125, Chania, Greece, 2012.
8. D. Rai, Y. Gong, and J. E. Beck. Using dirichlet priors to improve model parameter plausibility. In T. Barnes, M. Desmarais, C. Romero, and S. Ventura, editors, *Proceedings of the 2nd International Conference on Educational Data Mining*, Cordoba, Spain, 2009.
9. M. Walker, C. Kamm, and D. Litman. Towards developing general models of usability with PARADISE. *Natural Language Engineering*, 6(3):363–377, 2001.

# Authoring Tutors with Complex Solutions: A Comparative Analysis of Example Tracing and SimStudent

Christopher J. MacLellan[1], Erik Harpstead[1], Eliane Stampfer Wiese[1], Mengfan Zou[2], Noboru Matsuda[1], Vincent Aleven[1], and Kenneth R. Koedinger[1]

[1] Carnegie Mellon University, Pittsburgh PA, USA,
{cmaclell, eharpste, stampfer, noboru.matsuda, aleven, koedinger}@cs.cmu.edu,
[2] Tsinghua University, Beijing, China,
zmf11@mails.tsinghua.edu.cn

**Abstract.** Problems with many solutions and solution paths are on the frontier of what non-programmers can author with existing tutor authoring tools. Popular approaches such as Example Tracing, which allow authors to build tutors by demonstrating steps directly in the tutor interface. This approach encounters difficulties for problems with more complex solution spaces because the author needs to demonstrate a large number of actions. By using SimStudent, a simulated learner, it is possible to induce general rules from author demonstrations and feedback, enabling efficient support for complexity. In this paper, we present a framework for understanding solution space complexity and analyze the abilities of Example Tracing and SimStudent for authoring problems in an experimental design tutor. We found that both non-programming approaches support authoring of this complex problem. The SimStudent approach is 90% more efficient than Example Tracing, but requires special attention to ensure model completeness. Example Tracing, on the other hand, requires more demonstrations, but reliably arrives at a complete model. In general, Example Tracing's simplicity makes it good for a wide range problems, a reason for why it is currently the most widely used authoring approach. However, SimStudent's improved efficiency makes it a promising non-programmer approach, especially when solution spaces become more complex. Finally, this work demonstrates how simulated learners can be used to efficiently author models for tutoring systems.

**Keywords:** Tutor Authoring, Intelligent Tutoring Systems, Cognitive Modeling, Programming-by-Demonstration

## 1 Introduction

Intelligent Tutoring Systems (ITSs) are effective at improving student learning across many domains– from mathematics to experimental design [10, 13, 5]. ITSs

also employ a variety of pedagogical approaches for learning by doing, including intelligent novice [7], invention [12], and learning by teaching [9]. Many of these approaches require systems that can model complex solution spaces that accommodate multiple correct solutions to a problem and/or multiple possible paths to each solution. Further, modeling complex spaces can be desirable pedagogically: student errors during problem solving can provide valuable learning opportunities, and therefore may be desirable behaviors. Mathan and Koedingers spreadsheet tutor provides experimental support for this view– a tutor that allowed exploration of incorrect solutions led to better learning compared to one that enforced a narrower, more efficient solution path [7]. However, building tutoring systems for complex solution spaces has generally required programming. What options are available to the non-programmer? Authoring tools have radically reduced the difficulties and costs of tutor building [2, 6], and have allowed authoring without programming. Through the demonstration of examples directly in the tutor interface, an author can designate multiple correct solutions, and many correct paths to each solution. Yet, the capabilities of these tools for authoring problems with complex solution spaces has never been systematically analyzed.

In this paper, we define the concept of solution space complexity and, through a case study, explore how two authoring approaches deal with this complexity. Both approaches (Example Tracing and SimStudent) are part of the Cognitive Tutor Authoring Tools (CTAT) [1]. Our case study uses the domain of introductory experimental design, as problems in this area follow simple constraints (only vary one thing at a time), but solutions can be arbitrarily complex depending on how many variables are in the experiment and how many values each can take.

## 2 Solution Space Complexity

Solution spaces have varying degrees of complexity. Our framework for examining complexity considers both how many correct solutions satisfy a problem and how many paths lead to each solution. Within this formulation, we discuss how easily a non-programmer can author tutors that support many solutions and/or many paths to a solution.

How might this formulation of complexity apply to an experimental design tutor? Introductory problems in this domain teach the control of variables strategy (only manipulating a single variable between experimental conditions to allow for causal attribution) [3]. Due to the combinatorial nature of experiments (i.e., multiple conditions, variables, and variable values), the degree of complexity in a particular problem depends on how it is presented. To illustrate, imagine that students are asked to design an experiment to determine how increasing the heat of a burner affects the melting rate of ice in a pot (see Figure 1). The following tutor prompts (alternatives to the prompt in Figure 1) highlight how different problem framings will affect the solution complexity:

**One solution with one path** Design an experiment to determine how increasing the heat of a Bunsen burner affects the rate at which ice in a pot will

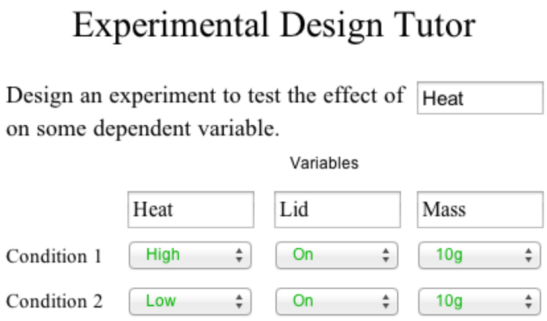



**Fig. 1.** Experimental design tutor interface

melt by assigning the first legal value to the variables in left to right, top down order as they appear in the table.

**One solution and many paths** Design an experiment to determine how increasing the heat of a Bunsen burner affects the rate at which ice in a pot will melt by assigning the first legal value to variables.

**Many solutions each with one path** Design an experiment to determine how increasing the heat of a Bunsen burner affects the rate at which ice in a pot will melt by assigning values to variables in left to right, top down order as they appear in the table.

**Many solutions with many paths** Design an experiment to determine how increasing the heat of a Bunsen burner affects the rate at which ice in a pot will melt.

While these examples show that solution space complexity can be qualitatively changed (i.e., one solution vs. many solutions) by reframing a problem, quantitative changes are also possible. For example, adding a fourth variable to the interface in Figure 1 would require two more steps per solution path (setting the variable for each condition), while adding another value to each variable increases the number of possible options at each step of the solution path. As this example illustrates, solution space complexity is not an inherent property of a domain, but rather arises from an authors design choices.

## 3 Tutor Authoring

Our analysis focuses on the Cognitive Tutor Authoring Tools (CTAT), as CTAT is the most widely used tutor authoring tool and the approaches it supports are representative of authoring tools in general [2]. CTAT supports non-programmers in building both tutor interfaces and cognitive models (for providing feedback). Cognitive models can be constructed with Example Tracing or SimStudent. In this section, we step through how Example-Tracing and SimStudent approaches would be applied by non-programmers to the experimental design task, using the

interface shown in Figure 1. Further, we discuss the features of each approach for handling solution space complexity in the context of this example.

### 3.1 Example Tracing

When building an Example-Tracing tutor in CTAT, the author demonstrates correct solutions directly in the tutoring interface. These demonstrated steps are recorded in a behavior graph. Each node in the behavior graph represents a state of the tutoring interface, and each link represents an action that moves the student from one node to another. In Example Tracing each link is produced as a result of a single action demonstrated directly in the tutor interface; many legal actions might be demonstrated for each state, creating branches in the behavior graph.

Figure 2 shows an example of our experimental design tutor interface and an associated behavior graph. The particular prompt chosen has 8 solutions and many paths to each solution. These alternative paths correspond to different orders in which the variables in the experimental design can be assigned. The Example-Tracing approach allows authors to specify that groups of actions can be executed in any order. In the context of our example, this functionality allows the author to demonstrate one path to each of the 8 unique solutions (these 8 paths are visible in Figure 2) and then specify that the actions along that path can be executed in any order. Unordered action groups are denoted in the behavior graph by colored ellipsoids.

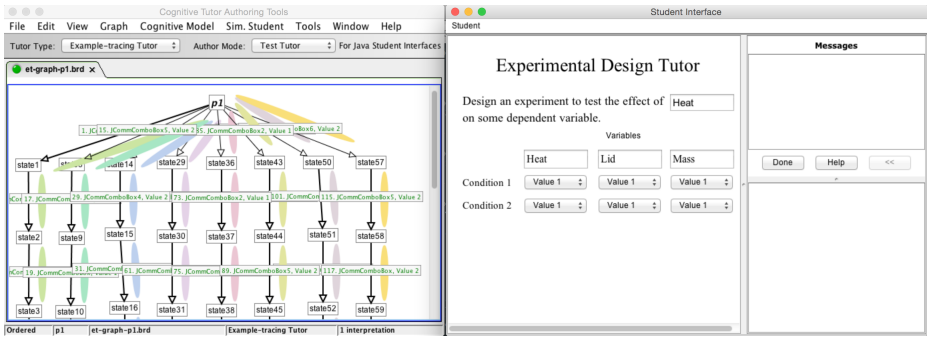



**Fig. 2.** An experimental design tutor (right) and its associated behavior graph (left). This tutor supports students in designing an experiment to test the effect of heat on a dependet variable. The correct answer is to pick two different values for the "Heat" variable and to hold the values constant for other variables.

Once a behavior graph has been constructed for a specific problem (e.g. determine the effect of heat on ice melting), that behavior graph can be generalized to other problems (e.g. determine the effect of sunlight on plant growth) using mass production. The mass production feature allows the author to replace specific values in the interface with variables and then to instantiate an arbitrary

number of behavior graphs with different values for the variables. This approach is powerful for supporting many different problems that have identical behavior graph structure, such as replacing all instances of "heat" with another variable, "sunlight". However, if a problem varies in the structure of its behavior graph, such as asking the student to manipulate a variable in the second column instead of the first (e.g., "lid" instead of "heat"), then a new behavior graph would need to be built to reflect the change in the column of interest.

How efficient is Example Tracing in building a complete cognitive model for the experimental design problem? The complete model consists of 3 behavior graphs (one for each of the three variable columns that could be manipulated). Each graph took 56 demonstrations and required 8 unordered action groups to be specified. Thus, the complete cognitive model required 168 demonstrations and 24 unordered group specifications. Using estimates from a previously developed Keystroke-Level Model [6], which approximates the time needed for an error-free expert to perform each interface action, we estimate that this model would take about 27 minutes to build using Example Tracing. Notably, the ability to specify unordered action groups offers substantial efficiency gains - without it, authoring would take almost 100 hours. Furthermore, with mass production, this model can generalize to any set of authored variables.

### 3.2 SimStudent

While the Example-Tracing behavior graph creates links from user demonstrations, the SimStudent system extends these capabilities by inducing production rule models from demonstrations and feedback (for details on this rule induction see [8]). In the experimental design tutor, SimStudent might learn a rule that sets one of the variables to an arbitrary value when no values for that variable have been assigned. Then, it might learn different rules for setting a variables second value based on whether or not it is being manipulated.

Authoring with SimStudent is similar to Example Tracing in that SimStudent asks for demonstrations when it does not know how to proceed. However, when SimStudent already has an applicable rule, it fires the rule and shows the resulting action in the tutor interface. It then asks the author for feedback on that action. If the feedback is positive, SimStudent may refine the conditions of its production rules before continuing to solve the problem. If the feedback is negative, SimStudent will try firing a different rule. When SimStudent exhausts all of its applicable rules, it asks the author to demonstrate a correct action. Figure 3 shows how SimStudent asks for demonstrations and feedback. When authoring with SimStudent, the author does not have to specify rule order - as long as a rule's conditions are satisfied, it is applicable. Authoring with SimStudent produces both a behavior graph (of the demonstrations and actions SimStudent took in the interface) and a production rule model.

To evaluate the efficiency of the SimStudent approach we constructed a complete model for the experimental design tutor. It can be difficult to determine when a SimStudent model is correct and complete from the authoring interactions alone. In most cases the SimStudent model is evaluated with set of held-out

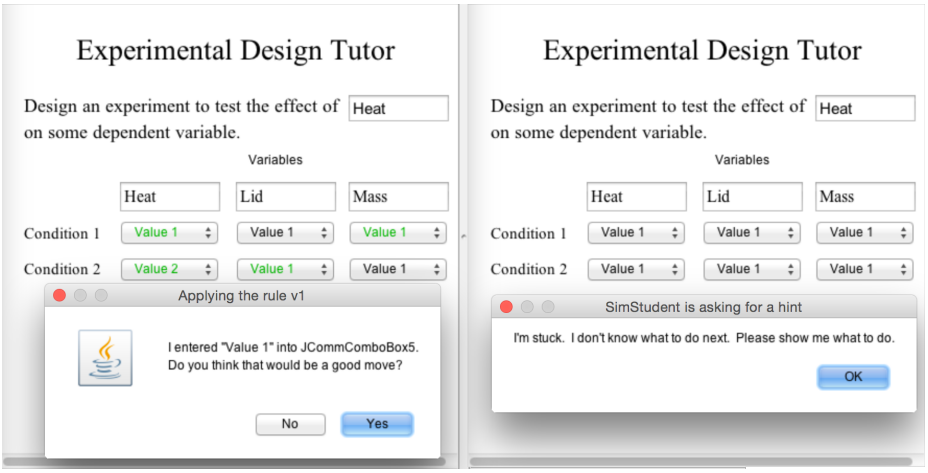



**Fig. 3.** SimStudent asking for feedback (left) and for a demonstration (right).

test problems (i.e., unit tests). However, in this case the learned rules were simple enough to evaluate by direct inspection. We noticed that SimStudent learned one correct strategy, but had not explored other solutions. This is typical of SimStudent - once it learns a particular strategy it applies it repeatedly. Therefore, authors must give it additional demonstrations of alternative paths. With the experimental design tutor, we noticed that SimStudent was always choosing the first value for non-manipulated variables, so we gave it additional demonstrations where non-manipulated variables took values besides those demonstrated on the initial run.

Ultimately, SimStudent acquired a complete model after 7 demonstrations and 23 feedback responses. Using the same Keystroke-Level Model from [6], we estimate that building a cognitive model using SimStudent would take an error-free expert about 2.12 minutes – much shorter than Example Tracing. Like Example Tracing, the model produced by SimStudent can work with arbitrary variables. Unlike Example Tracing, the learned model can work for unauthored variables; for example, students could define their own variables while using the tutor. This level of generality could be useful in inquiry-based learning environments [4]. Finally, if another variable column was added to the tutor, the SimStudent model would be able to function without modification. For Example Tracing, such a change would constitute a change to the behavior graph structure, so a completely new behavior graphs would need to be authored to support this addition.

## 4 Discussion

Both Example Tracing and SimStudent can create tutors for problems with complex solution spaces. However, our analysis shows that the two approaches

differ in terms of their efficiency and, as a result, how many solutions and paths they can handle in practice.

First, the Example-Tracing approach worked very well, even though the experimental design problems have a combinatorial structure. In particular, unordered action groups and mass production drastically reduced the number of demonstrations needed to cover the solution space, 168 vs. 40,362. The simplicity of Example Tracing combined with the power afforded by these features is likely why Example Tracing is the most widely used authoring approach today [2].

The SimStudent approach was more efficient than Example Tracing (approx. 2.12 vs. 27 minutes), but this comparison requires several caveats. The machine learning mechanisms of SimStudent generalize demonstrations and feedback into rules, which allows SimStudent to only model unique actions and the conditions under which they apply. However, this means SimStudent may not acquire a complete model. In the experimental design case study, SimStudent at first only learned that non-manipulated variables take their first value (rather than any value that is constant across conditions). In general, this problem arises when SimStudent acquires a model that can provide at least one correct solution for any problem. In these situations, it never prompts an author to provide alternative demonstrations; leading an unsuspecting author to create an incomplete model. A related complication is determining when the SimStudent model is complete. While determining the completeness of models in both Example Tracing and SimStudent can be difficult, authors must attempt to infer completeness from SimStudent's problem solving performance– a method that can be rather opaque at times. Thus, an open area for simulated learning systems is how best to evaluate the quality of learned models.

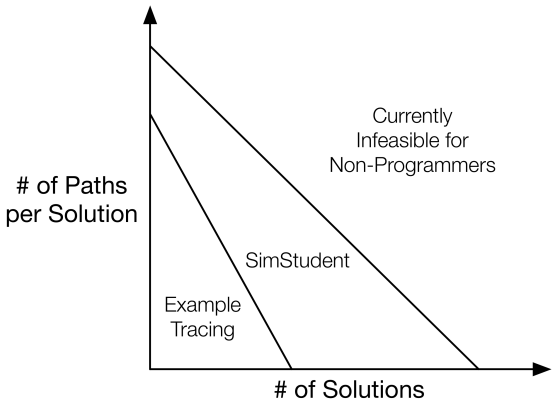



**Fig. 4.** How the space of solution space complexity is handled by existing non-programmer authoring approaches.

Overall our findings, when paired with those of previous work [6], suggest an interpretation depicted in Figure 4. In this figure the potential space of complexity is depicted in terms of number of unique solutions and number of paths per solution. The inner region denotes the area of the complexity space where we believe Example Tracing will maximize non-programmers' authoring utility. This region is skewed towards a higher number of paths, owing to Example Tracing's capacity to specify unordered actions. This portion of the complexity space contains many of the tutors that have already been built using Example Tracing [2]. As the complexity of a problem's solution space increases, Example Tracing becomes less practical (though still capable) and SimStudent becomes a more promising option, despite the caveats for using it. SimStudent's power of rule generalization gives it the ability to deal with more paths and unique solutions with less author effort, however, these capabilities come with the risk of producing incomplete models (without the author being aware).

Notably missing in the figure is any coverage of the upper right quadrant. This area would be a fruitful place to direct future work that supports non-programmers in authoring problems with many solutions with many paths. In particular, simulated learning systems might be extended to give non-programmers access to this portion of the space. One existing approach for dealing with highly complex solution spaces is to only model the aspects of the space that students are most likely to traverse. For example, work by Rivers and Koedinger [11] has explored the use of prior student solutions to seed a feedback model for introductory programming tasks. As it stands this area can only be reached using custom built approaches and would benefit from authoring tool research.

One limitation of our current approach is the assumption that there is a body of non-programmers that wants to build tutors for more complex problems. Our analysis here suggests that there is an open space for non-programming tools that support highly complex solution spaces, but it is less clear that authors have a desire to create tutors in this portion of the space. A survey of authors interested in building complex tutors without programming would help to shed light on what issues non-programmers are currently having in building their tutors. It is important that such a survey also include the perspective of those outside the normal ITS community to see if there are features preventing those who are interested from entering the space.

From a pedagogical point of view, it is unclear how much of the solution space needs to be modeled in a tutor. Waalkens et al. [16] have explored this topic by implementing three versions of an Algebra equation solving tutor, each with progressively more freedom in the number of paths that students can take to a correct solution. They found that the amount of freedom did not have an effect on students learning outcomes. However, there is evidence that the ability to use and decide between different strategies (i.e. solution paths) is linked with improved learning [14]. Further, subsequent work [15] has suggested that students only exhibit strategic variety if they are given problems that favor different strategies. Regardless of whether modeling the entire solution space is

pedagogically necessary, it is important that available tools support the ability to model complex spaces so that these research questions can be further explored.

## 5 Conclusion

The results of our analysis suggest that both the Example Tracing and SimStudent authoring approaches are promising methods for non-programmers to create tutors even for problems with many solutions with many paths. More specifically, we found that SimStudent was more efficient for authoring a tutor for experimental design, but authoring with SimStudent had a number of caveats related to ensuring that the authored model was complete. In contrast, Example Tracing was simple to use and it was clear that the authored models were complete. Overall, our analysis shows that Example Tracing is good for a wide range of problems that non-programmers might want to build tutors for (supported by its extensive use in the community [2]). However, the SimStudent approach shows great promise as an efficient authoring approach, especially when the solution space becomes complex. In any case, more research is needed to expand the frontier of non-programmers' abilities to author tutors with complex solution spaces.

Finally, this work demonstrates the feasibility and power of utilizing a simulated learning system (i.e., SimStudent) to facilitate the tutor authoring process. In particular authoring tutors with SimStudent took only 10% of the time that it took to author a tutor with Example-Tracing, a non-simulated learner approach. Educational technologies with increasingly complex solution spaces are growing in popularity (e.g. educational games and open-ended learning environments), but current approaches do not support non-programmers in authoring tutors for these technologies. Our results show that simulated learning systems are a promising tool for supporting these non-programmers. However, more work is needed to improve our understanding of how simulated learners can contribute to the authoring process and how the models learned by these systems can be evaluated.

## 6 Acknowledgements

We would like to thank Caitlin Tenison for her thoughtful comments and feedback on earlier drafts. This work was supported in part by a Graduate Training Grant awarded to Carnegie Mellon University by the Department of Education (#R305B090023) and by the Pittsburgh Science of Learning Center, which is funded by the NSF (#SBE-0836012). This work was also supported in part by National Science Foundation Awards (#DRL-0910176 and #DRL-1252440) and the Institute of Education Sciences, U.S. Department of Education (#R305A090519). All opinions expressed in this article are those of the authors and do not necessarily reflect the position of the sponsoring agency.

## References


1. Aleven, V., McLaren, B.M., Sewall, J., Koedinger, K.R.: The cognitive tutor authoring tools (CTAT): Preliminary evaluation of efficiency gains. In: Ikeda, M., Ashley, K.D., Tak-Wai, C. (eds.) ITS '06. pp. 61–70. Springer (2006)
2. Aleven, V., McLaren, B.M., Sewall, J., Koedinger, K.R.: A New Paradigm for Intelligent Tutoring Systems: Example-Tracing Tutors. IJAIED 19(2), 105–154 (2009)
3. Chen, Z., Klahr, D.: All Other Things Being Equal: Acquisition and Transfer of the Control of Variables Strategy. Child Development 70(5), 1098–1120 (1999)
4. Gobert, J.D., Koedinger, K.R.: Using Model-Tracing to Conduct Performance Assessment of Students' Inquiry Skills within a Microworld. Society for Research on Educational Effectiveness (2011)
5. Klahr, D., Triona, L.M., Williams, C.: Hands on what? The relative effectiveness of physical versus virtual materials in an engineering design project by middle school children. Journal of Research in Science Teaching 44(1), 183–203 (Jan 2007)
6. MacLellan, C.J., Koedinger, K.R., Matsuda, N.: Authoring Tutors with SimStudent: An Evaluation of Efficiency and Model Quality. In: Trausen-Matu, S., Boyer, K. (eds.) ITS '14 (2014)
7. Mathan, S.A., Koedinger, K.R.: Fostering the Intelligent Novice: Learning From Errors With Metacognitive Tutoring. Educational Psychologist 40(4), 257–265 (2005), http://www.tandfonline.com/doi/abs/10.1207/s15326985ep4004_7
8. Matsuda, N., Cohen, W.W., Koedinger, K.R.: Teaching the Teacher: Tutoring SimStudent Leads to More Effective Cognitive Tutor Authoring. IJAIED 25(1), 1–34 (2014)
9. Matsuda, N., Yarzebinski, E., Keiser, V., Cohen, W.W., Koedinger, K.R.: Learning by Teaching SimStudent – An Initial Classroom Baseline Study Comparing with Cognitive Tutor. IJAIED (2011)
10. Pane, J.F., Griffin, B.A., McCaffrey, D.F., Karam, R.: Effectiveness of Cognitive Tutor Algebra I at Scale. Tech. rep., RAND Corporation, Santa Monica, CA (2013)
11. Rivers, K., Koedinger, K.R.: Automating Hint Generation with Solution Space Path Construction. In: ITS '14, pp. 329–339. Springer (2014)
12. Roll, I., Aleven, V., Koedinger, K.R.: The Invention Lab : Using a Hybrid of Model Tracing and Constraint-Based Modeling to Offer Intelligent Support in Inquiry Environments. In: ITS '10. pp. 115–124 (2010)
13. Sao Pedro, M.A., Gobert, J.D., Heffernan, N.T., Beck, J.E.: Comparing Pedagogical Approaches for Teaching the Control of Variables Strategy. In: Taatgen, N., van Rijn, H. (eds.) CogSci '09. pp. 1–6 (2009)
14. Schneider, M., Rittle-Johnson, B., Star, J.R.: Relations among conceptual knowledge, procedural knowledge, and procedural flexibility in two samples differing in prior knowledge. Developmental Psychology 47(6), 1525–1538 (2011)
15. Tenison, C., MacLellan, C.J.: Modeling Strategy Use in an Intelligent Tutoring System: Implications for Strategic Flexibility. In: ITS '14, pp. 466–475. Springer (2014)
16. Waalkens, M., Aleven, V., Taatgen, N.: Computers & Education. Computers & Education 60(1), 159–171 (Jan 2013)

# Methods for Evaluating Simulated Learners: Examples from SimStudent

Kenneth R. Koedinger[1], Noboru Matsuda[1], Christopher J. MacLellan[1], and Elizabeth A. McLaughlin[1]

[1] Carnegie Mellon University, Pittsburgh, PA
koedinger@cmu.edu

**Abstract.** We discuss methods for evaluating simulated learners associated with four different scientific and practical goals for simulated learners. These goals are to develop a precise theory of learning, to provide a formative test of alternative instructional approaches, to automate authoring of intelligent tutoring systems, and to use as a teachable agent for students to learn by teaching. For each goal, we discuss methods for evaluating how well a simulated learner achieves that goal. We use SimStudent, a simulated learner theory and software architecture, to illustrate these evaluation methods. We describe, for example, how SimStudent has been evaluated as a theory of student learning by comparing, across four domains, the cognitive models it learns to the hand-authored models. The SimStudent-acquired models generally yield more accurate predictions of student data. We suggest future research into directly evaluating simulated learner predictions of the process of student learning.

**Keywords:** cognitive models, learning theory, instructional theory

## 1 Introduction

When is a simulated learner a success? We discuss different approaches to evaluating simulated learners (SLs). Some of these evaluation approaches are technical in nature, whether or how well a technical goal has been achieved, and some are empirical, whereby predictions from the SL are compared against data. These approaches can be framed with respect to four goals for developing SLs (see Table 1). These goals have been pursued in prior SL research, such as the use of "pseudo-students" [1] to test the quality of an instructional design (#2 in Table 1). Before describing different evaluation approaches appropriate for different goals, we first introduce SimStudent.

### 1.1 SimStudent: A Simulated Learner Theory and Software Architecture

SimStudent [2,3] is an SL system and theory in the class of adaptive production systems as defined by [4]. As such, it is similar to cognitive architectures such as ACT-R [5], Soar [6], and Icarus [7], however, it distinctly focuses on modeling inductive knowledge-level learning [8] of complex academic skills learning. SimStudent learns

from a few primary forms of instruction, including examples of correct actions, skill labels on similar actions, clues for what information in the interface to focus on to infer a next action, and finally yes-or-no feedback on actions performed by SimStudent.

**Table 1.** Scientific and Practical Goals for Simulated Learners (SLs)

1. *Precise Theory*. Use SLs to develop and articulate precise theory of learning.
   a. *Cognitive Model*. Create theories of domain expertise
   b. *Error Model*. Create theories of student domain misconceptions
   c. *Prior Knowledge*. Create theories of how prior knowledge changes learning
   d. *Learning Process*. Create theories of change in knowledge and performance
2. *Instructional Testing*. Use SLs as a "crash test" to evaluate instruction
3. *Automated Authoring*. Use SLs to automatically an intelligent tutoring system
4. *Teachable Agent*. Use SLs as a teachable agent or peer

To tutor SimStudent, a problem is entered in the tutoring interface (e.g., 2x = 8 in row 1 of Figure 1). SimStudent attempts to solve the problem by applying productions learned so far. If an applicable production is found, it is fired and problem interface is updated. The author then provides correctness *feedback* on SimStudent's step. If no correct production application is found, SimStudent asks the author to demonstrate the next step directly in the interface. When providing a demonstration, the author first specifies the *focus of attention* (i.e. input fields relevant to the current step) by double-clicking the corresponding interface elements (e.g., the cells containing 2x and 8 in Figure 1). The author takes action using the relevant information (e.g., entering divide 2 in Figure 1). Finally, the *author specifies a skill name* by clicking on the newly added edge of the behavior graph. This skill label is used to help guide SimStudent's learning and to make production rule names more readable.

SimStudent uses three machine-learning mechanisms (*how*, *where*, and *when*) to acquire production rules. When given a new demonstration (i.e., a positive example of a rule), SimStudent uses its *how* learner to produce a general composition of functions that replicate the demonstrated steps and ones like it. For example, in Figure 1, when given the demonstration "divide 2*"* for the problem 2x=8, SimStudent induces that the result of the "get-first-integer-without-sign" function when applied to left side of the problem and appended to the word "divide" explains the demonstration.

After an action sequence has been discovered, SimStudent uses its *where* learner to identify a generalized path to the focus of attention in the tutor interface. In Figure 1, the *where* learner discovers retrieval paths for the three cells in the first column. These paths are generalized as more positive examples and are acquired for a given rule. For example, when the author demonstrates the application of the divide rule shown in Figure 1 to the second row of the equation table, then the production retrieval path is generalized to work over any row in the equation table.

Finally, after learning an action sequence and general paths to relevant information, SimStudent uses its *when* learner to identify the conditions under which the learned production rule produces correct actions. For example, in Figure 1 SimStudent learns that this rule can only be correctly applied when one side of the equation

has a coefficient. In situations when SimStudent receives positive and negative feedback on its rule applications, it uses the *when* learner to update the conditions on the rules. Note, the *how* and *where* learners primarily use positive examples.

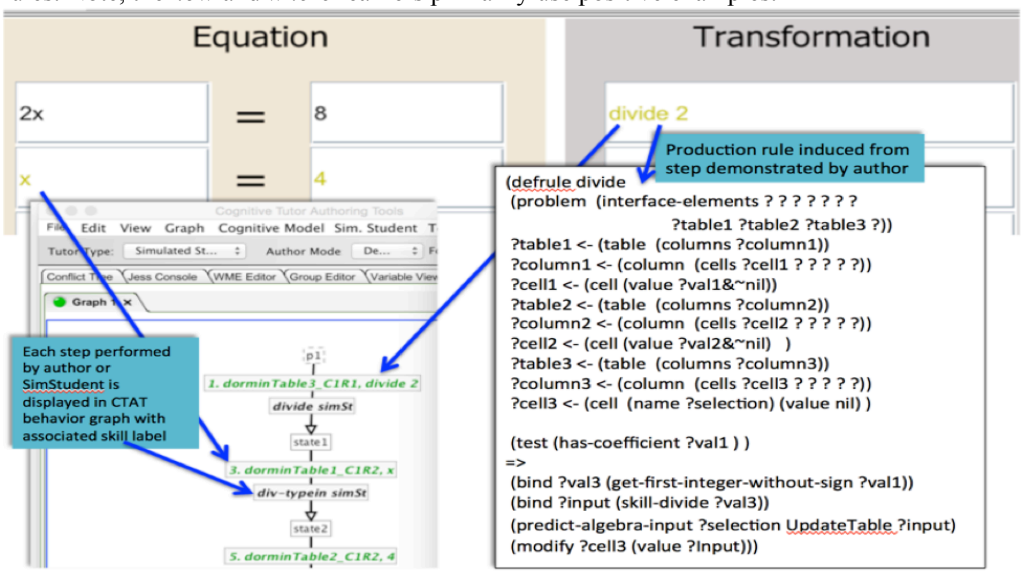



**Fig. 1.** After entering a problem, "2x=8" (top left), teaching of SimStudent occurs either by giving yes-or-no feedback when SimStudent attempts a step or by demonstrating a correct step when SimStudent cannot (e.g., "divide 2").

SimStudent is also capable of learning the representation of the chunks that make up the production system's working memory and are the informational basis on which productions are learned. It does so using an unsupervised grammar induction approach [3]. This feature sets it apart from other production rule learning systems.

## 2 Evaluating Simulated Learners as Theories of Learning

It is helpful to distinguish a general theory of learning from a theory of *student* learning. We focus on student learning because of the goals of AI in Education. However, it is worth mentioning evaluation criteria for a general learning theory, such as how quickly and independently learning takes place and how general and accurate is resulting performance. These criteria facilitate comparative evaluations. For instance, hierarchical Bayesian models are arguably better models of learning than other classification or neural network models because they learn as well with fewer examples [9].

### 2.1 Good Learning Theory Should Generate Accurate Cognitive Models

A student learning theory should produce the kind of expertise that human students acquire. In other words, the result of teaching an SL should be a cognitive model of

what human student's know after instruction. Thus, one way to evaluate an SL is to evaluate the quality of the cognitive models it produces. We proposed [10] six constraints to evaluate the quality of a cognitive model: 1) solution sufficiency, 2) step sufficiency, 3) choice matching, 4) computational parsimony, 5) acquirability, and 6) transfer. The first two are empirical and qualitative: Is the cognitive model that the SL acquires able to solve tasks and do so with steps that are consistent with human students? The third is quantitative: Does the frequency of strategy use and common error categories generated by the cognitive model on different tasks correspond with the same frequencies exhibited by human students? The last three are rational in character, involving inspection of the cognitive model.

These constraints were designed with hand-authored models in mind, so some, like the acquirability constraint (#5), appear trivial in the SL context. There is no question that the components of an SL-produced cognitive model are plausibly acquired because the SL does, in fact, acquire them. Similarly, the solution sufficiency constraint (#1) is straightforwardly achieved if the SL does not indeed succeed in learning the task domain. If the cognitive model that is produced solves problems using the kinds of intermediate steps used in student solutions, for example, it performs its solution in a step-based tutoring system interface, then the step sufficiency constraint (#2) is met.

How, then, can the remaining constraints be evaluated? In [11], we employed an educational data mining approach that evaluates the accuracy of a cognitive model by a "smooth learning curve" criteria [cf., 12,13]. Using a relatively simple statistical model of how instructional opportunities improve the accuracy of knowledge, this approach can measure and compare cognitive models in terms of their accuracy in predicting learning curve data. To employ the statistical model fit, the cognitive model is simplified into a "Q matrix", which maps each observed task performed (e.g., entering a step in a problem solving) to the knowledge components hypothesized to be needed to successfully perform that task. For any appropriate dataset uploaded into DataShop (learnlab.org/DataShop), the website allows users to edit and upload alternative cognitive models (in the Q matrix format), automatically performs statistical model fits, renders learning curve visualizations, and displays a rank ordering of the models in terms of their predictive accuracy [14].

We used this approach to evaluate the empirical accuracy of the cognitive models that SimStudent learns as compared to hand-authored cognitive models [11]. SimStudent was tutored in four domains: algebra, fractions, chemistry, and English grammar, in which we had existing human data and existing hand-authored cognitive models. In each domain SimStudent induced, from examples and from practice with feedback, both new chunk structures to represent the organization (or "grammar") of the perceptual input and new production rules that solve problems (e.g., add two fractions) or make decisions (e.g., select when to use "the" or "a" in English sentences). In each case, the production rules that SimStudent acquired were converted into the Q matrix format. Then the DataShop cognitive model comparison was employed to compare whether these models fit student learning curve data better than the hand-authored cognitive models do.

In all four domains, the SimStudent-acquired cognitive models made distinctions not present in the hand-authored models (e.g., it had two different production rules

across tasks for which the hand-authored model had one) and thus it tended to produce models with more knowledge components (as shown in Table 2). For example, SimStudent learned two different production rules for the typical last step in equation solving where one production covered typical cases (e.g., from 3x = 12 the student should "divide by 3") and another covered a perceptually distinct special case (e.g., from -x = 12 the student should divide by -1).

In all four domains, at least some of these distinctions improved the predictive fit to the learning curve data for the relevant tasks. For example, the SimStudent-acquired cognitive model in algebra leads to better accuracy because real students had a much higher error rate on tasks like -x=12 (where the coefficient, -1, is implicit) than on tasks like 3x=12 (where the coefficient, 3, is explicitly visible). In one domain (Fraction Addition,), the SimStudent-acquired cognitive model failed to make a key distinction present in the hand-authored model and thus, while better in some cases, its overall fit was worse. In the three other domains, the SimStudent-acquired cognitive models were found to be more accurate than the hand-authored cognitive models.

**Table 2.** A comparison of human-generated and SimStudent-discovered models.

| | Number of Production Rules | | Cross-Validated RMSE | |
|---|---|---|---|---|
| | Human-Generated Model | SimStudent Discovered Model | Human-Generated Model | SimStudent Discovered Model |
| Algebra | 12 | 21 | 0.4024 | **0.3999** |
| Stoichiometry | 44 | 46 | 0.3501 | **0.3488** |
| Fraction Addition | 8 | 6 | **0.3232** | 0.3343 |
| Article selection | 19 | 22 | 0.4044 | **0.4033** |

In other words, this "smooth learning curve" method of evaluation can provide evidence that an SL, SimStudent in this case, is a reasonable model of student learning in that it acquires knowledge at a grain size (as represented in the components of the cognitive model) that is demonstrably consistent with human data.

One limitation of this approach is that it *indirectly* compares an SL to human learners through the process of fitting a statistical model. In the case of algebra, for example, SimStudent's acquisition of two different productions for tasks of the form Nx=N versus tasks of the form -x=N gets translated into a prediction that student performance will be *different* in these situations, but the not direction of the difference. The parameter estimation in statistical model fit yields the prediction for which of these task categories (Nx=N or -x=N) is harder. A more *direct* comparison would not use an intermediate statistical model fit. It would require the SL to not only produce a relevant distinction, but to make a prediction of student performance differences, such as whether it takes longer to successfully learn some kinds of tasks than others. Such an evaluation approach is discussed in section 2.3.

### 2.2 Matching Student Errors and Testing Prior Knowledge Assumptions

As a model of student learning, a good SL should not only produce accurate performance with learning, but should also produce the kinds of errors that students produce [cf.,15]. Thus, comparing SL errors to student errors is another way to evaluate an SL.

One theory of student errors is that students learn incorrect knowledge (e.g., incorrect production rules or schemas) from *correct example-based instruction* due to the necessary fallibility of inductive learning processes. A further hypothesis is that inductive learning errors are more likely when students have "weak" (i.e., more domain general) rather than "strong" (i.e., more domain specific) prior knowledge. With weak prior knowledge, students may interpret examples shallowly, paying attention to more immediately perceived surface features, rather than more deeply, by making domain-relevant inferences from those surface features. Consider example-based instruction where a student is given the equation "3x+5 = 7" and told that "subtract 5" from both sides is a good next step. A novice student with weak prior knowledge might interpret this example shallowly, as subtracting a number (i.e., 5) instead of more deeply, as subtracting a term (i.e., +5). As a consequence, the student may induce knowledge that produces an error on a subsequent problem, such as "4x-2=5" where they subtract 2 from both sides. Indeed, this error is common among beginning algebra students.

We evaluated SimStudent by comparing induction errors it makes with human student errors [16]. More specifically, we evaluated the weak prior knowledge hypothesis expressed above. We conducted a simulation study by having multiple instances of SimStudent get trained by the Algebra Cognitive Tutor. We compared SimStudent behaviors with actual student data from the Cognitive Tutor's logs of student interactions with the system. When SimStudent starts with weak prior knowledge rather than strong prior knowledge, it learns more slowly, that is, the accuracy of learned skills is lower given the same amount of training. More importantly, SimStudent's ability to predict student errors increased significantly when given weak rather than strong prior knowledge. In fact, the errors generated by SimStudent with strong prior knowledge were almost never the same kinds of errors commonly made by real students.

In addition to illustrating how an SL can be evaluated by comparing its error generation to human errors, this example illustrates how an SL can be used to test assumptions about student prior knowledge. In particular, SimStudent provides a theoretical explanation of empirical results [17] showing correlations between tasks measuring prior knowledge (e.g., identify the negative terms in "3x-4 = -5-2x") and subsequent learning of target skills (e.g., solving algebra equations).

Some previous studies of students' errors focus primarily on a descriptive theory to explain why students made particular errors, for example, repair theory [15], the theory of bugs [18], and the theory of extrapolation technique [19]. With SLs, we can better understand the process of acquiring the incorrect skills that generate errors. The precise understanding that computational modeling facilitates provides us with insights into designing better learning environments that mitigate error formation.

### 2.3 Good Student Learning Theory Should Match Learning Process Data

Matching an SL's performance to learning process data is similar to the cognitive model evaluation discussed above in section 2.1. However, as indicated above, that approach has the limitation of being an indirect comparison with human data whereby there the fit to human data is, in a key sense, less challenging because it is mediated by a separate step parameter estimation of a statistical model. A more direct compari-

son is, in simple terms, to match the behavior of multiple instances of an SL (i.e., a whole simulated class) with the behavior of multiple students. The SLs interact with a tutoring system (like one shown in Figure 2) just as a class of human students would and their behavior is logged just as human student data is. Then the simulated and human student data logs can be compared, for example, by comparing learning curves that average across all (simulated and human) student participants.

## 3 Evaluating Simulated Learners as Instruction Testers

A number of projects have explored the use of an SL to compare different forms of instruction. VanLehn was perhaps the first to suggest such a use of a "pseudo student" [1]. A version of ACT-R's utility learning mechanism was used to show that the SL was more successful when given error feedback not only on target performance tasks (e.g., solving two-step equations), but also on shorter subtasks (e.g., one-step equations) [10]. A SimStudent study showed better learning from a combination of examples and problems to solve, than just giving it examples [2]. Another showed that interleaving problem types is better for learning than blocking problem types because interleaving provides better opportunities correcting over-generalization errors [20].

For a general theory of instruction, it is of scientific interest to understand the effectiveness of different forms of instruction for different kinds of SL systems even if the SL is not an accurate model of student learning. Such understanding is relevant to advancing applications of AI and is directly relevant to using an SL for automated ITS authoring (next section). Such theoretical demonstrations may also have relevance to a theory of *human* instruction as they may 1) provide theoretical explanations for instructional improvements that have been demonstrated with human learners or 2) generate predictions for what may work with human students.

These instructional conclusions can only be reliably extended to human learners when the SL is an accurate model of student learning. The most reliable evaluation of an SL as instructional tester is a follow-up random assignment experiment with human learners that demonstrates that the instructional form that was better for the SLs is also better for students. In the examples given above, there is some evidence that the SLs are accurate models of student learning (e.g., past relevant human experiments). However, in none of them was the ideal follow-up experiment performed.

## 4 Evaluating Simulated Learners as ITS Authoring Tools

In addition to their use as theories of learning and for testing instructional content, simulated learning systems can also be used to facilitate the authoring of Intelligent Tutoring Systems (ITS). In particular, once an SL has been sufficiently trained, the cognitive model it learns can then be used directly as an expert model. Previous work, such as Example Tracing tutor authoring [21], has explored how models can be acquired by demonstration. However, by using a simulated learning system to induce general rules form the demonstrations more general models can be acquired more efficiently. For example, the use of SimStudent as authoring tool is still experimental,

but there is evidence that it may accelerate the authoring process and produce more accurate cognitive models than hand authoring. One demonstration explored the benefits of a traditional programming by demonstration approach to authoring in SimStudent versus a programming by tutoring approach [2]. In the latter, SimStudent asks for demonstrations only at steps where it has no relevant productions. Otherwise, it performs a step and asks the author for feedback as to whether the step is correct or not. Programming by tutoring was found to be much faster than programming by demonstration (77 minutes vs. 238 minutes) and produced a more accurate cognitive model whereby there were fewer productions that produced over-generalization errors. Programming by tutoring is now the standard approach because of its improved efficiency and effectiveness. Better efficiency is obtained because many author demonstrations are replaced by SimStudent actions with a quick yes-or-no response. Better effectiveness is obtained because these actions expose over-generalization errors to which the author responds "no" and the system learns new if-part preconditions to more appropriately narrow the generality of the modified production rule.

A second demonstration of SimStudent as an authoring tool [22] compared authoring in SimStudent with authoring example-tracing tutors in CTAT. Tutoring SimStudent has considerable similarity with creating an example-tracing tutor except that SimStudent starts to perform actions for the author, which can be merely checked as desirable or not, saving the time it otherwise takes for an author to perform those demonstrations. This study reported a potential savings of 43% in authoring time.

## 5 Evaluating a Simulated Learner as a Teachable Agent

Simulated learner systems can be more directly involved in helping students learn when they are used as a teachable agent whereby students learn by teaching [cf., 23]. Evaluating the use of an SL in this form ideally involves multiple steps. One should start with an SL that has already received some positive evaluation as a good model of student learning (see section 2). Then incorporate it into a teachable agent architecture and, as early and often as possible, perform pilot students with individual students [cf., 24 on think aloud user studies) and revise the system design. Finally, for both formative and summative reasons, use random assignment experiments to compare student learning from the teachable agent with reasonable alternatives.

Using SimStudent, we built a teachable agent environment, called APLUS, in which students learn to solve linear equations by teaching SimStudent [25]. To evaluate the effectiveness of APLUS and advance the theory of learning by teaching, we conducted multiple *in vivo* experiments [25,26,27,28]. Each of the classroom studies have been randomized controlled trials with two conditions varying one instructional approach. In one study [25], the self-explanation hypothesis was tested. To do so, we developed a version of APLUS in which SimStudent occasionally asked "why" questions. For example, when a student provided negative feedback to a step SimStudent performed, SimStudent asked, "Why do you think adding 3 here on both sides is incorrect?" Students were asked to respond to SimStudent's questions either by selecting pre-specified menu items or entering a free text response. The results showed that

the amount and the level of elaboration of the response had a reliable correlation with students' learning measured by online pre- and post-tests.

## 6 Conclusion

We outlined four general purposes for simulated learners (see Table 1) and reviewed methods of evaluation that align with these purposes. To evaluate an SL as a precise theory of learning, one can evaluate the cognitive model that results from learning, evaluate the accuracy of error predictions as well as prior knowledge assumptions needed to produce those errors, or evaluate the learning process, that is, the changes in student performance over time. To evaluate an SL as an instructional test, one should not only evaluate the SL's accuracy as a theory of student learning, but should also perform human experiments to determine whether the instruction that works best for SLs also works best for human students. To evaluate an SL as an automated authoring tool, one can evaluate the speed and precision of rule production, the frequency of over-generalization errors and the fit of the cognitive models it produces. More ambitiously, one can evaluate whether the resulting tutor produces as good (or better!) learning than an existing tutor. Similarly, to evaluate an SL as a Teachable Agent, one can not only evaluate the system features, but also perform experiments on whether students learn better with that system than with reasonable alternatives.

Simulated learner research is still in its infancy so most evaluation methods have not been frequently used. We know of just one such study [29] that evaluated an SL as an instructional tester by following up a predicted difference in instruction with a random assignment experiment with real students. It used an extension of the ACT-R theory of memory to simulate positive learning effects of an optimized practice schedule over an evenly spaced practice schedule. The same experiment was then run with human students and it confirmed the benefits of the optimized practice schedule. Such experiments are more feasible when the instruction involved is targeting simpler learning processes, such as memory, but will be more challenging as they target more complex learning processes, such as induction or sense making [31].

The space of instructional choices is just too large, over 200 trillion possible forms of instruction [32], for a purely empirical science of learning and instruction to succeed. We need parallel and coordinated advances in theories of learning *and* instruction. Efforts to develop and evaluate SLs are fundamental to such advancement.

## References


1. VanLehn (1991). Two pseudo-students: Applications of machine learning to formative evaluation. In Lewis & Otsuki (Eds.), *Adv Res on Comp in Ed* (pp. 17-26). Amsterdm: Els.
2. Matsuda, Cohen, Koedinger (2015). Teaching the teacher. *Int J of AI in Ed, 25*, 1-34.
3. Li, Matsuda, Cohen, Koedinger (2015). Integrating representation learning and skill learning in a human-like intelligent agent. *AI, 219,* 67-91.
4. Anzai, Y. & Simon (1979). The theory of learning by doing. *Psych Rev, 86 (2)*, 124-140.
5. Anderson, J.R., & Lebiere, C. (1998). *The Atomic Components of Thought*. Hillsdale: Erl.

6. Laird, Newell, & Rosenbloom (1987). Soar. *AI, 33*(1), 1–64.
7. Langley & Choi (2006). A unified cognitive architecture for physical agents. In *Proc of AI*.
8. Newell, Allen. 1990. *Unified Theories of Cognition*. Cambridge, MA: Harvard U. Press.
9. Tenenbaum, J. B., Griffiths, T. L., & Kemp.C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*, 309-318.
10. MacLaren & Koedinger (2002). When and why does mastery learning work: Instructional experiments with ACT-R "SimStudents". In *Proc of ITS*, 355-366. Berlin: Spr-Ver.
11. Li, Stampfer, Cohen, & Koedinger (2013). General and efficient cognitive model discovery using a simulated student. In *Proc of Cognitive Science*. (pp. 894-9) Austin, TX.
12. Martin, Mitrovic, Mathan & Koedinger (2011). Evaluating and improving adaptive educational systems with learning curves. *User Modeling and User-Adapted Int, 21*(3), 249-283.
13. Stamper, J.C. & Koedinger, K.R. (2011). Human-machine student model discovery and improvement using data. In *Proc of AI in Ed*, pp. 353-360. Berlin: Springer.
14. Koedinger, Baker, Cunningham, Skogsholm, Leber, Stamper, (2010). A Data Repository for the EDM community: The PSLC DataShop. In *Hdbk of Ed Data Min*. Boca Rat: CRC.
15. Brown, J. S., & VanLehn, K. (1980). Repair theory. *Cognitive Science, 4,* 379-426.
16. Matsuda, Lee, Cohen, & Koedinger, (2009). A computational model of how learner errors arise from weak prior knowledge. In *Proc of Cognitive Science*. pp. 1288-1293.
17. Booth, J.L., & Koedinger, K.R. (2008). Key misconceptions in algebraic problem solving. In Love, McRae & Sloutsky (Eds.), *Proc of Cognitive Science,* pp. 571-576.
18. VanLehn, K. (1982). Bugs are not enough. *Journal of Mathematical Behavior, 3(2)*, 3-71.
19. Matz, M. (1980). Towards a process model for high school algebra errors. In Sleeman & Brown (Eds.), *Intelligent Tutoring Systems* (pp. 25-50). Orlando, FL: Academic Press.
20. Li, N., Cohen, W. W., & Koedinger, K. R. (2012). Problem Order Implications for Learning Transfer. In *Proceedings of Intelligent Tutoring Systems*, 185–194.
21. Aleven, V., McLaren, B., Sewall, J., & Koedinger, K. R. (2009). Example-tracing tutors: A new paradigm for intelligent tutoring systems. *Int J of AI in Education, 19,* 105-154.
22. MacLellan, Koedinger & Matsuda (2014). Authoring Tutors with SimStudent: An Evaluation of Efficiency and Model Quality. *Proc of Intelligent Tutoring Systems,* 551-560.
23. Biswas, G., Schwartz, D., Leelawong, K., Vye, N. (2005). Learning by Teaching: A New Agent Paradigm for Educational Software. *Applied Artificial Intelligence, 19*, 363-392.
24. Gomoll, K., (1990). Some techniques for observing users. In Laurel B. (ed.), *The Art of Human-Computer Interface Design*, Addison-Wesley, Reading, MA, pp. 85-90.
25. Matsuda, Yarzebinski, Keiser, Raizada, William, Stylianides & Koedinger (2013). Cognitive anatomy of tutor learning. *J of Ed Psy, 105(4)*, 1152-1163.
26. Matsuda, Cohen, Koedinger, Keiser, Raizada, Yarzebinski, Watson, Stylianides (2012). Studying the effect of tutor learning using a teachable agent. In *Proc of DIGITEL*, 25-32.
27. Matsuda, Griger, Barbalios, Stylianides, Cohen, & Koedinger (2014). Investigating the effect of meta-cognitive scaffolding for learning by teaching. In *Proc of ITS,* 104-113.
28. Matsuda, Keiser, Raizada, Tu, Stylianides, Cohen, Koedinger (2010). Learning by Teaching SimStudent: In *Proceedings of Intelligent Tutoring Systems,* 317-326.
29. Ritter, Anderson, Koedinger, & Corbett (2007). Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin &Review, 14(2)*, 249-255.
30. Pavlik & Anderson (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied, 14*, 101-117.
31. Koedinger, Corbett, Perfetti (2012). The KLI framework. *Cog Sci, 36 (5),* 757-798.
32. Koedinger, Booth, Klahr (2013). Instructional complexity. *Science, 342*, 935-937.

# Simulated learners in peers assessment for introductory programming courses

Alexandre de Andrade Barbosa[1,3] and Evandro de Barros Costa[2,3]

[1] Federal University of Alagoas - Arapiraca Campus, Arapiraca - AL, Brazil
[2] Federal University of Alagoas - Computer Science Institute, Maceio - AL, Brazil
[3] Federal University of Campina Grande, Campina Grande - PB, Brazil

{*alexandre.barbosa@arapiraca.ufal.br*, *ebc.academico@gmail.com*}

**Abstract.** Programming is one of the basic competences in computer science, despite its importance, it is easy to find students with difficulties to understand the concepts required to use this skill. Several researchers report that the impossibility to achieve a quick and effective feedback, is one of the motivators for the problematic scenario. The professor, even when helped by the TAs, is not able to perform the reviews quickly, for this activity requires a huge amount of time. Fast feedback is extremely important to enable the learning of any concept. Some researches suggest the use of peer assessment as a means of providing feedback. However, it is quite common that the feedback provided by peers is not adequate. In this paper, we propose the use of simulated learners in a peer assessment approach as part of the teaching and learning processes of programming. Currently a software tool is being developed to include the proposal described in this paper.

## 1 Introduction

Programming is one of the basic competences in computer science, it is the basis for the development of several other competences required for professionals in the area. However, despite its importance, it is easy to find students who are demotivated and with difficulties to understand the concepts required to use this skill [7]. These difficulties causes a large number of failures, dropouts or the approval of students without the required level of knowledge [14] [6] [5].

Many factors are identified in literature as causing the problematic scenario related to programming courses. Several researchers report that the impossibility to achieve a quick, effective and individualized feedback, is one of the motivators for the problematic scenario [10] [12]. An individual follow up is impossible due to many students enrolled in the courses. In addition, there is a great complexity involved in the evaluation of a program, for it is necessary to understand how the programmer has developed the algorithm, so the professor needs to comprehend the line of reasoning adopted by the student. In this way, the professor, even when helped by the TAs, cannot provide an adequate and fast feedback about the solutions created by the students. This activity will require a huge amount

of time to manually open the code, compile, run and verify the output of every students solution for programming assignment. If the grading depends on the structure and the quality of code, in addition to program output correctness, the situation is a lot worse. Traditionally the real comprehension state of the contents of a programming course is known only months after the beginning of the course, when an evaluation activity is performed. After an evaluation it may be too late to make any intervention.

Fast feedback is of extreme importance to enable the learning of any concept [12]. Thus, some researches have been developed with the aim to propose methods and tools to facilitate the monitoring of the activities of students in programming courses. Some of these researches, such as [9][11][13], suggests the use of peer assessment as a means of providing fast and effective feedback. This solution is broadly used in Massive Open Online Courses (MOOCs), as described in [3][8], where the courses are applied to hundreds or thousands of people enrolled in them, and just as occurs in the context of programming, it is impossible for the professor to evaluate each solution. However, the peer assessment approach as a means of providing feedback has some problems. Many times the feedback provided by peers is not adequate, because the results are often not similar to the analysis of an expert [8]. It is quite common to find comments that are summarized to a phrase of congratulation or critique.

The reasons related to lack of effectiveness of feedback provided are quite distinct, these may occur due to poor understanding of the content of the activity, because of the student's low motivation, or due to the short time that one has available for the activities.

In [2] paper, it was observed the impact of learning was observed when a student is influenced by the performance of their peers, the authors describe that some students are encouraged to perform better, but others experiencing the same situations end up discouraged to perform better.

In this paper is proposed the use of simulated learners in a peer assessment approach used as part of the teaching and learning processes of programming. Two concerns are explored in this proposal: the first is related to the search of methods that enable a positive influence between students; the second concern is related to an approach that allows a less costly way of testing any proposal of applicability of peer assessment approach.

This paper is divided in five sections. In Section 2 the concept of peer assessment is presented. Observations on the implementation of peer assessment in a programming course context are shown in Section 3. The proposal of using simulated learners in the context of peer assessment for introductory programming is presented in Section 4. Finally the conclusions and future work are shown in the last section.

## 2 Peer Assessment

Peer assessment, or peer review, is an evaluation method where students have responsibilities that traditionally belong to professors only. Among these respon-

sibilities there are the review and the critique of the solutions proposed by their peers. This way, they can experience the discipline as students and also from the perspective of a TA. Usually in a peer assessment environment, students also conduct self assessment. This way, they can reflect on their solution when compared to other solutions, develop their critical thinking skills and improve understanding of the concepts covered in the course.

In traditional approach, the professor, even when helped by TAs, can not provide fast and adequate feedback for each solution proposed by the students. The comments provided by the professor are generic observations based on observation of all students solutions.

In accordance with [9], peer review is a powerful pedagogical method, because once students need to evaluate the work of their peers, they begin to teach and learn from each other. Thus, the learning process becomes much more active, making the learning qualitatively better than the traditional approach. Sstudents can spend more time on analysis and construction of their comments, creating more particular descriptions on a given solution and enriching discussion about the topic studied.

Thus, the use of peer review can reduce the workload on the professor, permitting the professor to focus on other pedagogical activities [9][3]. This evaluation approach can also enable the evaluation of large-scale complex exercises, which can not be evaluated in a automatically or semi-automatic fashion [3][8].

The success of peer assessment approach is strongly influenced by the quality of feedback provided. However, this feedback if often not adequate, the results are often not similar to the analysis of an expert [8]. In [8] is described that in many cases the evaluations of the students are similar to the TAs evaluation, however there are situations where the evaluations are graded 10% higher than the TAs evaluation, in extreme cases the grades could be 70% higher than the TAs evaluation. In [3] is mentioned that in general, there is a high correlation between the grades provided by students and TAs, but often in the evaluations from students the grades are 7% higher than the grades given by TAs.

Thus, we can conclude that peer assessment approach is a promising evaluation method, however there are improvements and adjusts to be applied to obtain richer discussions and more accurate assessments.

## 3 Peer Assessment in introductory programing courses

Human interaction is described as an essential feature for learning in many domains, including the introductory programming learning [13]. In classroom programming courses the contact between students occurs on a daily basis, allowing, for example, the discussion of the problems presented in the exercise lists, the developed solutions and the formation of groups for the projects of the course. This contact is many times inexistent in online programming courses, interactions in this environment are the human-machine type. Thus, using the peer assessment approach may enable human interaction on online courses, or enhance the interaction between humans in presencial classroom courses.

To encourage the assimilation of the topics, the use of practical exercises is quite common in programming courses, the practice of programming skills is crucial for learning. Many researchers also argue that the programming learning involves the reading and understanding of third-party code. Through peer assessment approach both characteristics can be obtained. The professor can develop new exercises, or choose problems proposed by others, while students will have to observe, understand and evaluate the codes of their peers, as well to compare these codes with their solution.

In [11] the use of a peer assessment approach to the context of programming courses is described, this approach is supported by a web application. The results described on the paper have a high correlation between the evaluations of the TAs and students, the correlation is lowest when the complexity of the exercise is higher.

An approach of peer assessment evaluation for the context of programming learning, also supported by a web application is presented in [9]. Five activities where graded using peer assessment, the occurrence of conflicts ranged from 61 % to activity with a lower incidence of conflict, up to 80 % for the activity with the highest occurrence of conflicts. The system considers that a conflict occurs when the student does not agree with the assessment provided.

In [9] the authors describes that if the peer reviews are conducted in an inadequate way, the failure rates can increase. For the teaching approach used at the programming course described in [13] there are two types of activities that require assessment, quizzes and mini projects. Among these activities only the mini projects are evaluated through peer assessment. Thus, students are not overloaded and the approach can be used appropriately.

Another problem that can emerge with the use of peer review in a programming context, is the increase of plagiarism. Once the assessment activity will be distributed among the students, the similarities of self-identification of codes can become more complicated. However, solutions are widely used to carry out the detection automatically similarities, such as MOSS [1] e GPLAG [4].

## 4 Simulated learners as peers in a peer assessment environment for introductory programing courses

In previous sections the advantages and disadvantages associated with the use of peer assessment in a general context, and when applied to the context of programming courses have been described. In both cases, the success of the approach is strongly influenced by the quality of the feedback given. Therefore, it is necessary to identify situations where there is inadequate feedback as well as conflict situations. Situations where inadequate feedback occurs are when, for any reason, the feedback does not help in the learning process. Conflict situations occur when the student does not agree with the assessment provided, or when there are huge variations on the evaluations provided. To perform a validation of this proposal or of any proposal involving peer assessment, it is necessary to

allocate the resources of time, physical space and adequate human resources. Thus, it can be said that the test of this approach is a costly activity.

Two concerns are explored in this proposal: how we can achieve methods that enable a positive influence between students in peer assessment environments, in other words, how a student can give a high quality feedback to their peers; and how a peer assessment approach can be tested with a lower cost, since any validation of these assessment approaches requires a huge amount of resources.

### 4.1 A scenario of use of peer assessment with simulated learners

Traditionally in a peer assessment environment, the professor must create the assignment and a set of assessment criteria. Then students develop their solutions observing the assessment criteria and submitting the solution to be evaluated by their peers. Each student's evaluation must meet the assessment criteria. The students should provide comments to peers explaining the reasons associated to the outcome and a grade or an evaluation concept (eg. A-, B+, C). Each student will have their code evaluated by their peers, and should assess the codes of other students. In Figure 1, it is illustrated the scenario previously described. There are variations in ways peer assessment approach is used, the scenario just mentioned has many characteristics which are similar to all the variations.

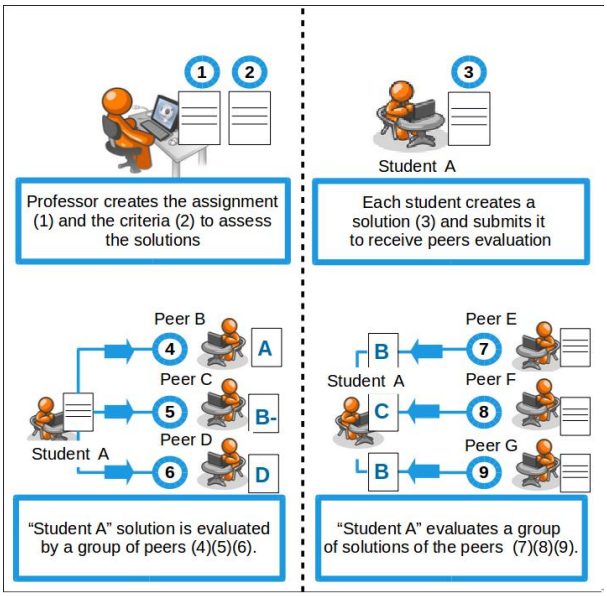



**Fig. 1.** A traditional peer assessment environment

In any peer assessment approach, it is possible to adopt pairing algorithms. Thereby, it is assured that evaluations are conducted by students with different

levels of knowledge. A student with low understanding of the subject will not be allocated to evaluate the work of another student in the same situation. Students with difficulties can clarify their doubts, while students with good understanding of the content should provide a good argumentation about their knowledge. However, it is not possible to ensure that a student evaluates the code that is the ideal for his/her learning and level of knowledge. As an example, in Figure 1, it is not possible to know if student "A" code is the best for peers "B", "C" and "D".

When a student does not agree with the evaluation provided by their peers, he/she will be able to request the intervention of the professor. This conflict situations are identified in [9]. However, in traditional peer assessment approach is not possible to identify incorrect evaluations provided by a student, or students that create biased evaluations only to help their fellows. As an example, in Figure 1, it is possible to see that different grades were given, but it is not possible to determine if the correct evaluations were given by peer "B", "C" or "D".

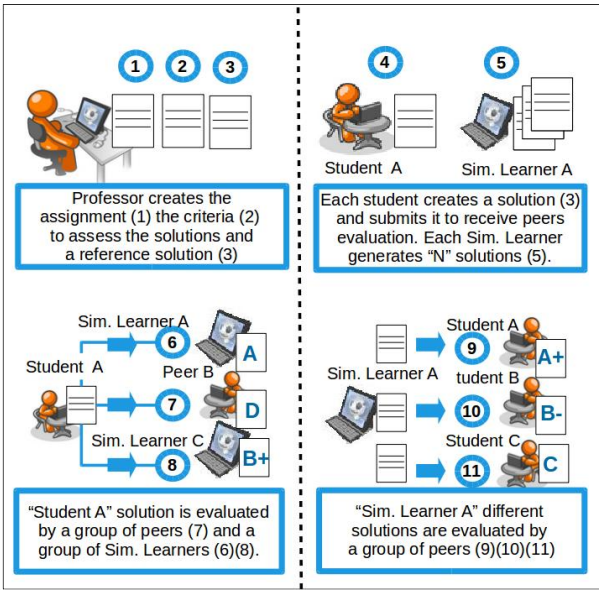



**Fig. 2.** A peer assessment environment using simulated learners

In a peer assessment environment that uses simulated learners, it is possible to solve the previous problems. As in traditional approach, the professor must create the assignment and a set of assessment criteria; in addition to that, he/she should provide a reference solution. Then, students develop their solutions, observing the assessment criteria and submitting the solution to be evaluated by their peers. At the same time, once a pairing algorithm can perform pairing of

the evaluators, each simulated learner must generate a code that is ideal for the learning and appropriate to the level of knowledge of each one of their peers, in this case, real students. Each student will have their code evaluated by their peers and by simulated students, and they should assess codes of other students, and codes of simulated students. In Figure 2 it is illustrated peer assessment environment with simulated learners. As an example, in Figure 2, it is possible to see that the simulated learner "A" generates a set of codes that are ideal for each student: "A", "B" and "C".

The identification of incorrect evaluations provided by a student, as well as students, who perform biased evaluations, could be carried out through the comparison of student's evaluations and the simulated student's evaluations. As an example, in Figure 2, it is possible to see that student "B", made an evaluation that is very different from the simulated learner's evaluations. In this way, it is possible to verify if the student did not understand the solution, or if the evaluation was created to help their fellows only.

**Providing useful solutions to student learning** A useful solution for the student learning does not always match the presentation of a correct and efficient code. Within the context of peer review may be more useful to display an incorrect code, as a way to make students to provide a set of review observations. To identify which type of code is best for a student; simulated learners can consult the representation of their cognitive status. In that way, it will be possible to the simulated learner identify the student misconceptions and errors in previous assignments, and generate variations of the reference solution that suits best for the student. Since multiple simulated students will be used, the codes that will be shown to students can range from efficient, correct and complete solutions to incorrect and/or incomplete solutions. Like that, it will be possible to check if students have different skills related to the content. To generate the variations from the reference solution, it is possible to combine testing techniques, such as mutant generation. Each code can be generated through the use of data related to the most common student's mistakes, emulating these behaviors and creating codes that are useful to learning. Once the research is in a preliminary stage, it is still not clear which artificial intelligence approaches should be used on the implementation of simulated students behaviors.

**Assessment of students solutions** Unlike what occurs in other contexts, for programming the evaluation of a solution can be automated or semi-automated. Typically a set of unit tests is applied to the code proposed by a student, who receives an indication that his/her code may be correct or incorrect, but no hint or comment is provided. Some researchers have investigated the use of different techniques to help assessment of codes and provide some guidance; these techniques usually employ software engineering metrics. Thus, simulated learners must be able to identify which subset of metrics can be used to perform the evaluation of the proposed solution for a student. The simulated learner should select the set of metrics that fits best to the objectives of the assignment and

the level of understanding that the student has at that moment. For each level of learning the same student can learn better if the set of metrics is properly selected. Each simulated student will use different strategies to evaluate the solutions provided by real students. Therefore, a variation between evaluations of simulated students is expected to occur. If an evaluation provided by a student has a very large variation in relation to the set of evaluations of simulated students, it will be necessary to investigate the motivation of this disparity. An acceptable variation threshold can be used to identify incorrect evaluations provided by students.

**Discussing assessment criterias** Once software engineering metrics were used in the evaluation, the explanation given by the simulated learner throughout the presentation of a set of metrics, is associated to the explanation of the metric choice and, possibly, of the snippet of the code where the observation is pertinent. Thereby, the simulated learner can help the professor to identify inadequate feedback, whenever an evaluation of a student is very different from the evaluation of a simulated learner, the professor and his tutors can then intervene.

### 4.2 Validation of peer assessent using simulated learners

Any validation of peer assessment approaches requires lots of physical space and a huge amount of human resources. As an example, if a validation of a pairing algorithm has to be done, it will be necessary to use a set of N students; this set must allow the creation of different profiles for evaluation of the pairing alternatives. The greater the possibilities of matching, the greater the amount of students required. Through the use of simulated learners any operational proposal of peer assessment can be tested at a much lower cost, since the physical space and human resources are drastically reduced. The researcher can determine how much of human resource will be available, replacing the students with simulated students. The researcher can also specify the desired behavior of students; the simulated students should emulate students with a high degree of understanding of the contents or with low understanding. After obtaining initial results with the use of simulated learners, the number of human individuals participating in an experiment can be increased, since it may be interesting to obtain a greater statistical power associated with the conclusions.

## 5 Conclusions and further work

In this paper, we have proposed the use of simulated learners in a peer assessment approach adopted as a support part of a programming course. The use of simulated learners as presented in this proposal aims to two goals: influence the students to provide better quality feedback; and allow for a less costly validation for peer assessment applied to programming contexts.

The research associated with the proposal presented in this paper is in a preliminary stage. Thus, the effectiveness of this proposal will be further evaluated in controlled experiments executed in the future. An open source software tool is being developed to include all aspects described throughout this proposal.

## References


1. Bowyer, K., Hall, L.: Experience using "moss" to detect cheating on programming assignments. In: Frontiers in Education Conference, 1999. FIE '99. 29th Annual. vol. 3, pp. 13B3/18–13B3/22 (Nov 1999)
2. Frost, S., McCalla, G.I.: Exploring through simulation the effects of peer impact on learning. In: Proc. of the Workshops at the 16th International Conference on Artificial Intelligence in Education AIED 2013, Memphis, USA, July 9-13, 2013 (2013), http://ceur-ws.org/Vol-1009/0403.pdf
3. Kulkarni, C., Wei, K.P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., Klemmer, S.R.: Peer and self assessment in massive online classes. ACM Trans. Comput. Hum. Interact. 20(6), 33:1–33:31 (Dec 2013)
4. Liu, C., Chen, C., Han, J., Yu, P.S.: Gplag: Detection of software plagiarism by program dependence graph analysis. In: Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 872–881. KDD '06, ACM, New York, USA (2006)
5. Mason, R., Cooper, G.: Introductory programming courses in australia and new zealand in 2013 - trends and reasons. In: Proc. of the Sixteenth Australasian Computing Education Conference - Volume 148. pp. 139–147. ACE, Darlinghurst, Australia (2014)
6. Mason, R., Cooper, G., de Raadt, M.: Trends in introductory programming courses in australian universities: Languages, environments and pedagogy. In: Proc. of the Fourteenth Australasian Computing Education Conference - Volume 123. pp. 33–42. ACE, Darlinghurst, Australia (2012)
7. McCracken, M., Almstrum, V., Diaz, D., Guzdial, M., Hagan, D., Kolikant, Y.B.D., Laxer, C., Thomas, L., Utting, I., Wilusz, T.: A multi-national, multi-institutional study of assessment of programming skills of first-year cs students. In: Working Group Reports from ITiCSE on Innovation and Technology in Computer Science Education. pp. 125–180. ITiCSE-WGR, New York, USA (2001)
8. Piech, C., Huang, J., Chen, Z., Do, C.B., Ng, A.Y., Koller, D.: Tuned models of peer assessment in moocs. CoRR abs/1307.2579 (2013)
9. de Raadt, M., Lai, D., Watson, R.: An evaluation of electronic individual peer assessment in an introductory programming course. In: Proc. of the Seventh Baltic Sea Conference on Computing Education Research - Volume 88. pp. 53–64. Koli Calling, Darlinghurst, Australia (2007)
10. Singh, R., Gulwani, S., Solar-Lezama, A.: Automated feedback generation for introductory programming assignments. In: Proc. of the 34th ACM SIGPLAN Conference on Programming Language Design and Implementation. pp. 15–26. PLDI, New York, USA (2013)
11. Sitthiworachart, J., Joy, M.: Computer support of effective peer assessment in an undergraduate programming class. Journal of Computer Assisted Learning 24(3), 217–231 (2008)
12. Stegeman, M., Barendsen, E., Smetsers, S.: Towards an empirically validated model for assessment of code quality. In: Proc. of the 14th Koli Calling Int. Conf. on Computing Education Research. pp. 99–108. Koli Calling, New York, USA (2014)

13. Warren, J., Rixner, S., Greiner, J., Wong, S.: Facilitating human interaction in an online programming course. In: Proc. of the 45th ACM Technical Symposium on Computer Science Education. pp. 665–670. SIGCSE, New York, USA (2014)
14. Yadin, A.: Reducing the dropout rate in an introductory programming course. ACM Inroads 2(4), 71–76 (2011)

# Simulated Learners for Testing Agile Teaming in Social Educational Games

Steeve Laberge and Fuhua Lin

School of Computing and Information Systems, Athabasca University, Edmonton, Canada
slaberge@acm.org, oscarl@athabascau.ca

**Abstract.** This paper proposes an approach for creating and testing an multiagent systems based adaptive social educational game (SEG), QuizMASter, using the concept of simulated learners to overcome experimentation complexity and unpredictable student availability, as is typical with online learning environments. We show that simulated learners can play two roles. First, it can be used for testing the game planning, scheduling and adaptive assessment algorithms. With some degree of success met with our initial experimentation with QuizMASter, advanced planning and coordination algorithms are now needed to allow the game-based assessment platform to realize its full potential. The multi-agent system approach is suitable for modeling and developing adaptive behaviour in SEGs. However, as we have found with our early prototypes, verifying and validating such a system is very difficult in an online context where students are not always available. MAS-based assessment game planning and coordination algorithms are complex and thus need simulated learners for testing purposes. Second, to overcome unpredictable student availability, we modeled QuizMASter as a new class of socio-technical system, human-agent collective (HAC). In the system, human learners and simulated learners (smart software agents) engage in flexible relationship in order to achieve both their individual and collective goals, while simulated learners are selected for serving as virtual team members.

**Keywords:** social educational agents, multiagent systems, simulated learners

## 1 Introduction

For decades, educational games have proven to be an effective means to motivate learners and enhance learning. Social (multi-player) educational games (SEGs) offer many opportunities to improve learning in ways that go beyond what a single-player game can achieve because SEGs allow players to be social, competitive, and collaborative in their problem solving. The presence of other players can be used to increase playability and to help teach team-work and social skills. SEGs promote intragroup cooperation and intergroup competition [1]. However, existing SEGs share many of the shortcomings of classroom role-playing. Setting

up existing SEGs is logistically challenging, expensive, and inflexible. Furthermore, players become bored after going through existing SEGs once or twice.

To test such a social educational game, we face two difficulties. One is how to test the planning and scheduling algorithms. Another is how to meet the need of agile team formation. In SEGs, group formation has big impact on group learning performance. Poor group formation in social games can result to homogeneity in student characteristic such that the peer learning is ineffective. Thus, there is a need to constitute a heterogeneous group SEGs that constitutes students with different collaborative competencies and knowledge levels. However, without empirical study it becomes difficult to conclude which group characteristics are desirable in the heterogeneity as different game-based learning needs may require different group orientations. Previous research has focused on various group orientation techniques and their impact on group performance like different learning styles in group orientation [2–4]. However, there is need to investigate the impact of other group orientation techniques on group performance like grouping students based on their collaboration competence levels. Furthermore, most of the previous research in group-formation focuses on classroom based learning. Also, it lacks the true experiment design methodology that is recommended when investigating learning outcomes from different game-based learning strategies. Simulated learners methodology [5] has shown a promising way to solve these challenges.

In this paper, we show that simulated learners can play two roles. First, it can be used for testing the game planning, scheduling and adaptive assessment algorithms. Second, working with human learners and forming human-agent collectives (HAC), simulated learners serve as virtual team members to enable asynchronous game-based learning in a context where student availability is unpredictable. This paper is structured as follows: In Section 2 we discuss recent advancements and related work. Section 3 describes QuizMASter. Section 4 presents the proposed architecture for development of QuizMASter. Section 5 explains how we intend to use simulated learners for testing QuizMASter. Finally, Section 6 concludes.

## 2 Related Work

Researchers have found that learning can be more attractive if learning experiences combine challenge and fun [6]. As social networks have become popular applications, they have given rise to social games. This kind of game is played by users of social networks as a way to interact with friends [7] and has become a part of the culture for digital natives. Social games have unique features that distinguish them from other video games. Those features are closely linked with the features of social networks [8]. Social games can make a contribution to social learning environments by applying game mechanics and other design elements, 'gamifying' social learning environments to make them more fun and engaging. For games to be effective as a learning tool, a delicate balance must be maintained between playability and educational value [9, 10], and between

game design and learning principles. Methods have been proposed for making valid inferences about what the student knows, using actions and events observed during gameplay. Such methods include evidence-centered-design (ECD) [11, 12]; the learning progressions model [13], the ecological approach to design of e-learning environments [14], stealth assessment [15], game analytics [16], and learning analytics [17]. Most of the new concepts target an ever-changing learning environment and learner needs, as today's education moves toward a digital, social, personalized, and fun environment. Moreover, as is the case for all competitive games, an equal match between players is essential to self-esteem and to maintain a high degree of player interest in the game. Hence, we need mechanisms and models that can aggregate the current performance and preferences of players, and accurately predict student performance in the game. Software agents have been used to implement consistent long-term intelligent behaviour in games [18], multi-agent collaborative team-based games [19], and adaptive and believable non-player character agents simulating virtual students [20]. The use of agent technologies leads to a system characterized by both autonomy and a distribution of tasks and control [21]. This trend has two aspects. First, game-based learning activities should be carefully orchestrated to be social and enjoyable. Second, game scheduling and coordination should be highly adaptive and flexible. However, nobody has yet developed models, algorithms, and mechanisms for planning, scheduling, and coordination that are suitable for creating and testing SEGs.

## 3 QuizMASter

QuizMASter is designed to be a formative assessment tool that enables students to be tested within a multi-player game [22]. Two or more students simultaneously log in remotely to the system via a Web-based interface. Each student is represented by one avatar in this virtual world. Students are able to view their own avatar as well as those of their opponents.

Each game has the game-show host who is also represented by an avatar visible to all contestants [22]. The game-show host poses each of the game questions to all the contestants. The students hear the voice of the host reading each question and view them displayed on their screens. They individually and independently from one another answer each question by, for instance, selecting an answer from available choices in a multiple-choice format. Each correct answer would receive one mark. Figure 1 shows a screen shot of QuizMASter.

### 3.1 Characteristics of QuizMASter

The environment for QuizMASter has the following characteristics:

**Flexibility.** The environment for QuizMASter needs flexibility for game enactment, to be able to cope with dynamic changes of user profiles, handle fragmentation of playing and learning time needed to accomplish activities and tasks,


**Fig. 1.** QuizMASter in Open Wonderland

adequately handle exceptional situations, predict changes due to external events, and offer sufficient interoperability with other software systems in educational institutions. Individual learners have particular interests, proficiency levels, and preferences that may result in conflicting learning goals.

**Social ability and interactivity.** The environment for QuizMASter should encourage interaction and collaboration among peers, and should be open to participation of students, teachers, parents, and experts on the subjects being taught. Web 2.0 has had a strong influence on the ways people learn and access information, and schools are taking advantage of this trend by adopting social learning environments. One way to engage learners in a collaborative production of knowledge is to promote social rewards.

**User control.** One of the most desirable features of social education games is to empower players with control over the problems that they solve. For example, in QuizMASter, students, parents, and teachers can design new rules to create their own games and modify the game elements to fit different knowledge levels.

**Customization.** Customization is a core principle that helps accommodate differences among learners [23]. Teachers could build a QuizMASter that has its own style and rules to determine the game's level of difficulty, to gear the game for specific goals or a specific group of learners. Some teachers may be interested in sharing collections of rules to fit the learning and play styles of their students. Like teachers, learners/players can be co-creators of their practice space through building new game scenarios, creating their own rules, sharing their strategies and making self-paced challenges [23].

## 4 The Proposed Architecture

Multi-agent technologies are considered most suitable for developing SEGs as it will lead to systems that operate in a highly dynamic, open, and distributed environment. In an MAS-based SEG, each learner/player is represented as an autonomous agent, called learner agent. MAS technologies, such as goal orientation and the Belief-Desire-Intention (BDI) paradigm, is used as the foundation for the agent architecture. These learner agents are able to reason about the learning goals, the strengths and weaknesses of learners and update the learner models.

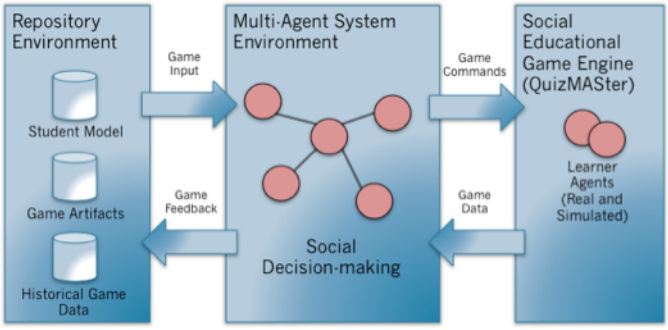



**Fig. 2.** Architecture for MAS-Based Social Educational Game Environment

Whenever a learner enters the system to play a social educational game, the learner agent will retrieve her/his learner model and acquire preferences about the current game-playing, and then send to a game management agent (GMA) of the system. The GMA is designed for setting up and maintaining teams for the system. The GMA will assign the learner to participate in a most suitable team that is undermanned according to the profile and preferences of the learner. The team will be configured in accordance with the game model by the GMA. Once the team has been completely formed, the GMA will create a game scheduling agent (GSA), a game host agent (GHA), and an assessment agent (AA) for each team. The GSA will continuously generate a game sequence dynamically adapted to the team's knowledge level (represented as a combined learner model [24]. The GHA will receive the game sequence from the scheduling agent and execute game sequence with the learners in the team. It will also be responsible for capturing data about learner/player performance. The AA will receive and interpret game events and communicate with the learner agents to update the learner model as necessary.

The GSA will dynamically schedule the game on the fly through interacting with other agents with a coordination mechanism, considering both the current world state and available resources, and solving conflicts in preferences and learning progression between the agents. The goal of the GSA is to optimize

the playability and educational values. We will model the game elements as resources. To solve the distributed constraint optimization problem, we are developing multiagent coordination mechanisms and scheduling algorithms to be used by the GSA.

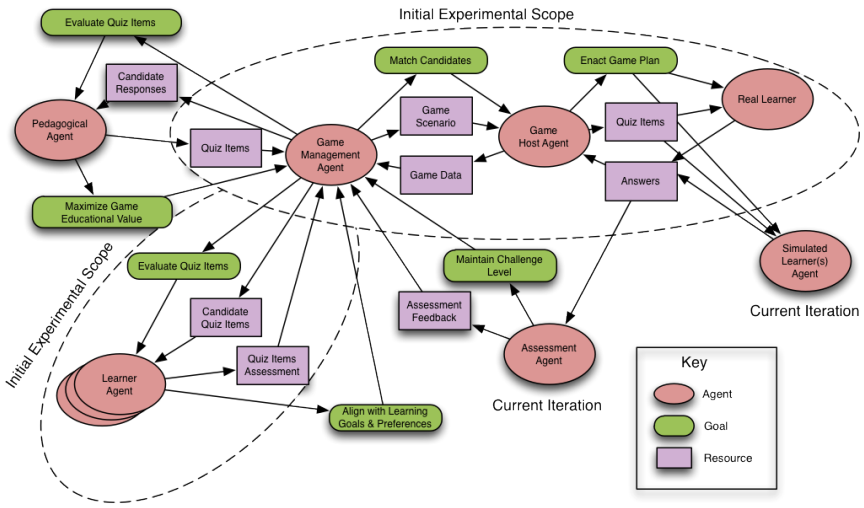



**Fig. 3.** MAS-Based SEG Agent Interaction Model

### 4.1 Planning and Scheduling Algorithms

The planning algorithms refer to the (local) planning algorithm of learner agents. To develop planning algorithms for learner agents, the following supporting models have been taken into consideration: (i) Learner models that accumulate and represent beliefs about the targeted aspects of skills. They are expressed as probability distributions for competency-model variables (called nodes) describing the set of knowledge and skills on which inferences are to be based. (ii) Evidence models that identify what the learner says or does, and provide evidence about those skills that express how the evidence depends on the competency-model variables in a psychometric model. (iii) Task/action models that express situations that can evoke required evidence. To design an action model, we adopt a model called Fuzzy Cognitive Goal Net [25] as the planning tool by combining the planning capability of Goal Net and reasoning ability of Fuzzy Cognitive Maps (FCMs). These FCMs give the learner agent a powerful reasoning ability for game context and player interactions, giving the task model accurate context awareness and learner awareness. We are developing coordination mechanisms

for the GMA and the GSA to solve the problem of team formation, scheduling and coordination in a highly flexible and dynamic manner. We considered the following concepts or methods:

(i) Contract-net protocols (CNPs) are used as a coordination mechanism by the GMA with a game model repository to timely form a team from all available players, using mutual selection and exchanging information in a structured way to converge on assignments. Each involved learner can delegate the negotiation process to its agent. These agents will strive to find a compromise team-joining decision obeying hard learning constraints while simultaneously resolving individual conflicts of interest.

(ii) The problem of scheduling and customizing a social educational game can be solved through social-choice-based customization. We view the SEG gameplay design as an optimization problem. Resources must be allocated through strategically scheduling, and coordinating a group of players according to their preferences and learning progressions. The constraints include key learning principles that inform the design of mechanics: challenge, exploration, risk taking, agency, and interactions [26-27]. The objective of the GSA is to maximize the learnability and engagement of the learners in the group. Social choice theory in MAS concerns the design and formal analysis of methods for aggregating preferences of multiple agents and collective decision-making and optimizing for preferences [28-29]. For example, we use a voting-based group decision-making approach such as Single Transferable Voting [30] to aggregate learner preferences and learning progression because it is computationally resistant to manipulation [31]. The purpose is to take information from individuals and combine it to produce the optimal result.

(iii) To support the need for dynamic decision making in the MAS-based SEG architecture, our current line of investigation is the concept of social choice Markov Decision Process (MDP) as recently proposed by Parkes and Procaccia [32]. In a social choice MDP, each state is defined by "preference profiles", which contain the preferences of all agents against a set of alternatives for a given scenario. The course of action from any given state is determined by a deterministic social choice function (the policy, in the context of the MDP) that takes into account the likelihood of transitions and their rewards. However, a preference profile is subject to change over time, especially in a live SEG context. For example, a learner that unexpectedly answers a question initially deemed beyond the learner's perceived level of comprehension would likely trigger a change of belief in the agents and potentially alter their ranking of alternatives. And since the number of alternatives in a SEG can be very large, the state space for any given SEG is huge, making the computation of optimal decision-making policies excessively difficult. We solve this problem by exploiting symmetries that exist in certain game types (e.g. in a quiz game SEG format, using a reduced set of question types that share common characteristics as a basis for alternatives as opposed to individual questions).

## 5 Simulated Learners

It is our view that the Belief-Desire-Intention (BDI) model is ideally suited for modeling and simulating learner behaviour. According to Jaques and Vicari (2007) [33], intelligent agents based on Bratman's Belief-Desire-Intention model, or BDI agents, are commonly used in modeling cognitive aspects, such as personality, affect, or goals. Píbil et al. (2012) claim BDI agent architecture is "a currently dominant approach to design of intelligent agents" [34]. Wong et al. (2012) describes the suitability of the BDI agent model for applications where both reactive behavior and goal-directed reasoning are required [35]. Soliman and Guetl (2012) suggest that BDI maps well onto models for pedagogically based selection of sub plans within a hierarchical planning strategy – "apprenticeship learning model" given as example [36]. They also talks about advantage of breaking plans down into smaller plans to allow for different "pedagogical permutations" allowing the agent to adapt to different learning styles, domain knowledge, and learning goals. Norling (2004) attributes the successful use of BDI agents for modeling human-like behavior in virtual characters to BDI's association to "folk psychology" [37]. This allows for an intuitive mapping of agent framework to common language that people use to describe the reasoning process. Of particular importance to this study is the way that implementations of the BDI architecture model long-term or interest goals. We have selected the Jason$^{TM}$ [38] platform for providing multi-agent BDI programming in AgentSpeak.

A shortcoming of the BDI paradigm is that although it is intended to be goal-driven, in most implementations this means/amounts to using goals to trigger plans, but does not support the concept of long-term goals or preferences [39], such as a student's long term learning goals, or the pedagogical goals of a CA. They feel that these types of goals are difficult to represent in most BDI systems because they signify an ongoing desire that must be maintained over a long period of time compared to relative short goal processing cycles. It is left to the developer to implement this type of preference goal through the belief system of the agent, modifications to the platform or environment, or other methods of simulating long-term goals.

Hübner, Bordini, and Wooldridge (2007) describe plan patterns for implementing declarative goals, with varying levels of commitment in AgentSpeak [40]. Bordini et al. (2007) expand on this in their chapter on advanced goal-based programming [38]. While AgentSpeak and Jason support achievement goals, these patterns are intended to address the lack of support for "richer goal structures", such as declarative goals, which they feel are essential to providing agents with rational behaviour. Pokahr et al. (2005) point out that the majority of BDI interpreters do not provide a mechanism for deliberating about multiple and possibly conflicting goals [41]. It is worth noting that there are "BDI inspired" systems that are more goal-oriented, such as Practionist and GOAL [42]. The Jason multi-agent platform for BDI agents was selected for this project because it is a well-established open-source project that is being actively maintained. It supports both centralized and distributed multi-agent environments. Píbil et

al. (2012) describes Jason as "one of the popular approaches in the group of theoretically-rooted agent-oriented programming languages" [34]. A major advantage of Jason is that it is easy to extend the language through Java based libraries and other components. Internal actions can allow the programmer to create new internal functionality or make use of legacy object-oriented code [38]. However, Píbil et al. (2012) caution that the use of such extensions, if used too heavily, can make the agent program difficult to comprehend without understanding the functionality of the Java code [34]. They raise the concern that novice programmers have few guidelines for choosing how much to program in AgentSpeak, and how much too program in Java. The usefulness of being able to extend Jason can be demonstrated by two examples of current research into integrating BDI with Bayesian Networks. Modeling of some student characteristics requires a probabilistic model; Bayesian Networks (BN) being a popular choice in recent years [43-44]. Recent work by Kieling and Vicari (2011) describes how they have extended Jason to allow a BDI agent to use a BN based probabilistic model. Similarly, Silva and Gluz (2011) extend the AgentSpeak(L) language to implement AgentSpeak(PL) by extending the Jason environment. AgentSpeak(PL) integrates probabilistic beliefs into BDI agents using Bayesian Networks [45]. Experimentation with QuizMASter to date has enabled the modelling of simulated learners in virtual worlds with an initial focus on their appearance, gestures, kinematics, and physical properties [46]. Recent related research work in that area has been on the creation of engaging avatars for 3D learning environments [47]. Employing the theory of Transformed Social Interaction (TSI) [48], simulated learners were designed with the following abilities:

(i) Self-identification: The self-identification dimension of TSI was implemented using facial-identity capture with a tool called FAtiMA. Each of the users' face were morphed with their default avatar agent's face to capitalize on human beings' disposition to prefer faces similar to their own and general preference of appearing younger (see Fig. 4).

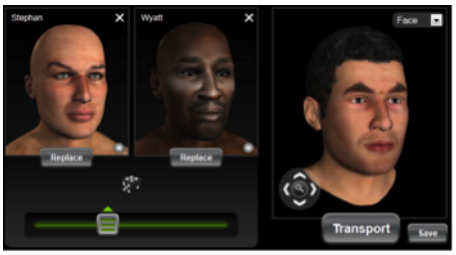



**Fig. 4.** Transformed Social Interaction – Image Morphing Technique

(ii) Sensory-abilities: Sensory-abilities dimension of TSI were implemented using a movement and visual tracking capability. The general challenge of sensory abilities implementation lies in two areas: the complexity of human senses and

the processing of sensory data of different modality and historicity. For the reason of simplicity, only visual tracking capability was exploited.

(iii) Situational-context: The situational-context dimension of TSI was implemented by using the best-view feature of Open Wonderland, whereby the temporal structure of a conversation can be altered.

The main idea of this research has been to explore the methodology for developing simulated learners for simulating and testing SEGs. That is, behind a simulated learner is an agent. Or we can say a simulated learner is an agent's avatar. All avatars, including real students' avatars and agent-based simulated learners, live in the virtual worlds, while the agents live in the multi-agent system. The integration of multi-agent systems with virtual worlds adds intelligence to the SEG platform and opens a number of extremely interesting and potentially useful research avenues concerning game-based learning. However, the advanced algorithms that support game planning, coordination and execution are difficult to test with real subjects considering the overhead involved is seeking authorization and the unpredictable availability of real life subjects in an online environment. This where an expanded view of simulated learners comes into play. The advantages of a simulated environment that closely approximates human behaviour include: (1) It allows for rapid and complete testing of advanced algorithms for game based adaptive assessment as well as SEG planning, coordination and execution in a simulated environment. The efficiency of the algorithms can be measured without first securing the availability of students; (2) With proper learner modeling and adaptive behaviour, simulated learners can engage with real life learners in friendly competitive games for the purpose of formative assessment, again working around the issue of availability of real students in an online learning environment.

## 6 Conclusions

As our recent experimentation suggests, many outstanding challenges must be addressed in developing intelligent SEGs. As we get closer to real world testing of our experimental game based assessment framework, we are faced with the complexity of enrolling real life learners in an e-learning environment and the variability that human interactions introduce in the measurement of adaptive algorithm efficiency. This is where we see the value of simulated learners. At this stage of our research, simulated learners have been rendered as Non Person Characters (NPCs) controlled by BDI agent running in the multi-agent system based virtual world. Our medium term goal is to extend the existing system to a particular learning subject (e.g., English language learning) to verify the effectiveness of the proposed virtual assessment environment and the benefit that students perceive from interacting with the proposed NPCs.

For simulated learners to be successful in our experimental framework, they must closely approximate the performance of real learners. The simple, pre-encoded behaviour we have implemented so far in the NPCs for QuizMASter will not suffice to demonstrate the efficiency of our adaptive algorithms and

allow for simulated learner agents to act as virtual players in our game based assessment framework. Current outstanding research questions within our group are:

1. How do we add intelligence and adaptive behaviour to the simulated learner agents while preserving our ability to obtain predictable and repeatable test results from our adaptive MAS framework?
2. How much autonomy can we afford to give to simulated learners in terms of independent thought and action, and to which degree should a simulated learner be able to adjust its behaviour as a function of its interactions with other agents, including real life learners?
3. How do we incorporate modern game, learning and assessment analytics in the supporting adaptive MAS framework in order to maximize the value of simulated learners as a means to perform non-intrusive, formative assessment?

## References


1. Romero, M., Usart, M., Ott, M., Earp, J., de Freitas, S., and Arnab, S.: Learning through playing for or against each other. Promoting Collaborative Learning in Digital Game Based Learning. ECIS 2012 Proceedings. Paper 93 (2012)
2. Alfonseca, E., Carro, R. M., Martin, E., Ortigosa, A., and Paredes, P.: The impact of learning styles on student grouping for collaborative learning: a case study. User Modeling and User-Adapted Interaction, 16(3–4), 377–401 (2006)
3. Deibel, K.: Team formation methods for increasing interaction during in-class group work. In ACM SIGCSE Bulletin (Vol. 37, 291–295). Caparica, Portugal (2005)
4. Grigoriadou, M., Papanikolaou, K. A., and Gouli, E.: Investigating How to Group Students based on their Learning Styles. In In ICALT, 2006 1139–1140 (2006)
5. McCalla, G. and Champaign J.: AIED Workshop on Simulated Learners. AIED 2013 Workshops Proceedings, Volume 4 (2013)
6. Vassileva, J.: Toward social learning environments. IEEE Transactions on Learning Technologies, 1(4), 199–213 (2008)
7. Klopfer, E., Osterweil, S., and Salen, K.: Moving learning games forward: obstacles, opportunities and openness, the education arcade. MIT (2009)
8. Jarvinen, A.: Game design for social networks: Interaction design for playful dispositions. In Stephen N. Spencer (Ed.), Proceedings of the 2009 ACM SIGGRAPH Symposium on Video Games (Sandbox '09), 95–102. ACM, New York, NY, USA (2009)
9. Van Eck, R.: Building Artificially Intelligent Learning Games. In V. Sugumaran (Ed.), Intelligent Information Technologies: Concepts, Methodologies, Tools and Applications, 793-825 (2008)
10. Augustin, T., Hockemeyer, C., Kickmeier-Rust, M. D., and Albert, D.: Individualized Skill Assessment in Digital Learning Games: Basic Definitions and Mathematical Formalism. IEEE Trans. on Learning Technologies, 4(2), 138-147 (2011)
11. Mislevy, R. J., and Haertel, G. D.: Implications of evidence-centered design for educational testing. Educational Measurement: Issues and Practice, 25(4), 6–20 (2006)
12. Mislevy, R. J., Steinberg, L. S., and Almond, R. G.: On the structure of educational assessments. Measurement: Interdisciplinary Research and Perspectives, 1, 3-67 (2003)

13. Corcoran, T., Mosher, F. A., and Rogat, A.: Learning Progressions in Science: An Evidence-Based Approach to Reform (Research Report No. RR-63). Center on Continuous Instructional Improvement, Teachers College—Columbia University (2009)
14. McCalla, G.: The Ecological Approach to the Design of E-Learning Environments: Purpose-based Capture and Use of Information About Learners. Journal of Interactive Media in Education. (7), 1-23 (2004)
15. Shute, V. J.: Stealth assessment in computer-based games to support learning. Computer games and instruction. Charlotte, NC: Information Age Publishers, 503-523 (2011)
16. Long, P., and Siemens, G.: Penetrating the fog: analytics in learning and education. Educause Review Online 46 (5): 31–40 (2011)
17. El-Nasr, M. S., Drachen, A., and Canossa, A.: Game Analytics: Maximizing the Value of Player Data, Springer (2013)
18. Oijen, J., and Dignum, F.: Scalable Perception for BDI-Agents Embodied in Virtual Environments, Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on, Vol. 2, 46-53, 22-27 (2011)
19. Patel, P., and Hexmoor, H.: Designing BOTs with BDI agents, Collaborative Technologies and Systems, 2009. CTS '09. International Symposium on, 180-186 (2009)
20. Lim, M., Dias, J., Aylett, R., and Paiva, A.: Creating adaptive affective autonomous NPCs, Autonomous Agents and Multi-Agent Systems (AAMAS), Springer, 24(2), 287-311 (2012)
21. Sterling, L. S. and Taveter, K.: The art of agent-oriented modeling, MIT Press (2009)
22. Dutchuk, M., Muhammadi, K. A., and Lin, F.: QuizMASter - A Multi-Agent Game-Style Learning Activity, EduTainment 2009, Aug 2009, Banff, Canada, Learning by Doing, (eds.), M Chang, et al., LNCS 5670, 263-272 (2009)
23. de Freitas, S. and Oliver, M.: How can exploratory learning with game and simulation within the curriculum be most effectively evaluated? Computers and Education. 46(3), 249-264 (2006)
24. Shabani, S., Lin, F. and Graf, S.: A Framework for User Modeling in QuizMASter. Journal of e-Learning and Knowledge Society, 8(3), 1826-6223 (2012)
25. Jennings N. R., L. Moreau, D. Nicholson, S. Ramchurn, S. Roberts, T. Rodden, and A. Rogers: Human-agent collectives. Commun. ACM 57, 12 (2014), 80-88 (2014)
26. Cai, Y., Miao, C., Tan, A.-H., and Shen, Z.: Fuzzy cognitive goal net for interactive storytelling plot design. In Proceedings of the 2006 ACM SIGCHI Int. conf. on Advances in comput. entertainment technology. ACM, NY, USA, Article 56 (2006)
27. Gee, James P.: Good Video Games + Good Learning. New York: Peter Lang (2008)
28. Gee, James P.: Learning by Design: Games as Learning Machines, Interactive Educational Multimedia, Vol. 8, April Ed. 2004, 13-15 (2004)
29. Brandt, F., Conitzer, V., and Endriss, U.: Computational Social Choice, Chapter 6 of book edited by Weiss, G., Multiagent Systems (2nd edition), 213-283 (2013)
30. Conitzer, V.: Making Decisions Based on the Preferences of Multiple Agents, Communications of the ACM, 53(3), 84-94 (2010)
31. Bartholdi, J. J. and Orlin, J. B: Single Transferable Vote Resists Strategic Voting. Social Choice and Welfare, 8, 341-354 (1991)
32. Parkes, D. C. and Procaccia, A. D.: Dynamic Social Choice with Evolving Preferences. In M. desJardins and M. L. Littman (eds.), AAAI : AAAI Press (2013)
33. Jaques, P. A., Vicari, R. M.: A BDI approach to infer student's emotions in an intelligent learning environment. Computers & Education, 49(2), 360–384 (2007)

34. Píbil, R., Novák, P., Brom, C., Gemrot, J.: Notes on Pragmatic Agent-Programming with Jason. In L. Dennis, O. Boissier, & R. H. Bordini (Eds.), Programming Multi-Agent Systems, 58–73. Springer Berlin Heidelberg (2012)
35. Wong, W., Cavedon, L., Thangarajah, J., Padgham, L.: Flexible Conversation Management Using a BDI Agent Approach. In Y. Nakano, M. Neff, A. Paiva, & M. Walker (Eds.), Intelligent Virtual Agents, 7502, 464–470). Springer (2012)
36. Soliman, M., Guetl, C.: Experiences with BDI-based design and implementation of Intelligent Pedagogical Agents. In 2012 15th International Conference on Interactive Collaborative Learning (ICL) (1–5). Presented at the 2012 15th International Conference on Interactive Collaborative Learning (ICL) (2012)
37. Norling, E.: Folk Psychology for Human Modelling: Extending the BDI Paradigm. In Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 1 (202–209). Washington, DC, USA: IEEE (2004)
38. Bordini, R. H., Hübner, J. F., Wooldridge, M.: Programming Multi-Agent Systems in AgentSpeak using Jason. John Wiley & Sons (2007)
39. Bellotti, F., Berta, R., De Gloria, A., Lavagnino, E.: Towards a conversational agent architecture to favor knowledge discovery in serious games. In Proc. of the 8th Int. Conf. on Advances in Comput. Entertainment Technology, 17:1–17:7 (2011)
40. Hübner, J. F., Bordini, R. H., Wooldridge, M.: Programming Declarative Goals Using Plan Patterns. In M. Baldoni & U. Endriss (Eds.), Proceedings on the Fourth International Workshop on Declarative Agent Languages and Technologies, held with AAMAS 2006 (123–140). Springer Berlin Heidelberg (2007)
41. Pokahr, A., Braubach, L., Lamersdorf, W. (2005). A BDI architecture for goal deliberation. In Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems (1295–1296). New York, NY, USA: ACM (2005)
42. Braubach, L., Pokahr, A.: Representing Long-Term and Interest BDI Goals. In L. Braubach, J.-P. Briot, & J. Thangarajah (Eds.), Programming Multi-Agent Systems (pp. 201–218). Springer Berlin Heidelberg (2010)
43. Conati, C., Maclaren, H.: Empirically building and evaluating a probabilistic model of user affect. User Modeling and User-Adapted Interaction, 19(3), 267–303 (2009)
44. Chrysafiadi, K., Virvou, M.: Student modeling approaches: A literature review for the last decade. Expert Systems with Applications, 40(11), 4715–4729 (2013)
45. Silva, D. G., Gluz, J. C.: AgentSpeak(PL): A New Programming Language for BDI Agents with Integrated Bayesian Network Model. In 2011 International Conference on Information Science and Applications (ICISA) (pp. 1–7). Presented at the 2011 International Conference on Information Science and Applications (ICISA) (2011)
46. McClure, G., Chang, M., Lin, F.: MAS Controlled NPCs in 3D Virtual Learning Environment, The International Workshop on Smart Learning Environments at the 9th International Conference on Signal-Image Technology and Internet-Based Systems, Japan, Kyoto, 2013-12-02, 1026 – 1033, DOI: 10.1109/SITIS.2013.166 (2013)
47. Walsh, T.: An Empirical Study of the Manipulability of Single Transferable Voting, Proceedings of the 2010 conference on ECAI, 257-262 (2010)
48. Leung, S., Virwaney, S., Lin, F., Armstrong, A J, Dubbelboer, A.: TSI-enhanced Pedagogical Agents to Engage Learners in Virtual Worlds", International Journal of Distance Education Technologies, 11(1), 1-13 (2013)

# Is this model for real? Simulating data to reveal the proximity of a model to reality

Rinat B. Rosenberg-Kima[1], Zachary A. Pardos[2]

[1] Tel-Aviv University
rinat.rosenberg.kima@gmail.com
[2] University of California, Berkeley
pardos@berkeley.edu

**Abstract.** Simulated data plays a central role in Educational Data Mining and in particular in Bayesian Knowledge Tracing (BKT) research. The initial motivation for this paper was to try to answer the question: given two datasets could you tell which of them is real and which of them is simulated? The ability to answer this question may provide an additional indication of the goodness of the model, thus, if it is easy to discern simulated data from real data that could be an indication that the model does not provide an authentic representation of reality, whereas if it is hard to set the real and simulated data apart that might be an indication that the model is indeed authentic. In this paper we will describe analyses of 42 GLOP datasets that were performed in an attempt to address this question. Possible simulated data based metrics as well as additional findings that emerged during this exploration will be discussed.

**Keywords:** Bayesian Knowledge Tracing (BKT), simulated data, parameters space.

## 1 Introduction

Simulated data has been increasingly playing a central role in Educational Data Mining [1] and Bayesian Knowledge Tracing (BKT) research [1, 4]. For example, simulated data was used to explore the convergence properties of BKT models [5], an important area of investigation given the identifiability issues of the model [3]. In this paper, we would like to approach simulated data from a slightly different angle. In particular, we claim that the question "*given two datasets could you tell which of them is real and which of them is simulated?"* is interesting as it can be used to evaluate the goodness of a model and may potentially serve as an alternative metric to RMSE, AUC, and others. In a previous work [6] we started approaching this problem by contrasting two real datasets with their corresponding two simulated datasets with Knowledge Tracing as the model. We found a surprising close to identity between the real and simulated datasets. In this paper we would like to continue this investigation by expanding the previous analysis to the full set of 42 Groups of Learning Opportunities (GLOPs) real datasets generated from the ASSISTments platform [7].

Knowledge Tracing (KT) models are widely used by cognitive tutors to estimate the latent skills of students [8]. Knowledge tracing is a Bayesian model, which assumes that each skill has 4 parameters: two knowledge parameters include initial (prior knowledge) and learn rate, and two performance parameters include guess and slip. KT in its simplest form assumes a single point estimate for prior knowledge and learn rate for all students, and similarly identical guess and slip rates for all students. Simulated data has been used to estimate the parameter space and in particular to answer questions that relate to the goal of maximizing the log likelihood (LL) of the model given parameters and data, and improving prediction power [7, 8, 9].

In this paper we would like to use the KT model as a framework for comparing the characteristics of simulated data to real data, and in particular to see whether it is possible to distinguish between the real and simulated datasets.

## 2 Data Sets

To compare simulated data to real data we started with 42 Groups of Learning Opportunities (GLOPs) real datasets generated from the ASSISTments platform[1] from a previous BKT study [7]. The datasets consisted of problem sets with 4 to 13 questions in linear order where all students answer all questions. The number of students per GLOP varied from 105 to 777. Next, we generated two synthetic, simulated datasets for each of the real datasets using the best fitting parameters that were found for each respective real datasets as the generating parameters. The two simulated datasets for each real one had the exact same number of questions, and same number of students.

## 3 Methodology

The approach we took to finding the best fitting parameters was to calculate LL with a grid search of all the parameters (prior, learn, guess, and slip). We hypothesized that the LL gradient pattern of the simulated data and real data will be different across the space. For each of the datasets we conducted a grid search with intervals of .04 that generated 25 intervals for each parameter and 390,625 total combinations of prior, learn, guess, and slip. For each one of the combinations LL was calculated and placed in a four dimensional matrix. We used fastBKT [12] to calculate the best fitting parameters of the real datasets and to generate simulated data. Additional code in Matlab and R was generated to calculate LL and RMSE and to put all the pieces together[2].

[1] Data can be obtained here: http://people.csail.mit.edu/zp/

[2] Matlab and R code will be available here: www.rinatrosenbergkima.com/AIED2015/

## 4 What are the Characteristics of the Real Datasets Parameters Space?

Before we explored the relationships between the real and sim datasets, we were interested to explore the BKT parameter profiles of the real datasets. We calculated the LL with a grid search of 0.04 granularity across the four parameters resulting in a maximum LL for each dataset and corresponding best prior, learn, guess, and slip. Figure 1 present the best parameters for each datasets, taking different views of the parameters space. The first observation to be made is that the best guess and slip parameters fell into two distinct areas (see figure 1, guess x slip).

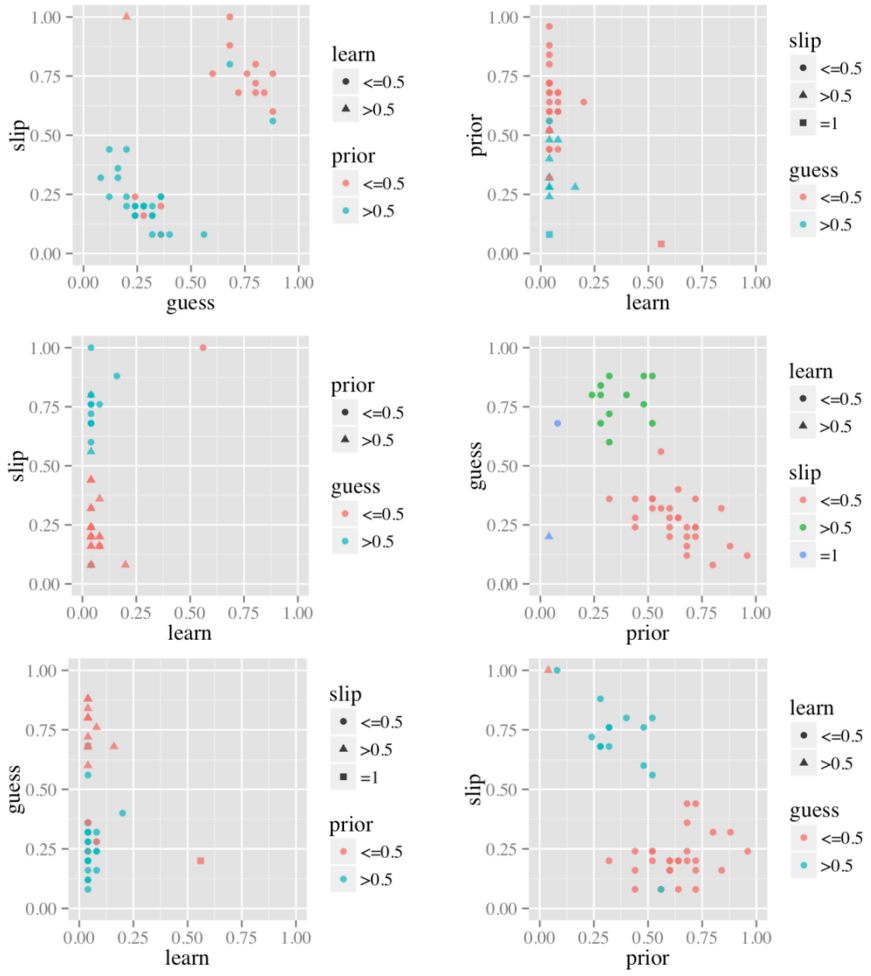



**Figure 1.** Best parameters across the 42 GLOP real datasets.

Much attention has been given to this LL space, which revealed the apparent co-linearity of BKT with two primary areas of convergence, the upper right area being a false, or "implausible" converging area as defined by [3]. What is interesting in this figure is that real data also converged to these two distinct areas. To further investigate this point, we looked for the relationships between the best parameters and the number of students in the dataset (see figure 2). We hypothesized that perhaps the upper right points were drawn from datasets with small number of students; nevertheless, as figure 2 reveals, that was not the case. Another interesting observation is that while in the upper right area (figure 1, guess x slip) most of the prior best values were smaller that 0.5, in the lower left area most of the prior best values were bigger than 0.5, thus revealing interrelationships between slip, guess, and prior that can be seen in the other views. Another observation is that while prior is widely distributed between 0 and 1, most of best learn values are smaller than 0.12.

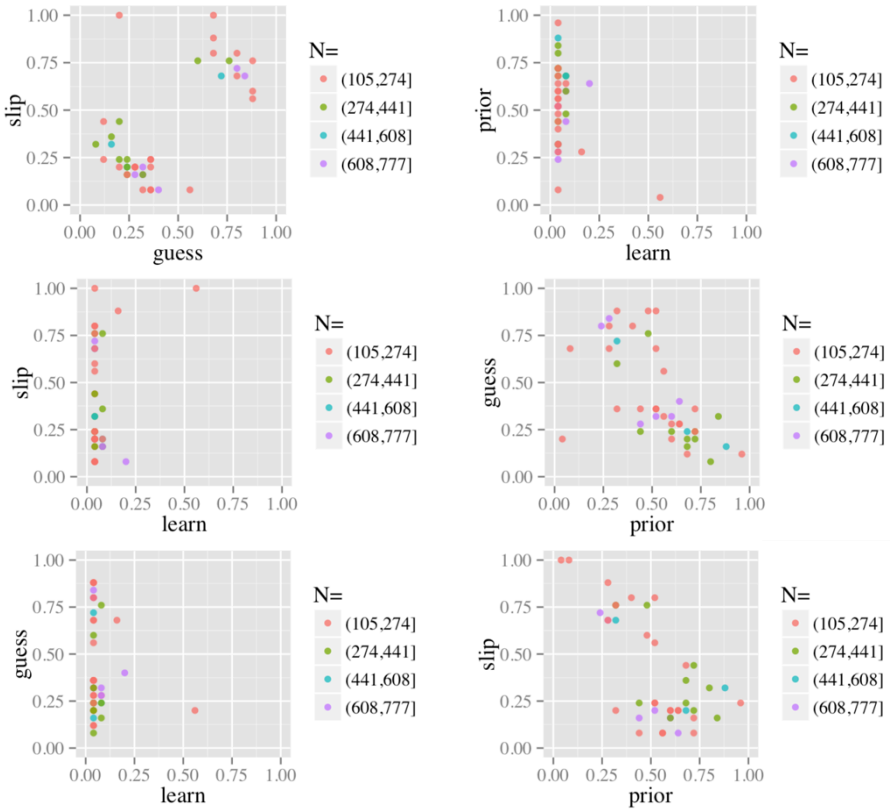



**Figure 2.** Best parameters across the 42 GLOP real datasets by number of students.

## 5 Does the LL of Sim vs. Real Datasets Look Different?

Our initial thinking was that as we are using a simple BKT model, it is not authentically reflecting reality in all its detail and therefore we will observe different patterns of LL across the parameters space between the real data and the simulated data. The LL space of simulated data in [5] was quite striking in its smooth surface but the appearance of real data was left as an open research question. First, we examined the best parameters spread across the 42 first set of simulated data we have generated. As can be seen in figure 3, the results are very similar (although not identical) to the results we received with the real data (see figure 1). This is not surprising, after all, the values of learn, prior, guess, and slip were inputs to the function generating the simulated data.

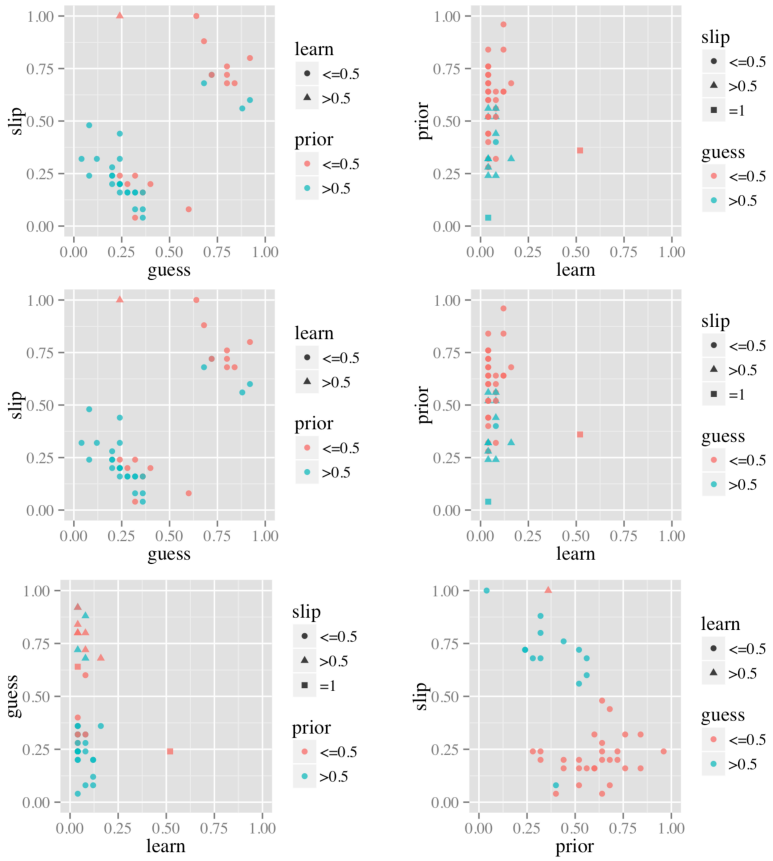



**Figure 3.** Best parameters across 42 GLOP simulated datasets.

In order to see if the differences between real and sim were more than just the difference between samples from the same distribution, we generated *two* simulated versions of each real dataset (sim1 and sim2) using the exact same number

of questions, number of students, generated with the best fitting parameters from the real dataset. We then visualized 2D LL heatmaps looking at two parameter plots at a time where the other two parameters were fixed to the best fitting values. For example, when visualizing LL heatmaps for the combination of guess and slip, we fixed learn and prior to be the best learn and the best prior from the real data grid search. To our surprise, when we plotted heatmaps of the LL matrices of the real data and the simulated data (the first column in figure 4 represents the real datasets, the second column represents the corresponding sim1, and the third column the corresponding sim2) we received what appears to be extremely similar heatmaps. Figure 4 and 5 displays a sample of 4 datasets, for each one displaying the real dataset heatmap and the corresponding two simulated datasets heatmaps.

The guess vs. slip heatmaps (see figure 4) prompted interesting observations. As mentioned above, the best guess and slip parameters across datasets fell into two areas (upper right and lower left). Interestingly, these two areas were also noticeable in the individual heatmaps. While in some of the datasets they were less clear (e.g., G5.198 in figure 4), most of the datasets appear to include two distinct global maxima areas. In some of the datasets the global maxima converged to the lower left expected area, as did the corresponding simulated datasets (e.g., G4.260 in figure 4), in other datasets the global maxima converged to the upper right "implausible" area, as did the corresponding simulated datasets (e.g., G6.208 in figure 4). Yet in some cases, one or more of the simulated dataset converged to a different area than that of the real dataset (e.g., G4.205 in figure 4). The fact that so many of the real datasets converged to the "implausible" area is surprising and may be due to small number of students or to other limitations of the model.

The learn vs. prior heatmaps were also extremely similar within datasets and exhibited a similar pattern also across datasets (see figure 5), although not all datasets had the exact pattern (e.g., G5.198 is quite different than the other 3 datasets in figure 5). While best learn values were low across the datasets, the values of best prior varied. As with guess vs. slip, in some cases the two simulated datasets were different (e.g., G4.205 had different best parameters also with respect to prior). Similar patterns of similarities within datasets and similarities with some clusters across datasets were also noticeable in the rest of the parameters space (learn vs. guess, learn vs. slip, prior vs. guess, prior vs. slip not displayed here due to space considerations).

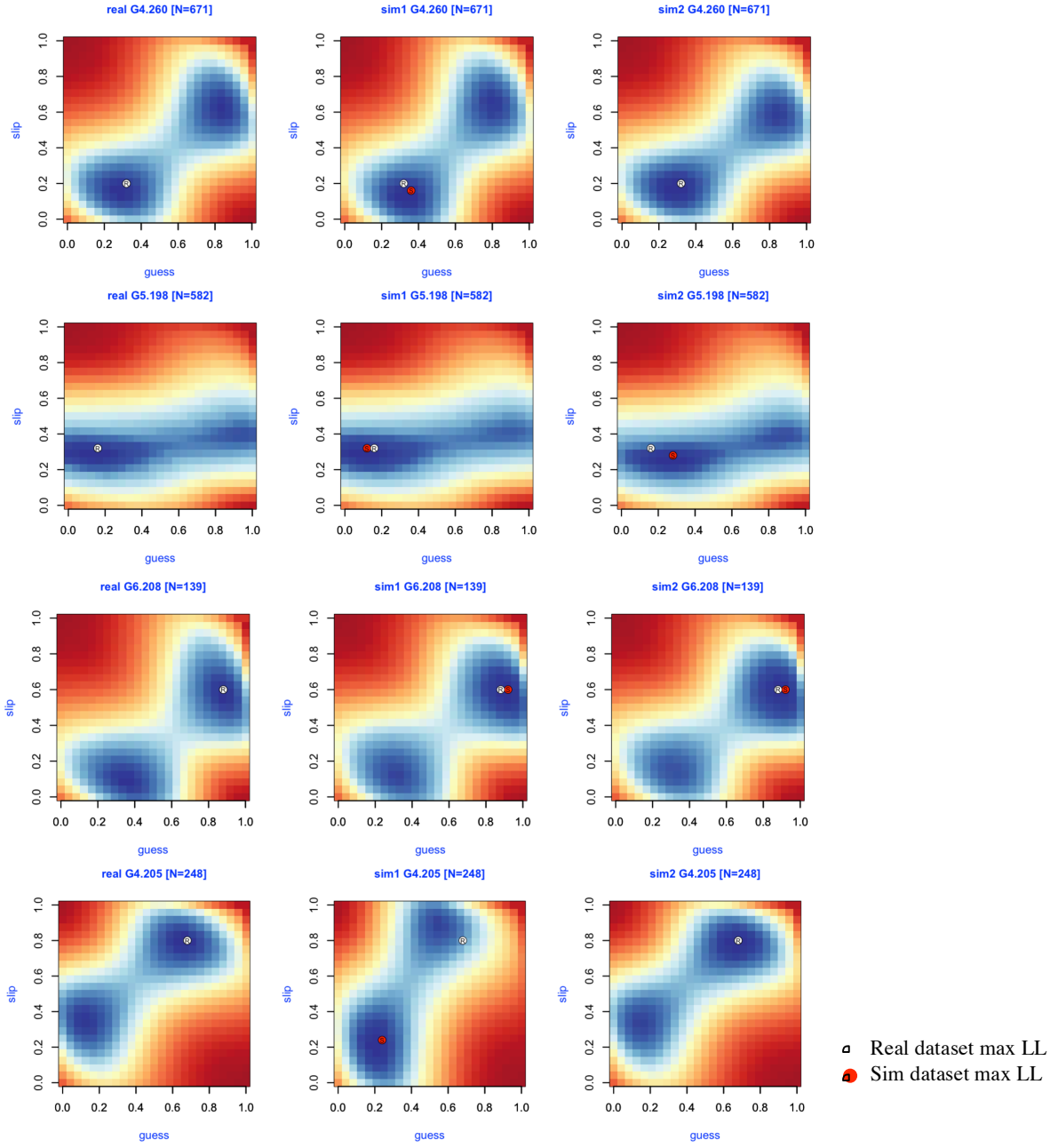



**Figure 4.** Heatmaps of (guess vs. slip) LL of 4 sample real GLOP datasets and the corresponding two simulated datasets that were generated with the best fitting parameters of the corresponding real dataset.

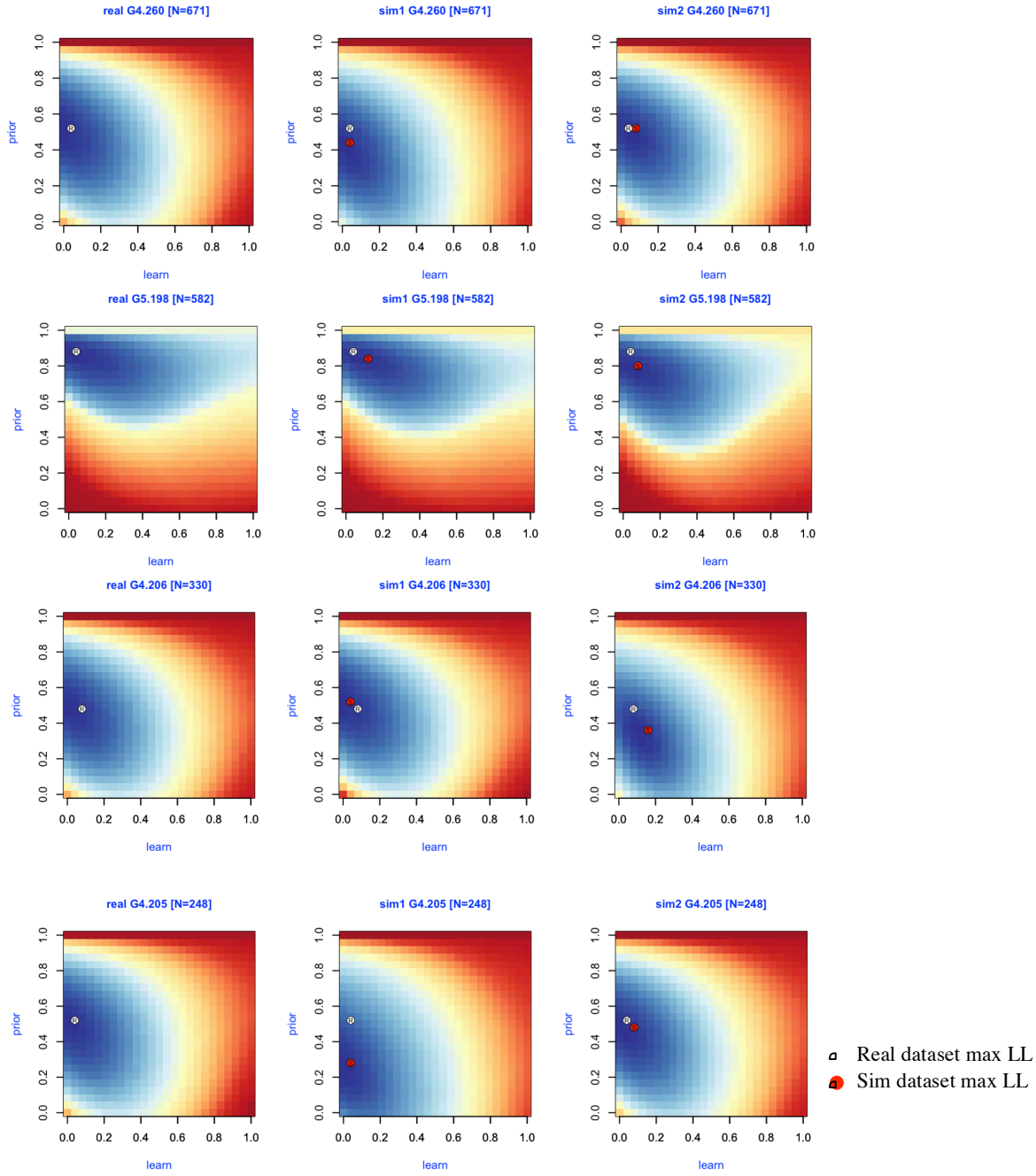



**Figure 5.** Heatmaps of (learn x prior) LL of 4 sample real GLOP datasets and the corresponding two simulated datasets that were generated with the best fitting parameters of the corresponding real dataset.

## 6 Exploring Possible Metrics Using the Real and Sim Datasets

In natural science domains, simulated data is often used as a mean to evaluate its underlying model. For example, simulated data is generated from a hypothesized model of the phenomena and if the simulated data appears to be similar to the real data observed in nature, it serves as evidence for the accuracy of the model. Then, if the underlying is validated, simulated data is used to make predictions (e.g., in the recent earthquake in Nepal a simulation was used to estimate the number of victims). Can this approach be used in education as well? What would be an indication of similarity between real and simulated data?

Figure 5 displays two preliminary approaches for comparing the level of similarity between the simulated and real data. First, the Euclidean distance between the real dataset parameters and the simulated data parameters was compared to the Euclidean distance between the two simulated datasets parameters. The idea is that if the difference between the two simulated datasets is smaller than the difference between the real and the simulated dataset this may be an indication that the model can be improved upon. Thus, points on the right side of the red diagonal indicate good fit of the model to the dataset. Interestingly, most of the points were on the diagonal and a few to the left of it. Likewise the max LL distance between the real and simulated datasets was compared to the max LL distance of the two simulated datasets. Interestingly, datasets with larger number of students did not result in higher similarity between the real and simulated dataset. Also, here we *did* find distribution of the points to the left and to the right of the diagonal.

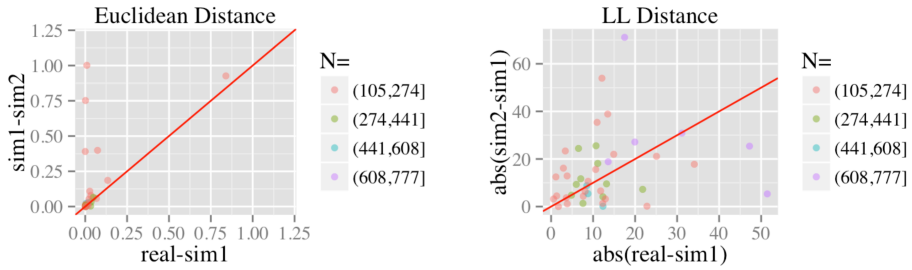



**Figure 5.** Using Euclidean distance and LL distance as means to evaluate the model.

## 7 Contribution

The initial motivation of this paper was to find whether it is possible to discern a real dataset from a simulated dataset. If for a given model it is possible to tell apart a simulated data from a real dataset then the authenticity of the model can be questioned. This line of thinking is in particular typical of simulation use in Science contexts, where different models are used to generate simulated data, and then if a simulated data has a good fit to the real phenomena at hand, then it may be possible to claim that the model provides an authentic explanation of the system [13]. We believe

that finding such a metric can serve as the foundation for evaluating the goodness of a model by comparing a simulated data from this model to real data and that such a metric could provide much needed substance in interpretation beyond that which is afforded by current RMSE and AUC measures. This can afford validation of the simulated data, which can than be used to make predictions on learning scenarios; decreasing the need to test them in reality, and at minimum, serving as an initial filter to different learning strategies.

## References


[1] R. S. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *J. Educ. Data Min.*, vol. 1, no. 1, pp. 3–17, 2009.

[2] M. C. Desmarais and I. Pelczer, "On the Faithfulness of Simulated Student Performance Data.," in *EDM*, 2010, pp. 21–30.

[3] J. E. Beck and K. Chang, "Identifiability: A fundamental problem of student modeling," in *User Modeling 2007*, Springer, 2007, pp. 137–146.

[4] Z. A. Pardos and M. V. Yudelson, "Towards Moment of Learning Accuracy," in *AIED 2013 Workshops Proceedings Volume 4*, 2013, p. 3.

[5] Z. A. Pardos and N. T. Heffernan, "Navigating the parameter space of Bayesian Knowledge Tracing models: Visualizations of the convergence of the Expectation Maximization algorithm.," in *EDM*, 2010, pp. 161–170.

[6] R. B. Rosenberg-Kima and Z. Pardos, "Is this Data for Real?," in *Twenty Years of Knowledge Tracing Workshop*, London, UK, pp. 141–145.

[7] Z. A. Pardos and N. T. Heffernan, "Modeling individualization in a bayesian networks implementation of knowledge tracing," in *User Modeling, Adaptation, and Personalization*, Springer, 2010, pp. 255–266.

[8] A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," *User Model. User-Adapt. Interact.*, vol. 4, no. 4, pp. 253–278, 1994.

[9] S. Ritter, T. K. Harris, T. Nixon, D. Dickison, R. C. Murray, and B. Towle, "Reducing the Knowledge Tracing Space.," *Int. Work. Group Educ. Data Min.*, 2009.

[10] R. S. d Baker, A. T. Corbett, S. M. Gowda, A. Z. Wagner, B. A. MacLaren, L. R. Kauffman, A. P. Mitchell, and S. Giguere, "Contextual slip and prediction of student performance after use of an intelligent tutor," in *User Modeling, Adaptation, and Personalization*, Springer, 2010, pp. 52–63.

[11] R. S. Baker, A. T. Corbett, and V. Aleven, "More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing," in *Intelligent Tutoring Systems*, 2008, pp. 406–415.

[12] Z. A. Pardos and M. J. Johnson, "Scaling Cognitive Modeling to Massive Open Environments (in preparation)," *TOCHI Spec. Issue Learn. Scale*.

[13] U. Wilensky, "GasLab—an Extensible Modeling Toolkit for Connecting Micro- and Macro-properties of Gases," in *Modeling and simulation in science and mathematics education*, Springer, 1999, pp. 151–178.

# Preface

This workshop, a follow-up to the successful first Simulated Learners workshop held at AIED 2013, is intended to bring together researchers who are interested in simulated learners, whatever their role in the design, development, deployment, or evaluation of learning systems. Its novel aspect is that it isn't simply a workshop about pedagogical agents, but instead focuses on the other roles for simulated learners in helping system designers, teachers, instructional designers, etc.

As learning environments become increasingly complex and are used by growing numbers of learners (sometimes in the hundreds of thousands) and apply to a larger range of domains, the need for simulated learners (and simulation more generally) is compelling, not only to enhance these environments with artificial agents, but also to explore issues using simulation that would be otherwise be too expensive, too time consuming, or even impossible using human subjects. While some may feel that MOOCs provide ample data for experimental purposes, it is hard to test specific hypotheses about particular technological features with data gathered for another purpose. Moreover, privacy concerns, ethics approval, attrition rates and platform constraints can all be barriers to this approach. Finally, with thousands of learners at stake, it is wise to test a learning environment as thoroughly as possible before deployment.

Since this is a follow-up to the 2013 workshop, we build on some of the ideas that emerged there (see proceedings at: http://goo.gl/12ODji).

The workshop explores these and other issues with the goal of further understanding the roles that simulated learners may play in advanced learning technology research and development, and in deployed learning systems.

John Champaign and Gord McCalla
Workshop Co-Chairs

# Second Workshop on Simulated Learners

*held in conjunction with*

Seventeenth International Conference on Artificial Intelligence in Education (AIED 2015)

Friday, June 26, 2015
Madrid, Spain

**Workshop Co-Chairs:**

John Champaign[1] and Gord McCalla[2]

[1] *University of Illinois at Springfield, Springfield, IL, 62703, USA*
[2] *University of Saskatchewan, Saskatoon, S7N 5C9, Canada*

https://sites.google.com/site/simulatedlearners/

# Table of Contents

# Preface

This workshop, a follow-up to the successful first Simulated Learners workshop held at AIED 2013, is intended to bring together researchers who are interested in simulated learners, whatever their role in the design, development, deployment, or evaluation of learning systems. Its novel aspect is that it isn't simply a workshop about pedagogical agents, but instead focuses on the other roles for simulated learners in helping system designers, teachers, instructional designers, etc.

As learning environments become increasingly complex and are used by growing numbers of learners (sometimes in the hundreds of thousands) and apply to a larger range of domains, the need for simulated learners (and simulation more generally) is compelling, not only to enhance these environments with artificial agents, but also to explore issues using simulation that would be otherwise be too expensive, too time consuming, or even impossible using human subjects. While some may feel that MOOCs provide ample data for experimental purposes, it is hard to test specific hypotheses about particular technological features with data gathered for another purpose. Moreover, privacy concerns, ethics approval, attrition rates and platform constraints can all be barriers to this approach. Finally, with thousands of learners at stake, it is wise to test a learning environment as thoroughly as possible before deployment.

Since this is a follow-up to the 2013 workshop, we build on some of the ideas that emerged there (see proceedings at: http://goo.gl/12ODji).

The workshop explores these and other issues with the goal of further understanding the roles that simulated learners may play in advanced learning technology research and development, and in deployed learning systems.

John Champaign and Gord McCalla
Workshop Co-Chairs

# An Approach to Developing Instructional Planners for Dynamic Open-Ended Learning Environments

Stephanie Frost and Gord McCalla

ARIES Lab, Department of Computer Science, University of Saskatchewan, Canada
stephanie.frost@usask.ca, mccalla@cs.usask.ca

**Abstract.** *Instructional planning* (IP) technology has begun to reach large online environments. However, many approaches rely on having centralized metadata structures about the learning objects (LOs). For *dynamic open-ended* learning environments (DOELEs), an approach is needed that does not rely on centralized structures such as prerequisite graphs that would need to be continually rewired as the LOs change. A promising approach is collaborative filtering based on learning sequences (CFLS) using the ecological approach (EA) architecture. We developed a CFLS planner that compares a given learner's most recent path of LOs (of length $b$) to other learners to create a neighbourhood of similar learners. The future paths (of length $f$) of these neighbours are checked and the most successful path ahead is recommended to the target learner, who then follows that path for a certain length (called $s$). We were interested in how well a CFLS planner, with access only to pure behavioural information, compared to a traditional instructional planner that used explicit metadata about LO prerequisites. We explored this question through simulation. The results showed that the CFLS planner in many cases exceeded the performance of the simple prerequisite planner (SPP) in leading to better learning outcomes for the simulated learners. This suggests that IP can still be useful in DOELEs that often won't have explicit metadata about learners or LOs.

**Keywords:** instructional planning, collaborative filtering, dynamic open-ended learning environments, simulated learning environments, simulated learners, ecological approach

## 1 Introduction

Online courses need to be able to personalize their interactions with their many learners not only to help each learner overcome particular impasses but also to provide a path through the learning objects (LOs) that is appropriate to that particular individual. This is the role of *instructional planning* (IP), one of the core AIED sub-disciplines. IP is particularly needed in open-ended learning environments (OELEs), where learners choose their own goals, because it has been shown that sometimes learners require an outside push to move forward

[11]. An added challenge is what we call a *dynamic open-ended* learning environment (DOELE), where both the learners and LOs are constantly changing. Some learners might leave before finishing the course, while others may join long after other learners have already begun. New material (LOs) may need to be added in response to changes in the course or the material, or to learner demand. Sometimes new material will be provided by the course developers, but the big potential is for this to be crowd sourced to anybody, including learners themselves. Other material may fade away over time.

Note that a DOELE is similar to, but not the same as, a "traditional" open-eded learning environment [8, 11]. A traditional open-ended environment also gives students choice, but mostly in the problems they solve and how they solve them, with the course itself fixed in its content, order and goals. In a DOELE everything is open-ended and dynamic, including even what is to be learned, how deeply, when it needs to be learned, and in what order.

An impediment to IP in a DOELE is that there is no centralized representation of knowledge about the content or the learners. Work has been done to make IP possible in online environments, such as [7], where authors showed that by extending the LO metadata, instructional plans could be improved to adapt based on individual learning styles as well as a resource's scheduling availability. But for IP to work in DOELEs, an approach to IP is needed where centralized course structures would not need to be continually revamped (by instructional designers, say) as learners and LOs change.

We wish to explore how IP can be done in a DOELE. We model a DOELE in the ecological approach (EA) architecture [14]. In the EA there is no overall course design. Instead, courses are conceived as collections of learning objects each of which captures usage data as learners interact with it. Over time this usage data accumulates and can be used for many pedagogical purposes, including IP [2]. Drawing inspiration from work like [1, 5], we propose a new IP algorithm based on collaborative filtering of learning sequences (CFLS). For a given learner our planner finds other learners who have traversed a similar sequence of learning objects with similar outcomes (i.e. similar paths). Then it suggests paths to the learner that were successful for these similar learners (peers) going forward.

To evaluate IP techniques in such an environment, one could implement a real course with thousands of learners using the EA to capture learner interactions with the various LOs in the course. However, after doing this it would take several years for enough learners to build up enough interactions with each LO to provide useful data to be used by an instructional planner. Also, in a course with thousands of learners, there is risk of causing confusion or inconvenience to a vast multitude if there are problems while the planner is under development. Finally, there are unanswered design questions such as the criteria to use for identifying an appropriate peer, how many LOs should be recommended for a learner before re-planning occurs, and appropriate values for many other parameters that would be used by the planner. In order to overcome these challenges and gain insight into these questions immediately, we have thus turned to simulation.

## 2 Simulation Environment

Before describing the CFLS planner and experiment in detail, we describe the simulation environment. The simulation is low-fidelity, using very simple abstractions of learners and LOs, as in our earlier work [6]. Each of the 40 LOs has a difficulty level and possible prerequisite relationships with other LOs. Each simulated learner has an attribute, *aptitude-of-learner*, a number between (0,1) representing a learner's basic capability for the subject and allows learners to be divided into groups: low ($\leq .3$), medium ($.4 - .7$) and high aptitude ($\geq .8$).

A number called *P[learned]* is used to represent the learning that occurred when a learner visits a LO, or the probability that the learner learned the LO. P[learned] is generated by an *evaluation function*, a weighted sum: 20% of the learner's score on a LO is attributed to *aptitude-of-learner*, 50% attributed to whether the learner has mastered all of the prerequisite LOs, 20% attributed to whether the learner had seen that LO previously, and 10% attributed to the difficulty level of the LO. We feel this roughly captures the actual influences on how likely it is that real learners would master a learning object.

The simulated learners move through the course by interacting with the LOs, one after another. After each LO is encountered by a simulated learner, the above evaluation function is applied to determine the learner's performance on the LO, the P[learned] for that learner on that LO. In the EA architecture, everything that is known about a learner at the time of an interaction with a LO (in this case, including P[learned]) is captured and associated with that LO. The order of the LOs visited can be set to random, or it can be determined by a planner such as the CFLS planner. To allow for the comparison of different planning approaches without advantaging one approach, each simulated learner halts after its 140th LO regardless of the type of planner being used.

## 3 Experiment

By default, the simulation starts with an empty history - no simulated learners have yet viewed any LOs. However, because the CFLS planner relies on having previous interaction data, it is necessary to initialize the environment. Thus, a simple prerequisite planner (SPP) was used to initialize the case base with a population of simulated learners. The SPP is privy to the underlying prerequisite structure and simply delivers LOs to learners in prerequisite order. As Table 1 shows, the SPP works much better than a random planner. The data from the 65 simulated learners who used the SPP thus was used to initialize the environment before the CFLS planner took over. This interaction data generated by the SPP also provides a baseline for comparison with the CFLS planner. Our simulation experiment was aimed at seeing if, with appropriate choices of $b$ and $f$ (described below) the CFLS planner could work as well or better than the SPP.

We emphasize that the CFLS planner has no knowledge about the underlying prerequisite structure of the learning objects. This is critical for CFLS planning to work in a DOELE. However, there are two places where clarification

**Table 1.** Baseline results for each group of simulated learners (high, medium and low aptitude) when visiting LOs randomly and following a simple prerequisite planner.

| Planning Type / Aptitude | low | medium | high |
|---|---|---|---|
| **Random** | N=21 | N=26 | N=18 |
| Average Score on Final Exam (P[learned]) | 0.107 | 0.160 | 0.235 |
| **Simple Prerequisite Planner (SPP)** | N=21 | N=26 | N=18 |
| Average Score on Final Exam (P[learned]) | 0.619 | 0.639 | 0.714 |

is required. First, while the SPP is running, the evaluation function will be used by the simulation to calculate P[learned] values for each LO visited. This usage data will contain implicit evidence of the prerequisite relationships. So, at a later time when the CFLS planner is given access to the same usage data, the CFLS planner could implicitly discover prerequisite relationships from the interaction data. Second, during the CFLS planner execution, the underlying prerequisite structure is still being consulted by the evaluation function. However, the CFLS planner knows nothing about such prerequisites, only the P[learned] outcome provided by the evaluation function. When simulated learners are replaced with real learners, the evaluation function would disappear and be replaced with a real world alternative, such as quizzes or other evidence to provide a value for P[learned]. Similarly, the CFLS planner does not require knowledge of the difficulty level of each LO, nor does it require knowledge of the aptitude of each learner; these are just stand-in values for real world attributes used by the simulation and would disappear when the planner is applied in a real world setting.

Different studies can use simulated student data in varying ways. In some cases, low fidelity modelling is not adequate. For example, in [4] it was found that the low fidelity method of generating simulated student data failed to adequately capture the characteristics of real data. As a result, when the simulated student dataset was used for training the cognitive diagnosis model, its predictive power was worse than when the cognitive diagnosis model was trained with a simulated student dataset that had been generated with a higher fidelity method. In our study, using a low fidelity model is still informative. We are less concerned with the exactness of P[learned] and are more interested in observing possible relative changes of P[learned] for certain groups of students, as different variations of the planner are tried on identical populations of simulated students.

The CFLS planner works as follows. For a given target learner the CFLS planner looks backward at the $b$ most recent learning objects traversed. Then, it finds other learners who have traversed the same $b$ learning objects with similar P[learned] values. These $b$ LOs can be in any order, a simplification necessary to create a critical mass of similar learners. These are learners in the target learner's "neighbourhood". The planner then looks forward at the $f$ next LOs traversed by each neighbour and picks the highest value path, where value is defined as the average P[learned] achieved on those $f$ LOs ahead. This path is then recommended to the learner, who must follow it for at least $s$ (for "sticky") LOs before replanning occurs. Of course, $s$ is always less than $f$. In our research

we explored various values of $b$ and $f$ to find which leads to the best results (we set $f = s$ for this experiment). "Best results" can be defined many ways, but we focused on two measurements that were taken for each learner at the end of each simulation: the percentage of LOs mastered, and the score on a final exam. A LO is considered to be mastered when a score of P[learned] = 0.6 or greater is achieved. The score on the final exam is taken as the average P[learned] on the LOs that are the leafs of the prerequisite graph (interpreted as the ultimate target concept, which in the real world might well be final exams).

There is still a cold start problem even after the simulation has been initialized with the interaction data from the SPP. This is because the simulated learners who are to follow the CFLS planner have not yet viewed any LOs themselves as they begin the course, so there is no history to match the $b$ LOs to create the plan. In this situation, the CFLS planner matches the learner with another arbitrary learner (from the interaction data from the SPP), and recommends whatever initial path that the other learner took when they first arrived in the course. While another solution to the cold start problem could be to start the new learner with the SPP, we did this to avoid any reliance whatsoever on knowing the underlying prerequisite structure.

The most computationally expensive part of the CFLS planner is finding the learners in the neighbourhood, which is at worst linear on the number of learners and linear on the amount of LO interaction history created by each learner. Each learner's LO interaction history must be searched to check for a match with $b$, with most learners being removed from the list during this process. The forward searching of $f$ is then executed using only the small resulting dataset.

## 4 Results

We ran the CFLS planner 25 different times with all pairings of the values of $b$ and $s$ ranging from 1 to 5, using a population of 65 simulated learners. This population had the same distribution of aptitudes as the population used to generate the baseline interaction data described above. The heat maps in Figs. 1 and 2 show the measurements for each of the 25 simulations, for each aptitude group, with the highest relative scores coloured red, mid-range scores coloured white, and the lowest scores coloured blue. In general, simulated learners achieved higher scores when following the CFLS planner than when given LOs randomly. The CFLS planner even exceeded the SPP in many cases.

A success triangle is visible in the lower left of each aptitude group. The success triangles can be interpreted to mean that if a path is going to be recommended, never send the learner any further ahead ($s$) than you have matched them in the past ($b$). For example if a learner's neighbourhood was created using their $b = 2$ most recent LOs, then never make the learner follow in a neighbour's steps further than $s = 2$ LOs. One reason for the eventual drop at high values of $b$ is that no neighbour could be found and a random match is used instead. However, the abrupt drop at $b > s$ was unexpected. To be sure the pattern was real, an extended series of simulations was run. We ran $b = 6$ and $s = 5$ to see

if there would be a drastic drop in performance, and indeed this was the case. We also ran another row varying $b$ with a fixed $s = 6$, and again found a drop at $b = 7$.

| LOW | | | | |
|---|---|---|---|---|
| b=1 s=1 | b=2 s=1 | b=3 s=1 | b=4 s=1 | b=5 s=1 |
| 100 | 21.9 | 32.5 | 31 | 37.7 |
| b=1 s=2 | b=2 s=2 | b=3 s=2 | b=4 s=2 | b=5 s=2 |
| 89.6 | 86 | 36.9 | 36.9 | 40.4 |
| b=1 s=3 | b=2 s=3 | b=3 s=3 | b=4 s=3 | b=5 s=3 |
| 72.1 | 68.6 | 62 | 43.21 | 40.1 |
| b=1 s=4 | b=2 s=4 | b=3 s=4 | b=4 s=4 | b=5 s=4 |
| 77.3 | 74.4 | 72.1 | 66.1 | 49.3 |
| b=1 s=5 | b=2 s=5 | b=3 s=5 | b=4 s=5 | b=5 s=5 |
| 68.1 | 70.7 | 67.5 | 67.4 | 63.5 |

| MEDIUM | | | | |
|---|---|---|---|---|
| b=1 s=1 | b=2 s=1 | b=3 s=1 | b=4 s=1 | b=5 s=1 |
| 100 | 50 | 45.8 | 42.2 | 44.1 |
| b=1 s=2 | b=2 s=2 | b=3 s=2 | b=4 s=2 | b=5 s=2 |
| 100 | 100 | 42.9 | 40.3 | 38 |
| b=1 s=3 | b=2 s=3 | b=3 s=3 | b=4 s=3 | b=5 s=3 |
| 100 | 99.4 | 98.6 | 42.4 | 43 |
| b=1 s=4 | b=2 s=4 | b=3 s=4 | b=4 s=4 | b=5 s=4 |
| 100 | 99.3 | 99.5 | 99.4 | 50.8 |
| b=1 s=5 | b=2 s=5 | b=3 s=5 | b=4 s=5 | b=5 s=5 |
| 100 | 100 | 100 | 100 | 100 |

| HIGH | | | | |
|---|---|---|---|---|
| b=1 s=1 | b=2 s=1 | b=3 s=1 | b=4 s=1 | b=5 s=1 |
| 100 | 75 | 39.3 | 39.7 | 41.1 |
| b=1 s=2 | b=2 s=2 | b=3 s=2 | b=4 s=2 | b=5 s=2 |
| 100 | 100 | 41.7 | 34.9 | 40 |
| b=1 s=3 | b=2 s=3 | b=3 s=3 | b=4 s=3 | b=5 s=3 |
| 100 | 100 | 100 | 42.5 | 50 |
| b=1 s=4 | b=2 s=4 | b=3 s=4 | b=4 s=4 | b=5 s=4 |
| 100 | 100 | 100 | 100 | 61 |
| b=1 s=5 | b=2 s=5 | b=3 s=5 | b=4 s=5 | b=5 s=5 |
| 100 | 100 | 100 | 100 | 100 |

**Fig. 1.** Average % Learning Objects Mastered by aptitude group

| LOW | | | | |
|---|---|---|---|---|
| b=1 s=1 | b=2 s=1 | b=3 s=1 | b=4 s=1 | b=5 s=1 |
| 0.6587 | 0.1036 | 0.1314 | 0.1283 | 0.146 |
| b=1 s=2 | b=2 s=2 | b=3 s=2 | b=4 s=2 | b=5 s=2 |
| 0.5178 | 0.4387 | 0.1398 | 0.1248 | 0.1363 |
| b=1 s=3 | b=2 s=3 | b=3 s=3 | b=4 s=3 | b=5 s=3 |
| 0.4051 | 0.266 | 0.2256 | 0.1586 | 0.132 |
| b=1 s=4 | b=2 s=4 | b=3 s=4 | b=4 s=4 | b=5 s=4 |
| 0.4138 | 0.2984 | 0.3016 | 0.2755 | 0.176 |
| b=1 s=5 | b=2 s=5 | b=3 s=5 | b=4 s=5 | b=5 s=5 |
| 0.357 | 0.2884 | 0.2859 | 0.2679 | 0.2249 |

| MEDIUM | | | | |
|---|---|---|---|---|
| b=1 s=1 | b=2 s=1 | b=3 s=1 | b=4 s=1 | b=5 s=1 |
| 0.6894 | 0.1851 | 0.2105 | 0.2099 | 0.2425 |
| b=1 s=2 | b=2 s=2 | b=3 s=2 | b=4 s=2 | b=5 s=2 |
| 0.7004 | 0.698 | 0.2058 | 0.22 | 0.1972 |
| b=1 s=3 | b=2 s=3 | b=3 s=3 | b=4 s=3 | b=5 s=3 |
| 0.6942 | 0.6761 | 0.6715 | 0.1944 | 0.2152 |
| b=1 s=4 | b=2 s=4 | b=3 s=4 | b=4 s=4 | b=5 s=4 |
| 0.6931 | 0.6867 | 0.6874 | 0.6856 | 0.2292 |
| b=1 s=5 | b=2 s=5 | b=3 s=5 | b=4 s=5 | b=5 s=5 |
| 0.6912 | 0.6884 | 0.6924 | 0.6965 | 0.6899 |

| HIGH | | | | |
|---|---|---|---|---|
| b=1 s=1 | b=2 s=1 | b=3 s=1 | b=4 s=1 | b=5 s=1 |
| 0.7641 | 0.2514 | 0.2805 | 0.2866 | 0.2702 |
| b=1 s=2 | b=2 s=2 | b=3 s=2 | b=4 s=2 | b=5 s=2 |
| 0.77 | 0.7694 | 0.2673 | 0.2738 | 0.2748 |
| b=1 s=3 | b=2 s=3 | b=3 s=3 | b=4 s=3 | b=5 s=3 |
| 0.7653 | 0.7638 | 0.7727 | 0.3019 | 0.3097 |
| b=1 s=4 | b=2 s=4 | b=3 s=4 | b=4 s=4 | b=5 s=4 |
| 0.768 | 0.7697 | 0.7633 | 0.7697 | 0.3431 |
| b=1 s=5 | b=2 s=5 | b=3 s=5 | b=4 s=5 | b=5 s=5 |
| 0.7601 | 0.7612 | 0.7591 | 0.7644 | 0.7636 |

**Fig. 2.** Average Score on Final Exam (P[learned]) by aptitude group

A hot spot of successful combinations of $b$ and $s$ appeared for each aptitude group. For low aptitude learners, it was best to only match on the $b = 1$ most recent learning objects, and to follow the selected neighbour for only $s = 1$ LOs ahead before replanning. This combination of $b$ and $s$ is the only time when the CFLS planner outperformed the SPP for the low aptitude group. However, for the medium and high aptitude groups, the CFLS planner outperformed the SPP in all cases within the success triangle. Looking at final exam scores (Fig. 2), medium aptitude learners responded well to being matched with neighbours using $b = 1$ or 2 and sticking with the chosen neighbour for the same distance ahead. The high aptitude group responded very well to using neighbourhoods created with $b = 3$ and recommending paths of $s = 3$.

Within the success triangles, the rows and columns of Fig. 2 were checked to see if there existed an ideal $b$ for a given $s$, and vice versa. Wherever there appeared to be a large difference, Student's t-test was used to check for statistical significance. We are able to use paired t-tests because the simulated learners have exactly the same characteristics in all the simulation runs, the only difference being the order in which LOs were interacted with. For example, learner #3 always has *aptitude-of-learner* = .4, so, there is no difference in that learner

between simulation runs. We used a two-tailed t-test because it was not certain whether one distribution was going to be higher or lower than the other.

Looking along the rows, when $s$ is held the same, there are some cases where one value of $b$ is better than another. For the low aptitude group, for the most part the lower the $b$, the better. For the medium aptitude group, there were no significant advantages to changing $b$. For the high aptitude group, when $s = 3$, the t-test was used to check if $b = 3$ was significantly more advantageous than using $b = 2$. The measurements for Score on the Final Exam for the high aptitude learners were compared between both simulation results, ($b = 2$ and $s = 3$) and ($b = 3$ and $s = 3$). With N=19 learners in this group, the calculated p-value was 0.009, indeed a statistically significant difference.

Looking along the columns, when $b$ is held the same there was a case where increasing $s$, i.e. sticking to a longer plan ahead, was statistically advantageous. In the medium aptitude group, when $b = 1$ it was statistically better to use $s = 2$ than to use $s = 1$ with a p-value of 0.011. None of the increases of $s$ with the same $b$ were significant for the high aptitude group, and there were no increases for the low aptitude group.

## 5 Analysis and Future Work

Through simulation, we have shown that a CFLS planner can be "launched" from an environment that has been conditioned with interaction data from another planner, such as an SPP, and operate successfully using only learner usage data kept by the EA and not needing centralized metadata such as a prerequisite graph. This is one of the key requirements for DOELEs. Like biological evolution, the EA is harsh in that it observes how learners succeed or fail as various paths are tried. Successful paths for particular types of learners, regardless of whether they follow standard prerequisites, is the only criterion of success. New learners or new learning objects will find their niche - some paths will work for some learners but not for others, and this is discovered automatically through usage.

More experiments are needed to explore the many possibilities of the simulation environment. While this experiment was not a true test of a DOELE because new learners and LOs were not inserted, this can be readily explored in future work. New additions could be matched randomly a few times in order to build enough data in the EA, and then automatically incorporated into neighbourhood matches or into future plans.

Given the evaluation function that was selected, we found that planning ahead and sticking to the plan worked best for high aptitude learners and a reactive approach (planning ahead but sticking to the plan for only a short time) worked best for the low aptitude learners. Would a different pattern emerge if a different evaluation function were chosen? Would a different threshold for mastery than P[learned] $> 0.6$ make any difference? In future work, would it be worthwhile to break down the aptitude groups into six: very-high, high, medium-high, medium-low, low, and very-low? This may assist with more easily tuning the weights of the evaluation function, as there was not much difference in our

results between the high and medium aptitude groups. In addition, more experiments where $s < f$ are needed to answer the question of whether the drop along the edge of each success triangle was because of $s$ or $f$. Also, in this work we did not look at the many different types of pedagogical interactions (ex. asking the student a question, giving a hint etc.) and focused on very abstract representations. More work is needed to explore this approach on systems later in the design process, when more detail about the content and the desired interactions with learners is known.

Future work could also investigate the usage of a differential planner, where different settings are tuned for different situations. For example, when creating a neighbourhood for a low aptitude learner, medium aptitude learners could be allowed into the neighbourhood if they have a matching $b$. Results could reveal situations where for example a low aptitude learner is helped by following in the steps of a medium aptitude learner. A differential planner could also dynamically choose the values of $b$ and $s$ for a given individual instead of using the same values for everyone at all times. For example, in a real world setting a CFLS planner may try to create a plan using a neighbourhood of $b = 3$, knowing it is optimal, but if for the specific case there is not enough data, it could change to $b = 2$ on the fly. Other aspects that could be changed are the criteria for creating the neighbourhood: rather than filtering by aptitude, another attribute could be chosen such as click behaviour or learning goals.

## 6 Conclusion

In this paper, we have described the need for instructional planning in DOELEs with many LOs aimed at large numbers of learners. Instructional planners such as [13] use AI planning technology that is based on states, actions and events, which are difficult to infer from an unstructured online environment. In recent years, instructional planning has been replaced by instructional design approaches such as [3]. Advanced instructional planners from the 1990s, such as PEPE and TOBIE [16] can blend different teaching strategies to appropriate situations. We have shown that instructional planning can still be done in the less rigid courses envisioned by the EA architecture and likely to be commonplace in the future, using only learner usage data kept by the EA and not needing centralized metadata about the course.

We have shown a specific planning technique, the CFLS planner, that is appropriate for DOELEs, and how to experiment in this domain. The simulation experiment revealed the number of LOs from a target learner's recent browsing history should be used for creating a neighbourhood ($b$), a question that has also been investigated by other researchers, such as in [18]. We have also found recommendations for settings for how far ahead to plan ($s$ and $f$) for different groups of learners, and identified questions for future work. As is the case with collaborative filtering and case-based approaches, the quality of the plans created is limited to the quality of LOs within the repository and the quality

of interactions that have previously occurred between learners and sequences of LOs.

The bottom-up discovery of prerequisite relationships has been investigated by others, such as [17]. When the need for centralized metadata about a course is discarded, and when the further step is taken that different paths can be found to work better for different learners, then a shift in thinking occurs. Each individual learner could effectively have a unique ideal (implicit) prerequisite graph. Whether or not a prerequisite relationship even exists between two LOs could vary from learner to learner. The notion of prerequisite can thus be viewed not only as a function of the content relationships, but also as a function of the individual learner.

Making recommendations of sequences has also been identified as a task in the recommender systems domain [9]. An approach such as a CFLS planner is a step in the direction of building recommender systems that can use sequence information to recommend sequences. This has also been accomplished with standards approaches such as [15]. Simulation with the EA provides another method for developing and testing such approaches.

Overall, the research we have done to date and the questions it raises, shows the value of exploring these complex issues using simulation. We were able to essentially generate some 25 different experiments exploring some issues in instructional planning, in a very short time when compared to what it would have taken to explore these same issues with real learners. Others have also used simulation for developing an educational planner, such as [10] for social assessment games. To be sure our simulation model was of low fidelity, but we suspect that there are some properties of the CFLS planner that we have uncovered that apply in the real world (the lower triangles seem to be very strong and consistent patterns). And, there are some very real issues that we can explore fairly quickly going forward that might reveal other strong patterns, as discussed. We believe that it isn't always necessary to have simulations with high cognitive fidelity (as in SimStudent [12]) to find out interesting things. Low fidelity simulations such as the ones we have used in this and our earlier work [6] (and those of [2]) have a role to play in AIED. Especially as we move into the huge questions of dynamic open-ended learning environments with thousands of learners and big privacy issues, the sharp minimalist modelling possible with low fidelity simulation should allow quick and safe experimentation without putting too many real learners at risk and without taking years to gain insights.

**Acknowledgements**
We would like to thank the Natural Sciences and Engineering Research Council of Canada for funding some aspects of this research.

## References

[1] Cazella, S., Reategui, E., and Behar, P.: Recommendation of Learning Objects Applying Collaborative Filtering and Competencies. IFIP Advances in Information and Communication Technology, 324, pp 35-43 (2010)

[2] Champaign, J.: Peer-Based Intelligent Tutoring Systems: A Corpus-Oriented Approach. Ph.D. Thesis, University of Waterloo, Waterloo, Canada (2012)
[3] Drachsler, H., Hummel, H. and Koper, R.: Using Simulations to Evaluate the Effects of Recommender Systems for Learners in Informal Learning Networks. SIRTEL Workshop (Social Information Retrieval for Technology Enhanced Learning) at the 3rd EC-TEL (European Conf. on Technology Enhanced Learning) Maastricht, The Netherlands: CEUR-WS.org, online CEUR-WS.org/Vol-382/paper2.pdf (2008)
[4] Desmarais, M., and Pelczer, I.: On the Faithfulness of Simulated Student Performance Data. In de Baker, R.S.J. et al. (Eds.), Proc. of the 3rd Int. Conf. on Educ. Data Mining, pp 21-30. Pittsburg USA (2010)
[5] Elorriaga, J. and Fernández-Castro, I.: Using Case-Based Reasoning in Instructional Planning: Towards a Hybrid Self-improving Instructional Planner. Int. Journal of Artificial Intelligence in Educ., 11(4), pp 416-449 (2000)
[6] Erickson, G., Frost, S., Bateman, S., and McCalla, G.: Using the Ecological Approach to Create Simulations of Learning Environments. In Lane, H.C. et al. (Eds), Proc. of the 16th Int. Con. on AIED, pp 411-420. Memphis USA: Springer (2013)
[7] Garrido, A. and Onaindia, E.: Assembling Learning Objects for Personalized Learning: An AI Planning Perspective. Intelligent Systems, IEEE, 28(2), pp 64-73 March/April (2013)
[8] Hannafin, M.J.: Learning in Open-Ended Environments: Assumptions, Methods and Implications. Educational Technology, 34(8), pp 48-55 (1994)
[9] Herlocker, J., Konstan, J., Terveen, L., and Riedl, J.: Evaluating Collaborative Filtering Recommender Systems. ACM Transactions on Information Systems (TOIS) 22(1), pp 5-53 (2004)
[10] Laberge, S., Lenihan, T., Shabani, S., and Lin, F.: Multiagent Coordination for Planning and Enacting an Assessment Game. Workshop on MultiAgent System Based Learning Environments of Int. Tutoring Systems (ITS) Honolulu, USA (2014)
[11] Land, S.: Cognitive Requirements for Learning with Open-Ended Learning Environments. Deuce. Technology Research and Development, 48(3), pp 61-78 (2000)
[12] Matsuda, N., Cohen, W. and Koedinger, K.: Teaching the Teacher: Tutoring SimStudent Leads to More Effective Cognitive Tutor Authoring. Int. Journal of Artificial Intelligence in Educ., 25(1), pp 1-34 (2014)
[13] Matsuda, N., and VanLehn, K.: Decision Theoretic Instructional Planner for Intelligent Tutoring Systems. In B. du Boulay (Ed.), Workshop Proc. on Modelling Human Teaching Tactics and Strategies, ITS 2000 pp 72-83. (2000)
[14] McCalla, G.: The Ecological Approach to the Design of e-Learning Environments: Purpose-based Capture and Use of Information about Learners. Journal of Interactive Media in Educ., http://jime.open.ac.uk/jime/article/view/2004-7-mccalla (2004)
[15] Shen, L. and Shen, R.: Learning Content Recommendation Service Based on Simple Sequencing Specification. In Liu W et al. (Eds.) Advances in Web-Based Learning - ICWL 2004 3rd Int. Conf. Web-based Learning, LNCS 3143, pp 363-370. Beijing, China:Springer (2004)
[16] Vassileva, J. and Wasson, B.: Instructional Planning Approaches: from Tutoring Towards Free Learning. Proceedings of Euro-AIED'96, Lisbon, Portugal (1996)
[17] Vuong, A., Nixon, T., and Towle, B.: A Method for Finding Prerequisites Within a Curriculum. In Pechenizkiy, M. et al. (Eds.) , Proc. of the 4th Int. Con. on Educ. Data Mining, pp 211-216. Eindhoven, the Netherlands (2011)
[18] Zhang, Y., and Cao, J.: Personalized Recommendation Based on Behavior Sequence Similarity Measures. In Cao, L. et al. (Eds.) Int. Workshop on Behaviour and Social Informatics / Behaviour and Social Informatics and Computing (BSI/BSIC 2013), Gold Coast QLD Australia / Beijing China, LNCS 8178, pp 165-177 (2013)

# Exploring the Issues in Simulating a Semi-Structured Learning Environment: the SimGrad Doctoral Program Design

David Edgar K. Lelei and Gordon McCalla

ARIES Laboratory, Department of Computer Science, University of Saskatchewan
davidedgar.lelei@usask.ca and mccalla@cs.usask.ca

**Abstract.** The help seeking and social integration needs of learners in a semi-structured learning environment require specific support. The design and use of educational technology has the potential to meet these needs. One difficulty in the development of such support systems is in their validation because of the length of time required for adequate testing. This paper explores the use of a simulated learning environment and simulated learners as a way of studying design validation issues of such support systems. The semi-structured learning environment we are investigating is a graduate school, with a focus on the doctoral program. We present a description of the steps we have taken in developing a simulation of a doctoral program. In the process, we illustrate some of the challenges in the design and development of simulated learning environments. Lastly, the expected contributions and our research plans going forward are described.

**Keywords:** Simulated learners, Simulated learning environment, Agent-based simulation, Help seeking, Doctoral learners, Multi-agent system.

## 1 Introduction

Artificial Intelligence in Education (AIED) is one of the research fields whose focus is the use of technology to support learners of all ages and across all domains[1]. Although, one shortcoming of AIED research is the limited research attention that very dynamic and semi-structured domains, such as a graduate school, have received. There is little research that investigates how technology can be used to help connect learners (help seeker and potential help givers) in the graduate school domain. Consequently, there is a gap in our understanding of how such technology may mitigate graduate learners' attrition rates and time-to-degree. We have suggested the use of reciprocal recommender technology to assist in the identification of a suitable helper [1]. However, the nature of graduate school means that validation of any education system designed to be used in a semi-structured environment would take a long time (measured in years). This paper aims to address this challenge by exploring the use of

[1] http://iaied.org/about/

simulated learning environment and simulated learners as a potential way of validating educational technologies designed to support doctoral learners.

In this paper, we first describe the nature and the metrics used by interested stakeholders to measure the success or lack therefore of a doctoral program. Following this, we briefly discuss the uses of simulation as it relates to learning environment. We then introduce the research questions we are interested in answering using simulation. We go on to describe the architectural design of our simulation model. Further, we show how data about the 'real world' target domain is used to inform the parameters and initial conditions for the simulation model. This provides the model with a degree of fidelity. Throughout this model development process, we illustrate some of the challenges in the design and development of simulated learning environments. We conclude the paper with a discussion of the expected contributions and our research plans going forward.

## 2 Understanding Doctoral Program

Graduate school is a very dynamic and complex social learning environment. A doctoral program in particular is a dynamic, semi-structured, and complex learning environment. Most doctoral programs have some structure in the sense that there are three distinct stages that doctoral learners must go through: admission stage, coursework stage, and dissertation stage. While coursework stage is fairly structured, the dissertation stage is not. Further, the dissertation stage have various milestones that include: comprehensive exam, thesis proposal, research, writing, and dissertation defense. As time passes, learners move from one stage to the next and their academic and social goals change. There is need for self-directed learning and individual doctoral learners are responsible for their own learning pace and choice of what to learn especially in the dissertation stage.

The dynamic nature of the program ensures that there is constant change; there are new learners joining the program, other learners leaving the program either through graduation or deciding to drop out, and still other learners proceeding from one stage to the next. There are two key aspects that influences learners to decide whether to persist or drop out of a learning institution: academic and social integration [2], [3] which are impacted by learner's initial characteristics and experiences during their duration in the program. The various stages of the doctoral program (e.g., coursework) and learning resources can be seen as factors that directly influence the academic integration of a doctoral learner. Peers and instructors/supervisors can be viewed as supporting the social aspects of the doctoral program and hence, directly impact the social integration of doctoral learners. As time passes, doctoral learners continually interact with both the academic and social facets of the doctoral program. As a result, there is constant change in learners' commitment to their academic goal and the social sides of the learning institution

Time-to-degree, completion rates, and attrition rates are important factors influencing the perception and experience of graduate education by interested stakeholders [4], [5]. Research on doctoral attrition and time-to-completion indicates that on aver-

age, the attrition rate is between 30% and 60% [5]–[8]. Long times to completion and a high attrition rate are costly in terms of money to the funding institution and the learning institution; and in terms of time and effort to the graduate student(s) and supervisor(s) [8]. Lack of both academic and social integration (isolation) have been shown to affect graduate learners decision to persist [2], [3], [9]. Learners facing academic and social integration challenges should be enabled to engage in a community of peers to foster interaction and hence, encourage peer help and personalized collaboration [10]. Understanding the nature of learner-institution interactions that foster doctoral learners' persistence to degree is important to both the learning institution and its learners. We use simulation to achieve this feat.

Simulation is an established third way of exploring research questions in addition to qualitative and quantitative methods [11], [12]. VanLehn [13] has identified three main uses of simulation in learning environments: 1) to provide an environment for human teachers to practise their teaching approaches; 2) to provide an environment for testing different pedagogical instructional design efforts; 3) to provide simulated learners who can act as companions for human learners. Our research is mainly focused on the first and the second uses – to enable deep insight into the complex interaction of the factors affecting doctoral learners' attrition and time-to-degree leading to a better design of an educational system. Therefore, our research questions are formulated around investigations of how various factors influence time-to-degree, completion rates, and dropout rates of doctoral students. We are interested in answering the following research questions:

1. How does the number of classes (as a platform for social integration with peers – potential helpers) offered by a program(s) or taken by a learner, influence learners' time-to-degree and their propensity to persist or drop out?
2. How does the average class size (as basis of learners' social integration) attended by learners, impact learners' time-to-degree and their inclination to persist or drop out? What is the optimum class size?
3. How does the overall population size of the learners (a few learners vs many learners) influence learners' time-to-degree and their likelihood to persist or drop out?
4. Does timely help affects doctoral learners' time-to-degree and their decision to persist or drop out? If so, how?
5. How does the level of reciprocation influence the formation of a 'helpful community' of learners and adaptive help seeking behavior of the learners?

Use of simulation enables us to explore the aforementioned issues in a fine-grained controlled environment. For example, it would be almost impossible in the 'real world' setting to examine the impact of different number of course to take or class size to attend. Two cohorts of learners will have different attributes. Simulation allows us to tweak the number of courses or class size without touching the other characteristics of learners. Hence, we are able to see the real impact of one variable at a time. Before any exploration and insight can be gained on these issues, there is need to design and implement the simulation model.

## 3 Building an Initial Prototype of *SimGrad*

In this section we demonstrate the steps we have taken in the development of our initial prototype of our simulated doctoral learning environment: *SimGrad*. We show how a designer of an educational technology can develop a model of their target learning environment and inform its initial condition with available 'real world' data.

### 3.1 *SimGrad* Design

We need to design a simulation model by addressing two key challenges. First, we need to consider issues related to the modeling of the learning environment: how do we design conceptual and computational models of a doctoral program and what stakeholders should be included in these models? The second concern is about modeling of simulated learners: what doctoral learners' features affect persistence and time-to-degree, what factors do we model, and can we inform these features with available 'real world' data?

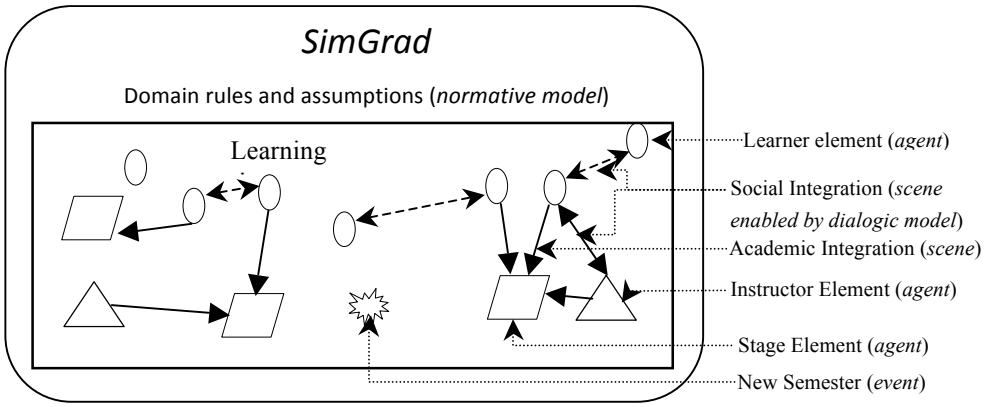



**Fig. 1.** SimGrad conceptual framework, its three elements, and the possible interaction between the elements

We have designed our conceptual model of the different aspects of simulated doctoral learners and doctoral learning environment based on the simulated learning environment specifications suggested by Koper et al. in [14], and features for building an electronic institution proposed by Esteva et al. [15]. We name our conceptual framework, *SimGrad*. Its core elements include: *normative model* - specifies requirements and constraints to guide agent actions and behavior; *dialogic model* – deals with interaction strategies and communication mechanism; *events* – refers to happenings in the model that trigger (re)action by agents; *scene* – description of an interaction between elements; *elements (agents)* – represent key stakeholders of the target domain that are modeled. Elements are modeled as agents. Each of the agents has attributes and behavior which are informed by our assumptions guided by our research questions and factors that influence learners' decision to persist. Every element of interest is to be

modeled within the learning environment and all possible interactions and operations within the learning setting is guided by domain rules represented by the normative model. See **Fig. 1**.

In our simulation model, we have chosen to model three types of elements: class, instructor, and learner. In this paper, in keeping with model simplicity, both the class and the instructor agents are passive while the learner agent is modeled to be active and reactive to its environment. Also, the only instructor's attributes we are interested in are related to classes (see **Table 1**). We modeled only one type of instructor agent. Another instructor type agent that can be modeled is the supervisor.

Each learner agent has the following properties: autonomy, social ability, reactivity, proactivity, and a degree of intentionality. We have also identified the following key attributes for our agent learner model: state – (busy, available), program, stages, course taken, peer interactions (pertaining challenges), academic integration, social integration, and motivation (see **Table 2**). In our model, peer interaction and state contribute to a learners' social integration, while research area, stage, course taken impact to their academic integration. Motivation combines both the social and academic integration and hence, is the main factor that determines whether an agent continues to persist or chooses to drop out of the program.

**Table 1.** Comparison of computed attributes of the three agent types

| *Attribute – data (value range)* | *Agent learner* | *Agent instructor* | *Agent class* |
|---|---|---|---|
| Total number of classes take, taught, or frequency of offering within 10 years – *numeric (0-20)* | X | X | X |
| Grade obtained, average awarded, or average obtained by learners – *numeric (0,12)* | X | X | X |
| Take classes from, teach classes in, or class offered in various programs – *textual (program id)* | X | X | X |
| Instructors teaching a class – *array list (instructor id)* | X | - | X |
| What is the class size – *numeric (1-5)* | X | | X |
| Number of classes taken or taught per year - *numeric (0,4)* | X | X | - |
| Which classes are taken or taught – *textual (class id)* | X | X | - |

The main intentions of each agent is to persist through doctoral requirements to graduation and to do so in a timely manner. However, each of these agents reacts to the different challenges at various stages of graduate school in divergent and autonomous ways. At the coursework stage, agents have the goal of taking courses that are relevant to their field and that they will perform well. When facing a course choice challenge or any other particular challenge, we have modeled our agents to proactively associate with peers to seek help. Each peer makes individual choice on whether to or not to respond to a request for help from others. The dialogic model

handles the agent to agent interaction and communication through a message passing mechanism [16].

**Table 2.** Attributes and parameters considered for an agent learner model for learners, their description and how each of them changes.

| *Attribute* | *Value - description* | *How it changes* |
|---|---|---|
| Enrolment | Date (MM/YYYY)<br>Indicate the month a year an agent enrolled in the program | Does not change |
| Graduation | Date (MM/YYYY)<br>Target graduation date | Evaluated whenever an agent completes a milestone |
| State | Textual (busy, available)<br>Indicates an agent availability to help others, assigned based on the smallest time unit model | Changes whenever an agent experiences a challenge |
| Program | Textual (program id)<br>Identify an agent's closer community within the larger community of learners | Does not change during a simulation run |
| Stage | Textual (admission, coursework, dissertation, timeout, dropout) | Admission stage is like an event. Learner move to the coursework immediately after admission. They more to dissertation after completing their course load. |
| Courses taken | Array [course, mark, instructor id](0-6)<br>Record courses taken by an agent and the marks obtain in each course | Every end of semester that the student took classes, this array is updated |
| Peer interaction | Array [learner id, challenge, result],<br>Keep track of an agent interactions with others and the outcome of the interaction | Changes whenever two agents interact |
| Academic integration | Numeric (-1,1)<br>Measures the academic satisfaction | Changes whenever an agent learner interacts with agent stage (i.e., completes a milestone or experience a challenge) |
| Social integration | Numeric (-1,1)<br>Measures a learners sense of belonging to the learning environment | Changes whenever an agent learner interacts with its peers or agent instructors |
| Motivation | Numeric (-1,1)<br>Measures the propensity of an agent to still want to persist. A motivation value above 0.3 indicates persistence. A value between -0.3 and 0.3 indicate help seeking needed. A value below -0.3 means the agent drops out | Whenever there is a change in the social and academic integration values. Its value is the average of the integration values. |

### 3.2 Informing *SimGrad* behavior and evaluation functions

Having identified the important agents and their key attributes, there are two sets of important functions for each element that need to be modelled: behaviour functions and evaluation functions [17]. Behaviour functions inform the decision making of the active elements and dictates the interaction patterns between them and the other modeled elements (e.g., how many classes a given agent takes). Evaluation functions indicate whether or not various interactions between the different agents in a simulation were successful (e.g., determine what grade a given agent attains in a class it took). Informing such functions with 'real world' data allows the simulation to behave in a way consistent with reality.

Simulation model fidelity is an issues that might arise when using simulation to study a target real world phenomenon. However, the most important issue to consider is the research question to be answered. While Champaign [18] used a very low fidelity model, Matsuda et al. [19] used a model with high cognitive fidelity to reach compelling conclusion. Further yet, Erickson et al. [17] also demonstrated that is possible to use a medium fidelity model and uncover interesting results. In some situations it might not be possible to have a high fidelity model because of lack of data. A case in point is our simulation scenario. Where possible, we inform our simulation functions with data received from the U of S on their doctoral program. An investigation into the U of S data showed that we will not be able to inform every aspect of our simulation model. It would be desirable to inform every initial aspects of our simulation model with 'real world' data but, we do not have data on the dissertation stage.

We are provided information on student id, years a student is registered, year of graduation (if graduated), student's program, classes taken and marks obtained, class instructor, and students instructional responsibilities. From this dataset we are able to inform the admission and coursework stages of our model (academic integration). However, there is no information concerning the dissertation stage and the social integration aspects. While it possible to inform various behaviour and evaluation functions for our simulation model, in this paper we focus on describing the steps we took to inform two functions of our simulation: learning environment admission behaviour function, and learners' class interactions behaviour function.

As already mentioned, admission is an important part of a doctoral program that contributes to it dynamic nature. The admission process is complex and involves a lot of stakeholders and processes, but we are concerned only with determining the year to year patterns in how many students are admitted. To provide some fidelity to our simulated learning environment admission, we analyzed data provided to us by the U of S University Data Warehouse[2]. The provided dataset contained information on doctoral learners registered in the 10 years 2005-2014. In this time there were 2291 doctoral learners with a total of 52850 data points on class registration. The 2005 registration included learners who had joined the program earlier than 2005. In order to get a clean admission pattern, we only considered learners who were registered from the year 2006 onwards. This reduced the population size to 1962.

[2] http://www.usask.ca/ict/services/ent-business-intelligence/university-data-warehouse.php

We were able to identify three admission periods, September, January, and May. We then obtained values for each of the admissions months for the years 2006-2014. This provided a distribution for each month that we used to generate a scatter plot of admission numbers. A sigmoidal pattern emerged. Next, we performed a non-linear curve fitting to the scatter plot so that the admission function can be represented in the form $Y = St*(c + x)$, where c is a constant, $St$ is a variable dependent on the admission period, and $x$ is the admission period. We then ran a regression to find values of each of these variables. This allowed us to model the admission patterns observed in the U of S dataset.

Next we derived the number of classes taken. To introduce some realism to the number classes taken behaviour, we had to further prune the data. We only considered data for students whose cohorts would have been registered for at least 3 years by the end of the year 2014 and hence, we considered class taking behaviour of 1466 U of S doctoral learners.

We obtained the number of classes each of the remaining learners we registered in and created a histogram. This histogram showed us the distribution of the number of students registered for a certain number of classes. Next, we transformed this distribution graph into a cumulative distribution function. We then took an inverse of the cumulative distribution function to achieve a quantile function. The quantile function, when run over many learners, assigns learners a class count that mimics the initial histogram. We use this quantile function to inform the number of classes a learner can take.

In this section we have described the importance of informing a simulation model with 'real world' data. We have described two functions that are informed with U of S dataset. Other examples of functions that can be informed using the U of S dataset include: class performance evaluation function, dropout behaviour function, time to degree behaviour function, and flow through behavior function (main as pertains to coursework stage). We have identified that missing data values is a major hindrance in this endeavor. There are possible ways of informing simulation attributes where there are no 'real world' data to derive from. A designer can either assign common sense values, generate and assign random values, or refer to the research literature to identify patterns that have been found by other researchers. Since we have the enrolment dates and the graduate dates for learners who graduate, we choose to derive common sense values with these two dates guiding the process and the value range.

## 4 Discussion, Expected Contributions, and Future Research Plans

Despite the growth in the use of simulation as a method for exploration and learning in many areas such as: engineering, nursing, medicine [20], and building design [21], research in the used of simulation within AIED is still at an early stage. There is need for more research to demonstrate that the outputs of simulation runs are desirable and informative to the AIED community. In this paper, we aim at contributing to this notion and by promoting the use of simulation in educational research and presenting

an agent based simulation conceptual framework for building simulated learning environment, with a focus on the semi-structured ones. Simulated learning environment and simulated learners are important in exploring and understanding a given learning domain. Further, it helps with the generation of system validation data.

The expected contributions to AIED include: providing a conceptual framework for simulated graduate school learning environment – an architecture that enables investigations into factors affecting doctoral learners progress through their program; shedding light on learner modeling issues in dynamic learning environments; and demonstrating the importance of simulation in exploring various AIED research domains, particularly semi-structured domains.

Current research work is focused on the implementation of the simulation model and the refinement of the various behaviour and evaluation functions. Once the implementation is done, we will validate our model against the dataset we have from the U of S before proceeding to explore the impact of various environmental factors. Since we are informing the simulation with both common sense assumptions and U of S dataset, the goal is to tweak the common sense assumptions such that when the model is run we get similar results as the U of S data in terms of class performance, dropout rate, and time-to-degree. Achieving this, would give us confidence that we have captured reality in some measurable way. We can then start exploring the various impact of measures we are interested in examining. As earlier indicated, we are interested in exploring the interactions of a number of variables: number of classes taken which will impact the availability of potential peer helpers, the effect of reciprocity on help seeking and help giving, and the effect of help seeking and other factors on doctoral learners' time-to-degree and attrition rates.

## Acknowledgement

We would like to thank University of Saskatchewan University Data Warehouse team for giving us access to 10 year dataset. Specifically, we would like to thank Mathew Zip for processing and explaining to the first author the nature of the dataset. We also wish to acknowledge and thank the Natural Science and Engineering Research Council of Canada (NSERC) for funding our research.

## References

[1] D. E. K. Lelei, "Supporting Lifelong Learning: Recommending Personalized Sources of Assistance to Graduate Students," in *Artificial Intelligence in Education*, 2013, pp. 912–915.

[2] V. Tinto, "Taking student success seriously: Rethinking the first year of college.," *Ninth Annu. Intersession Acad. Aff. Forum,* vol. 19, no. 2, pp. 1–8, 2005.

[3] V. Tinto, "Dropout from higher education: A theoretical synthesis of recent research," *Rev. Educ. Res.*, vol. 45, no. 1, pp. 89–125, 1975.

[4] H. Groenvynck, K. Vandevelde, and R. Van Rossem, "The Ph.D. Track: Who Succeeds, Who Drops Out?," *Res. Eval.*, vol. 22, no. 4, pp. 199–209, 2013.

[5] F. J. Elgar, “Phd Degree Completion in Canadian Universities,” Nova Scotia, Canada, 2003.
[6] M. Walpole, N. W. Burton, K. Kanyi, and A. Jackenthal, “Selecting Successful Graduate Students: In-Depth Interviews With GRE Users,” Princeton, NJ, 2002.
[7] L. Declou, “Linking Levels to Understand Graduate Student Attrition in Canada,” McMaster University, Hamilton, Ontario, Canada, 2014.
[8] B. E. Lovitts, *Leaving the Ivory Tower: The Causes and Consequences of Departure from Doctoral Study*, Illustrate., vol. 32. Rowman & Littlefield Publishers, 2001, p. 307.
[9] A. Ali and F. Kohun, “Dealing with isolation feelings in IS doctoral programs,” *Int. J. Dr. Stud.*, vol. 1, no. 1, pp. 21–33, 2006.
[10] Computing Research Association, “Grand research challenges in information systems,” Washington, D.C, 2003.
[11] R. Axelrod, “Advancing the Art of Simulation in the Social Sciences,” in *Proceedings of the 18th European Meeting on Cybernetics and Systems Research*, 2006, pp. 1–13.
[12] N. Gilbert and K. G. Troitzsch, “Simulation and Social Science,” in *Simulation for the Social Scientist*, McGraw-Hill International, 2005, pp. 1–14.
[13] K. VanLehn, S. Ohlsson, and R. Nason, “Applications of Simulated Students: An Exploration,” *J. Artif. Intell. Educ.*, vol. 5, no. 2, pp. 1–42, 1994.
[14] R. Koper and B. Olivier, “Representing the Learning Design of Units of Learning,” *Educ. Technol. Soc.*, vol. 7, no. 3, pp. 97–111, 2003.
[15] M. Esteva, J.-A. Rodriquez-Aguilar, C. Sierra, P. Garcia, and J. L. Arcos, “On the Formal Specification of Electronic Institutions,” in *Agent Mediated Eectronic Commerce*, 2001, pp. 126–147.
[16] H. J. C. Berendsen, D. Van Der Spoel, and R. Van Drunen, “GROMACS: A message-passing parallel molecular dynamics implementation,” *Comput. Phys. Commun.*, vol. 91, no. 1, pp. 43–56, 1995.
[17] G. Erickson, S. Frost, S. Bateman, and G. Mccalla, “Using the Ecological Approach to Create Simulations of Learning Environments,” in *In Artificial Intelligence in Education*, 2013, pp. 411–420.
[18] J. Champaign, “Peer-Based Intelligent Tutoring Systems: A Corpus-Oriented Approach,” University of Waterloo, 2012.
[19] N. Matsuda, W. W. Cohen, K. R. Koedinger, V. Keiser, R. Raizada, E. Yarzebinski, S. P. Watson, and G. Stylianides, “Studying the Effect of Tutor Learning Using a Teachable Agent that Asks the Student Tutor for Explanations,” in *2012 IEEE Fourth International Conference On Digital Game And Intelligent Toy Enhanced Learning*, 2012, pp. 25–32.
[20] A. L. Baylor and Y. Kim, “Simulating Instructional Roles Through Pedagogical Agents,” *Int. J. Artif. Intell. Educ.*, vol. 15, no. 2, pp. 95–115, 2005.
[21] G. Augenbroe, “Trends in Building Simulation,” *Build. Environ.*, vol. 37, no. 8, pp. 891–902, 2002.

# Exploring the Role of Small Differences in Predictive Accuracy using Simulated Data

Juraj Nižnan, Jan Papoušek, and Radek Pelánek

Faculty of Informatics, Masaryk University Brno
{niznan,jan.papousek,xpelanek}@mail.muni.cz

**Abstract.** Research in student modeling often leads to only small improvements in predictive accuracy of models. The importance of such improvements is often hard to assess and has been a frequent subject of discussions in student modeling community. In this work we use simulated students to study the role of small differences in predictive accuracy. We study the impact of such differences on behavior of adaptive educational systems and relation to interpretation of model parameters. We also point out a feedback loop between student models and data used for their evaluation and show how this feedback loop may mask important differences between models.

## 1 Introduction

In student modeling we mostly evaluate models based on the quality of their predictions of student answers as expressed by some performance metric. Results of evaluation often lead to small differences in predictive accuracy, which leads some researchers to question the importance of model improvements and meaningfulness of such results [1]. Aim of this paper is to explore the impact and meaning of small differences in predictive accuracy with the use simulated data. For our discussion and experiments in this work we use a single performance metric – Root Mean Square Error (RMSE), which is a common choice (for rationale and overview of other possible metrics see [15]). The studied questions and overall approach are not specific to this metric.

Simulated students provide a good way to study methodological issues in student modeling. When we work with real data, we can use only proxy methods (e.g., metrics like RMSE) to evaluate quality of models. With simulated data we know the "ground truth" so we can study the link between metrics and the true quality of models. This enables us to obtain interesting insight which may be useful for interpretation of results over real data and for devising experiments. Similar issues are studied and explored using simulation in the field of recommender systems [7, 17].

We use a simple setting for simulated experiments, which is based on an abstraction of a real system for learning geography [12]. We simulate an adaptive question answering system, where we assume items with normally distributed difficulties, students with normally distributed skills, and probability of correct

answer given by a logistic function of the difference between skill and difficulty (variant of a Rasch model). We use this setting to study several interrelated question.

### 1.1 Impact on Student Practice

*What is the impact of prediction accuracy (as measured by RMSE) on the behavior of an adaptive educational system and students' learning experience?*

Impact of small differences in predictive performance on student under-practice and over-practice (7-20%) has been demonstrated using real student data [18], but insight from a single study is limited. The relation of RMSE to practical system behavior has been analyzed also in the field of recommender systems [2] (using offline analysis of real data). This issue has been studied before using simulated data is several studies [5, 6, 10, 13]. All of these studies use very similar setting – they use Bayesian Knowledge Tracing (BKT) or its extensions and their focus is on mastery learning and student under-practice and over-practice. They differ only in specific aspects, e.g., focus on setting thresholds for mastery learning [5] or relation of moment of learning to performance metrics [13]. In our previous work [16] have performed similar kind of simulated experiments (analysis of under-practice and over-practice) both with BKT and with student models using logistic function and continuous skill.

In this work we complement these studies by performing simulated experiments in slightly different setting. Instead of using BKT and mastery learning, we use (variants of) the Rasch model and adaptive question answering setting. We study different models and the relation between their prediction accuracy and the set of items used by the system.

### 1.2 Prediction Accuracy and Model Parameters

*Can RMSE be used to identify good model parameters? What is the relation of RMSE to the quality of model parameters?*

In student modeling we often want to use interpretable models since we are interested not only in predictions of future answers, but also in reconstructing properties of students and educational domains. Such outputs can be used to improve educational systems as was done for example by Koedinger at al. [9]. When model evaluation shows that model A achieves better prediction accuracy (RMSE) then model B, results are often interpreted as evidence that model A better reflects "reality". Is RMSE a suitable way to find robust parameters? What differences in metric value are meaningful, i.e., when we can be reasonably sure that the better model really models reality in better way? Is statistical significance of differences enough? In case of real data it is hard to answer these question since we have no direct way to evaluate the relation of a model to reality. However, we can study these questions with simulated data, where we have access to the ground truth parameters. Specifically, in our experiments we study the relation of metric values with the accuracy of reconstructing the mapping between items and knowledge components.

### 1.3 Feedback between Data Collection and Evaluation

*Can the feedback loop between student models and adaptive choice of items influence evaluation of student models?*

We also propose novel use of simulated students to study a feedback loop between student models and data collection. The data that are used for model evaluation are often collected by a system which uses some student model for adaptive choice of items. The same model is often used for data collection and during model evaluation. Such evaluation may be biased – it can happen that the used model does not collect data that would show its deficiencies. Note that the presence of this feedback loop is an important difference compared to other forecasting domains. For example in weather forecasting models do not directly influence the system and cannot distort collected data. In student modeling they can.

So far this feedback has not been thoroughly studied in student modeling. Some issues related to this feedback have been discussed in previous work on learning curves [6, 11, 8]. When a tutoring system uses mastery learning, students with high skill drop out earlier from the system (and thus from the collected data), thus a straightforward interpretation of aggregated learning curves may be misleading. In this work we report experiment with simulated data which illustrate possible impact of this feedback loop on model evaluation.

## 2 Methodology

For our experiments we use a simulation of a simplified version of an adaptive question answering systems, inspired by our widely used application for learning geography [12]. Fig. 1 presents the overall setting of our experiments. System asks students about items, answers are dichotomous (correct/incorrect), each student answers each item at most once. System tries to present items of suitable difficulty. In evaluation we study both the prediction accuracy of models and also sets of used items. This setting is closely related to item response theory and computerized adaptive testing, specifically to simulated experiments with Elo-type algorithm reported by Doebler et al. [3].

**Simulated Students and Items** We consider a set of simulated students and simulated items. To generate student answers we use logistic function (basically the Rasch model, respectively one parameter model from item response theory): $P(correct|\theta_s, d_i) = 1/(1 + e^{-(\theta_s - d_i)})$, where $\theta_s$ is the skill of a student $s$ and $d_i$ is difficulty of an item $i$.

To make the simulated scenarios more interesting we also consider multiple knowledge components. Items are divided into disjoint knowledge components and students have different skill for each knowledge component. Student skills and item difficulties are sampled from a normal distribution. Skills for individual knowledge components are independent from one another.

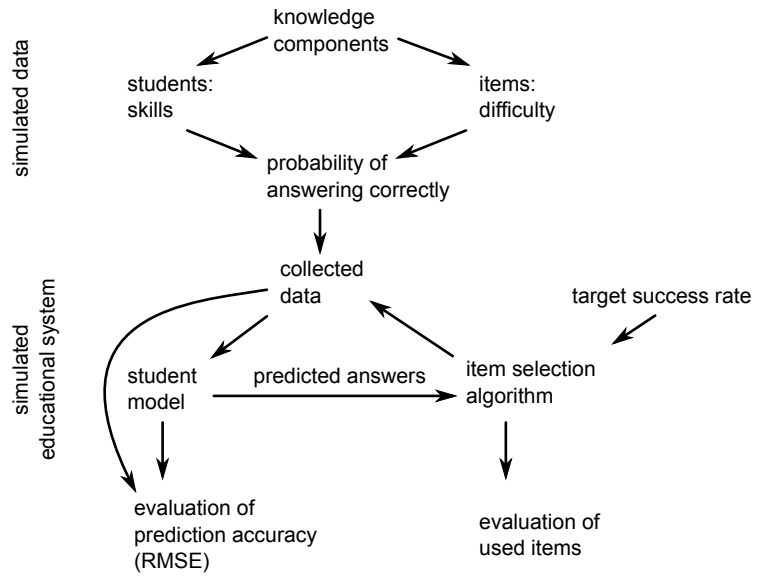



**Fig. 1.** Setting of our experiments

**Item Selection Algorithm** The item selection algorithm has as a parameter a target success rate $t$. It repeatedly presents items to a (simulated) student, in each step it selects an item which has the best score with respect to the distance of the predicted probability of correct answer $p$ and the target rate $t$ (illustrated by gray dashed line in Fig. 3). If there are multiple items with the same score, the algorithm randomly selects one of them.

**Student Models** Predictions used by the item selection algorithm are provided by a student model. For comparison we consider several simple student models:

- Optimal model – Predicts the exact probability that is used to generate the answer (i.e., a "cheating" model that has access to the ground truth student skill and item difficulty).
- Optimal with noise – Optimal model with added (Gaussian) noise to the difference $\theta_s - d_i$ (before we apply logistic function).
- Constant model – For all students and items it provides the same prediction (i.e., with this model the item selection algorithm selects items randomly).
- Naive model – Predicts the average accuracy for each item.
- Elo model – The Elo rating system [4, 14] with single skill. The used model corresponds to the version of the system as described in [12] (with slightly modified uncertainty function).
- Elo concepts – The Elo system with multiple skills with correct mapping of items to knowledge components.

– Elo wrong concepts – The Elo system with multiple skills with wrong mapping of items to knowledge components. The wrong mapping is the same as the correct one, but 50 (randomly choosen) items are classified incorrectly.

**Data** We generated 5,000 students and 200 items. Items are divided into 2 knowledge components, each user has 2 skills corresponding to the knowledge components and each item has a difficulty. Both skills and difficulties were sampled from standard normal distribution (the data collected from the geography application suggests that these parameters are approximately normally distributed). The number of items in a practice session is set to 50 unless otherwise noted.

## 3 Experiments

We report three types of experiments, which correspond to the three types of questions mentioned in the introduction.

### 3.1 Impact on Student Practice

Our first set of experiments studies differences in the behavior of the simulated system for different models. For the evaluation of model impact we compare the sets of items selected by the item selection algorithm. We make the assumption that the algorithm for item selection using the optimal model generates also the optimal practice for students. For each user we simulate practice of 50 items (each item is practiced at most once by each student). To compare the set of practiced items between those generated by the optimal model and other models we look at the size of the intersection. We assume that bigger intersection with the set of practiced items using the optimal model indicates better practice. Since the intersection is computed per user, we take the mean.

This is, of course, only a simplified measure of item quality. It is possible that an alternative model selects completely different set of items (i.e., the intersection with the optimal set is empty) and yet the items are very similar and their pedagogical contribution is nearly the same. However, for the current work this is not probable since we are choosing 50 items from a pool of only 200 items. For future work it would be interesting to try to formalize and study the "utility" of items.

**Noise Experiment** The optimal model with noise allows us to easily manipulate differences in predictive accuracy and study their impact on system behavior. Experiment reported in the left side of Fig. 2 shows both the predictive accuracy (measured by RMSE) and the impact on system behavior (measured by the size of the intersection with the optimal practiced set as described above) depending on the size of noise (we use Gaussian noise with a specified standard deviation). The impact of noise on RMSE is approximately quadratic and has a slow rise –

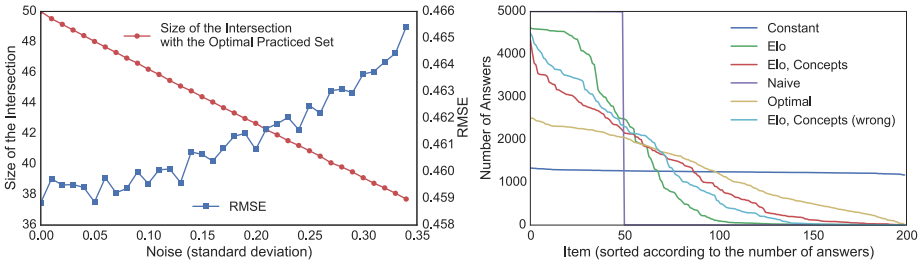



**Fig. 2.** Size of the intersection with the optimal practiced set of items and RMSE depending on Guassian noise in optimal model (left side). Distribution of answers over the items based on the given model (right side).

this is a direct consequence of the quadratic nature of the metric. The impact on used items is, however, approximately linear and rather steep. The most interesting part is for noise values in the interval $[0, 0.1]$. In this interval the rise in RMSE values is very small and unstable, but the impact on used items is already high.

**Model Comparison** Right side of the Fig. 2 shows the distribution of the number of answers per item for different models. The used models have similar predictive accuracy (specific values depend on what data we use for their evaluation, as discussed below in Section 3.3), yet the used model can dramatically change the form of the collected data.

When we use the optimal model, the collected data set covers almost fairly most items from the item pool. In the case of worse models the use of items is skewed (some items are used much more frequently than others). Obvious exception is the constant model for which the practice is completely random. The size of the intersection with the optimal practiced set for these models is – Constant: 12.5; Elo: 24.2; Elo, Concepts: 30.4; Elo, Concepts (wrong): 28.5; Naive: 12.0. Fig. 3 presents a distribution of answers according to the true probability of their correctness (given by the optimal model). Again there is a huge difference among the given models, especially between simple models and those based on Elo.

### 3.2 Prediction Accuracy and Model Parameters

Metrics of prediction accuracy (e.g., RMSE) are often used for model selection. Model that achieves lower RMSE is assumed to have better parameters (or more generally better "correspondence to reality"). Parameters of a selected model are often interpreted or taken into account in improvement of educational systems. We checked validity of this approach using experiments with knowledge components.

We take several models with different (random) mappings of items to knowledge components and evaluate their predictive accuracy. We also measure the

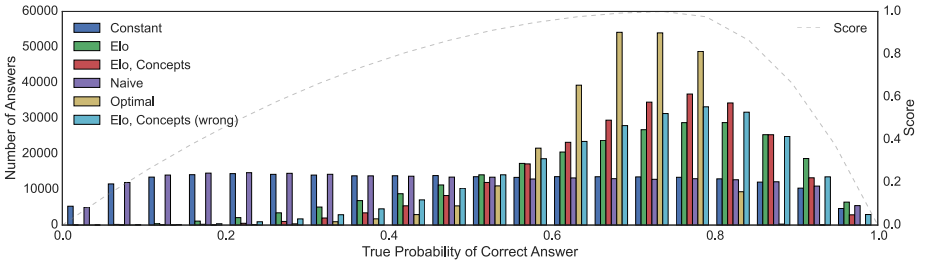



**Fig. 3.** Distribution of answers according to the true probability of correct answer. The gray dashed line stands for the score function used by the algorithm for item selection.

quality of the used mappings – since we use simulated data, we know the ground truth mapping and thus can directly measure the quality of each mapping. Quality is expressed as the portion of items for which the mapping agrees with the ground truth mapping. The names of the knowledge components are irrelevant in this setting. Therefore, we compute quality for each one-to-one mapping from the names of the components in the model to the names of the components in the ground truth. We select the highest quality as the quality of the model's item-to-component mapping. To focus only on quality of knowledge components, we simplify other aspects of evaluation, specifically each student answers all items and their order is selected randomly.

These experiments do not show any specific surprising result, so we provide only general summary. Experiments show that RMSE values correlate well with the quality of mappings. In case of small RMSE differences there may be "swaps", i.e., a model with slightly higher RMSE reflects reality slightly better. But such results occur only with insufficiently large data and are unstable. Whenever the differences in RMSE are statistically significant (as determined by t-test over different test sets), even very small differences in RMSE correspond to improvement in the quality of the used mappings. These results thus confirm that it is valid (at least in the studied setting) to argue that a model A better corresponds to reality than a model B based on the fact that the model A achieves better RMSE than the model B (as long as the difference is statistically significant). It may be useful to perform this kind of analysis for different settings and different performance metrics.

### 3.3 Feedback between Data Collection and Evaluation

To study feedback between the used student model and collected data (as is described in subsection 1.3) we performed the following experiment: We choose one student model and use it as an input for adaptive choice of items. At the same time we let all other models do predictions as well and log answers together with all predictions.

Fig. 4 shows the resulting RMSE for each model in individual runs (data collected using specific model). The figure shows several interesting results. When

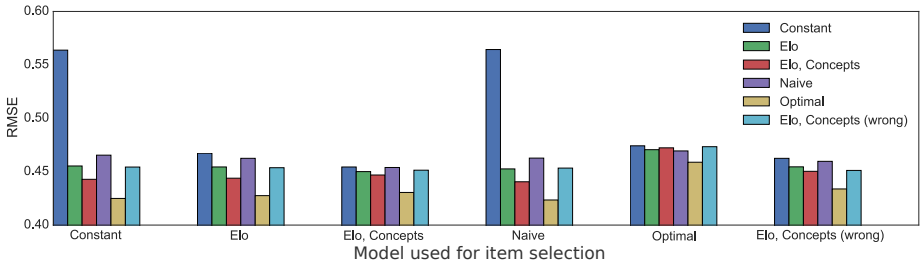



**Fig. 4.** RMSE comparison over data collected using different models.

the data are collected using the optimal model, the RMSE values are largest and closest together; even the ordering of models is different from other cases. In this case even the constant model provides comparable performance to other models – but it would be very wrong to conclude that "predictive accuracy of models is so similar that the choice of model does not matter". As the above presented analysis shows, different models lead to very different choice of items and consequently to different student experience. The reason for small differences in RMSE is not similarity between models, but characteristics of data ("good choice of suitable items"), which make predictions difficult and even a naive predictor comparatively good.

Another observation concerns comparison between the "Elo concepts" and "Elo concepts (wrong)" models. When data are collected by the "Elo concepts (wrong)" model, these two models achieve nearly the same performance, i.e., models seem to be of the same quality. But the other cases show that the "Elo concepts" model is better (and in fact it is by construction a better student model).

## 4 Conclusions

We have used simulated data to show that even small differences in predictive accuracy of student models (as measured by RMSE) may have important impact on behavior of adaptive educational systems and for interpretation of results of evaluation. Experiments with simulated data, of course, cannot demonstrate the practical impact of such small differences. We also do not claim that small differences in predictive accuracy are always important. However, experiments with simulated data are definitely useful, because they clearly illustrate mechanisms that could play role in interpretation of results of experiments with real student data. Simulated data also provide setting for formulation of hypotheses that could be later evaluated in experiments with real educational systems.

Simulated data also enable us to perform experiments that are not practical for realization with actual educational systems. For example in our experiment with the "feedback loop" we have used different student models as a basis for item selection. Our set of models includes even a very simple "constant model",

which leads to random selection of practiced item. In real setting we would be reluctant to apply such a model, as it is in contrary with the advertised intelligent behavior of our educational systems. However, experiments with this model in simulated setting provide interesting results – they clearly demonstrate that differences in predictive accuracy of models do not depend only on the intrinsic quality of used student models, but also on the way the data were collected.

Our analysis shows one particularly interesting aspect of student modeling. As we improve student models applied in educational systems, we should expect that evaluations of predictive accuracy performed over these data will show worse absolute values of performance metrics and smaller and smaller differences between models (even if models are significantly different), just because virtues of our models enable us to collect less predictable data.

## References


1. Joseph E Beck and Xiaolu Xiong. Limits to accuracy: How well can we do at student modeling. In *Proc. of Educational Data Mining*, 2013.
2. Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 39–46. ACM, 2010.
3. Philipp Doebler, Mohsen Alavash, and Carsten Giessing. Adaptive experiments with a multivariate elo-type algorithm. *Behavior research methods*, pages 1–11, 2014.
4. Arpad E Elo. *The rating of chessplayers, past and present*, volume 3. Batsford London, 1978.
5. Stephen E Fancsali, Tristan Nixon, and Steven Ritter. Optimal and worst-case performance of mastery learning assessment with bayesian knowledge tracing. In *Proc. of Educational Data Mining*, 2013.
6. Stephen E Fancsali, Tristan Nixon, Annalies Vuong, and Steven Ritter. Simulated students, mastery learning, and improved learning curves for real-world cognitive tutors. In *AIED Workshops*. Citeseer, 2013.
7. Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.
8. Tanja Käser, Kenneth R Koedinger, and Markus Gross. Different parameters-same prediction: An analysis of learning curves. In *Proceedings of 7th International Conference on Educational Data Mining. London, UK*, 2014.
9. Kenneth R Koedinger, John C Stamper, Elizabeth A McLaughlin, and Tristan Nixon. Using data-driven discovery of better student models to improve student learning. In *Artificial intelligence in education*, pages 421–430. Springer, 2013.
10. Jung In Lee and Emma Brunskill. The impact on individualizing student models on necessary practice opportunities. *International Educational Data Mining Society*, 2012.
11. R Charles Murray, Steven Ritter, Tristan Nixon, Ryan Schwiebert, Robert GM Hausmann, Brendon Towle, Stephen E Fancsali, and Annalies Vuong. Revealing the learning in learning curves. In *Artificial Intelligence in Education*, pages 473–482. Springer, 2013.

12. Jan Papoušek, Radek Pelánek, and Vít Stanislav. Adaptive practice of facts in domains with varied prior knowledge. In *Proc. of Educational Data Mining*, pages 6–13, 2014.
13. Zachary A Pardos and Michael V Yudelson. Towards moment of learning accuracy. In *AIED 2013 Workshops Proceedings Volume 4*, page 3, 2013.
14. Radek Pelánek. Application of time decay functions and Elo system in student modeling. In *Proc. of Educational Data Mining*, pages 21–27, 2014.
15. Radek Pelánek. Metrics for evaluation of student models. *Journal of Educational Data Mining*, 2015. To appear.
16. Radek Pelánek. Modeling student learning: Binary or continuous skill? In *Proc. of Educational Data Mining*, 2015.
17. Al Mamunur Rashid, Istvan Albert, Dan Cosley, Shyong K Lam, Sean M McNee, Joseph A Konstan, and John Riedl. Getting to know you: learning new user preferences in recommender systems. In *Proceedings of the 7th international conference on Intelligent user interfaces*, pages 127–134. ACM, 2002.
18. Michael V Yudelson and Kenneth R Koedinger. Estimating the benefits of student model improvements on a substantive scale. In *EDM 2013 Workshops Proceedings*, 2013.

# Using Data from Real and Simulated Learners to Evaluate Adaptive Tutoring Systems

José P. González-Brenes[1], Yun Huang[2]

[1] Pearson School Research & Innovation Network, Philadelphia, PA, USA
jose.gonzalez-brenes@pearson.com
[2] Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA
yuh43@pitt.edu

**Abstract.** Classification evaluation metrics are often used to evaluate adaptive tutoring systems— programs that teach and adapt to humans. Unfortunately, evidence suggests that existing convention for evaluating tutoring systems may lead to suboptimal decisions. In a companion paper, we propose Teal, a new framework to evaluate adaptive tutoring. In this paper we propose an alternative formulation of Teal using simulated learners. The main contribution of this novel formulation is that it enables approximate inference of Teal, which may useful on the cases that Teal becomes computationally intractable. We believe that this alternative formulation is simpler, and we hope it helps as a bridge between the student modeling and simulated learners community.

## 1 Introduction

Adaptive systems teach and adapt to humans and improve education by optimizing the subset of *items* presented to students, according to their historical performance [3], and on features extracted from their activities [6]. In this context, items are questions, or tasks that can be graded individually. Adaptive tutoring may be evaluated with randomized control trials. For example, in a seminal study [3] that focused on earlier adaptive tutors, a controlled trial measured the time students spent on tutoring, and their performance on post-tests. The study reported that the adaptive tutoring system enabled significantly faster teaching, while students maintained the same or better performance on post-tests

Unfortunately, controlled trials can become extremely expensive and time consuming to conduct: they require institutional review board approvals, experimental design by an expert, recruiting and often payment of enough participants to achieve statistical power, and data analysis. Automatic evaluation metrics improve the engineering process because they enable less expensive and faster comparisons between alternative systems.

The adaptive tutoring community has tacitly adopted conventions for evaluating tutoring systems [4]. Researchers often evaluate their models with classification evaluation metrics that assess the *student model* component of the tutoring system— student models are the subsystems that forecast whether a learner will answer the next item correctly. However, automatic evaluation metrics are

intended to measure an outcome of the end user. For example, the PARADISE [9] metric used in spoken dialogue systems correlates to user satisfaction scores. We are not aware of evidence that supports that classification metrics correlate with learning outcomes; yet there is a growing body of evidence [2, 5] that suggests serious problems with them. For example, classification metrics ignore that an adaptive system may not help learners— which could happen with a student model with a flat or decreasing learning curve [1, 8]. A decreasing learning curve implies that student performance decreases with practice; this curve is usually interpreted as a modeling problem, because it operationalizes that learners are better off with no teaching.

We study a novel formulation of the Theoretical Evaluation of Adaptive Learning Systems (Teal) [5] evaluation metric. The importance of evaluation metrics is that they help practitioners and researchers quantify the extent that a system helps learners.

## 2 Theoretical Evaluation of Adaptive Learning Systems

In this section, we just briefly summarize Teal and do not compare it with a related method called ExpOppNeed [7]. Teal assumes the adaptive tutoring system is built using a single-skill Knowledge Tracing Family model [3, 6]. Knowledge Tracing uses a Hidden Markov Model (HMM) per skill to model the student's knowledge as latent variables. It models whether a student applies a practice opportunity of a skill correctly. The latent variables are used to model the latent student proficiency, which is often modeled with a binary variable to indicated mastery of the skill.

To use Teal on data collected from students, we first train a model using an algorithm from the Knowledge Tracing family, then we use the learned parameters to calculate the effort and outcome for each skill.

- Effort: Quantifies how much practice the adaptive tutor gives to students. In this paper we focus on counting the number of items assigned to students but, alternatively, amount of time could be considered.
- Outcome: Quantifies the performance of students after adaptive tutoring. For simplicity, we operationalize performance as the percentage of items that students are able to solve after tutoring. We assume that the performance on solving items is aligned to the long-term interest of learners.

Algorithm 1 describes our novel formulation. Teal calculates the expected number of practice that an adaptive tutor gives to students. We assume that the tutor stops teaching a skill once the student is very likely to answer the next item correctly according to a model from the Knowledge Tracing Family [6]. The adaptive tutor teaches an additional item if two conditions hold: (i) it is likely that the student will get the next item wrong— in other words, the probability of answering correctly the next item is below a threshold $\tau$; and (ii) the tutor has not decided to stop instruction already.

The inputs of Teal are:

– Real student performance data from $m$ students practicing a skill. Data from each student is encoded into a sequence of binary observations of whether the student was able to apply correctly the skill at different points in time.
– A threshold $\tau \in \{0 \dots 1\}$ that indicates when to stop tutoring. We operationalize this threshold as the target probability that the student will apply the skill correctly.
– A parameter $T$ that indicates the number of practice opportunities each of the simulated students will practice the skill.

**Algorithm 1** Teal algorithm for models with one skill per item

**Require:** real student data $\mathbf{y}^{(1)} \dots \mathbf{y}^{(m)}$, threshold $\tau$, # of simulated time steps $T$
**function** TEAL
  $\theta \leftarrow$ Knowledge_Tracing($\mathbf{y}^{(1)} \dots \mathbf{y}^{(m)}$)
  $e \leftarrow \{\ \}$
  $s \leftarrow \{\ \}$
  **for** $\hat{\mathbf{y}} \in$ get_simulated_student($\theta, T$) **do**:
    $e \leftarrow$ calculate_effort($\hat{\mathbf{y}}, \theta, \tau$)
    **if** $e < T$ **then**
      $s \leftarrow$ calculate_score($\hat{\mathbf{y}}$, e)
    **else**
      $s \leftarrow$ imputed_value
  **return** mean($e$), mean($s$)

Teal learns a Knowledge Tracing model from the data collected from real students interacting with a tutor. Our new formulation uses simulated learners sampled from the Knowledge Tracing parameters. This enables us to decide how many simulated students to generate. Our original formulation required $2^m$ sequences to be generated, which can quickly become computationally intractable. If an approximate solution is acceptable, our novel formulation allows more efficient calculations of Teal. Teal quantifies the effort and outcomes of students in adaptive tutoring. Even though measuring effort and outcomes is not novel by itself, Teal's contribution is measuring both without a randomized trial. Teal quantifies effort as how much practice the tutor gives. For this, we count the number of items assigned to students. For a single simulated student, this is:

$$\text{calculate\_effort}(y_1, \dots, y_T, \theta, \tau) \equiv \arg\min_t p(y_t | y_1 \dots y_{t-1}, \theta) > \tau \qquad (1)$$

The threshold $\tau$ implies a trade-off between student effort and scores and responds to external expectations from the social context. Teal operationalizes the outcome as the performance of students after adaptive tutoring as the percentage of items that students are able to solve after tutoring:

$$\text{calculate\_score}(y_1, \dots, y_T, e) \equiv \sum_{t=e} \frac{\delta(\mathbf{y}_t, \text{correct})}{T - e} \qquad (2)$$

Here, $\delta(\cdot, \cdot)$ is the Kronecker function that returns 1 iff its arguments are equal.

## 3 Discussion

Simulation enables us to measure effort and outcome for a large population of students. Previously, we required Teal to be computed exhaustively on all student outcomes possibilities. We relax the prohibitively expensive requirement of calculating all student outcome combinations. Our contribution is that Teal can be calculated with a simulated dataset size that is large yet tractable.

## References


1. R. Baker, A. Corbett, and V. Aleven. More Accurate Student Modeling through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. In B. Woolf, E. Aïmeur, R. Nkambou, and S. Lajoie, editors, *Intelligent Tutoring Systems*, volume 5091 of *Lecture Notes in Computer Science*, pages 406–415. Springer Berlin / Heidelberg, 2008.
2. J. Beck and X. Xiong. Limits to accuracy: how well can we do at student modeling? In S. K. D'Mello, R. A. Calvo, and A. Olney, editors, *Proceedings of the 6th International Conference on Educational Data Mining, Memphis, Tennessee, USA, July 6-9, 2013*, pages 4–11. International Educational Data Mining Society, 2013.
3. A. Corbett and J. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278, 1995.
4. A. Dhanani, S. Y. Lee, P. Phothilimthana, and Z. Pardos. A comparison of error metrics for learning model parameters in bayesian knowledge tracing. Technical Report UCB/EECS-2014-131, EECS Department, University of California, Berkeley, May 2014.
5. González-Brenes and Y. José P., Huang. Your model is predictive— but is it useful? theoretical and empirical considerations of a new paradigm for adaptive tutoring evaluation. In J. G. Boticario, O. C. Santos, C. Romero, and M. Pechenizkiy, editors, *Proceedings of the 8th International Conference on Educational Data Mining*, Madrid, Spain, 2015.
6. J. P. González-Brenes, Y. Huang, and P. Brusilovsky. General Features in Knowledge Tracing: Applications to Multiple Subskills, Temporal Item Response Theory, and Expert Knowledge. In M. Mavrikis and B. M. McLaren, editors, *Proceedings of the 7th International Conference on Educational Data Mining*, London, UK, 2014.
7. J. I. Lee and E. Brunskill. The impact on individualizing student models on necessary practice opportunities. In K. Yacef, O. R. Zaïane, A. Hershkovitz, M. Yudelson, and J. C. Stamper, editors, *Proceedings of the 5th International Conference on Educational Data Mining*, pages 118–125, Chania, Greece, 2012.
8. D. Rai, Y. Gong, and J. E. Beck. Using dirichlet priors to improve model parameter plausibility. In T. Barnes, M. Desmarais, C. Romero, and S. Ventura, editors, *Proceedings of the 2nd International Conference on Educational Data Mining*, Cordoba, Spain, 2009.
9. M. Walker, C. Kamm, and D. Litman. Towards developing general models of usability with PARADISE. *Natural Language Engineering*, 6(3):363–377, 2001.

# Authoring Tutors with Complex Solutions: A Comparative Analysis of Example Tracing and SimStudent

Christopher J. MacLellan[1], Erik Harpstead[1], Eliane Stampfer Wiese[1], Mengfan Zou[2], Noboru Matsuda[1], Vincent Aleven[1], and Kenneth R. Koedinger[1]

[1] Carnegie Mellon University, Pittsburgh PA, USA,
{cmaclell, eharpste, stampfer, noboru.matsuda, aleven, koedinger}@cs.cmu.edu,
[2] Tsinghua University, Beijing, China,
zmf11@mails.tsinghua.edu.cn

**Abstract.** Problems with many solutions and solution paths are on the frontier of what non-programmers can author with existing tutor authoring tools. Popular approaches such as Example Tracing, which allow authors to build tutors by demonstrating steps directly in the tutor interface. This approach encounters difficulties for problems with more complex solution spaces because the author needs to demonstrate a large number of actions. By using SimStudent, a simulated learner, it is possible to induce general rules from author demonstrations and feedback, enabling efficient support for complexity. In this paper, we present a framework for understanding solution space complexity and analyze the abilities of Example Tracing and SimStudent for authoring problems in an experimental design tutor. We found that both non-programming approaches support authoring of this complex problem. The SimStudent approach is 90% more efficient than Example Tracing, but requires special attention to ensure model completeness. Example Tracing, on the other hand, requires more demonstrations, but reliably arrives at a complete model. In general, Example Tracing's simplicity makes it good for a wide range problems, a reason for why it is currently the most widely used authoring approach. However, SimStudent's improved efficiency makes it a promising non-programmer approach, especially when solution spaces become more complex. Finally, this work demonstrates how simulated learners can be used to efficiently author models for tutoring systems.

**Keywords:** Tutor Authoring, Intelligent Tutoring Systems, Cognitive Modeling, Programming-by-Demonstration

## 1 Introduction

Intelligent Tutoring Systems (ITSs) are effective at improving student learning across many domains– from mathematics to experimental design [10, 13, 5]. ITSs

also employ a variety of pedagogical approaches for learning by doing, including intelligent novice [7], invention [12], and learning by teaching [9]. Many of these approaches require systems that can model complex solution spaces that accommodate multiple correct solutions to a problem and/or multiple possible paths to each solution. Further, modeling complex spaces can be desirable pedagogically: student errors during problem solving can provide valuable learning opportunities, and therefore may be desirable behaviors. Mathan and Koedingers spreadsheet tutor provides experimental support for this view– a tutor that allowed exploration of incorrect solutions led to better learning compared to one that enforced a narrower, more efficient solution path [7]. However, building tutoring systems for complex solution spaces has generally required programming. What options are available to the non-programmer? Authoring tools have radically reduced the difficulties and costs of tutor building [2, 6], and have allowed authoring without programming. Through the demonstration of examples directly in the tutor interface, an author can designate multiple correct solutions, and many correct paths to each solution. Yet, the capabilities of these tools for authoring problems with complex solution spaces has never been systematically analyzed.

In this paper, we define the concept of solution space complexity and, through a case study, explore how two authoring approaches deal with this complexity. Both approaches (Example Tracing and SimStudent) are part of the Cognitive Tutor Authoring Tools (CTAT) [1]. Our case study uses the domain of introductory experimental design, as problems in this area follow simple constraints (only vary one thing at a time), but solutions can be arbitrarily complex depending on how many variables are in the experiment and how many values each can take.

## 2 Solution Space Complexity

Solution spaces have varying degrees of complexity. Our framework for examining complexity considers both how many correct solutions satisfy a problem and how many paths lead to each solution. Within this formulation, we discuss how easily a non-programmer can author tutors that support many solutions and/or many paths to a solution.

How might this formulation of complexity apply to an experimental design tutor? Introductory problems in this domain teach the control of variables strategy (only manipulating a single variable between experimental conditions to allow for causal attribution) [3]. Due to the combinatorial nature of experiments (i.e., multiple conditions, variables, and variable values), the degree of complexity in a particular problem depends on how it is presented. To illustrate, imagine that students are asked to design an experiment to determine how increasing the heat of a burner affects the melting rate of ice in a pot (see Figure 1). The following tutor prompts (alternatives to the prompt in Figure 1) highlight how different problem framings will affect the solution complexity:

**One solution with one path** Design an experiment to determine how increasing the heat of a Bunsen burner affects the rate at which ice in a pot will

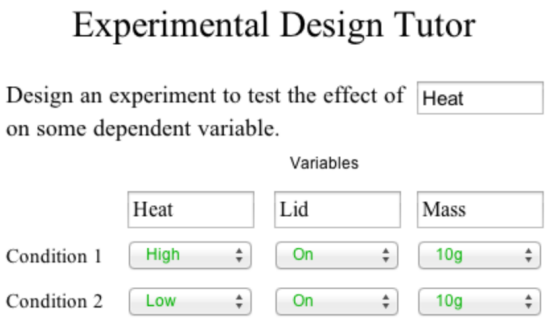



**Fig. 1.** Experimental design tutor interface

melt by assigning the first legal value to the variables in left to right, top down order as they appear in the table.

**One solution and many paths** Design an experiment to determine how increasing the heat of a Bunsen burner affects the rate at which ice in a pot will melt by assigning the first legal value to variables.

**Many solutions each with one path** Design an experiment to determine how increasing the heat of a Bunsen burner affects the rate at which ice in a pot will melt by assigning values to variables in left to right, top down order as they appear in the table.

**Many solutions with many paths** Design an experiment to determine how increasing the heat of a Bunsen burner affects the rate at which ice in a pot will melt.

While these examples show that solution space complexity can be qualitatively changed (i.e., one solution vs. many solutions) by reframing a problem, quantitative changes are also possible. For example, adding a fourth variable to the interface in Figure 1 would require two more steps per solution path (setting the variable for each condition), while adding another value to each variable increases the number of possible options at each step of the solution path. As this example illustrates, solution space complexity is not an inherent property of a domain, but rather arises from an authors design choices.

## 3 Tutor Authoring

Our analysis focuses on the Cognitive Tutor Authoring Tools (CTAT), as CTAT is the most widely used tutor authoring tool and the approaches it supports are representative of authoring tools in general [2]. CTAT supports non-programmers in building both tutor interfaces and cognitive models (for providing feedback). Cognitive models can be constructed with Example Tracing or SimStudent. In this section, we step through how Example-Tracing and SimStudent approaches would be applied by non-programmers to the experimental design task, using the

interface shown in Figure 1. Further, we discuss the features of each approach for handling solution space complexity in the context of this example.

### 3.1 Example Tracing

When building an Example-Tracing tutor in CTAT, the author demonstrates correct solutions directly in the tutoring interface. These demonstrated steps are recorded in a behavior graph. Each node in the behavior graph represents a state of the tutoring interface, and each link represents an action that moves the student from one node to another. In Example Tracing each link is produced as a result of a single action demonstrated directly in the tutor interface; many legal actions might be demonstrated for each state, creating branches in the behavior graph.

Figure 2 shows an example of our experimental design tutor interface and an associated behavior graph. The particular prompt chosen has 8 solutions and many paths to each solution. These alternative paths correspond to different orders in which the variables in the experimental design can be assigned. The Example-Tracing approach allows authors to specify that groups of actions can be executed in any order. In the context of our example, this functionality allows the author to demonstrate one path to each of the 8 unique solutions (these 8 paths are visible in Figure 2) and then specify that the actions along that path can be executed in any order. Unordered action groups are denoted in the behavior graph by colored ellipsoids.

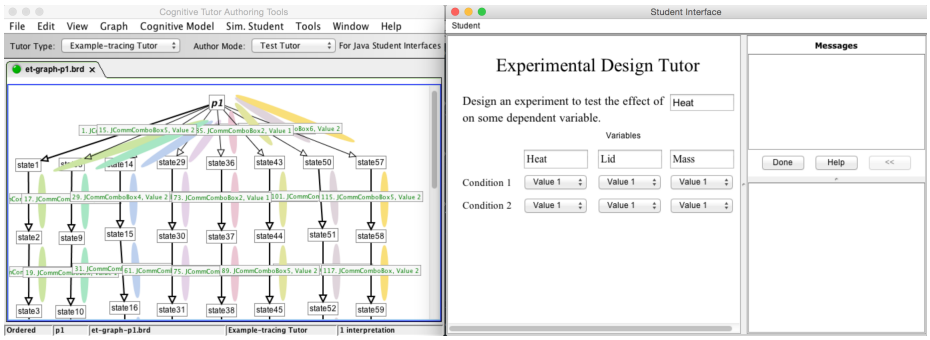



**Fig. 2.** An experimental design tutor (right) and its associated behavior graph (left). This tutor supports students in designing an experiment to test the effect of heat on a dependet variable. The correct answer is to pick two different values for the "Heat" variable and to hold the values constant for other variables.

Once a behavior graph has been constructed for a specific problem (e.g. determine the effect of heat on ice melting), that behavior graph can be generalized to other problems (e.g. determine the effect of sunlight on plant growth) using mass production. The mass production feature allows the author to replace specific values in the interface with variables and then to instantiate an arbitrary

number of behavior graphs with different values for the variables. This approach is powerful for supporting many different problems that have identical behavior graph structure, such as replacing all instances of "heat" with another variable, "sunlight". However, if a problem varies in the structure of its behavior graph, such as asking the student to manipulate a variable in the second column instead of the first (e.g., "lid" instead of "heat"), then a new behavior graph would need to be built to reflect the change in the column of interest.

How efficient is Example Tracing in building a complete cognitive model for the experimental design problem? The complete model consists of 3 behavior graphs (one for each of the three variable columns that could be manipulated). Each graph took 56 demonstrations and required 8 unordered action groups to be specified. Thus, the complete cognitive model required 168 demonstrations and 24 unordered group specifications. Using estimates from a previously developed Keystroke-Level Model [6], which approximates the time needed for an error-free expert to perform each interface action, we estimate that this model would take about 27 minutes to build using Example Tracing. Notably, the ability to specify unordered action groups offers substantial efficiency gains - without it, authoring would take almost 100 hours. Furthermore, with mass production, this model can generalize to any set of authored variables.

### 3.2 SimStudent

While the Example-Tracing behavior graph creates links from user demonstrations, the SimStudent system extends these capabilities by inducing production rule models from demonstrations and feedback (for details on this rule induction see [8]). In the experimental design tutor, SimStudent might learn a rule that sets one of the variables to an arbitrary value when no values for that variable have been assigned. Then, it might learn different rules for setting a variables second value based on whether or not it is being manipulated.

Authoring with SimStudent is similar to Example Tracing in that SimStudent asks for demonstrations when it does not know how to proceed. However, when SimStudent already has an applicable rule, it fires the rule and shows the resulting action in the tutor interface. It then asks the author for feedback on that action. If the feedback is positive, SimStudent may refine the conditions of its production rules before continuing to solve the problem. If the feedback is negative, SimStudent will try firing a different rule. When SimStudent exhausts all of its applicable rules, it asks the author to demonstrate a correct action. Figure 3 shows how SimStudent asks for demonstrations and feedback. When authoring with SimStudent, the author does not have to specify rule order - as long as a rule's conditions are satisfied, it is applicable. Authoring with SimStudent produces both a behavior graph (of the demonstrations and actions SimStudent took in the interface) and a production rule model.

To evaluate the efficiency of the SimStudent approach we constructed a complete model for the experimental design tutor. It can be difficult to determine when a SimStudent model is correct and complete from the authoring interactions alone. In most cases the SimStudent model is evaluated with set of held-out

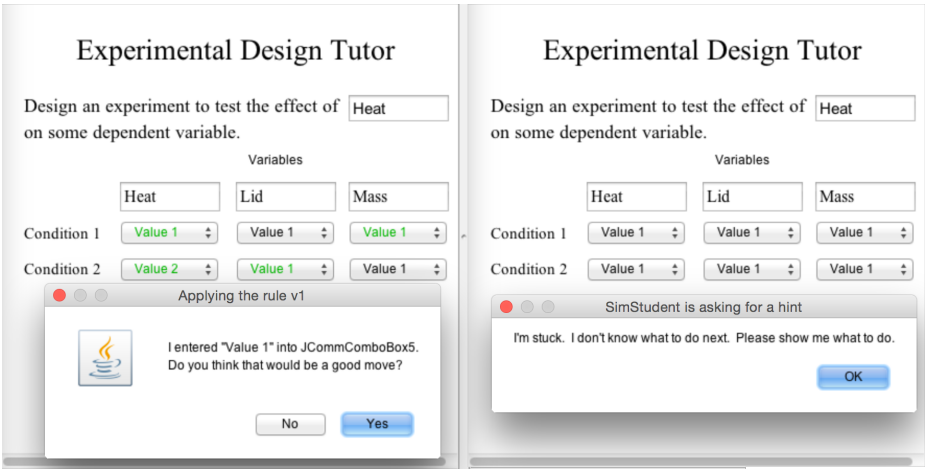



**Fig. 3.** SimStudent asking for feedback (left) and for a demonstration (right).

test problems (i.e., unit tests). However, in this case the learned rules were simple enough to evaluate by direct inspection. We noticed that SimStudent learned one correct strategy, but had not explored other solutions. This is typical of SimStudent - once it learns a particular strategy it applies it repeatedly. Therefore, authors must give it additional demonstrations of alternative paths. With the experimental design tutor, we noticed that SimStudent was always choosing the first value for non-manipulated variables, so we gave it additional demonstrations where non-manipulated variables took values besides those demonstrated on the initial run.

Ultimately, SimStudent acquired a complete model after 7 demonstrations and 23 feedback responses. Using the same Keystroke-Level Model from [6], we estimate that building a cognitive model using SimStudent would take an error-free expert about 2.12 minutes – much shorter than Example Tracing. Like Example Tracing, the model produced by SimStudent can work with arbitrary variables. Unlike Example Tracing, the learned model can work for unauthored variables; for example, students could define their own variables while using the tutor. This level of generality could be useful in inquiry-based learning environments [4]. Finally, if another variable column was added to the tutor, the SimStudent model would be able to function without modification. For Example Tracing, such a change would constitute a change to the behavior graph structure, so a completely new behavior graphs would need to be authored to support this addition.

## 4 Discussion

Both Example Tracing and SimStudent can create tutors for problems with complex solution spaces. However, our analysis shows that the two approaches

differ in terms of their efficiency and, as a result, how many solutions and paths they can handle in practice.

First, the Example-Tracing approach worked very well, even though the experimental design problems have a combinatorial structure. In particular, unordered action groups and mass production drastically reduced the number of demonstrations needed to cover the solution space, 168 vs. 40,362. The simplicity of Example Tracing combined with the power afforded by these features is likely why Example Tracing is the most widely used authoring approach today [2].

The SimStudent approach was more efficient than Example Tracing (approx. 2.12 vs. 27 minutes), but this comparison requires several caveats. The machine learning mechanisms of SimStudent generalize demonstrations and feedback into rules, which allows SimStudent to only model unique actions and the conditions under which they apply. However, this means SimStudent may not acquire a complete model. In the experimental design case study, SimStudent at first only learned that non-manipulated variables take their first value (rather than any value that is constant across conditions). In general, this problem arises when SimStudent acquires a model that can provide at least one correct solution for any problem. In these situations, it never prompts an author to provide alternative demonstrations; leading an unsuspecting author to create an incomplete model. A related complication is determining when the SimStudent model is complete. While determining the completeness of models in both Example Tracing and SimStudent can be difficult, authors must attempt to infer completeness from SimStudent's problem solving performance– a method that can be rather opaque at times. Thus, an open area for simulated learning systems is how best to evaluate the quality of learned models.

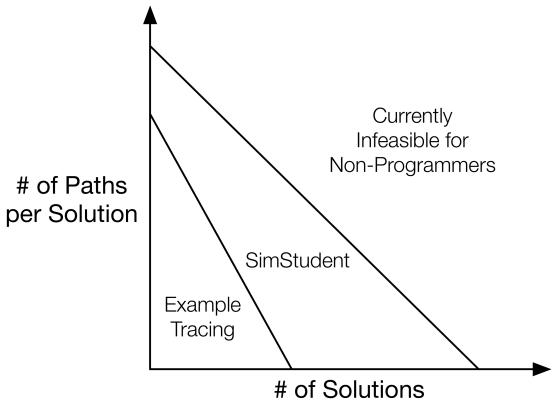



**Fig. 4.** How the space of solution space complexity is handled by existing non-programmer authoring approaches.

Overall our findings, when paired with those of previous work [6], suggest an interpretation depicted in Figure 4. In this figure the potential space of complexity is depicted in terms of number of unique solutions and number of paths per solution. The inner region denotes the area of the complexity space where we believe Example Tracing will maximize non-programmers' authoring utility. This region is skewed towards a higher number of paths, owing to Example Tracing's capacity to specify unordered actions. This portion of the complexity space contains many of the tutors that have already been built using Example Tracing [2]. As the complexity of a problem's solution space increases, Example Tracing becomes less practical (though still capable) and SimStudent becomes a more promising option, despite the caveats for using it. SimStudent's power of rule generalization gives it the ability to deal with more paths and unique solutions with less author effort, however, these capabilities come with the risk of producing incomplete models (without the author being aware).

Notably missing in the figure is any coverage of the upper right quadrant. This area would be a fruitful place to direct future work that supports non-programmers in authoring problems with many solutions with many paths. In particular, simulated learning systems might be extended to give non-programmers access to this portion of the space. One existing approach for dealing with highly complex solution spaces is to only model the aspects of the space that students are most likely to traverse. For example, work by Rivers and Koedinger [11] has explored the use of prior student solutions to seed a feedback model for introductory programming tasks. As it stands this area can only be reached using custom built approaches and would benefit from authoring tool research.

One limitation of our current approach is the assumption that there is a body of non-programmers that wants to build tutors for more complex problems. Our analysis here suggests that there is an open space for non-programming tools that support highly complex solution spaces, but it is less clear that authors have a desire to create tutors in this portion of the space. A survey of authors interested in building complex tutors without programming would help to shed light on what issues non-programmers are currently having in building their tutors. It is important that such a survey also include the perspective of those outside the normal ITS community to see if there are features preventing those who are interested from entering the space.

From a pedagogical point of view, it is unclear how much of the solution space needs to be modeled in a tutor. Waalkens et al. [16] have explored this topic by implementing three versions of an Algebra equation solving tutor, each with progressively more freedom in the number of paths that students can take to a correct solution. They found that the amount of freedom did not have an effect on students learning outcomes. However, there is evidence that the ability to use and decide between different strategies (i.e. solution paths) is linked with improved learning [14]. Further, subsequent work [15] has suggested that students only exhibit strategic variety if they are given problems that favor different strategies. Regardless of whether modeling the entire solution space is

pedagogically necessary, it is important that available tools support the ability to model complex spaces so that these research questions can be further explored.

## 5 Conclusion

The results of our analysis suggest that both the Example Tracing and SimStudent authoring approaches are promising methods for non-programmers to create tutors even for problems with many solutions with many paths. More specifically, we found that SimStudent was more efficient for authoring a tutor for experimental design, but authoring with SimStudent had a number of caveats related to ensuring that the authored model was complete. In contrast, Example Tracing was simple to use and it was clear that the authored models were complete. Overall, our analysis shows that Example Tracing is good for a wide range of problems that non-programmers might want to build tutors for (supported by its extensive use in the community [2]). However, the SimStudent approach shows great promise as an efficient authoring approach, especially when the solution space becomes complex. In any case, more research is needed to expand the frontier of non-programmers' abilities to author tutors with complex solution spaces.

Finally, this work demonstrates the feasibility and power of utilizing a simulated learning system (i.e., SimStudent) to facilitate the tutor authoring process. In particular authoring tutors with SimStudent took only 10% of the time that it took to author a tutor with Example-Tracing, a non-simulated learner approach. Educational technologies with increasingly complex solution spaces are growing in popularity (e.g. educational games and open-ended learning environments), but current approaches do not support non-programmers in authoring tutors for these technologies. Our results show that simulated learning systems are a promising tool for supporting these non-programmers. However, more work is needed to improve our understanding of how simulated learners can contribute to the authoring process and how the models learned by these systems can be evaluated.

## 6 Acknowledgements

We would like to thank Caitlin Tenison for her thoughtful comments and feedback on earlier drafts. This work was supported in part by a Graduate Training Grant awarded to Carnegie Mellon University by the Department of Education (#R305B090023) and by the Pittsburgh Science of Learning Center, which is funded by the NSF (#SBE-0836012). This work was also supported in part by National Science Foundation Awards (#DRL-0910176 and #DRL-1252440) and the Institute of Education Sciences, U.S. Department of Education (#R305A090519). All opinions expressed in this article are those of the authors and do not necessarily reflect the position of the sponsoring agency.

## References


1. Aleven, V., McLaren, B.M., Sewall, J., Koedinger, K.R.: The cognitive tutor authoring tools (CTAT): Preliminary evaluation of efficiency gains. In: Ikeda, M., Ashley, K.D., Tak-Wai, C. (eds.) ITS '06. pp. 61–70. Springer (2006)
2. Aleven, V., McLaren, B.M., Sewall, J., Koedinger, K.R.: A New Paradigm for Intelligent Tutoring Systems: Example-Tracing Tutors. IJAIED 19(2), 105–154 (2009)
3. Chen, Z., Klahr, D.: All Other Things Being Equal: Acquisition and Transfer of the Control of Variables Strategy. Child Development 70(5), 1098–1120 (1999)
4. Gobert, J.D., Koedinger, K.R.: Using Model-Tracing to Conduct Performance Assessment of Students' Inquiry Skills within a Microworld. Society for Research on Educational Effectiveness (2011)
5. Klahr, D., Triona, L.M., Williams, C.: Hands on what? The relative effectiveness of physical versus virtual materials in an engineering design project by middle school children. Journal of Research in Science Teaching 44(1), 183–203 (Jan 2007)
6. MacLellan, C.J., Koedinger, K.R., Matsuda, N.: Authoring Tutors with SimStudent: An Evaluation of Efficiency and Model Quality. In: Trausen-Matu, S., Boyer, K. (eds.) ITS '14 (2014)
7. Mathan, S.A., Koedinger, K.R.: Fostering the Intelligent Novice: Learning From Errors With Metacognitive Tutoring. Educational Psychologist 40(4), 257–265 (2005), http://www.tandfonline.com/doi/abs/10.1207/s15326985ep4004_7
8. Matsuda, N., Cohen, W.W., Koedinger, K.R.: Teaching the Teacher: Tutoring SimStudent Leads to More Effective Cognitive Tutor Authoring. IJAIED 25(1), 1–34 (2014)
9. Matsuda, N., Yarzebinski, E., Keiser, V., Cohen, W.W., Koedinger, K.R.: Learning by Teaching SimStudent – An Initial Classroom Baseline Study Comparing with Cognitive Tutor. IJAIED (2011)
10. Pane, J.F., Griffin, B.A., McCaffrey, D.F., Karam, R.: Effectiveness of Cognitive Tutor Algebra I at Scale. Tech. rep., RAND Corporation, Santa Monica, CA (2013)
11. Rivers, K., Koedinger, K.R.: Automating Hint Generation with Solution Space Path Construction. In: ITS '14, pp. 329–339. Springer (2014)
12. Roll, I., Aleven, V., Koedinger, K.R.: The Invention Lab : Using a Hybrid of Model Tracing and Constraint-Based Modeling to Offer Intelligent Support in Inquiry Environments. In: ITS '10. pp. 115–124 (2010)
13. Sao Pedro, M.A., Gobert, J.D., Heffernan, N.T., Beck, J.E.: Comparing Pedagogical Approaches for Teaching the Control of Variables Strategy. In: Taatgen, N., van Rijn, H. (eds.) CogSci '09. pp. 1–6 (2009)
14. Schneider, M., Rittle-Johnson, B., Star, J.R.: Relations among conceptual knowledge, procedural knowledge, and procedural flexibility in two samples differing in prior knowledge. Developmental Psychology 47(6), 1525–1538 (2011)
15. Tenison, C., MacLellan, C.J.: Modeling Strategy Use in an Intelligent Tutoring System: Implications for Strategic Flexibility. In: ITS '14, pp. 466–475. Springer (2014)
16. Waalkens, M., Aleven, V., Taatgen, N.: Computers & Education. Computers & Education 60(1), 159–171 (Jan 2013)

# Methods for Evaluating Simulated Learners: Examples from SimStudent

Kenneth R. Koedinger[1], Noboru Matsuda[1], Christopher J. MacLellan[1], and Elizabeth A. McLaughlin[1]

[1] Carnegie Mellon University, Pittsburgh, PA
koedinger@cmu.edu

**Abstract.** We discuss methods for evaluating simulated learners associated with four different scientific and practical goals for simulated learners. These purposes are to develop a precise theory of learning, to provide a formative test of alternative instructional approaches, to automate authoring of intelligent tutoring systems, and to use as a teachable agent for students to learn by teaching. For each purpose, we discuss methods for evaluating how well a simulated learner achieves that purpose. We use SimStudent, a simulated learner theory and software architecture, to illustrate these evaluation methods. We describe, for example, how SimStudent has been evaluated as a theory of student learning by comparing, across four domains, the cognitive models it learns to the hand-authored models. The SimStudent-acquired models yield more accurate predictions of student data in the three of the four domains. We suggest future research into more directly evaluating simulated learner predictions of the process of student learning.

**Keywords:** simulated learners, cognitive models, learning theory, instructional theory

## 1 Introduction

When is a simulated learner a success? We discuss different approaches to evaluating simulated learners. Some of these evaluation approaches are technical in nature, whether or how well a technical goal has been achieved, and some are empirical, whereby predictions from the simulated learner are compared against data. These approaches can be framed within the different goals and uses for simulated learners. Table 1 summarizes four purposes for developing simulated learners.

**Table 1.** Scientific and Practical Goals for Simulated Learners (SLs)

1. *Precise Theory*. Use SLs to develop and articulate precise theory of student learning in a replicable and unambiguous computational form.
   a. *Cognitive Model*. Create theories of domain expertise
   b. *Error Model*. Create theories of student domain misconceptions
   c. *Prior Knowledge*. Create theories of how different prior knowledge changes the nature and effectiveness of learning
   d. *Learning Process*. Create theories of change in student knowledge and performance
2. *Instructional Testing*. Use SLs as a "crash test" to evaluate and compare different instructional approaches on how well they facilitate learning.
3. *Automated Authoring*. Use SLs to automate the development of the expert component or cognitive model of an intelligent tutoring system.
4. *Teachable Agent*. Use SLs as a teachable agent or peer learner inside an instructional system to directly aid student learning.

Some of these goals have been pursued in prior simulated learner research. For example, [1] proposed the use of a simulated learner for instructional testing (#2 in Table 1). More specifically, he used "pseudo-students" during the design process for a formative evaluation of instruction to detect design defects.

Different evaluation approaches are appropriate for the different goals indicated in Table 1. In later sections, we discuss these evaluation approaches for each of the four main goals. But first, we introduce, SimStudent, the simulated learner system we have developed.

### 1.1 SimStudent: A Simulated Learner Theory and Software Architecture

We use SimStudent [2,3] as a running example to illustrate the evaluation techniques we discuss. SimStudent is a simulated learner system and theory in the class of adaptive production systems as defined by [4]. As such, it is similar to cognitive architectures such as ACT-R [5], Soar [6], and Icarus [7]. It is distinctive in its focus on modeling learning of complex academic topics, such as math, science, and language learning, and in its focus on inductive knowledge level learning [8]. SimStudent learns from a few primary forms of instruction, including examples of correct actions, skill labels on similar actions (which cue, but do not guarantee, learning and use of the same production rule), clues for what information in the interface to focus on to infer a next action, and finally yes-or-no feedback on actions performed by SimStudent.

To tutor SimStudent, the author first enters a problem in the tutoring interface (e.g., the "2x = 8" in the first row of Figure 1). SimStudent then attempts to solve the problem by applying productions learned so far. If an applicable production is found, the production application is visualized as a step in the behavior recorder represented as a new state-edge pair like the one shown in the bottom left of Figure 1. The author then provides correctness *feedback* on the step performed by SimStudent. When there are multiple applicable productions SimStudent shows the author all corresponding production applications and obtains correctness feedback on each.

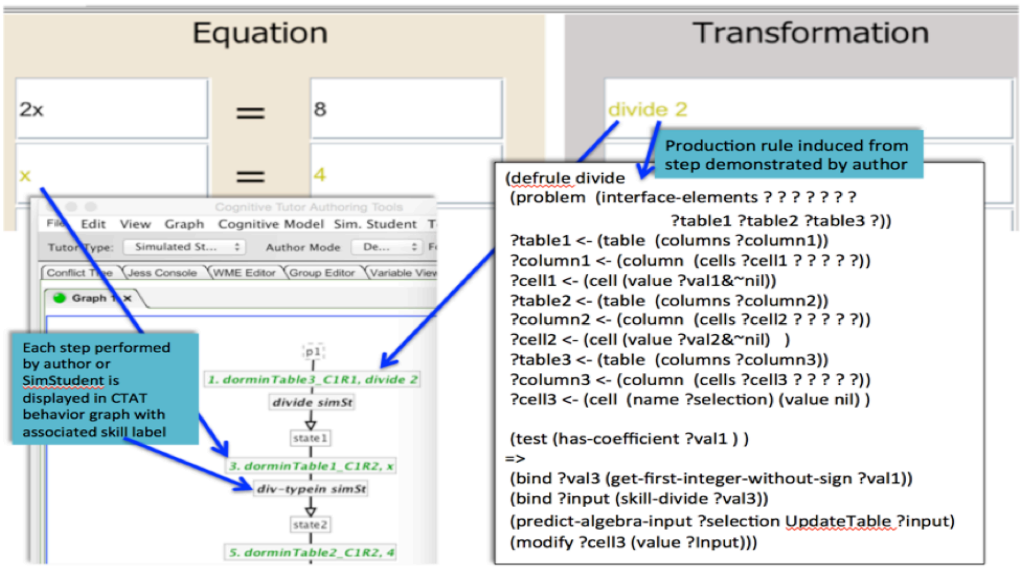



**Fig. 1.** After creating an interface (shown at top) and entering a problem ("2x=8"), teaching of SimStudent occurs either by giving yes-or-no feedback when SimStudent attempts a step or by demonstrating a correct step when SimStudent cannot (e.g., "divide 2"). SimStudent induces production rules from demonstrations (example shown on right) for each skill label (e.g., "divide" or "div-typein" shown on left). It refines productions based on subsequent positive (demo or yes feedback) or negative (no feedback) examples.

If no correct production application is found, then SimStudent asks the author to demonstrate the next step directly in the interface. When providing a demonstration, the author first specifies the *focus of attention* (i.e. input fields relevant to the current step) by double-clicking the corresponding interface elements, for example, the cells containing "2x" and "8" in Figure 1. The author then takes action using the relevant information (e.g., entering "divide 2" in Figure 1). The demonstrated step is visualized in the behavior graph. Finally, the *author specifies the skill name* by clicking on the newly added edge of the behavior graph. A small dialogue box appears to enter a skill name. This skill label is used to guide SimStudent's learning and to make the models acquired by SimStudent more interpretable (i.e., to give them human read-

able names). Generally speaking, authors could model a single observable step (i.e., an edge in the behavior graph) with a chain of production-rule applications. However, when using SimStudent to author an expert model, SimStudent generates a single production rule per each observable step.

SimStudent learns production rules from the author's demonstrations and feedback using three machine-learning mechanisms: *how*, *where*, and *when* learning. When given a new demonstration (i.e., a positive example of a rule), SimStudent uses its *how* learner to explain the demonstration and produces a general composition of functions that replicate the demonstrated steps and ones like it. For example, in Figure 1, when given the demonstration "divide 2" for the problem 2x=8, SimStudent induces (or "guesses") that the result of the "get-first-integer-without-sign" function when applied to left side of the problem and appended to the word "divide" explains the demonstration. This general sequence can then be used on novel problems (e.g., 4x=12 -> divide 4).

After an action sequence has been discovered, SimStudent uses its *where* learner to identify a generalized path to the focuses of attention in the tutor interface; e.g., to the left and right sides of the equation or to the next step input field. For the example in Figure 1, the *where* learner discovers retrieval paths for the three cells in the first column. These paths are generalized as more positive examples are acquired for a given rule. For example, when the author demonstrates the application of the divide rule shown in Figure 1 to the second row of the equation table, then the production's retrieval path might be generalized to function over any row in the equation table (rather than just the first two rows).

Finally, after learning an action sequence and general paths to relevant information, SimStudent uses its *when* learning to identify the conditions under which the learned production rule produces correct actions. For the example in Figure 1, SimStudent learns that this rule can only be correctly applied when one side of the equation has a coefficient. In situations when SimStudent receives positive and negative feedback on its rule applications, it uses the *when* learner to update the conditions on the rules. Thus, the *how* and *where* learners primarily use positive examples and the *when* learner uses both positive and negative examples.

SimStudent is also capable of learning the representation of the chunks (object attribute structures) that make up the production system working memory and are the informational basis on which productions are learned. It does so using an unsupervised grammar induction approach [3]. This feature particularly sets it apart from the other production rule learning systems mentioned above.

## 2 Evaluating Simulated Learners as Theories of Learning

It is helpful to distinguish a general theory of learning from a human-specific theory of *student* learning. We focus primarily on a theory of human student learning as it is most relevant to the education goals of the field of AI in Education. However, it is worth mentioning that there are evaluation criteria for a general learning theory, such as how quickly (e.g., in number of examples or time) and independently (e.g., with less supervision) learning takes and how general and accurate is the performance (e.g., problem solving or inference capability) of the resulting expert system. These criteria, then, provide guidance for comparative evaluations of general theories of learning. It is reasonable to consider as a better theory of learning one that produces improvements over a competing theory on any of speed, independence, generality, or accuracy without harming any of the others. In [9], for instance, Tenenbaum, Griffiths and Kemp have argued that hierarchical Bayesian models are better models of learning than other classification or neural network models because they can learn as well with fewer examples. (Note: If one suggests that a model is better because humans are able to learn with fewer examples, then one is moving into the realm of a human-specific theory of learning.)

While not necessary to evaluate general theories of learning, to evaluate the validity of theories of student learning, it is critical that the simulated learner be compared with student data. This data may involve student correct performance, incorrect performance, performance across tasks, and changes in performance over time. The data may be qualitative or quantitative.

### 2.1 Good Student Learning Theory Should Generate Accurate Cognitive Models

A student learning theory should produce the kind of expertise that human students acquire. In other words, the result of teaching a simulated learner should be a cognitive model of what a human student

knows after instruction and should behave as a human student does after instruction. For this kind of evaluation of a simulated learner, the issue reduces to the question of how to evaluate a cognitive model. In [10], we proposed six constraints to evaluate the quality of a cognitive model. These were labeled 1) solution sufficiency, 2) step sufficiency, 3) choice matching, 4) computational parsimony, 5) acquirability, and 6) transfer. The first two are empirical and qualitative: Is the cognitive model that the SL acquires able to solve tasks in the domain of interest and does it do so with steps that are consistent with human students? The third is quantitative: Does the frequency of strategy use and common error categories generated by the cognitive model on different tasks correspond with the frequency of strategy use and common error categories exhibited by human students on those tasks? The last three are rational in character, involving inspection of the cognitive model (or contrasting models) to judge whether it is (they are) not unnecessarily complex (#4), can be plausibly learned (#5), and implies transfer through overlap in knowledge components that apply across tasks (#6).

These constraints were designed to evaluate cognitive models developed by hand (e.g., by a scientist writing an expert system), but in the case that the model is generated by a simulated learner, the acquirability constraint (#5) is naturally achieved. The components of the cognitive models can be plausible because the SL does, in fact, learn them. If the cognitive model that is produced can solve tasks in the domain, for example, a simulated learner trained on algebra equations can solve equations, then the solution sufficiency constraint (#1) is met. If it solves them using the kinds of intermediate steps that match the kinds of steps in student solutions, for example, it performs its solution in a step-based tutoring system interface, then the step sufficient constraint (#2) is met.

How, then, can the remaining choice matching (#3), parsimony (#4), and transfer (#6) constraints be evaluated? In [11], we employed an approach that provides much of what is needed. This approach employs educational data mining and, in particular, evaluates the accuracy of a cognitive model by the so-called "smooth learning curve" criteria [cf., 12,13]. Using a relatively simple statistical model of how instructional opportunities improve the accuracy of knowledge, this approach can measure and compare cognitive models in terms of their accuracy in predicting learning curve data. To employ the statistical model fit, the cognitive model is simplified into a "Q matrix", which encodes a mapping from each observed task students perform (e.g., answering a question or entering a step in a problem solving) to the knowledge components that are hypothesized to be needed to successfully perform that task. Different cognitive models produce different Q matrices and different levels of predictive accuracy in fitting student learning curve data (measured, for example, by the root mean squared error on held-out data in cross validation). For any appropriate dataset uploaded into DataShop (learnlab.org/DataShop), the website allows users to edit and upload alternative cognitive models (in the Q matrix format), automatically performs statistical model fits, renders learning curve visualizations, and displays a ranking ordering of the models in terms of their predictive accuracy [14].

In [11], this approach was used to evaluate the empirical accuracy of the cognitive models that SimStudent learns as compared to hand-authored cognitive models. SimStudent was tutored in four domains: algebra, fractions, chemistry, and English grammar, in which we had existing human data and existing hand-authored cognitive models (see Figure 2). In each domain SimStudent induced, from examples and from practice with feedback, both new chunk structures to represent the organization (or "grammar") of the perceptual input in each domain and new production rules that solve problems (e.g., add two fractions) or make decisions (e.g., select when to use "the" or "a" in English sentences) in each domain. In each case, the production rules that SimStudent acquired were converted into the Q matrix format whereby a production (the columns of the Q matrix) is indicated as needed (entering a 1 rather than a 0 in the matrix) for a task (the rows of the Q matrix) if SimStudent uses that production to succeed on that task. Then the DataShop cognitive model comparison was employed to compare whether these models fit student learning curve data better than the hand-authored cognitive models do.

In all four domains, the SimStudent-acquired cognitive models made distinctions not present in the hand-authored models (e.g., it had two different production rules across tasks for which the hand-authored model had one) and thus it tended to produce models with more knowledge components (as shown in Table 2). For example, SimStudent learned two different production rules for the typical last step in equation solving where one production covered typical cases (e.g., from $3x = 12$ the student should "divide by 3") and another covered a perceptually distinct special case (e.g., from $-x = 12$ the student should divide by -1).

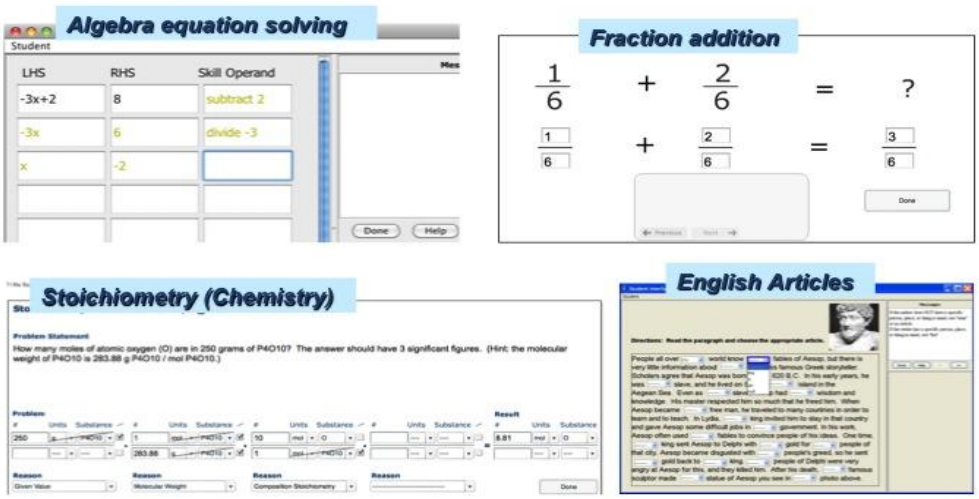



**Fig. 2.** Four domains in which SimStudent has been tutored and for which we have student learning data.

In all four domains, at least some of these distinctions improved the accuracy of fit to the learning curve data for the relevant tasks. Continuing the example, the SimStudent-acquired cognitive model in algebra leads to better accuracy because real students had a much higher error rate on tasks like -x=12 (where the coefficient, -1, is implicit) than on tasks like 3x=12 (where the coefficient, 3, is explicitly visible). More generally, in one domain (Fraction Addition, see Table 2), the SimStudent-acquired cognitive model failed to make a key distinction present in the hand-authored model and thus, while better in some cases, its overall fit was worse. In the three other domains (see Table 2), the SimStudent-acquired cognitive models were all found to be more accurate than the hand-authored cognitive models.

**Table 2.** A comparison of human-generated and SimStudent-discovered models. The columns on the left show the number of productions rules in the cognitive models and the columns on the right show the root mean squared error of the models for predicting held-out student learning data in cross validation.

| | Number of Production Rules | | Cross-Validated RMSE | |
|---|---|---|---|---|
| | Human-Generated Model | SimStudent Discovered Model | Human-Generated Model | SimStudent Discovered Model |
| Algebra | 12 | 21 | 0.4024 | **0.3999** |
| Stoichiometry | 44 | 46 | 0.3501 | **0.3488** |
| Fraction Addition | 8 | 6 | **0.3232** | 0.3343 |
| Article selection | 19 | 22 | 0.4044 | **0.4033** |

In other words, this "smooth learning curve" method of evaluation can provide evidence that a simulated learner, SimStudent in this case, is a reasonable model of student learning in that it acquires knowledge at a grain size (as represented in the components of the cognitive model) that is demonstrably consistent with human data.

One limitation of this approach is that it *indirectly* compares a simulated learner to human learners through the process of fitting a statistical model. In the case of algebra, for example, SimStudent's acquisition of two different productions for tasks of the form Nx=N versus tasks of the form -x=N gets translated into a prediction that student performance will be *different* in these situations, but the not direction of the difference. It is process of estimating the parameters of the statistical model that yields the prediction for which of these task categories (Nx=N or -x=N) will be harder. A more *direct* comparison would not use an intermediate statistical model fit. It would require the simulated learner to not only produce a relevant distinction, but to make a prediction of student performance differences, such as whether it takes longer to successfully learn some kinds of tasks than others. Such an evaluation approach is discussed in section 2.3.

### 2.2 Good Student Learning Theory Should Generate Errors that Match Student Errors and Test Prior Knowledge Assumptions

As a model of student learning, a good simulated learner should not only produce accurate performance with learning, but should also produce the kinds of errors that students produce [cf.,15]. Thus, another way to evaluate a simulated learner is compare the errors it generates with student errors.

One theory of student errors is that students learn incorrect knowledge (e.g., incorrect production rules or schemas) from *correct example-based instruction* due to the necessary fallibility of inductive learning processes. A further hypothesis is that inductive learning errors are more likely when students have "weak" (i.e., more domain general) rather than "strong" (i.e., more domain specific) prior knowledge. With weak prior knowledge, students may interpret examples shallowly, paying attention to more immediately perceived surface features, rather than more deeply, by making domain-relevant inferences from those surface features. Consider example-based instruction where a student is given the equation "3x+5 = 7" and told that "subtract 5" from both sides is a good next step. A novice student with weak prior knowledge might interpret this example shallowly, as subtracting a number (i.e., 5) instead of more deeply, as subtracting a term (i.e., +5). As a consequence, the student may induce knowledge that produces an error on a subsequent problem, such as "4x-2=5" where they subtract 2 from both sides. Indeed, this error is a common one among beginning algebra students.

In [16], we evaluated SimStudent by comparing induction errors it makes with human student errors. More specifically, we evaluated the weak prior knowledge hypothesis expressed above. We conducted a simulation study by having multiple instances of SimStudent get trained by Cognitive Tutor Algebra I. We compared SimStudent behaviors with actual student data from the Cognitive Tutor's logs of student interactions with the system. When SimStudent starts with weak prior knowledge rather than strong prior knowledge, we found that it learns more slowly, that is, the accuracy of learned skills is lower given the same amount of training or the amount of training needed to reach the same level of accuracy is greater. More importantly, we found that SimStudent's ability to predict student errors increased significantly when given weak rather than strong prior knowledge. In fact, the errors generated by SimStudent with strong prior knowledge were almost never the same kinds of errors commonly made by real students.

In addition to illustrating how a simulated learner can be evaluated by comparing its error generation to human errors, the above example illustrates how a simulated learner can be used to test assumptions about what prior knowledge students bring to their learning environments. The study above showed that novice algebra students do not have strong prior knowledge, particularly of the grammatical elements of equations, such as terms and coefficients. Thus, the SimStudent provides a theoretical explanation not only of common student error patterns, but also of empirical results (e.g., Booth and Koedinger, 2008) showing correlations between tasks measuring prior knowledge (e.g., identify the negative terms in "3x - 4 = -5 - 2x") and subsequent learning of target skills (e.g., solving algebra equations).

Some previous studies of students' errors focus primarily on a descriptive theory to explain why students made particular errors, for example, repair theory [15], the theory of bugs [18], and the theory of extrapolation technique [19]. With simulated learners [cf., 15], we can better understand the process of acquiring the incorrect skills that generate errors. The precise understanding that computational modeling of learning facilitates provides us with insights into designing better learning environments that anticipate and prevent or quickly remedy error formation.

### 2.3 Good Student Learning Theory Should Match Learning Process Data

Matching a simulated learners performance to learning process data is similar to the cognitive model evaluation discussed above in section 2.1. However, as indicated above, that approach has the limitation of being an indirect comparison with human data whereby there the fit to human data is, in a key sense, less challenging because it is mediated by a separate step parameter estimation of a statistical model. A more direct comparison is, in simple terms, to match the behavior of multiple instances of a simulated learner (i.e., a whole simulated class) with the behavior of multiple students. The simulated learners interact with a tutoring system (like one shown in Figure 2) just as a class of human students would and their behavior is logged just as human student data is. Then the simulated and human student data logs can be compared, for example, by comparing learning curves that average across all (simulated and human) student participants.

## 3 Evaluating Simulated Learners as Instruction Testers

A number of projects have explored the use of a simulated learner to compare different forms of instruction. VanLehn was perhaps the first to suggest such a use of a "pseudo student" [1]. MacLaren & Koedinger used a version of ACT-R's utility learning mechanism to show that the simulated learner was often successful when given error feedback not only on target performance tasks (e.g., solving two-step equations), but also on shorter subtasks (e.g., one-step equations) [10]. Matsuda, Cohen & Koedinger used SimStudent to show better learning from giving it a combination of examples and problems to solve, than just giving it examples [2]. Li, Cohen & Koedinger showed that interleaving problem types is as good, or better, for learning than blocking problem types because interleaving provides better opportunities for detecting and correcting generalization errors [20].

Recall the distinction mentioned above between general learning theory and human learning theory. This distinction can be extended to separate a general theory of instruction from a theory of instruction relevant to human students. For a general theory of instruction, it is of scientific interest to understand the effectiveness of different forms of instruction for different kinds of SL systems even if the SL is not (or not known to be) an accurate model of student learning. Such understanding may be relevant to advancing applications of AI and is directly relevant to the issue of how an SL can be easily trained for purposes of automated ITS authoring, the topic of the next section. Such theoretical demonstrations may also have relevance to a theory of *human* instruction as they may 1) provide theoretical explanations for instructional improvements that have been demonstrated with human learners or 2) generate predictions for what may (or may not) work (but has not yet been tried) with human students.

However, these instructional conclusions can only be reliably extended to human learners when there is existing evidence that the simulated learner is an accurate model of student learning (see the prior section 2). And ideally, the most reliable evaluation of a simulated learner as instructional tester is a follow-up random assignment experiment with human learners that demonstrates that the instructional form that was better for the simulated learners is also better for students. In the examples given above, there is some reasonable evidence that the simulated learners (or learning mechanisms) applied are accurate models of student learning. In many cases, there are past relevant human experiments. However, in none of these cases was the ideal follow-up experiment performed.

## 4 Evaluating Simulated Learners as ITS Authoring Tools

In addition to their use as theories of learning and for testing instructional content, simulated learning systems can also be used to facilitate the authoring of Intelligent Tutoring Systems (ITS). In particular, once a simulated learner has been sufficiently trained, the cognitive model it learns can then be used directly as an expert model. Previous work, such as Example Tracing tutor authoring [21], has explored how models can be acquired by demonstration. However, by using a simulated learning system to induce general rules form the demonstrations more general models can be acquired more efficiently. For example, the use of SimStudent as authoring tool is still experimental, but there is evidence that it may accelerate the authoring process and that it may produce more accurate cognitive models than hand authoring. In one demonstration, [2] explored the benefits of a traditional programming by demonstration approach to authoring in SimStudent versus a programming by tutoring approach, whereby SimStudent asks for demonstrations only at steps in a problem/activity where it has no relevant productions and otherwise it performs a step (firing a relevant production) and asks the author for feedback as to whether the step is correct/desirable or not. They found that programming by tutoring is much faster, 13 productions learned with 20 problems in 77 minutes versus 238 minutes in programming by demonstration. They also found that programming by tutoring produced a more accurate cognitive model whereby there were fewer productions that produced over-generalization errors. Programming by tutoring is now the standard approach used in SimStudent and its improved efficiency and effectiveness over programming by demonstration follow from having SimStudent start performing its own demonstrations. Better efficiency is obtained because the author need only respond to each of SimStudent's step demonstrations with a single click, on a yes or no button, which is much faster than demonstrating that step. Better effectiveness is obtained because these demonstrations better expose over-

generalization errors to which the author responds "no" and the system learns new IF-part preconditions to more appropriately narrow the generality of the modified production rule.

In a second demonstration of SimStudent as an authoring tool, [22] compared authoring in SimStudent (by tutoring) with authoring example-tracing tutors in CTAT. Tutoring SimStudent has considerable similarity with creating an example-tracing tutor except that SimStudent starts to perform actions for the author, which can be merely checked as desirable or not, saving the time it otherwise takes for an author to perform those demonstrations. That study reported a potential savings of 43% in authoring time by using SimStudent to aid in creating example-tracing tutors. As mentioned before, the work by [11] has shown that the models acquired using SimStudent better fit the student data. Thus, using the SimStudent system to author a tutoring system allows for the efficient construction of empirically better models.

## 5 Evaluating a Simulated Learner as a Teachable Agent

Simulated learner systems can be more directly involved in helping students learn when they are used as a teachable agent whereby students learn by teaching [cf., 23]. Evaluating the use of a simulated learner in this form ideally involves multiple steps. One should start with a simulated learner that has already received some positive evaluation as a good model of student learning (see section 2). Then incorporate it into a teachable agent architecture and, as early and often as possible, perform pilot students with individual students [cf., 24 on think aloud user studies) and revise the system design. Finally, for both formative and summative reasons, use random assignment experiments to compare student learning from the teachable agent with reasonable alternatives.

Using SimStudent, we built a teachable agent learning environment, called APLUS, in which students learn to solve linear equations by teaching SimStudent [25] . To evaluate the effectiveness of APLUS and advance the theory of learning by teaching, we conducted multiple *in vivo* experiments each with a specific hypotheses to test [25,26,27,28].

Each of the classroom studies have been randomized controlled trials with two conditions controlling a single study variable. For example, in one study [25], the self-explanation hypothesis was tested by having students justify their tutoring activities and decision making. To test this hypothesis, we developed a version of APLUS in which SimStudent occasionally asked "why" questions. For example, when a student provided negative feedback to a step SimStudent performed, SimStudent asked, "Why do you think adding 3 here on both sides is incorrect?" Students were asked to respond to SimStudent's questions either by selecting pre-specified menu items or entering a free text response. The results showed that the amount and the level of elaboration of the response had a reliable correlation with students' learning measured by online pre- and post-tests.

We also compared learning by teaching with other forms of instruction [28]. In this *in vivo* study, half of the students used APLUS and half used Cognitive Tutor Algebra I [29]. The results showed an aptitude-treatment interaction such that students scoring in the low half on the pre-test may not be ready to benefit from learning by teaching -- they learned better using the Cognitive Tutor than using APLUS -- whereas students scoring in the high half of the pre-test learned more from APLUS (i.e., by teaching) than from the Cognitive Tutor (i.e., by being tutored).

## 6 Conclusion

We outlined four general purposes for simulated learners (see Table 1) and reviewed methods of evaluation that align with these purposes. To evaluate a simulated learner as a precise theory of learning, one can evaluate the cognitive model that results from learning, evaluate the accuracy of error predictions as well as prior knowledge assumptions needed to produce those errors, or evaluate the learning process, that is, the opportunity by opportunity changes in student performance over time. To evaluate a simulated learner as an instructional test, one should not only evaluate the systems accuracy as a precise theory of student learning, but should also perform human experiment to determine whether the instruction that works best for simulated learners also works best for human students. To evaluate a simulated learner as an automated authoring tool, one can evaluate the speed and precision of rule production, the frequency of over-

generalization errors and the fit of the cognitive models it produces. More ambitiously, one can evaluate whether the resulting tutor produces as good (or better!) learning than an existing tutor. Similarly, to evaluate a simulated learner as a Teachable Agent, one can not only evaluate the features of a Teachable Agent system, but also perform experiments on whether students learn better with that system than with reasonable alternatives.

Simulated learner research is still in its infancy so most evaluation methods have not been frequently used. There have been very few studies that have evaluated a simulated learner as an instructional tester by following up a predicted difference in instruction with a random assignment experiment using the same forms of instruction with real students. We know of just one such study in which [29] first used an extension of the ACT-R theory of memory to simulate positive learning effects of an optimized practice schedule over a (highly recommended) spaced practice schedule. Next, he ran the same experiment with human students and confirmed the benefits of the optimized practice schedule. Such experiments are more feasible when the instruction involved is targeting simpler learning processes, such as memory, but will be more challenging as they target more complex learning processes, such as induction or sense making [cf., 31]

As far as we know, there have been no studies evaluating the learning process of a simulated learner as we recommended in section 2.3. Such evaluations would be particularly compelling demonstrations of the power of the simulated learner approach!

As we argued in [32], the space of instructional choices is just too large, over 200 trillion possible forms of instruction, for a purely empirical science of learning and instruction to succeed. We need parallel and coordinated advances in theories of learning *and* instruction and efforts to develop and evaluate simulated learners are fundamental to such advancement.

## References


1. VanLehn, K. (1991). Two pseudo-students: Applications of machine learning to formative evaluation. In R. Lewis & S. Otsuki (Eds.), *Advanced Research on Computers in Education* (pp. 17-26). Amsterdam: Elsevier.
2. Matsuda, N., Cohen, W. W., & Koedinger, K. R. (2015). Teaching the Teacher: Tutoring SimStudent leads to more Effective Cognitive Tutor Authoring. *International Journal of Artificial Intelligence in Education, 25*, 1-34.
3. Li, N., Matsuda, N., Cohen, W., & Koedinger, K.R. (2015). Integrating representation learning and skill learning in a human-like intelligent agent. *Artificial Intelligence, 219,* 67-91.
4. Anzai, Y. & Simon, H. A. (1979). The theory of learning by doing. *Psychological Review, 86 (2)*, 124-140.
5. Anderson, J. R., & Lebiere, C. (1998). *The Atomic Components of Thought*. Hillsdale, NJ: Erlbaum.
6. Laird, J. E., Newell, A., & Rosenbloom, P.S. (1987). Soar: an architecture for general intelligence. *Artificial Intelligence, 33*(1), 1–64.
7. Langley, P. & Choi, D. (2006). A unified cognitive architecture for physical agents. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, Boston.
8. Newell, Allen. 1990. *Unified Theories of Cognition*. Cambridge, MA: Harvard U. Press.
9. Tenenbaum, J. B., Griffiths, T. L., & Kemp.C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10, 309-318.
10. MacLaren, B. & Koedinger, K. R. (2002). When and why does mastery learning work: Instructional experiments with ACT-R "SimStudents". In S.A. Cerri, G. Gouarderes, & F. Paraguacu (Eds.), *Proceedings of the 6th International Conference on Intelligent Tutoring Systems*, 355-366. Berlin: Springer-Verlag.
11. Li, N., Stampfer, E., Cohen, W., & Koedinger, K.R. (2013). General and efficient cognitive model discovery using a simulated student. In M. Knauff, N. Sebanz, M. Pauen, I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. (pp. 894-9) Austin, TX: Cognitive Science Society.
12. Martin, B., Mitrovic, T., Mathan, S., & Koedinger, K.R. (2011). Evaluating and improving adaptive educational systems with learning curves. *User Modeling and User- Adapted Interaction: The Journal of Personalization Research (UMUAI), 21*(3), 249-283. [2011 James Chen Annual Award for Best UMUAI Paper]
13. Stamper, J.C. & Koedinger, K.R. (2011). Human-machine student model discovery and improvement using data. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Proceedings of the 15th International Conference on Artificial Intelligence in Education*, pp. 353-360. Berlin: Springer.
14. Koedinger, K.R., Baker, R.S.J.d., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J. (2010). A Data Repository for the EDM community: The PSLC DataShop. In Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (Eds.) *Handbook of Educational Data Mining*. Boca Raton, FL: CRC Press.

15. Brown, J. S., & VanLehn, K. (1980). Repair theory: A generative theory of bugs in procedural skills. *Cognitive Science, 4,* 379-426.
16. Matsuda, N., Lee, A., Cohen, W. W., and Koedinger, K. R. (2009). A computational model of how learner errors arise from weak prior knowledge. In *Proceedings of Conference of the Cognitive Science Society*. pp. 1288-1293. Amsterdam, The Netherlands.
17. Booth, J. L., & Koedinger, K. R. (2008). Key misconceptions in algebraic problem solving. In B. C. Love, K. McRae & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 571-576). Austin, TX: Cognitive Science Society.
18. VanLehn, K. (1982). Bugs are not enough: Empirical studies of bugs, impasses and repairs in procedural skills. *Journal of Mathematical Behavior, 3(2)*, 3-71.
19. Matz, M. (1980). Towards a process model for high school algebra errors. In D. Sleeman & J. S. Brown (Eds.), *Intelligent Tutoring Systems* (pp. 25-50). Orlando, FL: Academic Press.
20. Li, N., Cohen, W. W., & Koedinger, K. R. (2012). Problem Order Implications for Learning Transfer (pp. 185–194). Presented at the *Proceedings of the Eleventh International Conference on Intelligent Tutoring Systems*, Springer Berlin Heidelberg. doi:10.1007/978-3-642-30950-2_24
21. Aleven, V., McLaren, B., Sewall, J., & Koedinger, K. R. (2009). Example-tracing tutors: A new paradigm for intelligent tutoring systems. *International Journal of Artificial Intelligence in Education, 19,* 105-154.
22. MacLellan, C.J., Koedinger, K.R., Matsuda, N. (2014) Authoring Tutors with SimStudent: An Evaluation of Efficiency and Model Quality. *Proceedings of the 12th International Conference on Intelligent Tutoring Systems,* 551-560. Honolulu, HI.
23. Biswas, G., Schwartz, D., Leelawong, K., Vye, N. (2005) Learning by Teaching: A New Agent Paradigm for Educational Software. *Applied Artificial Intelligence, 19*, 363-392.
24. Gomoll, K., (1990). Some techniques for observing users. In Laurel B. (ed.), *The Art of Human-Computer Interface Design*, Addison-Wesley, Reading, MA, pp. 85-90.
25. Matsuda, N., Yarzebinski, E., Keiser, V., Raizada, R., William, W. C., Stylianides, G. J., & Koedinge, K. R. (2013). Cognitive anatomy of tutor learning: Lessons learned with SimStudent. *Journal of Educational Psychology, 105(4)*, 1152-1163. doi: 10.1037/a0031955
26. Matsuda, N., Cohen, W. W., Koedinger, K. R., Keiser, V., Raizada, R., Yarzebinski, E., Watson, S. P., & Stylianides, G. J. (2012). Studying the Effect of Tutor Learning using a Teachable Agent that asks the Student Tutor for Explanations. In M. Sugimoto, V. Aleven, Y. S. Chee & B. F. Manjon (Eds.), *Proceedings of the International Conference on Digital Game and Intelligent Toy Enhanced Learning* (DIGITEL 2012) (pp. 25-32). Los Alamitos, CA: IEEE Computer Society.
27. Matsuda, N., Griger, C. L., Barbalios, N., Stylianides, G., Cohen, W.W., & Koedinger, K. R. (2014). Investigating the Effect of Meta-Cognitive Scaffolding for Learning by Teaching. In S. Trausen-Matu, K. Boyer, M. Crosby & K. Panourgia (Eds.), *Proceedings of the International Conference on Intelligent Tutoring Systems* (pp. 104-113). Switzerland: Springer.
28. Matsuda, N., Keiser, V., Raizada, R., Tu, A., Stylianides, G. J., Cohen, W. W., & Koedinger, K. R. (2010). Learning by Teaching SimStudent: Technical Accomplishments and an Initial Use with Students. In V. Aleven, J. Kay & J. Mostow (Eds.), *Proceedings of the International Conference on Intelligent Tutoring Systems* (pp. 317-326). Heidelberg, Berlin: Springer.
29. Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin &Review, 14(2)*, 249-255.
30. Pavlik, P.I. & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied, 14*, 101-117.
31. Koedinger, K.R., Corbett, A.C., & Perfetti, C. (2012). The Knowledge-Learning-Instruction (KLI) framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science, 36 (5),* 757-798. ISSN: 0364-0213 print / 1551-6709 online DOI: 10.1111/j.1551-6709.2012.01245.x
32. Koedinger, K.R., Booth, J.L., & Klahr, D. (2013). Instructional complexity and the science to constrain it. *Science, 342*, 935-937.

# Simulated learners in peers assessment for introductory programming courses

Alexandre de Andrade Barbosa[1,3] and Evandro de Barros Costa[2,3]

[1] Federal University of Alagoas - Arapiraca Campus, Arapiraca - AL, Brazil
[2] Federal University of Alagoas - Computer Science Institute, Maceio - AL, Brazil
[3] Federal University of Campina Grande, Campina Grande - PB, Brazil

{*alexandre.barbosa@arapiraca.ufal.br*, *ebc.academico@gmail.com*}

**Abstract.** Programming is one of the basic competences in computer science, despite its importance, it is easy to find students with difficulties to understand the concepts required to use this skill. Several researchers report that the impossibility to achieve a quick and effective feedback, is one of the motivators for the problematic scenario. The professor, even when helped by the TAs, is not able to perform the reviews quickly, for this activity requires a huge amount of time. Fast feedback is extremely important to enable the learning of any concept. Some researches suggest the use of peer assessment as a means of providing feedback. However, it is quite common that the feedback provided by peers is not adequate. In this paper, we propose the use of simulated learners in a peer assessment approach as part of the teaching and learning processes of programming. Currently a software tool is being developed to include the proposal described in this paper.

## 1 Introduction

Programming is one of the basic competences in computer science, it is the basis for the development of several other competences required for professionals in the area. However, despite its importance, it is easy to find students who are demotivated and with difficulties to understand the concepts required to use this skill [7]. These difficulties causes a large number of failures, dropouts or the approval of students without the required level of knowledge [14] [6] [5].

Many factors are identified in literature as causing the problematic scenario related to programming courses. Several researchers report that the impossibility to achieve a quick, effective and individualized feedback, is one of the motivators for the problematic scenario [10] [12]. An individual follow up is impossible due to many students enrolled in the courses. In addition, there is a great complexity involved in the evaluation of a program, for it is necessary to understand how the programmer has developed the algorithm, so the professor needs to comprehend the line of reasoning adopted by the student. In this way, the professor, even when helped by the TAs, cannot provide an adequate and fast feedback about the solutions created by the students. This activity will require a huge amount

of time to manually open the code, compile, run and verify the output of every students solution for programming assignment. If the grading depends on the structure and the quality of code, in addition to program output correctness, the situation is a lot worse. Traditionally the real comprehension state of the contents of a programming course is known only months after the beginning of the course, when an evaluation activity is performed. After an evaluation it may be too late to make any intervention.

Fast feedback is of extreme importance to enable the learning of any concept [12]. Thus, some researches have been developed with the aim to propose methods and tools to facilitate the monitoring of the activities of students in programming courses. Some of these researches, such as [9][11][13], suggests the use of peer assessment as a means of providing fast and effective feedback. This solution is broadly used in Massive Open Online Courses (MOOCs), as described in [3][8], where the courses are applied to hundreds or thousands of people enrolled in them, and just as occurs in the context of programming, it is impossible for the professor to evaluate each solution. However, the peer assessment approach as a means of providing feedback has some problems. Many times the feedback provided by peers is not adequate, because the results are often not similar to the analysis of an expert [8]. It is quite common to find comments that are summarized to a phrase of congratulation or critique.

The reasons related to lack of effectiveness of feedback provided are quite distinct, these may occur due to poor understanding of the content of the activity, because of the student's low motivation, or due to the short time that one has available for the activities.

In [2] paper, it was observed the impact of learning was observed when a student is influenced by the performance of their peers, the authors describe that some students are encouraged to perform better, but others experiencing the same situations end up discouraged to perform better.

In this paper is proposed the use of simulated learners in a peer assessment approach used as part of the teaching and learning processes of programming. Two concerns are explored in this proposal: the first is related to the search of methods that enable a positive influence between students; the second concern is related to an approach that allows a less costly way of testing any proposal of applicability of peer assessment approach.

This paper is divided in five sections. In Section 2 the concept of peer assessment is presented. Observations on the implementation of peer assessment in a programming course context are shown in Section 3. The proposal of using simulated learners in the context of peer assessment for introductory programming is presented in Section 4. Finally the conclusions and future work are shown in the last section.

## 2 Peer Assessment

Peer assessment, or peer review, is an evaluation method where students have responsibilities that traditionally belong to professors only. Among these respon-

sibilities there are the review and the critique of the solutions proposed by their peers. This way, they can experience the discipline as students and also from the perspective of a TA. Usually in a peer assessment environment, students also conduct self assessment. This way, they can reflect on their solution when compared to other solutions, develop their critical thinking skills and improve understanding of the concepts covered in the course.

In traditional approach, the professor, even when helped by TAs, can not provide fast and adequate feedback for each solution proposed by the students. The comments provided by the professor are generic observations based on observation of all students solutions.

In accordance with [9], peer review is a powerful pedagogical method, because once students need to evaluate the work of their peers, they begin to teach and learn from each other. Thus, the learning process becomes much more active, making the learning qualitatively better than the traditional approach. Sstudents can spend more time on analysis and construction of their comments, creating more particular descriptions on a given solution and enriching discussion about the topic studied.

Thus, the use of peer review can reduce the workload on the professor, permitting the professor to focus on other pedagogical activities [9][3]. This evaluation approach can also enable the evaluation of large-scale complex exercises, which can not be evaluated in a automatically or semi-automatic fashion [3][8].

The success of peer assessment approach is strongly influenced by the quality of feedback provided. However, this feedback if often not adequate, the results are often not similar to the analysis of an expert [8]. In [8] is described that in many cases the evaluations of the students are similar to the TAs evaluation, however there are situations where the evaluations are graded 10% higher than the TAs evaluation, in extreme cases the grades could be 70% higher than the TAs evaluation. In [3] is mentioned that in general, there is a high correlation between the grades provided by students and TAs, but often in the evaluations from students the grades are 7% higher than the grades given by TAs.

Thus, we can conclude that peer assessment approach is a promising evaluation method, however there are improvements and adjusts to be applied to obtain richer discussions and more accurate assessments.

## 3 Peer Assessment in introductory programing courses

Human interaction is described as an essential feature for learning in many domains, including the introductory programming learning [13]. In classroom programming courses the contact between students occurs on a daily basis, allowing, for example, the discussion of the problems presented in the exercise lists, the developed solutions and the formation of groups for the projects of the course. This contact is many times inexistent in online programming courses, interactions in this environment are the human-machine type. Thus, using the peer assessment approach may enable human interaction on online courses, or enhance the interaction between humans in presencial classroom courses.

To encourage the assimilation of the topics, the use of practical exercises is quite common in programming courses, the practice of programming skills is crucial for learning. Many researchers also argue that the programming learning involves the reading and understanding of third-party code. Through peer assessment approach both characteristics can be obtained. The professor can develop new exercises, or choose problems proposed by others, while students will have to observe, understand and evaluate the codes of their peers, as well to compare these codes with their solution.

In [11] the use of a peer assessment approach to the context of programming courses is described, this approach is supported by a web application. The results described on the paper have a high correlation between the evaluations of the TAs and students, the correlation is lowest when the complexity of the exercise is higher.

An approach of peer assessment evaluation for the context of programming learning, also supported by a web application is presented in [9]. Five activities where graded using peer assessment, the occurrence of conflicts ranged from 61 % to activity with a lower incidence of conflict, up to 80 % for the activity with the highest occurrence of conflicts. The system considers that a conflict occurs when the student does not agree with the assessment provided.

In [9] the authors describes that if the peer reviews are conducted in an inadequate way, the failure rates can increase. For the teaching approach used at the programming course described in [13] there are two types of activities that require assessment, quizzes and mini projects. Among these activities only the mini projects are evaluated through peer assessment. Thus, students are not overloaded and the approach can be used appropriately.

Another problem that can emerge with the use of peer review in a programming context, is the increase of plagiarism. Once the assessment activity will be distributed among the students, the similarities of self-identification of codes can become more complicated. However, solutions are widely used to carry out the detection automatically similarities, such as MOSS [1] e GPLAG [4].

## 4 Simulated learners as peers in a peer assessment environment for introductory programing courses

In previous sections the advantages and disadvantages associated with the use of peer assessment in a general context, and when applied to the context of programming courses have been described. In both cases, the success of the approach is strongly influenced by the quality of the feedback given. Therefore, it is necessary to identify situations where there is inadequate feedback as well as conflict situations. Situations where inadequate feedback occurs are when, for any reason, the feedback does not help in the learning process. Conflict situations occur when the student does not agree with the assessment provided, or when there are huge variations on the evaluations provided. To perform a validation of this proposal or of any proposal involving peer assessment, it is necessary to

allocate the resources of time, physical space and adequate human resources. Thus, it can be said that the test of this approach is a costly activity.

Two concerns are explored in this proposal: how we can achieve methods that enable a positive influence between students in peer assessment environments, in other words, how a student can give a high quality feedback to their peers; and how a peer assessment approach can be tested with a lower cost, since any validation of these assessment approaches requires a huge amount of resources.

### 4.1 A scenario of use of peer assessment with simulated learners

Traditionally in a peer assessment environment, the professor must create the assignment and a set of assessment criteria. Then students develop their solutions observing the assessment criteria and submitting the solution to be evaluated by their peers. Each student's evaluation must meet the assessment criteria. The students should provide comments to peers explaining the reasons associated to the outcome and a grade or an evaluation concept (eg. A-, B+, C). Each student will have their code evaluated by their peers, and should assess the codes of other students. In Figure 1, it is illustrated the scenario previously described. There are variations in ways peer assessment approach is used, the scenario just mentioned has many characteristics which are similar to all the variations.

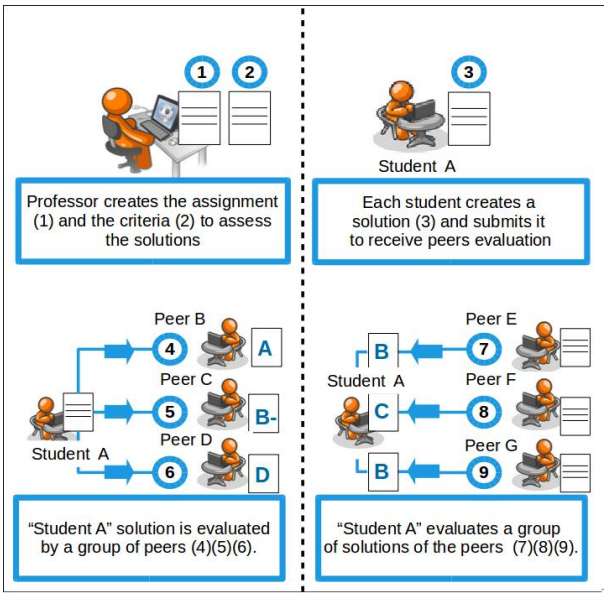



**Fig. 1.** A traditional peer assessment environment

In any peer assessment approach, it is possible to adopt pairing algorithms. Thereby, it is assured that evaluations are conducted by students with different

levels of knowledge. A student with low understanding of the subject will not be allocated to evaluate the work of another student in the same situation. Students with difficulties can clarify their doubts, while students with good understanding of the content should provide a good argumentation about their knowledge. However, it is not possible to ensure that a student evaluates the code that is the ideal for his/her learning and level of knowledge. As an example, in Figure 1, it is not possible to know if student "A" code is the best for peers "B", "C" and "D".

When a student does not agree with the evaluation provided by their peers, he/she will be able to request the intervention of the professor. This conflict situations are identified in [9]. However, in traditional peer assessment approach is not possible to identify incorrect evaluations provided by a student, or students that create biased evaluations only to help their fellows. As an example, in Figure 1, it is possible to see that different grades were given, but it is not possible to determine if the correct evaluations were given by peer "B", "C" or "D".

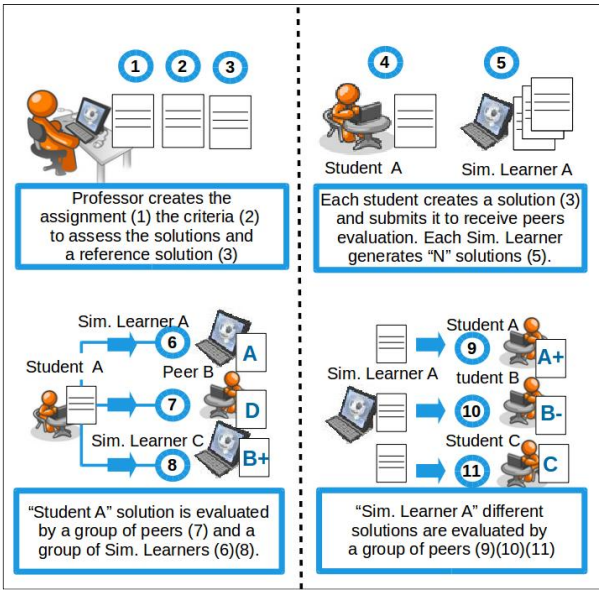



**Fig. 2.** A peer assessment environment using simulated learners

In a peer assessment environment that uses simulated learners, it is possible to solve the previous problems. As in traditional approach, the professor must create the assignment and a set of assessment criteria; in addition to that, he/she should provide a reference solution. Then, students develop their solutions, observing the assessment criteria and submitting the solution to be evaluated by their peers. At the same time, once a pairing algorithm can perform pairing of

the evaluators, each simulated learner must generate a code that is ideal for the learning and appropriate to the level of knowledge of each one of their peers, in this case, real students. Each student will have their code evaluated by their peers and by simulated students, and they should assess codes of other students, and codes of simulated students. In Figure 2 it is illustrated peer assessment environment with simulated learners. As an example, in Figure 2, it is possible to see that the simulated learner "A" generates a set of codes that are ideal for each student: "A", "B" and "C".

The identification of incorrect evaluations provided by a student, as well as students, who perform biased evaluations, could be carried out through the comparison of student's evaluations and the simulated student's evaluations. As an example, in Figure 2, it is possible to see that student "B", made an evaluation that is very different from the simulated learner's evaluations. In this way, it is possible to verify if the student did not understand the solution, or if the evaluation was created to help their fellows only.

**Providing useful solutions to student learning** A useful solution for the student learning does not always match the presentation of a correct and efficient code. Within the context of peer review may be more useful to display an incorrect code, as a way to make students to provide a set of review observations. To identify which type of code is best for a student; simulated learners can consult the representation of their cognitive status. In that way, it will be possible to the simulated learner identify the student misconceptions and errors in previous assignments, and generate variations of the reference solution that suits best for the student. Since multiple simulated students will be used, the codes that will be shown to students can range from efficient, correct and complete solutions to incorrect and/or incomplete solutions. Like that, it will be possible to check if students have different skills related to the content. To generate the variations from the reference solution, it is possible to combine testing techniques, such as mutant generation. Each code can be generated through the use of data related to the most common student's mistakes, emulating these behaviors and creating codes that are useful to learning. Once the research is in a preliminary stage, it is still not clear which artificial intelligence approaches should be used on the implementation of simulated students behaviors.

**Assessment of students solutions** Unlike what occurs in other contexts, for programming the evaluation of a solution can be automated or semi-automated. Typically a set of unit tests is applied to the code proposed by a student, who receives an indication that his/her code may be correct or incorrect, but no hint or comment is provided. Some researchers have investigated the use of different techniques to help assessment of codes and provide some guidance; these techniques usually employ software engineering metrics. Thus, simulated learners must be able to identify which subset of metrics can be used to perform the evaluation of the proposed solution for a student. The simulated learner should select the set of metrics that fits best to the objectives of the assignment and

the level of understanding that the student has at that moment. For each level of learning the same student can learn better if the set of metrics is properly selected. Each simulated student will use different strategies to evaluate the solutions provided by real students. Therefore, a variation between evaluations of simulated students is expected to occur. If an evaluation provided by a student has a very large variation in relation to the set of evaluations of simulated students, it will be necessary to investigate the motivation of this disparity. An acceptable variation threshold can be used to identify incorrect evaluations provided by students.

**Discussing assessment criterias** Once software engineering metrics were used in the evaluation, the explanation given by the simulated learner throughout the presentation of a set of metrics, is associated to the explanation of the metric choice and, possibly, of the snippet of the code where the observation is pertinent. Thereby, the simulated learner can help the professor to identify inadequate feedback, whenever an evaluation of a student is very different from the evaluation of a simulated learner, the professor and his tutors can then intervene.

### 4.2 Validation of peer assessent using simulated learners

Any validation of peer assessment approaches requires lots of physical space and a huge amount of human resources. As an example, if a validation of a pairing algorithm has to be done, it will be necessary to use a set of N students; this set must allow the creation of different profiles for evaluation of the pairing alternatives. The greater the possibilities of matching, the greater the amount of students required. Through the use of simulated learners any operational proposal of peer assessment can be tested at a much lower cost, since the physical space and human resources are drastically reduced. The researcher can determine how much of human resource will be available, replacing the students with simulated students. The researcher can also specify the desired behavior of students; the simulated students should emulate students with a high degree of understanding of the contents or with low understanding. After obtaining initial results with the use of simulated learners, the number of human individuals participating in an experiment can be increased, since it may be interesting to obtain a greater statistical power associated with the conclusions.

## 5 Conclusions and further work

In this paper, we have proposed the use of simulated learners in a peer assessment approach adopted as a support part of a programming course. The use of simulated learners as presented in this proposal aims to two goals: influence the students to provide better quality feedback; and allow for a less costly validation for peer assessment applied to programming contexts.

The research associated with the proposal presented in this paper is in a preliminary stage. Thus, the effectiveness of this proposal will be further evaluated in controlled experiments executed in the future. An open source software tool is being developed to include all aspects described throughout this proposal.

## References


1. Bowyer, K., Hall, L.: Experience using "moss" to detect cheating on programming assignments. In: Frontiers in Education Conference, 1999. FIE '99. 29th Annual. vol. 3, pp. 13B3/18–13B3/22 (Nov 1999)
2. Frost, S., McCalla, G.I.: Exploring through simulation the effects of peer impact on learning. In: Proc. of the Workshops at the 16th International Conference on Artificial Intelligence in Education AIED 2013, Memphis, USA, July 9-13, 2013 (2013), http://ceur-ws.org/Vol-1009/0403.pdf
3. Kulkarni, C., Wei, K.P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., Klemmer, S.R.: Peer and self assessment in massive online classes. ACM Trans. Comput. Hum. Interact. 20(6), 33:1–33:31 (Dec 2013)
4. Liu, C., Chen, C., Han, J., Yu, P.S.: Gplag: Detection of software plagiarism by program dependence graph analysis. In: Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 872–881. KDD '06, ACM, New York, USA (2006)
5. Mason, R., Cooper, G.: Introductory programming courses in australia and new zealand in 2013 - trends and reasons. In: Proc. of the Sixteenth Australasian Computing Education Conference - Volume 148. pp. 139–147. ACE, Darlinghurst, Australia (2014)
6. Mason, R., Cooper, G., de Raadt, M.: Trends in introductory programming courses in australian universities: Languages, environments and pedagogy. In: Proc. of the Fourteenth Australasian Computing Education Conference - Volume 123. pp. 33–42. ACE, Darlinghurst, Australia (2012)
7. McCracken, M., Almstrum, V., Diaz, D., Guzdial, M., Hagan, D., Kolikant, Y.B.D., Laxer, C., Thomas, L., Utting, I., Wilusz, T.: A multi-national, multi-institutional study of assessment of programming skills of first-year cs students. In: Working Group Reports from ITiCSE on Innovation and Technology in Computer Science Education. pp. 125–180. ITiCSE-WGR, New York, USA (2001)
8. Piech, C., Huang, J., Chen, Z., Do, C.B., Ng, A.Y., Koller, D.: Tuned models of peer assessment in moocs. CoRR abs/1307.2579 (2013)
9. de Raadt, M., Lai, D., Watson, R.: An evaluation of electronic individual peer assessment in an introductory programming course. In: Proc. of the Seventh Baltic Sea Conference on Computing Education Research - Volume 88. pp. 53–64. Koli Calling, Darlinghurst, Australia (2007)
10. Singh, R., Gulwani, S., Solar-Lezama, A.: Automated feedback generation for introductory programming assignments. In: Proc. of the 34th ACM SIGPLAN Conference on Programming Language Design and Implementation. pp. 15–26. PLDI, New York, USA (2013)
11. Sitthiworachart, J., Joy, M.: Computer support of effective peer assessment in an undergraduate programming class. Journal of Computer Assisted Learning 24(3), 217–231 (2008)
12. Stegeman, M., Barendsen, E., Smetsers, S.: Towards an empirically validated model for assessment of code quality. In: Proc. of the 14th Koli Calling Int. Conf. on Computing Education Research. pp. 99–108. Koli Calling, New York, USA (2014)

13. Warren, J., Rixner, S., Greiner, J., Wong, S.: Facilitating human interaction in an online programming course. In: Proc. of the 45th ACM Technical Symposium on Computer Science Education. pp. 665–670. SIGCSE, New York, USA (2014)
14. Yadin, A.: Reducing the dropout rate in an introductory programming course. ACM Inroads 2(4), 71–76 (2011)

# Simulated Learners for Testing Agile Teaming in Social Educational Games

Steeve Laberge and Fuhua Lin

School of Computing and Information Systems, Athabasca University, Edmonton, Canada
slaberge@acm.org, oscarl@athabascau.ca

**Abstract.** This paper proposes an approach for creating and testing an multiagent systems based adaptive social educational game (SEG), QuizMASter, using the concept of simulated learners to overcome experimentation complexity and unpredictable student availability, as is typical with online learning environments. We show that simulated learners can play two roles. First, it can be used for testing the game planning, scheduling and adaptive assessment algorithms. With some degree of success met with our initial experimentation with QuizMASter, advanced planning and coordination algorithms are now needed to allow the game-based assessment platform to realize its full potential. The multi-agent system approach is suitable for modeling and developing adaptive behaviour in SEGs. However, as we have found with our early prototypes, verifying and validating such a system is very difficult in an online context where students are not always available. MAS-based assessment game planning and coordination algorithms are complex and thus need simulated learners for testing purposes. Second, to overcome unpredictable student availability, we modeled QuizMASter as a new class of socio-technical system, human-agent collective (HAC). In the system, human learners and simulated learners (smart software agents) engage in flexible relationship in order to achieve both their individual and collective goals, while simulated learners are selected for serving as virtual team members.

**Keywords:** social educational agents, multiagent systems, simulated learners

## 1 Introduction

For decades, educational games have proven to be an effective means to motivate learners and enhance learning. Social (multi-player) educational games (SEGs) offer many opportunities to improve learning in ways that go beyond what a single-player game can achieve because SEGs allow players to be social, competitive, and collaborative in their problem solving. The presence of other players can be used to increase playability and to help teach team-work and social skills. SEGs promote intragroup cooperation and intergroup competition [1]. However, existing SEGs share many of the shortcomings of classroom role-playing. Setting

up existing SEGs is logistically challenging, expensive, and inflexible. Furthermore, players become bored after going through existing SEGs once or twice.

To test such a social educational game, we face two difficulties. One is how to test the planning and scheduling algorithms. Another is how to meet the need of agile team formation. In SEGs, group formation has big impact on group learning performance. Poor group formation in social games can result to homogeneity in student characteristic such that the peer learning is ineffective. Thus, there is a need to constitute a heterogeneous group SEGs that constitutes students with different collaborative competencies and knowledge levels. However, without empirical study it becomes difficult to conclude which group characteristics are desirable in the heterogeneity as different game-based learning needs may require different group orientations. Previous research has focused on various group orientation techniques and their impact on group performance like different learning styles in group orientation [2–4]. However, there is need to investigate the impact of other group orientation techniques on group performance like grouping students based on their collaboration competence levels. Furthermore, most of the previous research in group-formation focuses on classroom based learning. Also, it lacks the true experiment design methodology that is recommended when investigating learning outcomes from different game-based learning strategies. Simulated learners methodology [5] has shown a promising way to solve these challenges.

In this paper, we show that simulated learners can play two roles. First, it can be used for testing the game planning, scheduling and adaptive assessment algorithms. Second, working with human learners and forming human-agent collectives (HAC), simulated learners serve as virtual team members to enable asynchronous game-based learning in a context where student availability is unpredictable. This paper is structured as follows: In Section 2 we discuss recent advancements and related work. Section 3 describes QuizMASter. Section 4 presents the proposed architecture for development of QuizMASter. Section 5 explains how we intend to use simulated learners for testing QuizMASter. Finally, Section 6 concludes.

## 2 Related Work

Researchers have found that learning can be more attractive if learning experiences combine challenge and fun [6]. As social networks have become popular applications, they have given rise to social games. This kind of game is played by users of social networks as a way to interact with friends [7] and has become a part of the culture for digital natives. Social games have unique features that distinguish them from other video games. Those features are closely linked with the features of social networks [8]. Social games can make a contribution to social learning environments by applying game mechanics and other design elements, 'gamifying' social learning environments to make them more fun and engaging. For games to be effective as a learning tool, a delicate balance must be maintained between playability and educational value [9, 10], and between

game design and learning principles. Methods have been proposed for making valid inferences about what the student knows, using actions and events observed during gameplay. Such methods include evidence-centered-design (ECD) [11, 12]; the learning progressions model [13], the ecological approach to design of e-learning environments [14], stealth assessment [15], game analytics [16], and learning analytics [17]. Most of the new concepts target an ever-changing learning environment and learner needs, as today's education moves toward a digital, social, personalized, and fun environment. Moreover, as is the case for all competitive games, an equal match between players is essential to self-esteem and to maintain a high degree of player interest in the game. Hence, we need mechanisms and models that can aggregate the current performance and preferences of players, and accurately predict student performance in the game. Software agents have been used to implement consistent long-term intelligent behaviour in games [18], multi-agent collaborative team-based games [19], and adaptive and believable non-player character agents simulating virtual students [20]. The use of agent technologies leads to a system characterized by both autonomy and a distribution of tasks and control [21]. This trend has two aspects. First, game-based learning activities should be carefully orchestrated to be social and enjoyable. Second, game scheduling and coordination should be highly adaptive and flexible. However, nobody has yet developed models, algorithms, and mechanisms for planning, scheduling, and coordination that are suitable for creating and testing SEGs.

## 3 QuizMASter

QuizMASter is designed to be a formative assessment tool that enables students to be tested within a multi-player game [22]. Two or more students simultaneously log in remotely to the system via a Web-based interface. Each student is represented by one avatar in this virtual world. Students are able to view their own avatar as well as those of their opponents.

Each game has the game-show host who is also represented by an avatar visible to all contestants [22]. The game-show host poses each of the game questions to all the contestants. The students hear the voice of the host reading each question and view them displayed on their screens. They individually and independently from one another answer each question by, for instance, selecting an answer from available choices in a multiple-choice format. Each correct answer would receive one mark. Figure 1 shows a screen shot of QuizMASter.

### 3.1 Characteristics of QuizMASter

The environment for QuizMASter has the following characteristics:

**Flexibility.** The environment for QuizMASter needs flexibility for game enactment, to be able to cope with dynamic changes of user profiles, handle fragmentation of playing and learning time needed to accomplish activities and tasks,


**Fig. 1.** QuizMASter in Open Wonderland

adequately handle exceptional situations, predict changes due to external events, and offer sufficient interoperability with other software systems in educational institutions. Individual learners have particular interests, proficiency levels, and preferences that may result in conflicting learning goals.

**Social ability and interactivity.** The environment for QuizMASter should encourage interaction and collaboration among peers, and should be open to participation of students, teachers, parents, and experts on the subjects being taught. Web 2.0 has had a strong influence on the ways people learn and access information, and schools are taking advantage of this trend by adopting social learning environments. One way to engage learners in a collaborative production of knowledge is to promote social rewards.

**User control.** One of the most desirable features of social education games is to empower players with control over the problems that they solve. For example, in QuizMASter, students, parents, and teachers can design new rules to create their own games and modify the game elements to fit different knowledge levels.

**Customization.** Customization is a core principle that helps accommodate differences among learners [23]. Teachers could build a QuizMASter that has its own style and rules to determine the game's level of difficulty, to gear the game for specific goals or a specific group of learners. Some teachers may be interested in sharing collections of rules to fit the learning and play styles of their students. Like teachers, learners/players can be co-creators of their practice space through building new game scenarios, creating their own rules, sharing their strategies and making self-paced challenges [23].

## 4 The Proposed Architecture

Multi-agent technologies are considered most suitable for developing SEGs as it will lead to systems that operate in a highly dynamic, open, and distributed environment. In an MAS-based SEG, each learner/player is represented as an autonomous agent, called learner agent. MAS technologies, such as goal orientation and the Belief-Desire-Intention (BDI) paradigm, is used as the foundation for the agent architecture. These learner agents are able to reason about the learning goals, the strengths and weaknesses of learners and update the learner models.

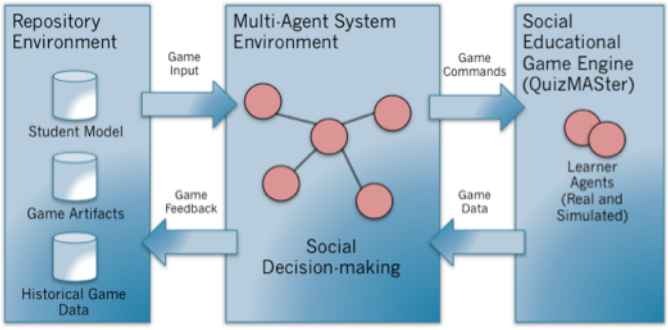



**Fig. 2.** Architecture for MAS-Based Social Educational Game Environment

Whenever a learner enters the system to play a social educational game, the learner agent will retrieve her/his learner model and acquire preferences about the current game-playing, and then send to a game management agent (GMA) of the system. The GMA is designed for setting up and maintaining teams for the system. The GMA will assign the learner to participate in a most suitable team that is undermanned according to the profile and preferences of the learner. The team will be configured in accordance with the game model by the GMA. Once the team has been completely formed, the GMA will create a game scheduling agent (GSA), a game host agent (GHA), and an assessment agent (AA) for each team. The GSA will continuously generate a game sequence dynamically adapted to the team's knowledge level (represented as a combined learner model [24]. The GHA will receive the game sequence from the scheduling agent and execute game sequence with the learners in the team. It will also be responsible for capturing data about learner/player performance. The AA will receive and interpret game events and communicate with the learner agents to update the learner model as necessary.

The GSA will dynamically schedule the game on the fly through interacting with other agents with a coordination mechanism, considering both the current world state and available resources, and solving conflicts in preferences and learning progression between the agents. The goal of the GSA is to optimize

the playability and educational values. We will model the game elements as resources. To solve the distributed constraint optimization problem, we are developing multiagent coordination mechanisms and scheduling algorithms to be used by the GSA.

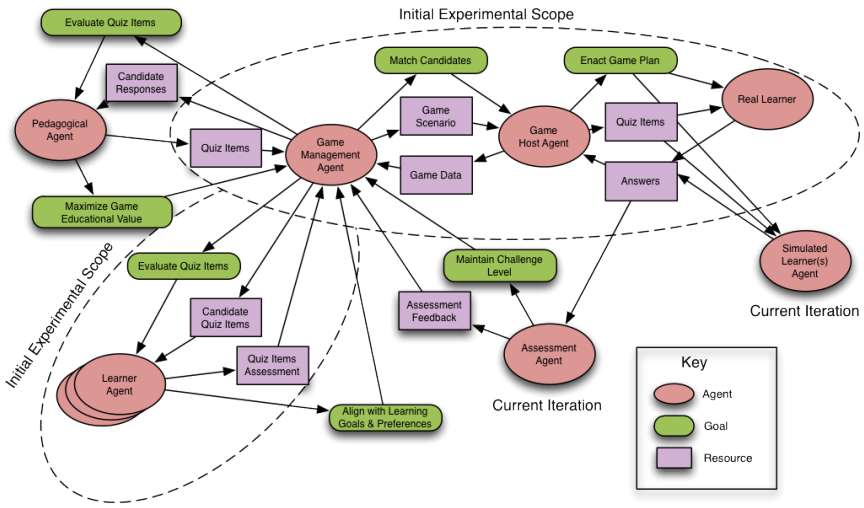



**Fig. 3.** MAS-Based SEG Agent Interaction Model

### 4.1 Planning and Scheduling Algorithms

The planning algorithms refer to the (local) planning algorithm of learner agents. To develop planning algorithms for learner agents, the following supporting models have been taken into consideration: (i) Learner models that accumulate and represent beliefs about the targeted aspects of skills. They are expressed as probability distributions for competency-model variables (called nodes) describing the set of knowledge and skills on which inferences are to be based. (ii) Evidence models that identify what the learner says or does, and provide evidence about those skills that express how the evidence depends on the competency-model variables in a psychometric model. (iii) Task/action models that express situations that can evoke required evidence. To design an action model, we adopt a model called Fuzzy Cognitive Goal Net [25] as the planning tool by combining the planning capability of Goal Net and reasoning ability of Fuzzy Cognitive Maps (FCMs). These FCMs give the learner agent a powerful reasoning ability for game context and player interactions, giving the task model accurate context awareness and learner awareness. We are developing coordination mechanisms

for the GMA and the GSA to solve the problem of team formation, scheduling and coordination in a highly flexible and dynamic manner. We considered the following concepts or methods:

(i) Contract-net protocols (CNPs) are used as a coordination mechanism by the GMA with a game model repository to timely form a team from all available players, using mutual selection and exchanging information in a structured way to converge on assignments. Each involved learner can delegate the negotiation process to its agent. These agents will strive to find a compromise team-joining decision obeying hard learning constraints while simultaneously resolving individual conflicts of interest.

(ii) The problem of scheduling and customizing a social educational game can be solved through social-choice-based customization. We view the SEG gameplay design as an optimization problem. Resources must be allocated through strategically scheduling, and coordinating a group of players according to their preferences and learning progressions. The constraints include key learning principles that inform the design of mechanics: challenge, exploration, risk taking, agency, and interactions [26-27]. The objective of the GSA is to maximize the learnability and engagement of the learners in the group. Social choice theory in MAS concerns the design and formal analysis of methods for aggregating preferences of multiple agents and collective decision-making and optimizing for preferences [28-29]. For example, we use a voting-based group decision-making approach such as Single Transferable Voting [30] to aggregate learner preferences and learning progression because it is computationally resistant to manipulation [31]. The purpose is to take information from individuals and combine it to produce the optimal result.

(iii) To support the need for dynamic decision making in the MAS-based SEG architecture, our current line of investigation is the concept of social choice Markov Decision Process (MDP) as recently proposed by Parkes and Procaccia [32]. In a social choice MDP, each state is defined by "preference profiles", which contain the preferences of all agents against a set of alternatives for a given scenario. The course of action from any given state is determined by a deterministic social choice function (the policy, in the context of the MDP) that takes into account the likelihood of transitions and their rewards. However, a preference profile is subject to change over time, especially in a live SEG context. For example, a learner that unexpectedly answers a question initially deemed beyond the learner's perceived level of comprehension would likely trigger a change of belief in the agents and potentially alter their ranking of alternatives. And since the number of alternatives in a SEG can be very large, the state space for any given SEG is huge, making the computation of optimal decision-making policies excessively difficult. We solve this problem by exploiting symmetries that exist in certain game types (e.g. in a quiz game SEG format, using a reduced set of question types that share common characteristics as a basis for alternatives as opposed to individual questions).

## 5 Simulated Learners

It is our view that the Belief-Desire-Intention (BDI) model is ideally suited for modeling and simulating learner behaviour. According to Jaques and Vicari (2007) [33], intelligent agents based on Bratman's Belief-Desire-Intention model, or BDI agents, are commonly used in modeling cognitive aspects, such as personality, affect, or goals. Píbil et al. (2012) claim BDI agent architecture is "a currently dominant approach to design of intelligent agents" [34]. Wong et al. (2012) describes the suitability of the BDI agent model for applications where both reactive behavior and goal-directed reasoning are required [35]. Soliman and Guetl (2012) suggest that BDI maps well onto models for pedagogically based selection of sub plans within a hierarchical planning strategy – "apprenticeship learning model" given as example [36]. They also talks about advantage of breaking plans down into smaller plans to allow for different "pedagogical permutations" allowing the agent to adapt to different learning styles, domain knowledge, and learning goals. Norling (2004) attributes the successful use of BDI agents for modeling human-like behavior in virtual characters to BDI's association to "folk psychology" [37]. This allows for an intuitive mapping of agent framework to common language that people use to describe the reasoning process. Of particular importance to this study is the way that implementations of the BDI architecture model long-term or interest goals. We have selected the Jason$^{TM}$ [38] platform for providing multi-agent BDI programming in AgentSpeak.

A shortcoming of the BDI paradigm is that although it is intended to be goal-driven, in most implementations this means/amounts to using goals to trigger plans, but does not support the concept of long-term goals or preferences [39], such as a student's long term learning goals, or the pedagogical goals of a CA. They feel that these types of goals are difficult to represent in most BDI systems because they signify an ongoing desire that must be maintained over a long period of time compared to relative short goal processing cycles. It is left to the developer to implement this type of preference goal through the belief system of the agent, modifications to the platform or environment, or other methods of simulating long-term goals.

Hübner, Bordini, and Wooldridge (2007) describe plan patterns for implementing declarative goals, with varying levels of commitment in AgentSpeak [40]. Bordini et al. (2007) expand on this in their chapter on advanced goal-based programming [38]. While AgentSpeak and Jason support achievement goals, these patterns are intended to address the lack of support for "richer goal structures", such as declarative goals, which they feel are essential to providing agents with rational behaviour. Pokahr et al. (2005) point out that the majority of BDI interpreters do not provide a mechanism for deliberating about multiple and possibly conflicting goals [41]. It is worth noting that there are "BDI inspired" systems that are more goal-oriented, such as Practionist and GOAL [42]. The Jason multi-agent platform for BDI agents was selected for this project because it is a well-established open-source project that is being actively maintained. It supports both centralized and distributed multi-agent environments. Píbil et

al. (2012) describes Jason as "one of the popular approaches in the group of theoretically-rooted agent-oriented programming languages" [34]. A major advantage of Jason is that it is easy to extend the language through Java based libraries and other components. Internal actions can allow the programmer to create new internal functionality or make use of legacy object-oriented code [38]. However, Píbil et al. (2012) caution that the use of such extensions, if used too heavily, can make the agent program difficult to comprehend without understanding the functionality of the Java code [34]. They raise the concern that novice programmers have few guidelines for choosing how much to program in AgentSpeak, and how much too program in Java. The usefulness of being able to extend Jason can be demonstrated by two examples of current research into integrating BDI with Bayesian Networks. Modeling of some student characteristics requires a probabilistic model; Bayesian Networks (BN) being a popular choice in recent years [43-44]. Recent work by Kieling and Vicari (2011) describes how they have extended Jason to allow a BDI agent to use a BN based probabilistic model. Similarly, Silva and Gluz (2011) extend the AgentSpeak(L) language to implement AgentSpeak(PL) by extending the Jason environment. AgentSpeak(PL) integrates probabilistic beliefs into BDI agents using Bayesian Networks [45]. Experimentation with QuizMASter to date has enabled the modelling of simulated learners in virtual worlds with an initial focus on their appearance, gestures, kinematics, and physical properties [46]. Recent related research work in that area has been on the creation of engaging avatars for 3D learning environments [47]. Employing the theory of Transformed Social Interaction (TSI) [48], simulated learners were designed with the following abilities:

(i) Self-identification: The self-identification dimension of TSI was implemented using facial-identity capture with a tool called FAtiMA. Each of the users' face were morphed with their default avatar agent's face to capitalize on human beings' disposition to prefer faces similar to their own and general preference of appearing younger (see Fig. 4).

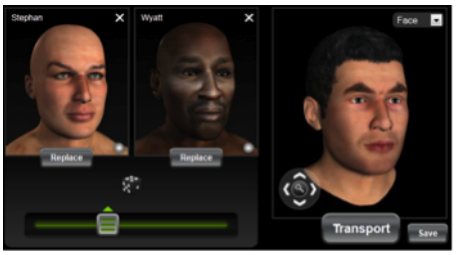



**Fig. 4.** Transformed Social Interaction – Image Morphing Technique

(ii) Sensory-abilities: Sensory-abilities dimension of TSI were implemented using a movement and visual tracking capability. The general challenge of sensory abilities implementation lies in two areas: the complexity of human senses and

the processing of sensory data of different modality and historicity. For the reason of simplicity, only visual tracking capability was exploited.

(iii) Situational-context: The situational-context dimension of TSI was implemented by using the best-view feature of Open Wonderland, whereby the temporal structure of a conversation can be altered.

The main idea of this research has been to explore the methodology for developing simulated learners for simulating and testing SEGs. That is, behind a simulated learner is an agent. Or we can say a simulated learner is an agent's avatar. All avatars, including real students' avatars and agent-based simulated learners, live in the virtual worlds, while the agents live in the multi-agent system. The integration of multi-agent systems with virtual worlds adds intelligence to the SEG platform and opens a number of extremely interesting and potentially useful research avenues concerning game-based learning. However, the advanced algorithms that support game planning, coordination and execution are difficult to test with real subjects considering the overhead involved is seeking authorization and the unpredictable availability of real life subjects in an online environment. This where an expanded view of simulated learners comes into play. The advantages of a simulated environment that closely approximates human behaviour include: (1) It allows for rapid and complete testing of advanced algorithms for game based adaptive assessment as well as SEG planning, coordination and execution in a simulated environment. The efficiency of the algorithms can be measured without first securing the availability of students; (2) With proper learner modeling and adaptive behaviour, simulated learners can engage with real life learners in friendly competitive games for the purpose of formative assessment, again working around the issue of availability of real students in an online learning environment.

## 6 Conclusions

As our recent experimentation suggests, many outstanding challenges must be addressed in developing intelligent SEGs. As we get closer to real world testing of our experimental game based assessment framework, we are faced with the complexity of enrolling real life learners in an e-learning environment and the variability that human interactions introduce in the measurement of adaptive algorithm efficiency. This is where we see the value of simulated learners. At this stage of our research, simulated learners have been rendered as Non Person Characters (NPCs) controlled by BDI agent running in the multi-agent system based virtual world. Our medium term goal is to extend the existing system to a particular learning subject (e.g., English language learning) to verify the effectiveness of the proposed virtual assessment environment and the benefit that students perceive from interacting with the proposed NPCs.

For simulated learners to be successful in our experimental framework, they must closely approximate the performance of real learners. The simple, pre-encoded behaviour we have implemented so far in the NPCs for QuizMASter will not suffice to demonstrate the efficiency of our adaptive algorithms and

allow for simulated learner agents to act as virtual players in our game based assessment framework. Current outstanding research questions within our group are:

1. How do we add intelligence and adaptive behaviour to the simulated learner agents while preserving our ability to obtain predictable and repeatable test results from our adaptive MAS framework?
2. How much autonomy can we afford to give to simulated learners in terms of independent thought and action, and to which degree should a simulated learner be able to adjust its behaviour as a function of its interactions with other agents, including real life learners?
3. How do we incorporate modern game, learning and assessment analytics in the supporting adaptive MAS framework in order to maximize the value of simulated learners as a means to perform non-intrusive, formative assessment?

## References


1. Romero, M., Usart, M., Ott, M., Earp, J., de Freitas, S., and Arnab, S.: Learning through playing for or against each other. Promoting Collaborative Learning in Digital Game Based Learning. ECIS 2012 Proceedings. Paper 93 (2012)
2. Alfonseca, E., Carro, R. M., Martin, E., Ortigosa, A., and Paredes, P.: The impact of learning styles on student grouping for collaborative learning: a case study. User Modeling and User-Adapted Interaction, 16(3–4), 377–401 (2006)
3. Deibel, K.: Team formation methods for increasing interaction during in-class group work. In ACM SIGCSE Bulletin (Vol. 37, 291–295). Caparica, Portugal (2005)
4. Grigoriadou, M., Papanikolaou, K. A., and Gouli, E.: Investigating How to Group Students based on their Learning Styles. In In ICALT, 2006 1139–1140 (2006)
5. McCalla, G. and Champaign J.: AIED Workshop on Simulated Learners. AIED 2013 Workshops Proceedings, Volume 4 (2013)
6. Vassileva, J.: Toward social learning environments. IEEE Transactions on Learning Technologies, 1(4), 199–213 (2008)
7. Klopfer, E., Osterweil, S., and Salen, K.: Moving learning games forward: obstacles, opportunities and openness, the education arcade. MIT (2009)
8. Jarvinen, A.: Game design for social networks: Interaction design for playful dispositions. In Stephen N. Spencer (Ed.), Proceedings of the 2009 ACM SIGGRAPH Symposium on Video Games (Sandbox '09), 95–102. ACM, New York, NY, USA (2009)
9. Van Eck, R.: Building Artificially Intelligent Learning Games. In V. Sugumaran (Ed.), Intelligent Information Technologies: Concepts, Methodologies, Tools and Applications, 793-825 (2008)
10. Augustin, T., Hockemeyer, C., Kickmeier-Rust, M. D., and Albert, D.: Individualized Skill Assessment in Digital Learning Games: Basic Definitions and Mathematical Formalism. IEEE Trans. on Learning Technologies, 4(2), 138-147 (2011)
11. Mislevy, R. J., and Haertel, G. D.: Implications of evidence-centered design for educational testing. Educational Measurement: Issues and Practice, 25(4), 6–20 (2006)
12. Mislevy, R. J., Steinberg, L. S., and Almond, R. G.: On the structure of educational assessments. Measurement: Interdisciplinary Research and Perspectives, 1, 3-67 (2003)

13. Corcoran, T., Mosher, F. A., and Rogat, A.: Learning Progressions in Science: An Evidence-Based Approach to Reform (Research Report No. RR-63). Center on Continuous Instructional Improvement, Teachers College—Columbia University (2009)
14. McCalla, G.: The Ecological Approach to the Design of E-Learning Environments: Purpose-based Capture and Use of Information About Learners. Journal of Interactive Media in Education. (7), 1-23 (2004)
15. Shute, V. J.: Stealth assessment in computer-based games to support learning. Computer games and instruction. Charlotte, NC: Information Age Publishers, 503-523 (2011)
16. Long, P., and Siemens, G.: Penetrating the fog: analytics in learning and education. Educause Review Online 46 (5): 31–40 (2011)
17. El-Nasr, M. S., Drachen, A., and Canossa, A.: Game Analytics: Maximizing the Value of Player Data, Springer (2013)
18. Oijen, J., and Dignum, F.: Scalable Perception for BDI-Agents Embodied in Virtual Environments, Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on, Vol. 2, 46-53, 22-27 (2011)
19. Patel, P., and Hexmoor, H.: Designing BOTs with BDI agents, Collaborative Technologies and Systems, 2009. CTS '09. International Symposium on, 180-186 (2009)
20. Lim, M., Dias, J., Aylett, R., and Paiva, A.: Creating adaptive affective autonomous NPCs, Autonomous Agents and Multi-Agent Systems (AAMAS), Springer, 24(2), 287-311 (2012)
21. Sterling, L. S. and Taveter, K.: The art of agent-oriented modeling, MIT Press (2009)
22. Dutchuk, M., Muhammadi, K. A., and Lin, F.: QuizMASter - A Multi-Agent Game-Style Learning Activity, EduTainment 2009, Aug 2009, Banff, Canada, Learning by Doing, (eds.), M Chang, et al., LNCS 5670, 263-272 (2009)
23. de Freitas, S. and Oliver, M.: How can exploratory learning with game and simulation within the curriculum be most effectively evaluated? Computers and Education. 46(3), 249-264 (2006)
24. Shabani, S., Lin, F. and Graf, S.: A Framework for User Modeling in QuizMASter. Journal of e-Learning and Knowledge Society, 8(3), 1826-6223 (2012)
25. Jennings N. R., L. Moreau, D. Nicholson, S. Ramchurn, S. Roberts, T. Rodden, and A. Rogers: Human-agent collectives. Commun. ACM 57, 12 (2014), 80-88 (2014)
26. Cai, Y., Miao, C., Tan, A.-H., and Shen, Z.: Fuzzy cognitive goal net for interactive storytelling plot design. In Proceedings of the 2006 ACM SIGCHI Int. conf. on Advances in comput. entertainment technology. ACM, NY, USA, Article 56 (2006)
27. Gee, James P.: Good Video Games + Good Learning. New York: Peter Lang (2008)
28. Gee, James P.: Learning by Design: Games as Learning Machines, Interactive Educational Multimedia, Vol. 8, April Ed. 2004, 13-15 (2004)
29. Brandt, F., Conitzer, V., and Endriss, U.: Computational Social Choice, Chapter 6 of book edited by Weiss, G., Multiagent Systems (2nd edition), 213-283 (2013)
30. Conitzer, V.: Making Decisions Based on the Preferences of Multiple Agents, Communications of the ACM, 53(3), 84-94 (2010)
31. Bartholdi, J. J. and Orlin, J. B: Single Transferable Vote Resists Strategic Voting. Social Choice and Welfare, 8, 341-354 (1991)
32. Parkes, D. C. and Procaccia, A. D.: Dynamic Social Choice with Evolving Preferences. In M. desJardins and M. L. Littman (eds.), AAAI : AAAI Press (2013)
33. Jaques, P. A., Vicari, R. M.: A BDI approach to infer student's emotions in an intelligent learning environment. Computers & Education, 49(2), 360–384 (2007)

34. Píbil, R., Novák, P., Brom, C., Gemrot, J.: Notes on Pragmatic Agent-Programming with Jason. In L. Dennis, O. Boissier, & R. H. Bordini (Eds.), Programming Multi-Agent Systems, 58–73. Springer Berlin Heidelberg (2012)
35. Wong, W., Cavedon, L., Thangarajah, J., Padgham, L.: Flexible Conversation Management Using a BDI Agent Approach. In Y. Nakano, M. Neff, A. Paiva, & M. Walker (Eds.), Intelligent Virtual Agents, 7502, 464–470). Springer (2012)
36. Soliman, M., Guetl, C.: Experiences with BDI-based design and implementation of Intelligent Pedagogical Agents. In 2012 15th International Conference on Interactive Collaborative Learning (ICL) (1–5). Presented at the 2012 15th International Conference on Interactive Collaborative Learning (ICL) (2012)
37. Norling, E.: Folk Psychology for Human Modelling: Extending the BDI Paradigm. In Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 1 (202–209). Washington, DC, USA: IEEE (2004)
38. Bordini, R. H., Hübner, J. F., Wooldridge, M.: Programming Multi-Agent Systems in AgentSpeak using Jason. John Wiley & Sons (2007)
39. Bellotti, F., Berta, R., De Gloria, A., Lavagnino, E.: Towards a conversational agent architecture to favor knowledge discovery in serious games. In Proc. of the 8th Int. Conf. on Advances in Comput. Entertainment Technology, 17:1–17:7 (2011)
40. Hübner, J. F., Bordini, R. H., Wooldridge, M.: Programming Declarative Goals Using Plan Patterns. In M. Baldoni & U. Endriss (Eds.), Proceedings on the Fourth International Workshop on Declarative Agent Languages and Technologies, held with AAMAS 2006 (123–140). Springer Berlin Heidelberg (2007)
41. Pokahr, A., Braubach, L., Lamersdorf, W. (2005). A BDI architecture for goal deliberation. In Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems (1295–1296). New York, NY, USA: ACM (2005)
42. Braubach, L., Pokahr, A.: Representing Long-Term and Interest BDI Goals. In L. Braubach, J.-P. Briot, & J. Thangarajah (Eds.), Programming Multi-Agent Systems (pp. 201–218). Springer Berlin Heidelberg (2010)
43. Conati, C., Maclaren, H.: Empirically building and evaluating a probabilistic model of user affect. User Modeling and User-Adapted Interaction, 19(3), 267–303 (2009)
44. Chrysafiadi, K., Virvou, M.: Student modeling approaches: A literature review for the last decade. Expert Systems with Applications, 40(11), 4715–4729 (2013)
45. Silva, D. G., Gluz, J. C.: AgentSpeak(PL): A New Programming Language for BDI Agents with Integrated Bayesian Network Model. In 2011 International Conference on Information Science and Applications (ICISA) (pp. 1–7). Presented at the 2011 International Conference on Information Science and Applications (ICISA) (2011)
46. McClure, G., Chang, M., Lin, F.: MAS Controlled NPCs in 3D Virtual Learning Environment, The International Workshop on Smart Learning Environments at the 9th International Conference on Signal-Image Technology and Internet-Based Systems, Japan, Kyoto, 2013-12-02, 1026 – 1033, DOI: 10.1109/SITIS.2013.166 (2013)
47. Walsh, T.: An Empirical Study of the Manipulability of Single Transferable Voting, Proceedings of the 2010 conference on ECAI, 257-262 (2010)
48. Leung, S., Virwaney, S., Lin, F., Armstrong, A J, Dubbelboer, A.: TSI-enhanced Pedagogical Agents to Engage Learners in Virtual Worlds", International Journal of Distance Education Technologies, 11(1), 1-13 (2013)

# Is this model for real? Simulating data to reveal the proximity of a model to reality

Rinat B. Rosenberg-Kima[1], Zachary A. Pardos[2]

[1] Tel-Aviv University
rinat.rosenberg.kima@gmail.com
[2] University of California, Berkeley
pardos@berkeley.edu

**Abstract.** Simulated data plays a central role in Educational Data Mining and in particular in Bayesian Knowledge Tracing (BKT) research. The initial motivation for this paper was to try to answer the question: given two datasets could you tell which of them is real and which of them is simulated? The ability to answer this question may provide an additional indication of the goodness of the model, thus, if it is easy to discern simulated data from real data that could be an indication that the model does not provide an authentic representation of reality, whereas if it is hard to set the real and simulated data apart that might be an indication that the model is indeed authentic. In this paper we will describe analyses of 42 GLOP datasets that were performed in an attempt to address this question. Possible simulated data based metrics as well as additional findings that emerged during this exploration will be discussed.

**Keywords:** Bayesian Knowledge Tracing (BKT), simulated data, parameters space.

## 1 Introduction

Simulated data has been increasingly playing a central role in Educational Data Mining [1] and Bayesian Knowledge Tracing (BKT) research [1, 4]. For example, simulated data was used to explore the convergence properties of BKT models [5], an important area of investigation given the identifiability issues of the model [3]. In this paper, we would like to approach simulated data from a slightly different angle. In particular, we claim that the question "*given two datasets could you tell which of them is real and which of them is simulated?"* is interesting as it can be used to evaluate the goodness of a model and may potentially serve as an alternative metric to RMSE, AUC, and others. In a previous work [6] we started approaching this problem by contrasting two real datasets with their corresponding two simulated datasets with Knowledge Tracing as the model. We found a surprising close to identity between the real and simulated datasets. In this paper we would like to continue this investigation by expanding the previous analysis to the full set of 42 Groups of Learning Opportunities (GLOPs) real datasets generated from the ASSISTments platform [7].

Knowledge Tracing (KT) models are widely used by cognitive tutors to estimate the latent skills of students [8]. Knowledge tracing is a Bayesian model, which assumes that each skill has 4 parameters: two knowledge parameters include initial (prior knowledge) and learn rate, and two performance parameters include guess and slip. KT in its simplest form assumes a single point estimate for prior knowledge and learn rate for all students, and similarly identical guess and slip rates for all students. Simulated data has been used to estimate the parameter space and in particular to answer questions that relate to the goal of maximizing the log likelihood (LL) of the model given parameters and data, and improving prediction power [7, 8, 9].

In this paper we would like to use the KT model as a framework for comparing the characteristics of simulated data to real data, and in particular to see whether it is possible to distinguish between the real and simulated datasets.

## 2 Data Sets

To compare simulated data to real data we started with 42 Groups of Learning Opportunities (GLOPs) real datasets generated from the ASSISTments platform[1] from a previous BKT study [7]. The datasets consisted of problem sets with 4 to 13 questions in linear order where all students answer all questions. The number of students per GLOP varied from 105 to 777. Next, we generated two synthetic, simulated datasets for each of the real datasets using the best fitting parameters that were found for each respective real datasets as the generating parameters. The two simulated datasets for each real one had the exact same number of questions, and same number of students.

## 3 Methodology

The approach we took to finding the best fitting parameters was to calculate LL with a grid search of all the parameters (prior, learn, guess, and slip). We hypothesized that the LL gradient pattern of the simulated data and real data will be different across the space. For each of the datasets we conducted a grid search with intervals of .04 that generated 25 intervals for each parameter and 390,625 total combinations of prior, learn, guess, and slip. For each one of the combinations LL was calculated and placed in a four dimensional matrix. We used fastBKT [12] to calculate the best fitting parameters of the real datasets and to generate simulated data. Additional code in Matlab and R was generated to calculate LL and RMSE and to put all the pieces together[2].

[1] Data can be obtained here: http://people.csail.mit.edu/zp/

[2] Matlab and R code will be available here: www.rinatrosenbergkima.com/AIED2015/

## 4 What are the Characteristics of the Real Datasets Parameters Space?

Before we explored the relationships between the real and sim datasets, we were interested to explore the BKT parameter profiles of the real datasets. We calculated the LL with a grid search of 0.04 granularity across the four parameters resulting in a maximum LL for each dataset and corresponding best prior, learn, guess, and slip. Figure 1 present the best parameters for each datasets, taking different views of the parameters space. The first observation to be made is that the best guess and slip parameters fell into two distinct areas (see figure 1, guess x slip).

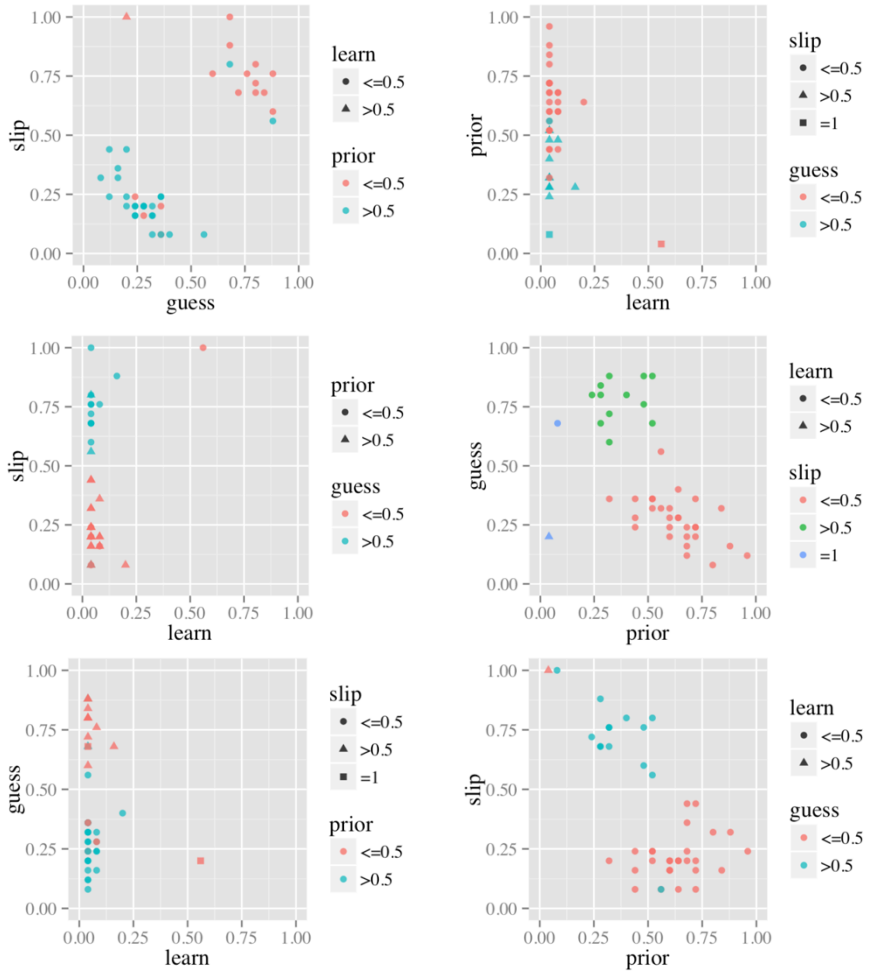



**Figure 1.** Best parameters across the 42 GLOP real datasets.

Much attention has been given to this LL space, which revealed the apparent co-linearity of BKT with two primary areas of convergence, the upper right area being a false, or "implausible" converging area as defined by [3]. What is interesting in this figure is that real data also converged to these two distinct areas. To further investigate this point, we looked for the relationships between the best parameters and the number of students in the dataset (see figure 2). We hypothesized that perhaps the upper right points were drawn from datasets with small number of students; nevertheless, as figure 2 reveals, that was not the case. Another interesting observation is that while in the upper right area (figure 1, guess x slip) most of the prior best values were smaller that 0.5, in the lower left area most of the prior best values were bigger than 0.5, thus revealing interrelationships between slip, guess, and prior that can be seen in the other views. Another observation is that while prior is widely distributed between 0 and 1, most of best learn values are smaller than 0.12.

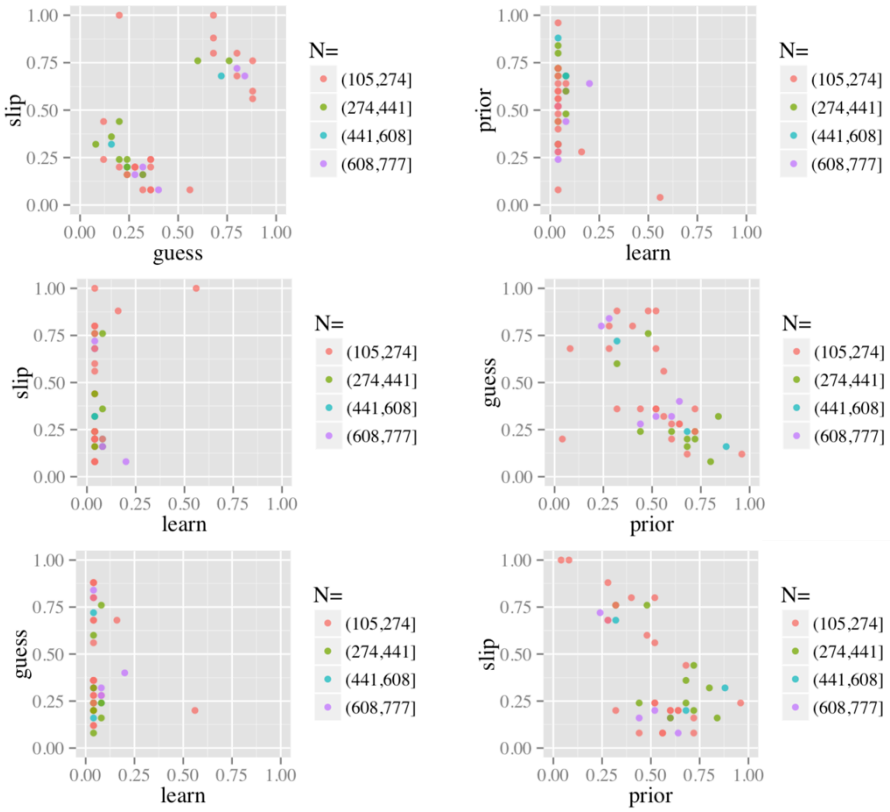



**Figure 2.** Best parameters across the 42 GLOP real datasets by number of students.

## 5 Does the LL of Sim vs. Real Datasets Look Different?

Our initial thinking was that as we are using a simple BKT model, it is not authentically reflecting reality in all its detail and therefore we will observe different patterns of LL across the parameters space between the real data and the simulated data. The LL space of simulated data in [5] was quite striking in its smooth surface but the appearance of real data was left as an open research question. First, we examined the best parameters spread across the 42 first set of simulated data we have generated. As can be seen in figure 3, the results are very similar (although not identical) to the results we received with the real data (see figure 1). This is not surprising, after all, the values of learn, prior, guess, and slip were inputs to the function generating the simulated data.

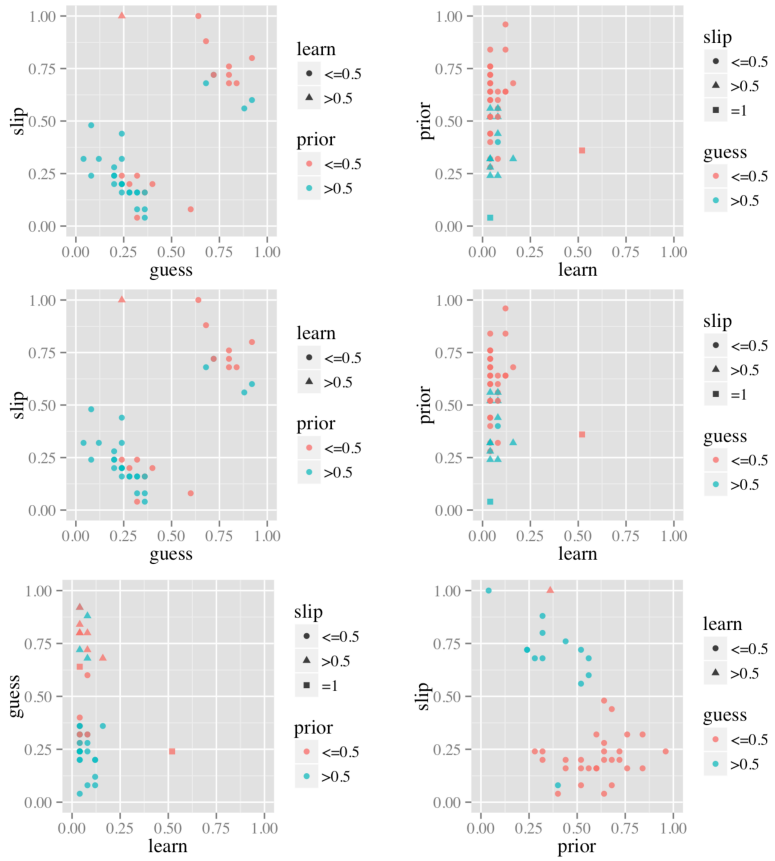



**Figure 3.** Best parameters across 42 GLOP simulated datasets.

In order to see if the differences between real and sim were more than just the difference between samples from the same distribution, we generated *two* simulated versions of each real dataset (sim1 and sim2) using the exact same number

of questions, number of students, generated with the best fitting parameters from the real dataset. We then visualized 2D LL heatmaps looking at two parameter plots at a time where the other two parameters were fixed to the best fitting values. For example, when visualizing LL heatmaps for the combination of guess and slip, we fixed learn and prior to be the best learn and the best prior from the real data grid search. To our surprise, when we plotted heatmaps of the LL matrices of the real data and the simulated data (the first column in figure 4 represents the real datasets, the second column represents the corresponding sim1, and the third column the corresponding sim2) we received what appears to be extremely similar heatmaps. Figure 4 and 5 displays a sample of 4 datasets, for each one displaying the real dataset heatmap and the corresponding two simulated datasets heatmaps.

The guess vs. slip heatmaps (see figure 4) prompted interesting observations. As mentioned above, the best guess and slip parameters across datasets fell into two areas (upper right and lower left). Interestingly, these two areas were also noticeable in the individual heatmaps. While in some of the datasets they were less clear (e.g., G5.198 in figure 4), most of the datasets appear to include two distinct global maxima areas. In some of the datasets the global maxima converged to the lower left expected area, as did the corresponding simulated datasets (e.g., G4.260 in figure 4), in other datasets the global maxima converged to the upper right "implausible" area, as did the corresponding simulated datasets (e.g., G6.208 in figure 4). Yet in some cases, one or more of the simulated dataset converged to a different area than that of the real dataset (e.g., G4.205 in figure 4). The fact that so many of the real datasets converged to the "implausible" area is surprising and may be due to small number of students or to other limitations of the model.

The learn vs. prior heatmaps were also extremely similar within datasets and exhibited a similar pattern also across datasets (see figure 5), although not all datasets had the exact pattern (e.g., G5.198 is quite different than the other 3 datasets in figure 5). While best learn values were low across the datasets, the values of best prior varied. As with guess vs. slip, in some cases the two simulated datasets were different (e.g., G4.205 had different best parameters also with respect to prior). Similar patterns of similarities within datasets and similarities with some clusters across datasets were also noticeable in the rest of the parameters space (learn vs. guess, learn vs. slip, prior vs. guess, prior vs. slip not displayed here due to space considerations).

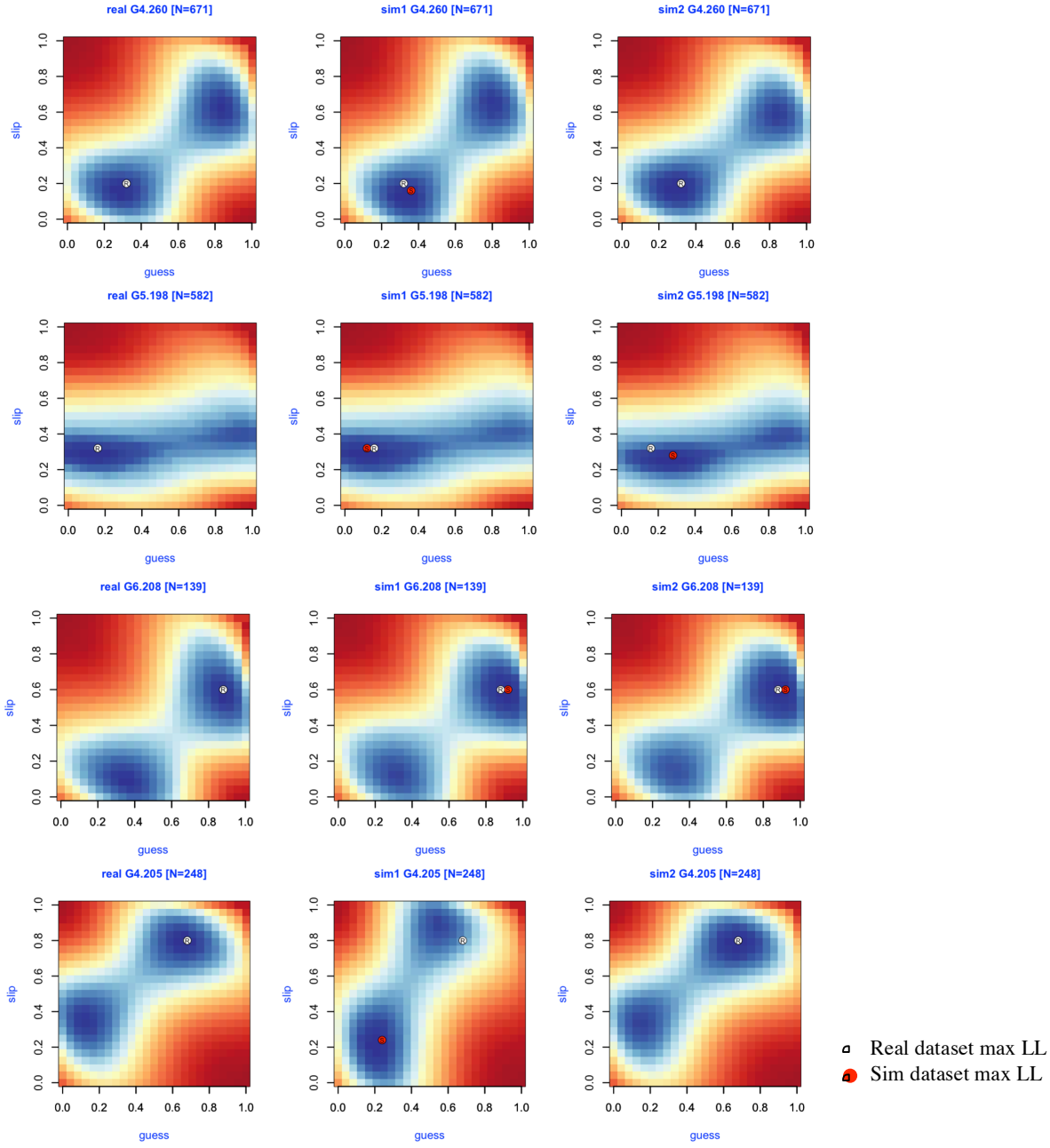



**Figure 4.** Heatmaps of (guess vs. slip) LL of 4 sample real GLOP datasets and the corresponding two simulated datasets that were generated with the best fitting parameters of the corresponding real dataset.

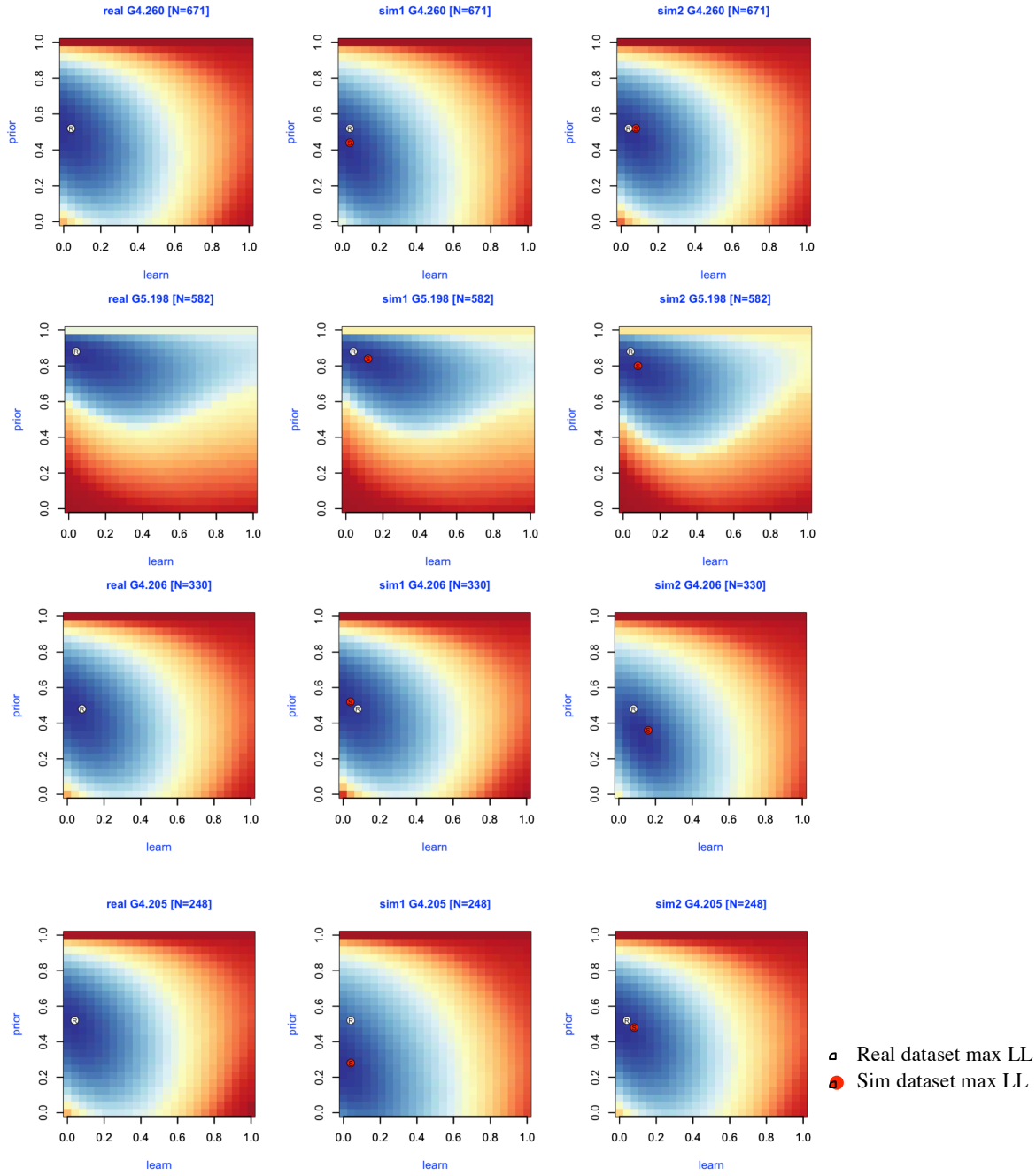



**Figure 5.** Heatmaps of (learn x prior) LL of 4 sample real GLOP datasets and the corresponding two simulated datasets that were generated with the best fitting parameters of the corresponding real dataset.

## 6 Exploring Possible Metrics Using the Real and Sim Datasets

In natural science domains, simulated data is often used as a mean to evaluate its underlying model. For example, simulated data is generated from a hypothesized model of the phenomena and if the simulated data appears to be similar to the real data observed in nature, it serves as evidence for the accuracy of the model. Then, if the underlying is validated, simulated data is used to make predictions (e.g., in the recent earthquake in Nepal a simulation was used to estimate the number of victims). Can this approach be used in education as well? What would be an indication of similarity between real and simulated data?

Figure 5 displays two preliminary approaches for comparing the level of similarity between the simulated and real data. First, the Euclidean distance between the real dataset parameters and the simulated data parameters was compared to the Euclidean distance between the two simulated datasets parameters. The idea is that if the difference between the two simulated datasets is smaller than the difference between the real and the simulated dataset this may be an indication that the model can be improved upon. Thus, points on the right side of the red diagonal indicate good fit of the model to the dataset. Interestingly, most of the points were on the diagonal and a few to the left of it. Likewise the max LL distance between the real and simulated datasets was compared to the max LL distance of the two simulated datasets. Interestingly, datasets with larger number of students did not result in higher similarity between the real and simulated dataset. Also, here we *did* find distribution of the points to the left and to the right of the diagonal.

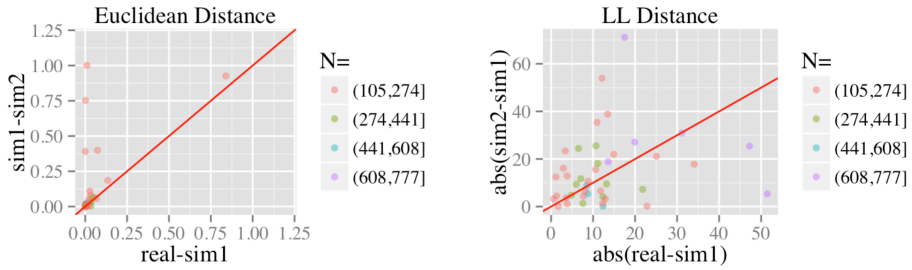



**Figure 5.** Using Euclidean distance and LL distance as means to evaluate the model.

## 7 Contribution

The initial motivation of this paper was to find whether it is possible to discern a real dataset from a simulated dataset. If for a given model it is possible to tell apart a simulated data from a real dataset then the authenticity of the model can be questioned. This line of thinking is in particular typical of simulation use in Science contexts, where different models are used to generate simulated data, and then if a simulated data has a good fit to the real phenomena at hand, then it may be possible to claim that the model provides an authentic explanation of the system [13]. We believe

that finding such a metric can serve as the foundation for evaluating the goodness of a model by comparing a simulated data from this model to real data and that such a metric could provide much needed substance in interpretation beyond that which is afforded by current RMSE and AUC measures. This can afford validation of the simulated data, which can than be used to make predictions on learning scenarios; decreasing the need to test them in reality, and at minimum, serving as an initial filter to different learning strategies.

## References


[1] R. S. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *J. Educ. Data Min.*, vol. 1, no. 1, pp. 3–17, 2009.

[2] M. C. Desmarais and I. Pelczer, "On the Faithfulness of Simulated Student Performance Data.," in *EDM*, 2010, pp. 21–30.

[3] J. E. Beck and K. Chang, "Identifiability: A fundamental problem of student modeling," in *User Modeling 2007*, Springer, 2007, pp. 137–146.

[4] Z. A. Pardos and M. V. Yudelson, "Towards Moment of Learning Accuracy," in *AIED 2013 Workshops Proceedings Volume 4*, 2013, p. 3.

[5] Z. A. Pardos and N. T. Heffernan, "Navigating the parameter space of Bayesian Knowledge Tracing models: Visualizations of the convergence of the Expectation Maximization algorithm.," in *EDM*, 2010, pp. 161–170.

[6] R. B. Rosenberg-Kima and Z. Pardos, "Is this Data for Real?," in *Twenty Years of Knowledge Tracing Workshop*, London, UK, pp. 141–145.

[7] Z. A. Pardos and N. T. Heffernan, "Modeling individualization in a bayesian networks implementation of knowledge tracing," in *User Modeling, Adaptation, and Personalization*, Springer, 2010, pp. 255–266.

[8] A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," *User Model. User-Adapt. Interact.*, vol. 4, no. 4, pp. 253–278, 1994.

[9] S. Ritter, T. K. Harris, T. Nixon, D. Dickison, R. C. Murray, and B. Towle, "Reducing the Knowledge Tracing Space.," *Int. Work. Group Educ. Data Min.*, 2009.

[10] R. S. d Baker, A. T. Corbett, S. M. Gowda, A. Z. Wagner, B. A. MacLaren, L. R. Kauffman, A. P. Mitchell, and S. Giguere, "Contextual slip and prediction of student performance after use of an intelligent tutor," in *User Modeling, Adaptation, and Personalization*, Springer, 2010, pp. 52–63.

[11] R. S. Baker, A. T. Corbett, and V. Aleven, "More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing," in *Intelligent Tutoring Systems*, 2008, pp. 406–415.

[12] Z. A. Pardos and M. J. Johnson, "Scaling Cognitive Modeling to Massive Open Environments (in preparation)," *TOCHI Spec. Issue Learn. Scale*.

[13] U. Wilensky, "GasLab—an Extensible Modeling Toolkit for Connecting Micro- and Macro-properties of Gases," in *Modeling and simulation in science and mathematics education*, Springer, 1999, pp. 151–178.

# Workshop on Developing a Generalized Intelligent Framework for Tutoring (GIFT): Informing Design through a Community of Practice

*held in conjunction with*

Seventeenth International Conference on Artificial Intelligence in Education (AIED 2015)

Friday, June 26, 2015
Madrid, Spain

**Workshop Co-Chairs:**

Benjamin Goldberg[1], Robert Sottilare[1], Anne Sinatra[1], Keith Brawner[1], Scott Ososky[1,2]

[1] *U.S. Army Research Laboratory, Orlando, FL 32826*
[2] *Oak Ridge Associated Universities, Oak Ridge, TN 37830*

https://gifttutoring.org/news/42

# Table of Contents

# Preface

The purpose of this workshop is to examine current research within the AIED community focused on improving adaptive tools and methods for authoring, automated instruction and evaluation associated with the Generalized Intelligent Framework for Tutoring (GIFT). As GIFT is an open-source architecture used to build and deliver adaptive functions in computer-based learning environments (Sottilare, Brawner, Goldberg & Holden, 2013), this workshop aids in gathering feature requirements from the field and addressing issues to better support future users.

The topics of interest highlight current research conducted within the GIFT community (i.e., 400+ users in 30+ countries) across three themes: (1) modeling across affect, metacogntion, teams, and experts; (2) tutorial intervention through communication, guidance, and sequencing; and (3) persistence functions of intelligent tutoring associated with competency modeling and social media. Each theme will be comprised of short papers describing capability enhancements to the GIFT architecture, the motivation behind the described work, and considerations associated with its implementation. Paper presentations are organized to provide attendees with an interactive experience through hands-on demonstrations.

For attendees unfamiliar with GIFT and its project goals, this workshop exposes those individuals to the GIFT architectural structure, enabling participants to learn how to construct original functions, and how the framework can be applied to their own research. The intent is to engage the AIED community in an in-depth exploration of the various research topics being investigated and the potential leveraging and collaboration that a community framework such as GIFT affords.

Benjamin Goldberg, Robert Sottilare, Anne Sinatra, Keith Brawner, Scott Ososky
The GIFT 2015 Co-Chairs

## References

1. Sottilare, R., Brawner, K. W., Goldberg, B., & Holden, H. (2013). The Generalized Intelligent Framework for Tutoring (GIFT). In C. Best, G. Galanis, J. Kerry & R. Sottilare (Eds.), *Fundamental Issues in Defense Training and Simulation* (pp. 223-234). Burlington, VT: Ashgate Publishing Company.

# Challenges in Moving Adaptive Training & Education from State-of-Art to State-of-Practice

Robert A. Sottilare[1]

[1]U.S. Army Research Laboratory, Orlando, FL 32826
Robert.a.sottilare.civ@mail.mil

**Abstract.** Adaptive training and education (ATE) systems are the convergence of intelligent tutoring system (ITS) technologies and external training and education capabilities (e.g., serious games, virtual humans and simulations). Like ITSs, ATEs provide instructional experiences that are tailored to the learner and may be more effective than the training or educational systems alone. ATEs also leverage existing environments, content and domain knowledge to reduce the authoring workload. The Generalized Intelligent Framework for Tutoring (GIFT) is an open-source ATE architecture with the primary goal to support easy authoring, automated instructional management during ATE experiences, and a testbed to evaluate the effect of ATE tools and methods. While this paper addresses challenges and goals in bringing ATE solutions from state-of-art to state-of-practice within GIFT, it also highlights generalized challenges in making ITS technologies ubiquitous and practical on a large scale across a broader variety of domains.

**Keywords:** adaptive training and education (ATE), intelligent tutoring system (ITS), authoring, instructional management, domain modeling

## 1 Introduction

An adaptive training and education (ATE) system is the convergence of Intelligent Tutoring Systems (ITS) technologies and what might normally be standalone training and educational capabilities (e.g., serious games, virtual humans, and virtual, mixed, and augmented-reality simulations). The resulting integration provides intelligently-tailored, computer-guided learning experiences for both individual learners and teams which leverages and enhances the capabilities of existing training and educational infrastructure.

ATE research is focused on optimizing performance, efficiency (e.g., reduced time to competency) deep learning (e.g. higher retention and reduced need for refresher training), and transfer of skills to the operational environment (on the job). The Generalized Intelligent Framework for Tutoring (GIFT) is an open-source, modular architecture whose goals include reducing the cost and skill for authoring ATE systems, automating instructional management, and tools for the evaluation of ATE technologies [1]. GIFT was created to capture best instructional practices and the results of

enabling ATE research objectives including ITS design, data analytics, human-system interaction, automated authoring, and the application of learning theory.

Several ATE integration tools and prototypes have been created and are being evaluated. The Game-based Architecture for Mentor-Enhanced Training Environments (GAMETE), is a middleware tool to integrate serious games (e.g., Virtual Medic) and tutors (e.g., GIFT-based tutors and AutoTutor Lite tutors) [2]. The Student Information Model for Intelligent Learning Environments (SIMILE) is a tool for linking actions in games to ITS learning measures [3]. Newtonian Talk is the integration of Physics Playground, AutoTutor, and GIFT [4] to support interactive discovery learning of key physics principles. Virtual Battle Space 2, a serious military training game, has also been integrated with GIFT [5]. As a result of developing and evaluating these prototype ATE tools and systems, lessons-learned and several challenge areas have been identified.

## 2 Challenges, Goals, and Objectives

The idea of generalizing the authoring of ITSs for broad application across task domains (cognitive, affective, psychomotor, and social) ranging from simple to complex, and from well-defined to ill-defined is not a new goal [6, 7]. However, there remain several challenges in realizing a generalized tutoring architecture to produce standalone ITSs and integrated ATE systems. We have identified seven challenge areas or barriers to adoption of ATE technologies: affordability and efficiency; adaptability and persistence; accuracy and validity; relevance and generalizability; accessibility; credibility; and effectiveness.

Each of these challenges could also be considered a desired characteristic or end state. While all of the seven challenges may be considered on the critical path to practical ATE systems, the challenges which impact authoring and learner modeling are most critical. The authoring process is critical to affordability and is therefore the most significant barrier to adoption.

Accurate learner modeling is critical to the whole instructional decision process for ATE systems. To fully understand the learner's states and adapt instruction to optimize learning and mitigate barriers to learning, ATE systems (and ITSs) need to meet two challenges: low cost, unobtrusive methods to acquire learner behavioral and physiological data; and highly accurate, near real-time classification methods for learner states based on behavioral and physiological data. The effect of adaptive instruction on learner states and specifically critical learning moderators [8] (e.g., engagement, motivation) is illustrated in Figure 1.

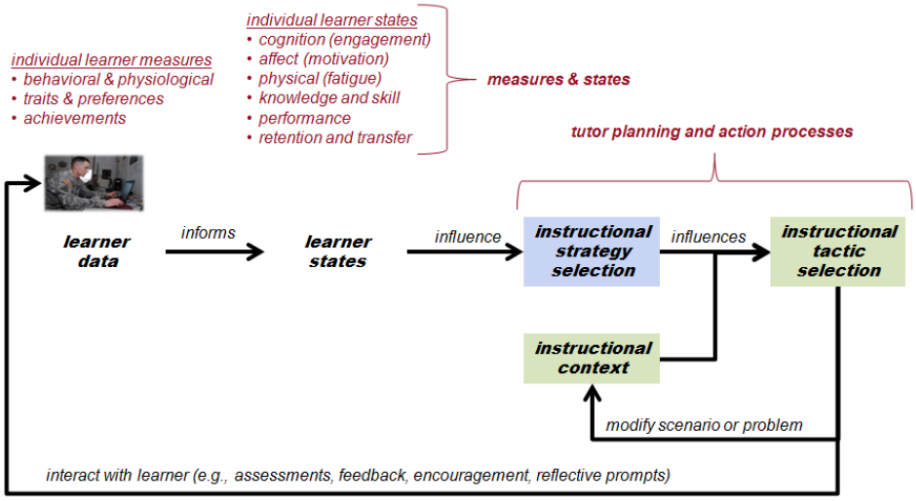



**Fig. 1.** Updated Learning Effect Model

Inaccurate modeling of learner states can lead to the selection of less than optimal strategies and tactics. Negative outcomes include the selection of instructional strategies which either confuse or frustrate the learner to the point of withdrawal or provide negative training effects because the strategy selected is in opposition to the learner's actual state.

The following is a discussion of the seven challenges and their associated goals and objectives along with a projected impact on adoption in the context of associated ATE/ITS processes: authoring, maintenance, individual learner and team modeling, instructional management, domain modeling, user interface design, and architecture.

### 2.1 Challenge: Affordable, Efficient, and Effective Adaptive Systems

Due to high authoring costs, the investment in ITS development is only practical for high density courses, those with a high student population. ITS and ATE system developers be able to define what a pound of adaptive training and education is worth in comparison to alternative methods, and they must be able to quantify a return-on-investment and associated breakeven points for these investments [9]. Adaptive systems by their nature require the authoring of additional content and domain knowledge.

To make ATE technologies affordable, we must first examine the authoring and maintenance processes. Aleven, McLaren, Sewall and Koedinger [10] assert that it takes approximately 200-300 hours of development time to author one hour of adaptive instruction. This assertion is based on well-defined, cognitive (e.g., problem solving and decision-making) domains. Research is needed to define the authoring time for more complex, ill-defined domains. A goal for GIFT designers is to reduce authoring time for any domain to just a few hours. This would make it practical for teachers,

course managers, and other domain experts to rapidly develop adaptive content and make courses agile and adaptive to learner needs.

However, in the case of ATE systems, we are looking at a broader definition of domain complexity with ill-defined domains and non-cognitive tasks and factors. So given we are developing more complex instruction, our goal is not just to reduce the time and cost to author ATE systems, but also to reduce the skills required to develop and maintain standalone ITSs and integrated ATE systems. To meet this goal, we must improve interoperability and reuse of ITS components and domain knowledge, automate authoring processes wherever possible to take humans out of the loop, improve curation (search, retrieval, management) of domain knowledge, and improve user interfaces to enhance authoring efficiency (ease of use) where human-in-the-loop authoring is required.

What will it take to make ATE authoring available to the masses? A goal is for domain experts to be able to author ATE systems without knowledge of computer programming, instructional design principles, or learning theory. These would be integral to ATE design along with automated authoring tools and artificially intelligent job aids which will guide authors efficiently through the end-to-end development process in the future. As part of the authoring process, we advocate standards to make integration of external training and education systems with ITS easier. Fixing the authoring process is a "must do" to make ATE systems practical (affordable, effi-cient, and effective).

### 2.2 Challenge: Enhance Adaptability and Persistence

The adaptability of ATE systems is limited when compared to expert human tutors. Our goal is to enhance the ability of ATE systems to provide unique learning experiences for each and every learner. ATE systems by their nature require additional content and associated domain knowledge to support a broad population of learners. This fact drives the cost of ATE systems and limits options for tailoring of ATE experiences for individual learners and teams of learners. By finding tools and methods to reduce the time/cost and skills required to author ATE systems, we can provide more tailoring options in the same or less development time.

Another area for improvement in ATE systems design is in individual learner and team modeling. Our objectives are to enhance short-term and long-term learner modeling to improve the adaptability of ATE systems. Research is needed to understand the relationship between desired outcomes (e.g., learning, performance, retention, and transfer) and the learner's behaviors, transient states (e.g., goals, affect), trends and cumulative states (e.g., domain competency and prior knowledge), and their enduring traits (e.g., personality, gender, and first language) in order to facilitate efficient learner modeling, optimized instructional decisions, and thereby authoring of ATE systems. Adaptive instruction based on long-term modeling of the learner will offer persistence not present in today's ITSs. We can enhance adaptability by making learner and team modeling central to instructional decisions made by ATE systems.

### 2.3 Challenge: Enhance Accuracy and Validity of Instructional Decisions

In order to make appropriate adaptive instructional decisions, ATE systems need to improve their perception of learner states. Research is needed to develop low cost, unobtrusive methods to acquire learner data to support state classification. In turn, research is also needed to improve the accuracy of real-time classification for both individuals and units [11].

To insure the validity (suitability) of instructional decisions based on sound learning theory, domain-independent instructional strategies (e.g., metacognitive prompts) may be selected based on the accurate classification of the learner's states. For example, imagine a learner whose state is classified as "confusion" by an ATE. If the accuracy of this classification is less than 80 percent, then a metacognitive prompt to have the learner reflect on a recent decision could clarify any ambiguity of the "confusion" classification.

Similarly, domain-specific actions (tactics) based on a selected instructional strategy and context (conditions within the domain). Research is needed to develop methods to optimally select the best possible strategies and tactics given the learners states, the conditions within the training or educational domain, and the availability of options provided by the author of the ATE. Within GIFT, the learning effect model for individual learners [11, 12, 13], as updated in Figure 1, describes the interaction between the learner and the ITS.

### 2.4 Challenge: Enhance Task Relevance & Implement Generalized Solutions

In order to be practical, ATE systems must be able to represent domain knowledge in relevant task domains. Today, the most popular ITS domains are mathematics, physics, and computer programming. The characteristics of other domains may not be as well defined or as simple. For example, tasks involving psychomotor and perceptual measures (e.g., sports, laparoscopic surgery, and marksmanship) are not well-represented in the ITS community.

Research is needed to expand the dimensions of domain knowledge to support a broader variety of task domains. One objective is to develop standards to represent domain knowledge beyond the cognitive task domain (e.g., affective, psychomotor, perceptual, social, ill-defined, and complex domains). Once the domain can be represented, authoring tools and instructional strategies, tactics, and policies should be tailored to support adaptive interaction with the learner.

As mentioned previously, it will be critical to be able to easily integrate external training and educational environments to reduce the authoring burden, but also to enhance the experiences that are familiar to learners. Representing the domain knowledge of relevant task domains and integrating with other systems will provide the basis for an ATE architecture which we are currently prototyping as GIFT.

### 2.5 Challenge: Support Tutoring at the Point-of-Need

To be effective, ATE must be accessible at the user's point-of-need. The ATE architecture must develop services to allow access anyplace and anytime (24/7/365). To meet this goal we have formulated two primary objectives. The first is to move GIFT, an adaptive training and education architecture, to the cloud. We are developing a cloud-based architecture that allows real-time access for learners and units to support individual, collaborative (social), and team training and education. Since learners, authors, and other ATE system users may find themselves in areas of degraded communications, we are also developing cloud-based services to download virtual machine versions of GIFT to allow local development and synch with the cloud as needed.

### 2.6 Challenge: Enhance the Credibility and Supportiveness of the Tutor

To enhance the learner's perception of ATEs as credible training and educational tools (e.g., domain experts, trusted advisors, teachers), we are closely emulating best practices of expert tutors and learning theory. To this end we have implemented component display theory [14] as our default pedagogical module, the engine for managing adaptive pedagogy or eMAP.

To capture and maintain the attention of learners, we are developing methods to evaluate the suitability of user interfaces (e.g., virtual humans) and domain knowledge (e.g., content) to enhance the learner's perception of ATE systems with respect to domain expertise and learner support. To be efficient, we are developing user inter-faces for various roles in the ATE environment (e.g., learners, authors, and power-users). These interfaces will allow users to construct their own mental models and interact in a manner that is conducive to learning.

### 2.7 Challenge: Continuously Evaluate Effectiveness

As with many systems, we anticipate that ATE systems will be deployed with implementations of *best known practices*. ATE systems must not only provide adaptive instruction, but be adaptive to continuously improve. The challenge is to collect and analyze large datasets on a regular basis to identify trends and issues, and to evaluate the effectiveness of current tools and methods against alternative tools and methods. The ATE architecture must be able to support continuous evaluation of its tools and methods, and be modular in order to support rapid change.

We are developing tools and methods within GIFT to evaluate the effectiveness of the authoring and instructional management processes. Our goal is to support the continuous improvement of ATE technologies. To this end we are developing tools and methods to reduce the time/cost and skill required to evaluate the effectiveness of ITS technologies. We are also developing data analytic methods to evaluate user-generated content (social media) to maintain cognizance of the primary users (learners and authors) and to enable them as change agents.

## 3 Conclusions

This paper reviewed several challenges to adoption of ATE systems as practical tools for learning. We noted that several ongoing research initiatives and identified several more which are needed to support changes to the authoring and maintenance, instructional management, learner modeling, and domain modeling processes along with underlying services provided by the architecture through the user interface.

We also noted that ATE systems have a long-term focus as well as a short-term learning focus. Big data collected continuously on both the learner populations and the ATE system may be analyzed to provide insight on both effective and ineffective instructional methods and user interfaces for both authoring and instruction. Research is still needed to fully understand the effect of combining ITSs with existing training and education systems in order to quantify a return-on-investment.

We recommend additional research emphasis on the following challenge problems: methods to automate the authoring process to the maximum extent possible; enhanced job aids and user interfaces for the authoring process where automation is not possible yet; methods to automate integration of existing training and education systems with ITSs; methods to increase the accuracy of learner state classification and optimize instructional decisions by the tutor; methods to evaluate the effectiveness of ATE system tools and methods; and methods to evaluate user-generated content (e.g., social media) to enhance learner experiences in ATE systems.

We also note the need to expand ITSs beyond the existing well-defined domains (e.g., mathematics, physics, and computer programming) to include more ill-defined and dynamic domains (e.g. psychomotor domains including sports). Finally, we advocate the development of collective level models (e.g., shared states, team behaviors, and team cohesion) for unit-level tasks and collective learning environments [15].

## References


1. R. Sottilare, K. Brawner, B. Goldberg & H. Holden. The Generalized Intelligent Framework for Tutoring (GIFT). Orlando, FL: U.S. Army Research Laboratory – Human Research & Engineering Directorate (ARL-HRED), 2012.
2. J. Engimann, T. Santarelli, W. Zachary, X. Hu, Z. Cai, H. Mall & B. Goldberg, "Game-based Architecture for Mentor-Enhanced Training Environments (GAMETE)", In R. Sottilare (Ed.) *2nd Annual GIFT Users Symposium* (GIFTSym2), Pittsburgh, Pennsylvania, 12-13 June 2014. Army Research Laboratory, Orlando, Florida. ISBN: 978-0-9893923-4-1
3. H. Mall & B. Goldberg, "SIMILE: An Authoring and Reasoning System for GIFT", In R. Sottilare (Ed.) *2nd Annual GIFT Users Symposium* (GIFTSym2), Pittsburgh, Pennsylvania, 12-13 June 2014. Army Research Laboratory, Orlando, Florida. ISBN: 978-0-9893923-4-1
4. M. Ventura, X. Hu, B. Nye & W. Zhao, "NewtonianTalk: Integration of Physics Playground and AutoTutor using GIFT", In Proceedings of the "Developing a Generalized Intelligent Framework for Tutoring (GIFT): Informing Design through a Community of Practice" Workshop at the *17th International Conference on Artificial Intelligence in Education* (AIED 2015), Madrid, Spain, June 2015.

5. B. Goldberg, R. Sottilare, K. Brawner & H. Holden, “Adaptive Game-Based Tutoring: Mechanisms for Real-Time Feedback and Adaptation”, *International Defense & Homeland Security Simulation Workshop in Proceedings of the I3M Conference*. Vienna, Austria, September 2012.
6. T. Murray. Authoring intelligent tutoring systems: An analysis of the state of the art. *International Journal of Artificial Intelligence in Education*, 1999, 10(1):98–129.
7. T. Murray. “An Overview of Intelligent Tutoring System Authoring Tools: Updated analysis of the state of the art”, *Authoring tools for advanced technology learning environments*. 2003, 491-545.
8. M. R. Lepper, M. Drake, & T. M. O'Donnell-Johnson. Scaffolding techniques of expert human tutors. In K. Hogan & M. Pressley (Eds), Scaffolding student learning: Instructional approaches and issues. New York: Brookline Books, 1997, 108-144.
9. J. Fletcher & R. Sottilare. Cost Analysis for Training & Educational Systems. In R. Sottilare, A. Graesser, X. Hu, & B. Goldberg (Eds.) Design Recommendations for Intelli-gent Tutoring Systems: Volume 2 - Instructional Management. Army Research Laborato-ry, Orlando, Florida, 2014. ISBN: 978-0-9893923-2-7.
10. V. Aleven, B. McLaren, J. Sewall & K. Koedinger, “The Cognitive Tutor Authoring Tools (CTAT): Preliminary Evaluation of Efficiency Gains”, Proceedings of the *8th International Conference on Intelligent Tutoring Systems*, 2006, 61-70.
11. R. Sottilare, C. Ragusa, M. Hoffman & B. Goldberg, “Characterizing an adaptive tutoring learning effect chain for individual and team tutoring”. In Proceedings of the *Interservice/Industry Training Simulation & Education Conference*, Orlando, Florida, De-cember 2013.
12. J.D. Fletcher & R. Sottilare. Shared Mental Models of Cognition for Intelligent Tutoring of Teams. In R. Sottilare, A. Graesser, X. Hu, & H. Holden (Eds.) *Design Recommendations for Intelligent Tutoring Systems: Volume 1- Learner Modeling*. Army Research Laboratory, Orlando, Florida, 2013. ISBN 978-0-9893923-0-3.
13. R. Sottilare. Considerations in the Development of an Ontology for a Generalized Intelligent Framework for Tutoring. *International Defense & Homeland Security Simulation Workshop* in Proceedings of the *I3M Conference*. Vienna, Austria, September 2012.
14. D. Merrill, B. Reiser, M. Ranney, & J. Trafton, “Effective Tutoring Techniques: A Comparison of Human Tutors and Intelligent Tutoring Systems”. *The Journal of the Learning Sciences,* 1992, 2(3), 277-305.
15. R. Sottilare. Special Report: Adaptive Intelligent Tutoring System (ITS) Research in Support of the Army Learning Model - Research Outline. *Army Research Laboratory* (ARL-SR-0284), December 2013.

# Learning Ecosystems Using the Generalized Intelligent Framework for Tutoring (GIFT) and the Experience API (xAPI)

Michael Hruska[1], Ashley Medford[1], Jennifer Murphy[2]

[1]Problem Solutions 1407 Eisenhower Blvd., Johnstown, PA 15904
[2]Quantum Improvements Consulting 812 Oak Chase Dr., Orlando, FL 32828

**Abstract.** Learning ecosystems provide a combination of technologies and support resources available to help individuals learn within an environment [1]. The Experience API (xAPI) is an enabling specification for learning ecosystems, which provides a method for producing interoperable data that can be shared within a learning ecosystem [2]. Version 4.1 of the Generalized Intelligent Framework for Tutoring (GIFT) provides support to both produce and consume xAPI data. A number of use cases are enabled by this support. This paper will explore the use cases, functionality enabled, setup and design guidance in addition to exploring practical applications for using GIFT and xAPI within learning ecosystems.

**Keywords:** adaptation, Experience API, intelligent tutoring systems, learning, xAPI, GIFT, computer-based tutoring systems, learning ecosystems

## 1 Introduction

Organizations in the U.S. alone invested approximately $164.2 billion on employee training and development in 2012 [3], and in 2013, an average of over $1,200 per employee was spent for direct learning [4]. With 38% of this training being delivered using technology [4], this investment is increasingly being spent on non-traditional training methods and technologies. As learning ecosystems continue to grow in complexity, so too do the challenges faced by education and training professionals.

Personalizing education and assessing student learning are grand, educational challenges being faced today [5]. Recent efforts on learning ecosystems reflect this movement towards adaptive and tailored learning [5,6]. In general, the goal in a learning ecosystem is to leverage performance data in order to assess and adapt learning and in turn, increase training effectiveness and lower associated training time and costs [6]. By capturing the massive amount of learning data tied to each individual and bound within a learning ecosystem, the ability to meet these educational challenges by intelligently tailoring learning and assessing performance is possible.

Research and development efforts by the Advanced Distributed Learning (ADL) initiative of the Department of Defense (DoD) and the U.S. Army's Research Laboratory (ARL) are striving to meet these complex challenges. The Experience API specification (xAPI), developed by ADL, provides an interoperable means to describe and track learning in various learning ecosystem components [7]. ARL's work on interoperability of performance data and intelligent tutors, specifically the Generalized Intelligent Framework for Tutoring (GIFT), along with xAPI provide a basis for this paper. The use of xAPI in conjunction with intelligent tutoring (e.g., GIFT) permits the creation of a reference architecture and provides functionality for a number of use cases. Installation and configuration of open source software components enable testing and experimentation around these use cases. This paper outlines the technical information, reference architecture, use cases, configuration, and expected behaviors of the technology components surrounding this work.

### 1.1 Existing Efforts

The ARL effort on Interoperable Performance Assessment (IPA) focuses on uniformly defining and describing learning experiences [8]. IPA defines methods for encoding human performance data using xAPI statements [9]. The goal of such encoding is to create data with *inter-system data value* to support adaptation in learning ecosystems. Additionally, interoperable encoding can provide rich data analytics and visualizations.

ARL's IPA research works primarily toward the goal of defining uniform performance measures in simulation and providing summative assessments towards these measures from multiple sources. Additional IPA efforts, focused on using small group and team data, also indicate the potential of such approaches to adapt and even drive team formation [10]. Overall, IPA efforts aim to address the following use cases: show a historical view of proficiency; show a live view of performance; enable macro and micro training adaptation, and; collect Big Data for trends analysis.

### 1.2 Experience API and Learning Ecosystems

The xAPI is a supporting specification for learning ecosystems. The xAPI specification defines an interface for a common and interoperable data store for xAPI statements, known as a Learning Record Store (LRS). The LRS provides a single storage point in a learning ecosystem. Systems within a learning ecosystem either act as a "producer" of xAPI statements or as a "consumer". [7]

### 1.3 The Generalized Intelligent Frameworks for Tutoring (GIFT)

GIFT, developed by ARL's Human Research and Engineering Directorate (HRED), provides a service-oriented framework of tools, methods and standards to make it easier to author computer-based tutoring systems (CBTS), manage instruction, and assess the effect of CBTS, components and methodologies [11]. GIFT was enhanced

to interoperate with xAPI in Version 3.02 to provide a *consumer* functionality and in version 4.1 to provide *producer* functionality.

### 1.4 Reference Architecture

The Figure below (Fig. 1) shows a reference architecture for a learning ecosystem using GIFT.

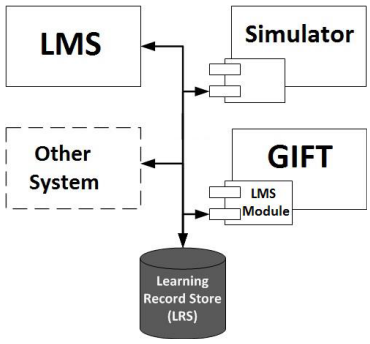



**Fig. 1.** A reference architecture for a learning ecosystem is shown [12]. The architecture shows a *Learning Record Store (LRS)* where data is stored and retrieved by elements of the ecosystem. A simulator or other system(s) may produce or consume data that is stored in the LRS. *GIFT* uses the *LMS Module*, which is enabled to both produce and consume xAPI data via the LRS submodule. GIFT is thus able to provide interoperability between these other systems using their xAPI data.

The architecture is composed of components that might comprise the learning ecosystem like a Learning Management System (LMS), a Simulator, GIFT, and other systems such as games or virtual worlds. In the example, the use of xAPI data as a common data format enables the LMS, GIFT, and other systems to be interoperable. The xAPI data created by the systems is stored in the LRS. In turn, xAPI data pulled from the LRS may be consumed by any of the systems within the ecosystem. Notably, GIFT provides both consumer and producer functionality as it (a) produces xAPI statements for other elements in the ecosystem and (b) consumes xAPI statements [12].

### 1.5 Use Cases

A number of use cases for learning ecosystems are supported by GIFT and its xAPI functionality. GIFT may be used in conjunction with an LRS and other systems to demonstrate and test these use cases. The following are some potential use cases that may be built upon GIFT and the xAPI functionality:

1. **Multiple System Performance Assessment.** Multiple systems including live scenarios using observer based tools, simulations, LMS, and games can be utilized to assess performance and produce xAPI data. Multiple systems can be used to assess

a singular competency or set of competencies across multiple delivery modalities to demonstrate performance over time. This data can be employed to drive adaptation as GIFT acts as a consumer.

2. **Using Simulation for Assessment.** A simulation may be used for performance assessment. The simulation produces xAPI data. This data may also be used to drive adaptation as GIFT acts as a consumer.
3. **GIFT-Driven Data Production.** xAPI data about course content and concepts contained within a course can be created and stored in an LRS. This data provides granular evidence of a user's interaction with a course and its corresponding concepts.
4. **Macro-Adaptation.** GIFT can provide macro adaptation or outer loop adaptation based upon the data it consumes. Performance deficiencies produced by GIFT or other systems that are stored as xAPI data can be used to intelligently navigate or recommend courses or other learning experiences. For example, a learner uses a simulator for marksmanship training and is found deficient in breathing techniques. The next time the learner logs into GIFT, he/she would then receive training recommendations such as courses or additional simulator training to improve their breathing techniques. In other words, GIFT leverages xAPI data about a user's deficiencies that is produced within a single learning event and then provides recommendations or adapts the individual's overall learning path to address these deficiencies.
5. **Inter-System Driven Micro-Adaptation.** GIFT can provide micro-adaptation within a scenario based upon data it consumes from other systems. For example, a learner participates in several marksmanship simulations and is found deficient in breathing techniques. Leveraging this xAPI data from one or multiple learning events, a future marksmanship simulator adapts within its scenario by providing additional guidance for breathing techniques. In other words, GIFT is able to leverage past xAPI data produced by other systems to drive micro-adaptation within future learning events in other systems.

## 2 Using GIFT and xAPI

GIFT (Version 4.1) is capable of both producing and consuming xAPI statements. Minimal configuration is required to setup this functionality in GIFT. Version 4.1 natively supports use cases 1, 2, 3, and 4 outlined in Section 1.5. Additional programming related to content development is required to support use case 5.

### 2.1 GIFT LMS/LRS Module

The LMS module within GIFT, responsible for retrieving and storing training and assessment history, enables xAPI support. The LMS module has been enhanced by creating an LRS submodule within which it allows both polling of and writing to the LRS.

### 2.2 Setting up GIFT with xAPI Support

In order to enable xAPI functionality for GIFT, an LRS must be available and connected to the network which GIFT is installed on. The following steps need to be completed to enable xAPI support in GIFT:

1. Install GIFT framework (refer to www.gifttutoring.org)
2. Install an LRS (see below)
3. Configure GIFT to communicate with the LRS end point

Several open source LRS options exist as well as commercial options. The following open source LRS solutions are currently available:

- Open source LRS from ADL - https://github.com/adlnet/ADL_LRS
- Hosted LRS from ADL- https://lrs.adlnet.gov/xapi/
- Open source LRS from learning locker - http://learninglocker.net/

**Configuration of xAPI End Point.** Once GIFT and the LRS are installed, GIFT must be configured to communicate with the LRS endpoint. The following steps must be undertaken to allow GIFT to communicate with the LRS:

1. Open the LMSConnections.xml file located in the `<GIFT Root>\GIFT\config\lms` directory
2. Select edit, and add a new connection entry under the `<LMSConnections>` root using the following information format and entering the `username`, `password`, and `URL` for the LRS installation between the XML elements:

```
<Connection>
  <enabled>true</enabled>
  <impl>lms.impl.Lrs</impl>
  <name>LRS Name</name>
  <Parameters>
    <networkAddress>https://lrs.url</networkAddress>
    <username>username</username>
    <password>password</password>
</Parameters>
</Connection>
```

### 2.3 GIFT as a Producer of Interoperable Data

Once configured, GIFT is enabled to act as a producer of xAPI data. As a producer, once a training scenario is completed, the course records and scores are passed to the LMS module for storage. This data is then passed to the LMS database as well as the LRS sub-module. An xAPI statement is generated for each level of the graded score nodes, and each statement is linked to their parent statement. The figure (Fig. 2.) be-

low outlines an example of data that is created and defined for the elements in the xAPI format.

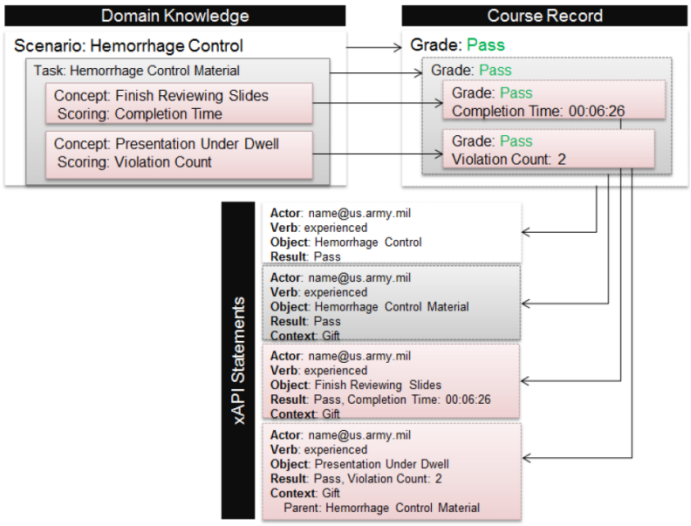



**Fig. 2.** An example of data from a *Domain Knowledge File, Course Record,* and *xAPI Statements* is shown. The example outlines the scenario, tasks, concept, and grades that are used to define the xAPI data elements. [12]

**Editing Domain Knowledge File.** In order for GIFT to produce xAPI data, the concepts that are represented within a course must be added to the XML file that represents the course. The following steps must be taken to update the file:

1. Edit the XML file for the course located at `<GIFT Root>\Domain`
2. Add a `<concepts>` section under the `<Course>` root. Below is an example of the addition of the `<concepts>` elements:

```
<Course name="Course Example"...>
    ...<concepts>
       <concept>Skill 1</concept>
       <concept>Skill 2</concept>
       <concept>Skill 3</concept>
    </concepts>...
</Course>
```

### 2.4 GIFT as a Consumer of Interoperable Data

The LMS module of GIFT also provides consumer functionality. The consumer function allows GIFT, via the LRS submodule, to poll the LRS end point. xAPI statements are used to extend GIFT's course suggestion capabilities. The LMS polling function retrieves a user's history, using their email address as an identifier when the user logs

into GIFT. The LMS module examines available course metadata definitions to find courses with concepts that match the user's deficiencies. The LMS module then recommends concepts matching deficiencies noted in xAPI statements for which the user is "below" concept proficiency. Dynamic filtering of course suggestions is presented through the "Recommended Courses" (See Fig 3).

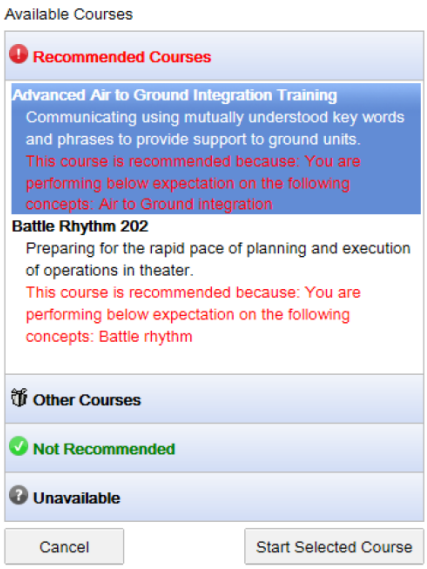



**Fig. 3.** A screen shot of GIFT *Available Courses* is shown. The example outlines recommended courses as determined by the LMS module by examining course metadata and deficiencies stored in xAPI statements within the LRS. [12]

## 3 Conclusions

GIFT allows enhanced functionality via its LMS module to integrate external data sources in a learning ecosystem. GIFT also enables data created within GIFT to be stored in an interoperable way that supports learning ecosystems via xAPI in an LRS. This functionality enables GIFT and other systems to evaluate incoming student competencies in order to better inform instructional strategy. Systems in the learning ecosystem are also enabled to make recommendations for the next training events based on performance data.

Using this functionality, researchers may test a number of different use cases and functions of adaptive learning in learning ecosystems. Usage of xAPI data in learning ecosystems with GIFT and other producers will allow consumers in learning ecosystems to assess and tailor learning and ultimately, to leverage Big Data analytics to discover trends over time.

The ability to leverage xAPI data in GIFT enables the investigation of a number of research questions. For example, the Army's current training modernization goals call for the development of persistent representations of Soldier performance in order to

support a culture of lifelong learning. In order to develop these complex student models, Soldier performance must be tracked across multiple training environments (e.g., events, simulators, courses). By producing and consuming xAPI statements, GIFT can support interoperable student models. However, while research is ongoing in this area, demonstrating interoperable performance data across multiple platforms through GIFT has yet to be accomplished. Further, the question of how best to remediate student performance using xAPI data through GIFT has yet to be investigated. A major question remains about the specific level of granularity of these xAPI statements that is most appropriate for adapting training through GIFT. It is very likely that as independent researchers develop their own solutions for adapting training based on xAPI data, the level of detail required will depend upon the specific domain and application. For the Army to reach its goal of tracking performance across a Soldier's career, however, there must be some consensus on how to standardize the granularity of xAPI statements. These, and other research questions, provide possibilities for research going forward.

## References


1. Kelly, D.: What Is a Learning Ecosystem? The eLearning Guild, http://twist.elearningguild.net/2013/11/what-is-a-learning-ecosystem/ (2013)
2. Advanced Distributed Learning: xAPI Specification. GitHub, https://github.com/adlnet/xAPI-Spec (2015)
3. American Society for Training & Development: 2013 State of the Industry. ASTD DBA Association for Talent Development (ATD), Alexandria (2013)
4. Association for Talent Development: 2014 State of the Industry. ASTD DBA Association for Talent Development (ATD), Alexandria (2014)
5. Woolf, B.P: A Roadmap for Education Technology. National Science Foundation #0637190 (2010)
6. Foreman, S. & Rosenburg, M. J.: Learning and Performance Ecosystems: Strategy, Technology, Impact, and Challenges. Santa Rosa: The eLearning Guild, http://www.elearningguild.com (2014)
7. Advanced Distributed Learning: Training and Learning Architecture (TLA): Experience API (xAPI), http://www.adlnet.gov/tla/experience-api (2014)
8. Poeppelman, T., Ayers, J., Hruska, Long, R., Amburn, C., Bink, M.: Interoperable Performance Assessment using the Experience API. In: The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC), vol. 2013. (2013)
9. Hruska, M., et al. (2013)
10. Hruska, M., Long, R., Amburn, C., Kilcullen, T., Poeppelman, T.: Experience API and Team Evaluation: Evolving Interoperable Performance Assessment. In: The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC), vol. 2014. (2014)
11. Sottilare, R.A., Brawner, K.W., Goldberg, B.S., & Holden, H.K.: The Generalized Intelligent Framework for Tutoring (GIFT). U.S. Army Research Laboratory – Human Research & Engineering Directorate (ARL-HRED), Orlando (2012)
12. Hruska, M., Long, R., Amburn, C.: Human Performance Interoperability via xAPI: Current Military Outreach Efforts. In: Fall Simulation Interoperability Workshop, 14F-SIW-035, Orlando (2014)

# Demonstration: Using GIFT to Support Students' Understanding of the UrbanSim Counter Insurgency Simulation

James. R. Segedy[1], John S. Kinnebrew[1], & Gautam Biswas[1]

[1]Institute of Software Integrated Systems, Department of Electrical Engineering and Computer Science, Vanderbilt University, 1025 16th Avenue South, Nashville, TN, 37212, U.S.A.
{james.segedy,john.s.kinnebrew,gautam.biswas}@vanderbilt.edu

**Abstract.** This paper presents our recent work with the Generalized Intelligent Framework for Tutoring (GIFT) for authoring tutors and training systems in concert with already developed external applications that provide a wide variety of educational experiences. In this paper, we describe our efforts to extend the GIFT system to develop metacognitive tutoring support for UrbanSim, a turn-based simulation environment for learning about counterinsurgency operations. We discuss specific extensions to GIFT as well as the links we have developed between GIFT and UrbanSim to track player activities. Additionally, we discuss a conversational approach that we are designing to interpret players' strategies and provide feedback when they adopt suboptimal approaches for their counter-insurgency operations.

**Keywords:** GIFT, UrbanSim, Scaffolding, Adaptive Support

## 1 Introduction

The Generalized Intelligent Framework for Tutoring (GIFT) provides a software platform and authoring system for designing, developing, and implementing online and in-class educational programs [1-2]. An important aspect of GIFT that makes it different from a number of conventional tutoring systems is its emphasis on interoperability across a variety of existing training applications (TAs). The overall goals are to reduce the high design and development costs of building computer-based tutors and to increase the reusability of educational applications while also creating engaging and adaptive learning spaces that students can access as needed.

While this is a significant advantage of GIFT, it introduces challenges in the number of use cases that must be considered in order to fully leverage and develop a general framework that is compatible with different forms of available educational resources. In this paper, we present our work in exploiting the GIFT platform to develop a metacognitive tutoring environment for the UrbanSim TA [3], a counter-insurgency (COIN) command simulation developed by the Institute for Creative Technologies at the University of Southern California. We describe the steps involved in developing generalized connectors that are currently tailored to support communi-

cation from UrbanSim to GIFT. Our work illustrates the flexibility of the GIFT platform to accommodate dynamic tracking of student activities in the UrbanSim COIN environment. Our overall goals are to simultaneously model student problem solving performance, behavior, and strategies, so that the developed GIFT tutor will provide dynamic support when students are involved in training episodes. Our experiences in developing GIFT to support cognitive and metacognitive tutoring lead to a set of design recommendations for further increasing the capabilities, adaptability, and flexibility of developing a variety of tutor-supported TAs with GIFT. We hope that our experiences and development efforts will help future GIFT developers working with other TAs.

## 2 UrbanSim

UrbanSim [3] (Figure 1) is a turn-based simulation environment in which users assume command of a COIN operation in a fictional Middle-Eastern country. Users have access to a wealth of information about the area of operation they have been assigned to. This includes: intelligence reports on key individuals, groups, and structures; information about the stability of each district and region in the area of operation; economic, military, and political ties between local groups in the region; the commanding team's current level of population support; and the team's progress in achieving six primary lines of effort. The actions that users take are scenario-specific, but they generally involve increasing the area's stability by making progress along the different lines of effort: (1) improving civil security; (2) improving governance; (3) improving economic stability; (4) strengthening the host nation's security forces; (5) developing and protecting essential services and infrastructure; and (6) gaining the trust and cooperation of the area's population.

Students conduct their operations by assigning orders to available units under their command (*e.g.*, *E CO b* and *G CO a* in Figure 1). To commit their orders, they press the *COMMIT FRAGOS* (FRAGmentary OrderS) button to complete one turn in the simulation environment. The simulation then executes the user's orders; simultaneously, it has access to a sociocultural model and complementary narrative engine that determine the actions of non-player characters in the game, which also affects the simulation results. For example, a friendly police officer may accidentally be killed during a patrol through a dangerous area. These *significant activities* and *situational reports* are communicated to the user, and the results of all activities may result in net changes to the user's population support and line of effort scores (see bottom right of Figure 1).

UrbanSim provides documentation and tutorials that should help students gain an appreciation for the challenges inherent in managing COIN operations. For example, they should learn the importance of maintaining situational awareness, managing trade-offs, and anticipating 2$^{nd}$- and 3$^{rd}$-order effects of their actions, especially as the game evolves [3]. They should also understand that their actions themselves produce intelligence (through their consequences as observed in the simulation environment), and, therefore, the need to continually "*learn and adapt*" in such complex domains

where the available information is often overwhelming, but at the same time may be incomplete. In other words, students should realize that their decisions produce intelligence that may be critical for decision making and planning during the next set of turns. Students can learn about the effects of their actions by viewing causal graphs provided by their security officer (S2). Users who adopt strategies to better understand the area of operation and its culture by viewing and interpreting the effects of their actions using these causal graphs should progressively make better decisions in the simulation environment as the COIN scenario evolves.


**Fig. 1.** UrbanSim

## 3 Developing an Application to Connecting UrbanSim to GIFT

Connecting a TA to the GIFT environment involves creating an interoperability interface. This interface is responsible for reporting the actions performed in the TA (and the resulting TA state) to GIFT while also handling control messages sent by GIFT to the TA to keep the two systems in alignment. The various components and their interactions necessary for connecting UrbanSim and GIFT are shown in Figure 2. UrbanSim produces log files that include information on the actions taken by actors in UrbanSim (and the effects of those actions). To report this information to GIFT, we have authored a Java application that monitors the log files and transmits the data to the interoperability interface, which passes the information to GIFT in a predefined struc-

tured format. GIFT can then use this data to tutor the student through a web-based interface.

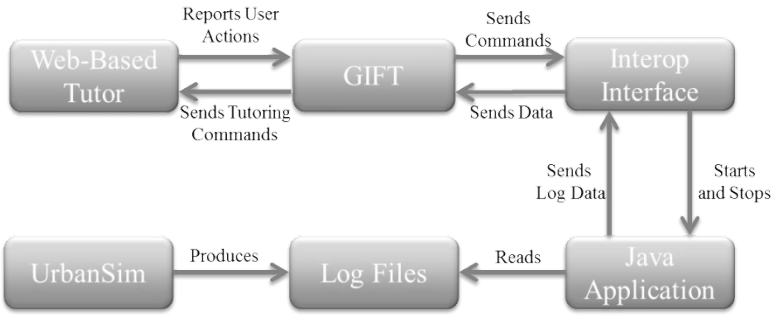



**Fig. 2.** Communication between GIFT and UrbanSim

The first step in developing this infrastructure required us to create the log parsing application. This involved completing the following steps:

1. Representing the complex set of data models used by UrbanSim.
2. Representing the actions taken by users and the contexts in which they occurred.
3. Monitoring the UrbanSim log directory and translating the log data into the representations created during steps 1 and 2.
4. Implementing code to establish a socket connection with the interop interface and publish the information obtained from the UrbanSim log files.

To represent the data models used by UrbanSim, we reverse engineered the plain-text save files generated by the program, extracted the data objects, their properties, and relationships to other objects and then created 22 Java classes to represent these data models. We then analyzed UrbanSim to extract the set of 38 measurable actions available to students in the program. Finally, we analyzed the set of 19 measurable contexts in which actions could occur. In this instance, a context can be considered to be equivalent to an interface configuration. For example, the configuration shown in Figure 1 shows a map of the area of operation. By tracking the actions and contexts logged by UrbanSim, we were able to create a detailed understanding of students' behaviors in the program. Once these objects had been defined, we focused on developing the algorithm for detecting changes in a log file, extracting the new information, processing it effectively, and then communicating it to the GIFT environment.

Once our log parser application had been written and tested, we turned our attention to writing the GIFT interoperability interface that would connect to the log parser, receive data, and report it to GIFT. To test this functionality, we configured a GIFT performance assessment condition. A condition receives data from the interoperability interface and uses it to assess a student's current level of performance with respect to a concept. In GIFT, a learner model is defined as a set of named concepts that are assessed continually while students are interacting with designated course

materials. At any time, each concept may be assessed as being below, at, or above expectation. The data representation is similar to the sampling of a stream: GIFT monitors the student's task performance over time and updates the concept assessments based on the student's most recent performance. Thus, a student may perform above expectation on one concept at some point in the simulated scenario, but fall below expectation on the next turn because they missed a critical piece of information (situational awareness). A history of these assessments is maintained for feedback purposes during a particular learning session and also across multiple sessions. In the tutor we are developing for UrbanSim, the condition we created detects when a student commits their orders and then presents them with a survey through GIFT's tutor user interface, as shown in Figure 3. We expect that the data collected through this survey will provide valuable insight into how students analyze situations in UrbanSim and learn from them as the simulation progresses.

## 4 Design Recommendations

Our goal in the work is to develop a tutor for UrbanSim using the GIFT framework that can analyze users' understanding of the current COIN scenario, and determine what strategies the user is adopting (if any) in determining their next moves. As we have moved toward this goal, our experiences in coupling UrbanSim and GIFT by authoring a log parsing tool and implementing an interoperability plugin resulted in the following design recommendations to facilitate tutor development:

1. **Expand Instructional Triggers**: GIFT is designed such that all tutoring decisions are bound to changes in a student's concept assessments (below, at, or above expectation). This makes it difficult to author instructional interventions based on non-performance factors. For example, to configure GIFT to show the survey in Figure 3, we had to create a performance assessment condition that detected when the student committed orders and assessed the *committed orders* concept as above expectation (instead of at expectation). The survey was then triggered by a change in the assessment of the committed orders concept. It may be desirable to expand these triggers such that instructional decisions may be directly bound to elapsed time or the occurrence of an event of interest. This could lead to more straightforward authoring of such instruction.
2. **Allow for Contextualized Conversational Instruction and Assessment**: GIFT allows a course author to develop mid-lesson surveys and uses the AutoTutor Lite [4] conversations to administer instructional interventions in appropriate situations. However, the content of these surveys and conversations must be determined ahead of time and may not be parameterized by variables derived from student performance and the state of the system. For example, question 1 in Figure 3 cannot be modified to ask the student about a specific FRAGO that they just committed. Additionally, GIFT does not allow many of these student responses on surveys and in conversations to serve as on-line assessments of their understanding (the exception is that specific answers to multiple choice questions may be linked to assessments of specific concepts). Thus, a student may, in their interactions with surveys and

conversations, reveal information about their understanding that is not utilized in future GIFT interactions. Contextualized conversational feedback has been shown to positively affect learner behavior [5], and so we recommend that such feedback capabilities be incorporated into future versions of GIFT.

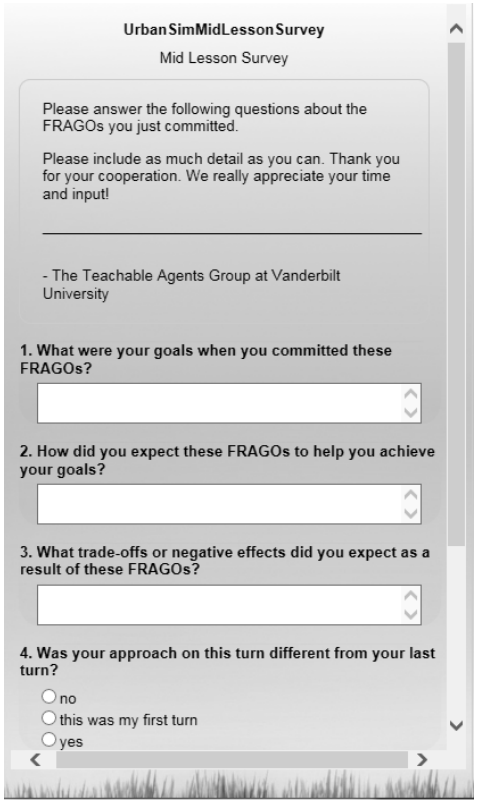



**Fig. 3.** UrbanSim survey presented through GIFT

3. **Expand Configurability of Dynamic Course Flow**: Currently, the primary structure of a GIFT course is fixed and specified in configuration files. Thus, even if concept assessments show that the student lacks pre-requisite skills, it is difficult to dynamically reconfigure the GIFT course to provide tutorial interventions that help the student develop that skill. In recent versions of GIFT, a system called eMAP [1] has been implemented which allows for dynamic assessment and instruction with regard to mastering a set of domain concepts. While this provides some dynamic capabilities in terms of course flow, we recommend that this system be expanded in the future. In particular, the potential of dynamic GIFT courses could be greatly enhanced with the ability to configure additional aspects of a course or instructional intervention to adapt to the needs of learners. For example, a future version of GIFT could support dynamic flow between *multiple* training applications if

a student's performance in one training application proves that they need training in pre-requisite skills before they are ready to succeed at their task.

## 5 Conclusions and Future Work

In this paper, we have presented our experiences in creating an application to synchronize the UrbanSim counter-insurgency command simulation with the Generalized Intelligent Framework for Tutoring (GIFT). We provided an overview of the process and potential in employing GIFT to augment a training application with new capabilities for learner modeling and support. The work presented here is part of a larger project aimed at developing metacognitive tutoring functionalities for GIFT to enhance students' future learning and problem-solving abilities. Our future work includes collecting data from students using UrbanSim, performing a systematic study of the strategies they employ and their sources of confusion, and using the insight obtained from this study to identify opportunities for providing feedback and scaffolding in our GIFT tutor for UrbanSim. A study of strategies at the cognitive and metacognitive levels may require us to build an extended task model of the COIN operations that are relevant to the UrbanSim scenario. We will also work toward implementing the design recommendations that we discussed in the previous section.

**Acknowledgements.** This work has been supported by an ARL research grant through funding from ONR grant #W911NF-14-2-0050. We wish to thank our ARL sponsors Drs. Robert Sottilare and Benjamin Goldberg, and Lt. Col. Kenric Smith with their help in understanding and interpreting the GIFT and UrbanSim COIN environments. We would also like to thank Nate Blomberg and Mike Hoffman from Dignitas Technologies, who provided significant technical assistance as we developed the technology necessary for connecting UrbanSim and GIFT.

## References

1. Nye, B.D., Sottilare, R.A., Ragusa, C., Hoffman, M.: Defining Instructional Challenges, Strategies, and Tactics for Adaptive Intelligent Tutoring Systems. In Sottilare, R., Graesser, A., Hu, X., Goldberg, B. (eds.): Design Recommendations for Intelligent Tutoring Systems (Vol. 2). U.S. Army Research Laboratory, Orlando, FL (2014)
2. Sottilare, R., Graesser, A., Hu, X., Holden, H. (eds): Design Recommendations for Intelligent Tutoring Systems (Vol. 1). U.S. Army Research Laboratory, Orlando, FL (2013)
3. McAlinden, R., Durlach, P., Lane, H., Gordon, A., Hart, J.: UrbanSim: A game-based instructional package for conducting counterinsurgency operations. In: Proceedings of the 26th Army Science Conference, Orlando, FL (2008)
4. Hu, X., Cai, Z., Craig, S.D., Wang, T., Graesser, A.C.: AutoTutor Lite. In: Proceedings of the 14th International Conference of Artificial Intelligence in Education, pp. 802 (2009).
5. Segedy, J.R., Kinnebrew, J.S., Biswas, G.: The Effect of Contextualized Conversational Feedback in a Complex Open-Ended Learning Environment. Educational Technology Research and Development, 61, 71-89 (2013)

# Opportunities and Challenges in Generalizable Sensor-Based Affect Recognition for Learning

Jonathan P. Rowe[1], Bradford W. Mott[1], and James C. Lester[1]

[1]Center for Educational Informatics, North Carolina State University, Raleigh, NC 27695
{jprowe, bwmott, lester}@ncsu.edu

**Abstract.** Recent years have witnessed major research advances in sensor-based affect recognition. Alongside these advances, there are many open questions about how effectively current affective recognition techniques generalize to new populations and domains. We conducted a study of learner affect with a population of cadets from the U.S. Military Academy using a serious game about tactical combat casualty care. Using the study data, we sought to reproduce prior affect recognition findings by inducing models that leveraged posture-based predictor features that had previously been found to predict affect in other populations and learning environments. Our findings suggest that features and techniques, drawn from the literature but adapted to our setting, did not yield comparably effective models of affect recognition. Several of our affect recognition models performed only marginally better than chance, and one model actually performed worse than chance, despite using principled features and methods. We discuss the challenges of devising generalizable models of affect recognition using sensor data, as well as opportunities for improving the accuracy and generalizability of posture-based affect recognition.

**Keywords:** Affect Recognition, Posture, Microsoft Kinect, GIFT

## 1 Introduction

Affect is instrumental to learning. Students' affective experiences shape their learning behaviors and outcomes, and vice versa. Growing recognition of this relationship has led to the emergence of work on affect-enabled learning technologies, which endow educational software with the ability to recognize, understand, and express affect. Several affect-enabled learning technologies have been developed in recent years, spanning a broad range of domains, including computer science education [1], reading comprehension [2], mathematics [3], and computer literacy [4]. Although these bespoke affect-sensitive systems have yielded promising results, there are many open questions about whether existing affect recognition techniques generalize to new domains, populations, and settings.

Recent work on sensor-based affect recognition holds promise for yielding generalizable models. Because sensor-based models typically do not rely on features that are specific to particular learning environments, in principle, they should port across domains and settings. Sensor-based affect recognition models have been devised for a

range of modalities, including facial recognition, gaze tracking, speech analysis, physiological signals (e.g., heart rate, electrodermal activity), hand gesture, and posture [5]. In this work, we focus on posture-based affect recognition, which has shown promise for its capacity to predict student affect [1, 3, 4]. Motion sensors, such as Microsoft Kinect, can be used to gather rich data streams about posture, they are relatively low-cost, and they are increasingly getting integrated into mainstream computers [6]. By modeling these rich data streams with machine learning techniques, posture-based affect recognition models have been induced that can effectively predict participants' affective self-reports, as well as expert judgments of affect gleaned from freeze-frame video analyses [1, 3, 4].

In this paper, we summarize our work on posture-based affect recognition with the Generalized Intelligent Framework for Tutoring (GIFT). In collaboration with Teachers College Columbia University and the U.S. Army Research Laboratory, we conducted a study of learner affect with cadets from the U.S. Military Academy (USMA) using a serious game for learning tactical combat casualty care skills. Using this study data, we sought to reproduce prior affect recognition findings, leveraging posture-based predictor features that had previously been found to predict affect in other populations and learning environments. However, our results indicated that the same features and techniques, adapted to our setting, did not yield comparably effective models. Our affect recognition models performed only marginally better than chance, and in fact, one model actually performed worse than chance. We discuss the challenges of devising generalizable models of affect recognition using sensor data, and describe opportunities for improving the predictive accuracy of posture-based affect recognition models.

## 2 Posture Sensor-Based Affect Recognition

Several research labs have investigated multimodal affect recognition in learning environments over the past decade. Our research on generalizable sensor-based affect recognition is strongly influenced by this work. To date, posture-based affect recognition models have been induced with data from pressure-sensitive chairs [3, 4], as well as motion sensors, such as Microsoft Kinect [1]. These two data streams, drawing from distinct types of sensors, are superficially different, but can be distilled into analogous predictor features that have similar relationships with affective states such as engagement, boredom, frustration, and confusion. Features can be distilled from both types of data to indicate leaning forward, leaning backward, sitting upright, and fidgeting. We summarize several representative studies that have utilized these types of features to recognize learner affect, and that have influenced our own work.

D'Mello and Graesser utilized posture data from the Body Pressure Measurement System (BPMS) to predict judgments of student affect during learning with AutoTutor [4]. The BPMS is a pressure-sensitive system that is comprised of a grid of sensing elements placed across a chair's seat and back. In their study, participants were video recorded, and several judges analyzed the video using freeze frame analysis in order to code participants' affective states retrospectively. Using this data, D'Mello

and Graesser induced a series of emotion-specific binary logistic regression models, each distinguishing a particular affective state from neutral, using 16 posturebased features as predictors. Their findings indicated that the models, averaged across judges, explained approximately 11% of the variance in affective state, with findings in line with an attentive-arousal theoretical framework. Specifically, affect such as delight and flow coincided with forward leaning, boredom coincided with a tendency to lean back, and states such as confusion and frustration coincided with an upright posture.

Cooper et al. used a suite of sensors to collect data on student affect in Wayang Outpost, an ITS for high school geometry [3]. The sensors included a skin conductance bracelet, pressure sensitive mouse, pressure sensitive chair, and mental state camera, which provided data on student posture, movement, grip tension, arousal, and facial expression. The pressure sensitive chair was a simplified version of the sensing system utilized by D'Mello & Graesser [4], utilizing a series of six forcesensitive resistors distributed across the seat and back of a seat cover cushion. Data from these channels was distilled into predictor features to predict students' emotion self-reports, which were queried every five minutes throughout the learning interaction. The posture-based features included net change in seat and back pressure between the current timestep and previous timestep, and a feature indicating whether the student was leaning forward or not. Step-wise linear regression models were induced to predict students' emotion self-reports. Results indicated that posture-based features were significantly predictive of self-reported excitement during learning, although they were not part of the best-performing models for other emotional states.

Grafsgaard et al. have investigated postured-based affect prediction using Microsoft Kinect sensors with an intelligent tutoring system for introductory programming [1]. Posture features were distilled from depth image recordings by tracking the distance between the depth camera and the participant's head, upper torso, and lower torso. The features included discretized distance indicators, such as near, mid, and far head positions, each determined by whether the tracked head point was closer or farther from the median head position by one standard deviation. In addition, a postural movement feature was distilled to label occasions where the average amount of acceleration of the head tracking point was greater than the population average over a one-second window. The posture-based predictor features were combined with features distilled from other multimodal streams to induce multiple regression models for predicting students' retrospective self-reports of engagement and frustration. Findings indicated that posture features were predictive of both self-reported affective states: leaning forward was predictive of both higher engagement and higher frustration, and postural movement was associated with increased frustration and reduced learning.

Building upon this foundation, we set out to distill similar predictor features from the data collected at USMA, and apply similar machine learning methods, to produce affect recognition models for predicting field observations of affect.

## 3 Kinect-Driven Affect Recognition in GIFT

We collected learning and affect data from 119 USMA cadets as they used the vMedic serious game environment for learning tactical combat casualty care skills. In vMedic, the learner adopts the role of a combat medic who must properly treat and evacuate one (or several) of her injured fellow soldiers by following standard medical procedures within the game environment. All participants completed the same training module, which was managed by GIFT. The training module consisted of a pre-test, a brief PowerPoint on tactical combat casualty care, four training scenarios in vMedic, and a post-test.

Each participant was assigned to a research station that consisted of an Alienware laptop, a Microsoft Kinect for Windows sensor, an Affectiva Q Sensor, and a mouse and pair of headphones. As participants completed the study materials, a pair of field observers regularly recorded participants' physical displays of emotion. The field observers followed an observation protocol, BROMP, developed by Baker et al. [7], in which observers walked around the perimeter of the study room, discreetly recording observations of each participant's affect in a round robin sequence. The field observers coded for seven affective states: concentration, confusion, boredom, surprise, frustration, contempt, and other.

The study produced several parallel data streams, including vMedic trace data, Kinect position tracking data, electrodermal activity data, pre- and post-test response data, and field observation data. In this work, we focus on analysis of the Kinect and field observation data, which were fused into a single time-synchronized dataset. The dataset was cleaned and filtered in order to remove any Kinect-tracking glitches, as well as non-essential vertex data. Afterward, 73 predictor features were distilled, which characterized participants' postural positions and dynamics, inspired by similar features from the research literature on posture-based affect recognition. The features included summary statistics for three points tracked by the Kinect: head, top_skull, and center_shoulder. Specifically, we computed features for the current distance and depth of each vertex; the minimum, maximum, median, and variance in distance of each vertex observed thus far; the same statistics for 5, 10, and 20-second windows; several features that characterized net changes in vertex distance, analogous to the net_change features reported in [3, 4]; and sit_forward, sit_back, and sit_mid features analogous to those reported in [1, 3].

Using this feature data, we induced separate affect detectors for each emotional state using a range of machine learning techniques in RapidMiner 5.3, inclu ing J48 decision trees, naïve Bayes, support vector machines, logistic regression, and JRip [8]. The detectors were cross-validated using 10-fold participant-level cross validation. Oversampling was used to balance class frequency by cloning minority class instances in the training sets. Forward feature selection was performed to reduce the number of predictor features used in the models. We calculated Kappa and A' to assess the models' performance.

Across all of the emotions, our posture-based affect recognition models achieved an average Kappa of 0.064, and 0.521 for A' [8]. The best performing model was for boredom, which showed Kappa=0.109, A'=0.528 using logistic regression. Overall,

the models performed slightly better than chance, with the exception of the surprise detector, which actually performed worse than chance, Kappa=-0.001, A'=0.493.

These results were surprisingly modest, despite our best efforts to run a carefully designed study and reproduce previously reported methods. There are several possible explanations. It is possible that BROMP labels, which are based on holistic judgments of affect over 20-second windows, are ill matched for methods that leverage low-level postural features as predictors. Previous work utilized self-reports and freeze frame video analysis, which have different tradeoffs than BROMP. Additionally, much of the work on posture-based affect recognition has taken place in laboratory settings with a single participant at a time. In our study, up to 10 participants were present, with each research station having a slightly different sensor position and orientation. This variation may have introduced additional noise to the data, which could have been problematic for the methods reported here. Further, the population of learners we used in the study, USMA cadets, showed considerable restraint in their physical expressions of affect. As such, the displays of affect via body language may have been different than those encountered in prior work, making them ill matched for the predictor features that we engineered. These findings underscore the challenges to be overcome in efforts to devise generalizable models of affect recognition.

We draw several lessons for our continued work on sensor-based affect recognition with GIFT. First, orienting Kinect sensors' position and orientation to track points on participants' lower torso could prove important for posture detection. In the present study, our sensor configuration enabled us to track only vertices on participants' upper torso and head, which may have limited the features we could distill.

Second, it would be useful to validate the Kinect vertex data recorded by GIFT against the sensor's raw depth video data. Prior work on Kinect-based posture detection directly leveraged raw depth channel data, but this method is memoryintensive and requires custom implementation of posture tracking algorithms [1]. While vertex data produced by Kinect should in principle provide the same information about posture as raw depth data, validating this fact would ensure that our findings relate to the generalizability of affect recognition techniques, and not assumptions about underlying data sources.

Third, investigating alternate machine learning techniques could prove useful for enhancing the predictive ability of posture-based predictor features. It is possible that temporal models, such as dynamic Bayesian networks, which explicitly model shifts in posture and affect, could yield improved results. Furthermore, recent work on deep learning techniques may show promise, given their capacity to perform automated representation learning. Although additional work is merited to manually engineer high-level features to match the holistic encodings of affect provided by BROMP, it would be ideal to automate this manual feature engineering process, as is one of the promises of representation learning techniques such as deep learning.

## 4 Conclusions

We have described work investigating the generalizability of posture sensor-based affect recognition. We collected a multimodal dataset on affect and learning with a group of USMA cadets using a serious game for tactical combat casualty care. Leveraging techniques from the affective computing research literature, we distilled a range of posture-based predictor features for modeling participants' affective states with machine learning. Our results indicated that posture-based features and models, which had previously been found to yield effective affect recognition systems, did not work as effectively on our data as had been found with other populations and learning environments. In fact, most of our affect recognition models performed only marginally better than chance, despite the use of principled features and models. Although there are several directions to investigate for enhancing our posture-based affect recognition models, the failure of existing techniques to generalize to our data is notable. These findings underscore the challenges, and opportunities, in research on affect recognition and generalizable approaches to intelligent tutoring.

**Acknowledgments.** The authors wish to thank the U.S. Army Research Laboratory for supporting this research. We are grateful to Vasiliki Georgoulas, Michael Matthews, and James Ness for facilitating the study at the United States Military Academy. Additionally, we wish to thank our collaborators, Ryan Baker, Jeanine DeFalco, and Luc Paquette at Teacher's College Columbia University, and Keith Brawner, Benjamin Goldberg, and Bob Sottilare from ARL.

## References

1. Grafsgaard, J. F., Wiggins, J. B., Vail, A. K., Boyer, K. E., Wiebe, E. N., Lester, J. C.: The Additive Value of Multimodal Features for Predicting Engagement, Frustration, and Learning during Tutoring. In: Proceedings of the 16th ACM International Conference on Multimodal Interaction, pp. 42–49. (2014)
2. Mills, C., Bosch, N., Graesser, A., Mello, S. D.: To Quit or Not to Quit: Predicting Future Behavioral Disengagement from Reading Patterns. In: Proceedings of the 12th International Conference on Intelligent Tutoring Systems, pp. 19–28. (2014)
3. Cooper, D. G., Arroyo, I., Woolf, B. P., Muldner, K., Burleson, W., Christopherson, R.: Sensors model student self concept in the classroom. In: Proceedings of the 17th International Conference on User Modeling, Adaptation, and Personalization, pp. 30–41. (2009)
4. Mello, S. D., Graesser, A.: Mining Bodily Patterns of Affective Experience During Learning. In: Proceedings of the 3rd International Conference on Educational Data Mining, pp. 31–40. (2010)
5. D'Mello, S. K., Kory, J.: A Review and Meta-Analysis of Multimodal Affect Detection Systems. ACM Computing Surveys, 47(3), 43 (2015)
6. Intel. (2015, March 20). Intel RealSense. Retrieved from http://www.intel.com/realsense.
7. Baker, R. S. J. d., D'Mello, S. K., Rodrigo, M. M. T., Graesser, A. C.: Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive– affective

states during interactions with three different computer-based learning environments. International Journal of Human-Computer Studies, 68(4), 223–241 (2010)
8. Paquette, L., Rowe, J., Baker, R., Mott, B., Lester, J., DeFalco, J., Brawner, K., Sottilare, R., Georgoulas, V.: Sensor-Free or Sensor-Full: A Comparison of Data Modalities in Multi-Channel Affect Detection. Under review for the 8th International Conference on Educational Data Mining, (under review)

# The Development of a Testbed to Assess an Intelligent Tutoring System for Teams

Desmond Bonner[1], Jamiahus Walton[1], Michael C. Dorneich[1], Stephen B. Gilbert[1], Eliot Winer[2], Robert A. Sottilare[3]

[1]Industrial and Manufacting Systems Engineering, 2Mechanical Engineering
[2]Iowa State University, 3004 Black Engineering, Ames, IA, 50011, USA.
[3]U.S. Army Research Laboratory- Human Research and Engineering Directorate
Orlando, Florida
dbonner, jwalton, dorneich, gilbert, ewiner}@iastate.edu,
robert.sottilare@us.army.mil

**Abstract.** Work has been ongoing to develop an Intelligent Tutoring System (ITS) for teams. As part of this work, we are developing a flexible, scalable, military-based set of collaborative team tasks that can serve as a "testbed" to exercise various aspects of a team ITS architecture.Warfighting teams are a core part of any operation as individual soldiers combine their skill sets and plan, coordinate and act as one entity to accomplish assigned objectives. The team ITS test bed presented in this paper uses simple team tasks to train soldiers on basic functions including observation, target detection, target identification, communication within the team and decision making under stress. The testbed allows for manipulation of various dimensions of tutor feedback, learner work-load, and team size. The testbed enables researchers to systematically evaluate the effectiveness of different types of feedback on militarily-relevant training tasks.

**Keywords:** Team Tutoring, Team Training, Intelligent Tutoring Systems (ITSs), Generalized Intelligent Framework for Tutoring (GIFT)

## 1 Introduction

Work has been ongoing to develop Intelligent Tutoring Systems (ITSs) to support tailored, guided learning experiences for teams conducting collaborative tasks [1-3]. As part of this work, we have been developing a flexible, scalable, militarily-relevant set of collaborative team tasks that can serve as "testbed" to exercise various aspects of a team ITS architecture. This paper focuses on the development of a generic testbed and an effective implementation of an ITS for training team tasks which can serve as a model for future ITSs. While work has been previously conducted in this area (see section 2), the work which is described in this paper differs as it attempts to remove humans from the tutor role completely, seeks to encourage proper performance while learners are performing several sub-tasks within a larger one, and ac-

complish both goals while simultaneously applying them to two or more individual learners concurrently within a collaborative team setting.

There is a need for effective team training in the military to match the tasks conducted by military teams in the operational environment. It is important that tailored training be easy to distribute while minimizing cost [4]. Tailored training through the convergence of ITSs and Virtual Reality (VR) training (e.g., serious games and virtual simulations) is emerging to become part of the Army's plan for the 21st Century soldier competencies [4,5]. VR can simulate a combat zone and allow inexperienced soldiers to learn how to react to high-stress situations without exposure to actual harm. In a virtual environment, random events can occur by the trainer's design, which mimic events such as sniper attacks, improvised explosive devices (IEDs), and hostile civilian environments. The goal for the military application of VR is not only to expose soldiers to a broad spectrum of potential environments, but also effectively train soldiers by providing tailored instruction and feedback [5]. The result is more efficient training and shorter time to reach competency.

An ITS is a computerized learning environment that incorporates content from a specific domain (e.g. military training) to provide instruction through the use of feedback and immediate interaction based on an individual learner's rate of comprehension [6]. ITSs attempt to play the role of a trainer or instructor in a training simulation. However, capturing the expertise of a human trainer is difficult. The most crucial element in training is the experience of the trainer, usually a Non-Commissioned Officer (NCO), which is shared with soldiers [7]. Beyond individual training, the military trains teams of soldiers to work together to accomplish mission goals. Military teams are capable of achieving goals that cannot be accomplished by an individual warfighter alone. Thus, the trainer is responsible for enhancing the performance and learning of multiple soldiers.

A human trainer is most effective when giving one-on-one training or tutoring [8]. The goal of ITS development was to find a tutor that was just as effective as one-to-one tutoring as it is the most effective form of education. Students who receive one-to-one tutoring perform better than those who receive conventional group education [9]. Most students have the potential to reach a high level of learning and human one-to-one tutoring allows them the opportunity to reach their potential. However, only until recently, ITS's were solely focused on individual tutoring [10]. The challenge is to make ITS training effective for teams. Developing and testing ITS for effective team training is vital to the success of military operations. Due to the increasing complexity of missions which include specific tasks, the timing and characteristics of feedback that teams receive during training is crucial to understanding a tutor's effectiveness in addition to its development [3].

Development of a Team ITS will extend an existing individual (or one-to-one) authoring architecture to small groups. Our goal is to develop an architecture for authoring team ITSs using VR and the authoring capabilities of the General Intelligent Framework for Tutoring (GIFT) [11]. This will require a test bed to assess the effectiveness of the tutor. The testbed needs to be flexible and scalable so that it can be adapted to explore different teaming variables, such as the elements and dimensions of team-based feedback [2, 12].

To develop a team training testbed, the collaborative team task of joint reconnaissance ("recon") was chosen based on its ability to test various dimensions of feedback, and its scalability with respect to workload and team size. The next section describes related work that informed the development of the testbed. The subsequent section details the generic Recon Task Testbed developed to exercise a team tutoring architecture. Finally, an initial implementation is described that tests two of the many dimensions of feedback: public vs. private, and team vs. individual feedback.

## 2 Related Work

Several areas of research informed the development of the Recon Task Testbed. Team training in the military and the development of individual ITSs has formed the basis of the collaborative tasks included within the Testbed. Research on the types of feedback in training scenarios was reviewed extensively. Finally, the authoring tool that is being extended from individuals to team tutors is briefly introduced. This research supports U.S. Army training objectives [5].

One of the goals for the Army is to maintain a tactical edge over potential threats through the ability to learn faster [5]. In order for teams to learn faster it is necessary for their training to be adaptive. The military is headed towards more effective training by becoming less dependent on lengthy PowerPoint slides for soldier comprehension [5]. When using an excess of PowerPoint slides to present important information students will be less engaged and unlikely to grasp material [13]. When the time comes to apply the material in field training, the learner's earlier low engagement may reflect performance. With VR training, students can be exposed to material and apply it simultaneously.

Applying VR with an ITS has been explored in previous work [4,14]. ITSs have been more effective for learning than traditional training which takes place in classrooms [6]. It reduces the time required for learning and in some cases is less costly than conventional learning. ITSs such as SimStudent predict future behavior from students by looking at previous behavioral patterns and therefore can reduce learning time [15]. It has been difficult to successfully apply what works in individual ITSs to a Team ITS [10]. Team training requires a higher expenditure of flexibility and energy in regards to authoring ITSs in addition to the human trainer. Some tutors have been created in order to assist human trainers with facilitating collaborative learning and team training such as the Advanced Embedded Training System (AETS) [16]. With AETS, the workload for the human trainer required for successful tactical team training was reduced [16].

Teams are usually made up of individuals who differ in competency, content comprehension, and skill levels. Also, team interaction is another factor which individual tutors do not have to consider. Work from Suh and Lee address the complexities of team collaborative work through an asynchronous text system called the Extensible Collaborative learning Agent (ECOLA) [17]. In their work, they go on to describe challenges such as complex educational elements which exist in collaborative systems. Specifically, feedback and the method which it is distributed can influence a

team. According to Billings, feedback generally improves performance [18]. Additional characteristics of a team including how the team reacts to feedback may determine its success or failure before an assessment task even begins [1]. Team feedback has many dimensions [2]: subject (individual, team), target (public, private), timing (immediate, after), type (proactive, reactive), specificity (generic, specific), tone (positive, negative), and style (collaborative, competitive). These aspects can be effectively tested in an ITS authoring environment by using GIFT.

GIFT is a modular computer-based ITS which has three primary functions which include authoring, instructional management and evaluation of ITSs. GIFT's authoring goals are to decrease effort for creating tutors by providing aid in organizing knowledge, supporting good design principles, and leveraging open source solutions [19]. Instructional manager goals for GIFT are to integrate pedagogical best practices in ITS created from the platform. The effectiveness evaluation construct's purpose is to allow researchers to evaluate whole ITSs or their component tools and methods of ITS technologies [19]. GIFT was developed for use with individual training. The project on which this paper is based has the goal to extend GIFT to team ITSs. A team architecture has been proposed [3]. The Recon Test Bed has been developed to test that architecture.

## 3 Testbed Development

The Recon Testbed is based on the collaborative team task of reconnaissance, and requires several military skills. In the military, communication is key to mission success, especially for security purposes. There are four types of security operations. They include Screening, Guarding, Covering, and Area Security [20]. The Recon Scenario is derived from Area Security as it involves reconnaissance in support of various assets. Specifically it resembles aspects of patrolling. In patrolling, Observation posts are used to provide security to a platoon [7]. Within the task, users perform the five fundamentals of all security related missions. These include: orient the main body, perform continuous reconnaissance, provide early and accurate warnings, provide reaction time and maneuver space, and maintain enemy contact [7]. How well users execute these fundamentals during the task will partially determine the feedback that is received.

Feedback in teams has many dimensions (see Section 2). It is the goal of the testbed to enable experimenters to vary these dimensions as needed to test the effectiveness of team feedback. In addition, the testbed must allow the experimenter to manipulate the task load (workload) of the participant. This can be done by changing the rate at which events occur.

The recon task itself, built in VBS2, is meant to serve as the testbed for these dimensions. In conducting the task, users are exposed to various military scenarios such as observation, fields of fire, and communication within a fire team element. The team members (two minimum) are assigned sectors to watch. For instance, if there are four teammates on the top of a building, each may be assigned one quarter of the 360-degree field of view. Each is tasked with scanning (observing) their sector by con-

stantly panning to see the extent of activity (target detection) in their sector. Each trainee must identity (target identification) any opposing force member that is spotted. If the threat is moving into a teammate's sector, the learner then must transfer responsibility by communicating the position to that teammate. The teammate must then acknowledge the change of responsibility back to the first teammate, thus accepting responsibility.

In the example of four team members, the initial condition of scanning is based on the 90-degree sector given to each team member. The team member must scan this sector continuously for the purpose of mimicking the actual field task and to effectively participate in the other conditions of the recon scenario. The team is given feedback according to how effectively they cover their entire area. This is relative to fields of fire and reconnaissance strategies outlined in the Army Field Manual for Infantry Platoons and Squads [7].

Figure 1 illustrates two teammates (BLUFOR) each monitoring a 90-degree sector. Participants are responsible for tracking all targets (OPFOR) and ignoring any distractors (civilians). When a target approaches the sector border in the center, the participant must alert the team member who has responsibility for that sector. Workload can manipulated by changing the number of enemies/civilians, the speed by which they move, the similarity of their appearance, and the rate by which they appear/disappear.

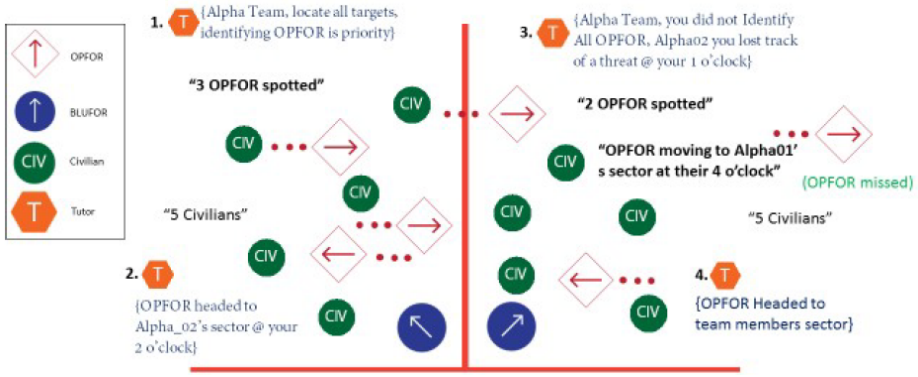



**Fig. 1.** Example of a recon task in which two team members scan a 180-degree field.

The dimensions of feedback can be varied in the task by changing the content or delivery of the ITS feedback. Table 1 describes how feedback dimensions can be manipulated in the Recon Testbed to test the effectiveness of team feedback.

**Table 1.** Dimensions of Feedback

| Dimension | Levels | How realized in Recon Testbed |
|---|---|---|
| Subject | *Individual, Team* | Tutor provides feedback about an **individual** team member or **entire** team |
| Target | *Public, Private* | Tutor provides feedback to either a single person (**private**) or team (**public**) |
| Timing | *immediate, after, omitted* | Feedback occurs based on patterns or task effectiveness **during** the task, or **after** overall the grade or rating is given. Feedback is **omitted** when an error is committed, but is not sufficiently important to interrupt training to provide immediate feedback or to be included in the After Action Review. |
| Type | *Proactive, reactive* | **Proactive**: feedback before a learner makes error,<br>**Reactive**: Feedback after a learner makes an error |
| Specificity | *Generic, specific* | **Generic:** "*Good Job Soldier*"<br>**Specific**: "*You missed an OPFOR located at 7 o'clock*" |
| Tone | *Positive, negative* | **Positive:** "*…you might want to try…*"<br>**Negative**: "*…your poor performance is hurting the team*" |
| Style | *Collaborative, Competitive* | **Collaborative**: "Slow down scanning to help team…"<br>**Competitive**: "Your performance is worse than Joe." |

## 4 Initial Implementation and Future Work

The first implementation will study two dimensions of feedback: Access (public vs. private) to feedback, and target (group vs. individual) feedback. For example, the feedback is given to a single person in the private condition while the entire team is given feedback in the public setting. Individual and Group feedback refers to whom the feedback is about (one person's actions or the team's efforts). Table 1 describes the tasks of each learner when monitoring their sector. The team tutor will be the basis of experiments to test the effectiveness of different types of team ITS feedback.

**Table 2.** Tasks performed in the initial Recon Testbed by each learner.

| Task | Description |
|---|---|
| Scanning | The Learner rotates their viewpoint within the 180 degree sector. Learner must cover the entire 180 continuously throughout the task |
| Identify | The learner presses a key whenever they spot a new OPFOR avatar. This must be done quickly with 10 seconds of the OPFOR becoming visible |
| Transfer (informing) | When an OPFOR avatar is close to moving into a teammate's assigned sector, the learner must inform the team member. |
| Transfer (confirming) | Learner must confirm transfer of responsibility for the OPFOR moving into their sector from the teammate who initiated the transfer process. |

Beyond the initial study, we plan to expand the Recon Testbed significantly. Currently, the testbed allows for the manipulation of feedback dimensions that enables researchers to systematically test the effectiveness of different types of feedback on training. The testbed is scalable and flexible, allowing for different sizes of teams, and varying levels of task load, which can be altered in the future. By including these features, the testbed will provide a platform to study several aspects of military-relevant team training.

## References


1. D. Bonner, S. Gilbert, M.C. Dorneich, S. Burke, J. Walton, C. Ray, & E. Winer, (2015, February). Taxonomy of Teams, Team Tasks, and Tutors. In *Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium*.
2. J. Walton, M.C. Dorneich, S. Gilbert, D. Bonner, E. Winer, & C. Ray (2015, February). Modality and Timing of Team Feedback: Implications for GIFT. In *Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium*.
3. E. Winer, J. Holub, T. Richardson, M. Hoffman, S. Gilbert, & M. Dorneich (2015). "Characteristics of a Multi-User Tutoring Architecture software architecture," In. R. Sottilare (Ed) *2nd Annual GIFT User Symposium*. Pittsburgh, PA, June 12-13. *Army Research Laboratory*, Orlando, Florida, 2015. ISBN: 978-0-9893923-4-1
4. J. Stevens, P. Kincaid, & R. Sottilare (2015). Visual modality research in virtual and mixed reality simulation. *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*.
5. M. E. Dempsey, (2011). The US Army learning concept for 2015 (TRADOC Pam 525-8-2). Washington, DC: Department of the Army HQ, US Training and Doctrine Command.
6. A. C. Graesser, M. W. Conley, & A. Olney, (2012). Intelligent tutoring systems.

# Developing an Experiment with GIFT: Then and Now

Anne M. Sinatra, Ph.D.[1]

[1]U.S. Army Research Laboratory – Human Research and Engineering Directorate
anne.m.sinatra.civ@mail.mil

**Abstract.** The Generalized Intelligent Framework for Tutoring (GIFT) is a domain-independent open-source intelligent tutoring framework. In the past new versions of GIFT were released every 6 months, and currently, officially tested versions of GIFT are released every 9 months. Each new version of GIFT includes additional capabilities and functionalities. In the current paper and presentation, the "Logic Puzzle Tutorial" course that was developed in GIFT 2.5, and has been included with releases of GIFT since version 4.0 will be discussed. The presentation will describe the rationale and methods behind the course's development, and discuss different approaches that might have been used with the features that are present in GIFT today.

Keywords: Adaptive Tutoring, Intelligent Tutoring, Experimental Design, Course Development, Generalized Intelligent Framework for Tutoring, Psychology

## 1 Introduction

The Generalized Intelligent Framework for Tutoring (GIFT) is an open-source domain-independent intelligent tutoring framework [1]. Since GIFT is domain independent it offers great flexibility in the types of tutors and experiments that can be developed with it. While the development of adaptive tutoring systems is a primary objective of GIFT, it was also designed to be used as a testbed and for analysis purposes. Experiments can and have been developed and run using GIFT [2,3]. GIFT provides opportunities to create experiments that use adaptive feedback/assessment, and experiments that do not. In fact, GIFT is very useful as a mechanism to run traditional experiments in the area of psychology [4]. One such experiment was run as part of a Post Doctoral fellowship with Army Research Laboratory to investigate the impact of self-reference and context personalization on computer-based tutoring [3,5]. The skill that was taught to individuals was deductive reasoning, which was done through an interactive logic puzzle tutorial. The current paper discusses the development of the logic puzzle tutorial, and the different approaches that may have been taken had the current features of GIFT been available at the time.

## 1.1 Logic Puzzle Tutorial Experiment

The "Logic Puzzle Tutorial" course has been included with GIFT software releases since GIFT 4.0 in November 2013. This tutorial was originally developed for use in an experiment to examine the impact of self-reference on learning deductive reasoning skills and completing logic puzzles. The description and results of the original experiment are available in the form of an Army Research Laboratory technical report [5]. In the full experiment there were 3 versions of the logic puzzle tutorial. All of the versions were identical except for the names that were included in the puzzles and learning material. The names that were included were determined based on the condition, and the names that the participants were asked to type into the program. In the self-reference condition, the participant entered his or her name, and the names of 2 friends. In the popular culture condition, the participants were prompted to enter specific names of characters from the Harry Potter series. In the baseline condition, participants were asked to enter 3 provided names that were not common for their age group (based on birth name data). See Figure 1 for a screenshot comparison of the popular culture and baseline conditions.

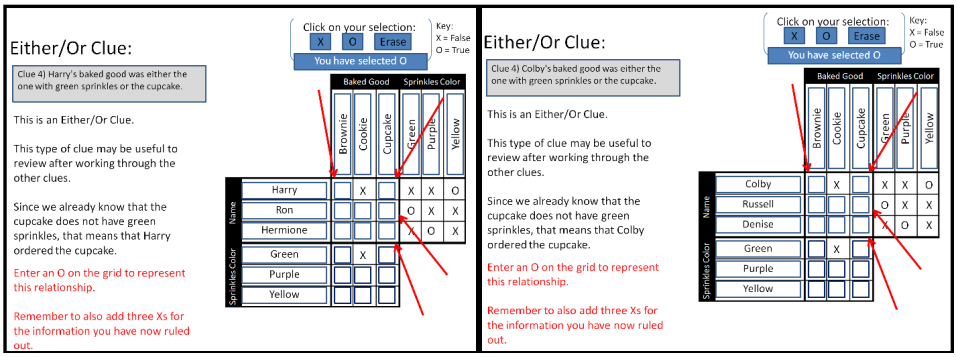



**Fig. 1.** Screenshots that demonstrate the manipulation of interest in the original logic puzzle tutorial experiment. Note that the names present in the puzzles and clues are different, with the Popular Culture condition on the left, and the Baseline condition on the right.

In the experiment, adaptive tutoring and feedback were provided to participants through a logic puzzle tutorial created in PowerPoint with Visual Basic for Applications (VBA). GIFT provided the interface that participants used for the study, presented surveys, opened and closed the PowerPoint based tutorial, launched web-page based questionnaires, and connected to a Q-sensor for physiological data collection. The original course was developed in GIFT 2.5, which was an experiment-based version of the November 2012 release of GIFT 2.0. Despite being developed in GIFT 2.5, the course is still compatible with current versions of GIFT (at the time of writing the most recent version is GIFT 2014-3X, which was released in December 2014).

### 1.2 Version of the "Logic Puzzle Tutorial" course included with GIFT

The released version of the "Logic Puzzle Tutorial" is a slightly modified version of the baseline condition tutorial from the original experiment and includes the names that were used in the experimental version. In this version, the tutorial automatically has the names present in it as opposed to prompting the user to enter them as in the experiment. The released version of the tutorial course includes a subset of the questionnaires and question based knowledge assessments that the participants answered. In the full experiment, after the completion of the tutorials the participants answered multiple-choice assessments, engaged in solving an "easy" puzzle and then a "difficult" puzzle. The released version of the course only includes the "easy" puzzle. See Figure 2 for a screenshot of the "easy" puzzle that is included with GIFT releases. Unlike the tutorial portion of the course, the "easy" puzzle does not include any adaptive feedback to the participants. However, the answers that are provided by the participant are saved to an external excel file for future analysis. There are two output files of interest for the researcher: 1) output of the surveys/questionnaires that can be accessed through GIFT's Event Reporting Tool (ERT), and 2) Excel output of the puzzle which is saved in the Domain folder associated with the "Logic Puzzle Tutorial" course.

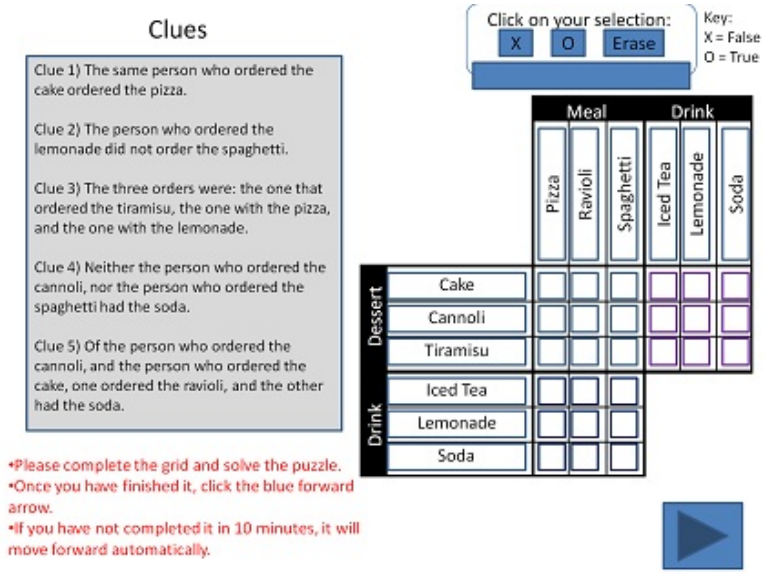



**Fig. 2.** Screenshot of the "easy" logic puzzle that is included with the "Logic Puzzle Tutorial" course in GIFT.

## 2 Tools used in Course Development: Then and Now

GIFT contains a suite of Authoring Tools that can be used for course development. The tools that were used in the development of the original "Logic Puzzle Tutorial" course/experiment were the Course Authoring Tool (CAT), Sensor Configuration Authoring Tool (SCAT), and Survey Authoring System (SAS). Since the adaptive feedback occurred within the PowerPoint tutorial, a placeholder Domain Knowledge File (DKF) was used that did not result in adaptive feedback provided directly by GIFT. As GIFT has continued to develop, many of GIFT's tools have been updated and have new functionalities in their current versions.

### 2.1 Course Authoring Tool (CAT) and GIFT Authoring Tool: Then and Now

The primary tool used for the development of the "Logic Puzzle Tutorial" was the CAT. The CAT allows the author to create a course flow that includes the order of guidance, training applications (e.g., PowerPoint), and surveys that the participant receives. Once design decisions have been made about the course and the components have been created, the CAT is where the transitions and flow of the course are specified. Figure 3 is a screenshot of the original "Logic Puzzle Tutorial" course loaded in the CAT. Note the linear structure of the elements, and the nodes that can expand to provide more detail.

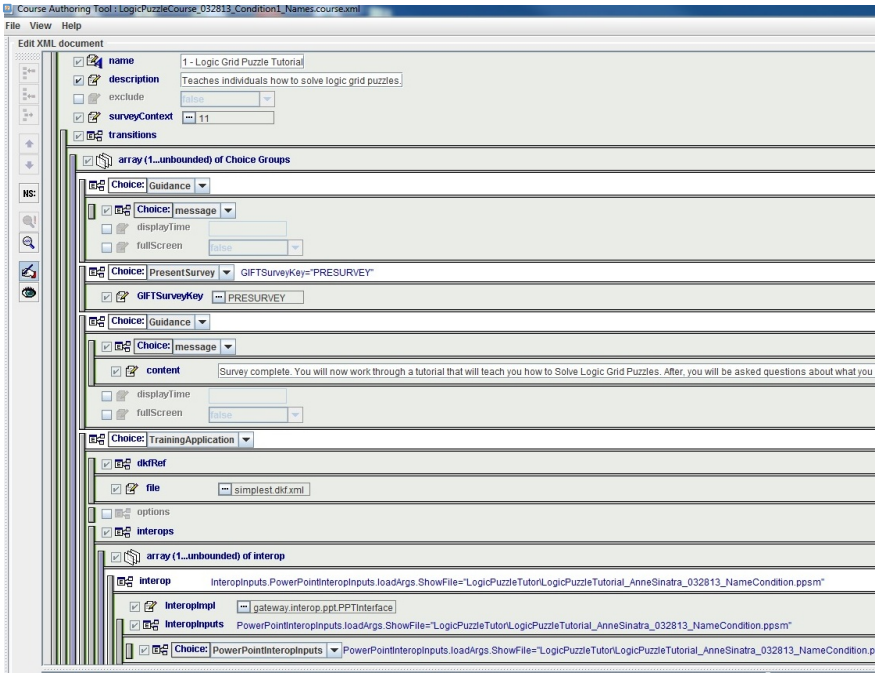



**Fig. 3.** Screenshot of the Course Authoring Tool that was used to create the "Logic Puzzle Tutorial" course.

The original XML (Extensible Markup Language) editor based CAT is still included with current releases of GIFT. However, an additional GIFT Authoring Tool (GAT) has been designed to allow an author to perform the same functionality in a more user-friendly interface. The same functionalities and course elements can be created using the GAT, but the interface is more straightforward and uses drop down menus that are closer matches for a general user's mental model than an XML editor based tool. A screenshot of the GAT with the "Logic Puzzle Tutorial" course loaded in it can be seen in Figure 4. While the redesign of this tool would not have impacted the design of the original course, it is expected that it would have led to a faster understanding of how to create the GIFT course.

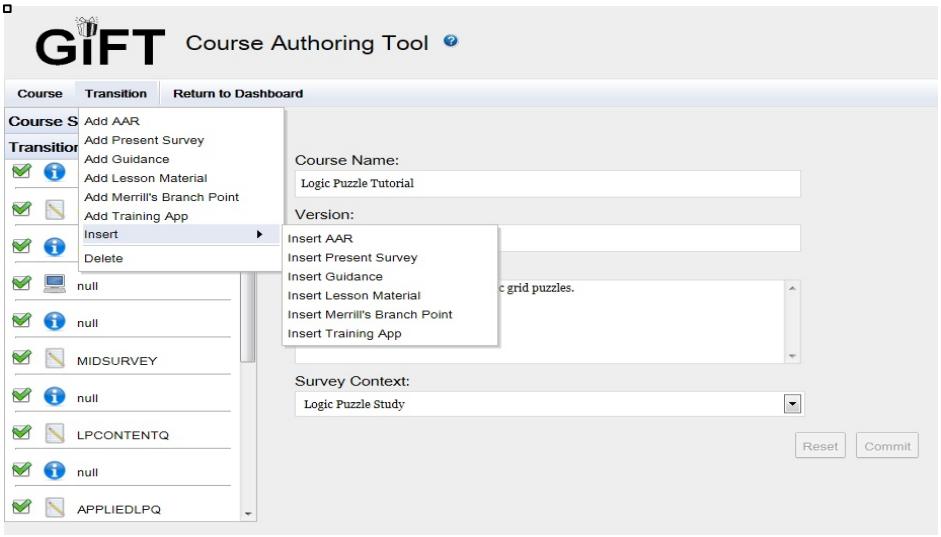



**Fig. 4.** The "Logic Puzzle Tutorial" course loaded in the GIFT Authoring Tool.

## 2.2 Survey Authoring System (SAS): Then and Now

The SAS was heavily used in the design of the "Logic Puzzle Tutorial" course. Many surveys including multiple-choice and multiple answer questions were created for use in the course. All of these surveys are available in releases of GIFT from version 3.0 to present, and many of these surveys are referenced within the "Logic Puzzle Tutorial" course. See Figure 5 for a screenshot of the SAS.

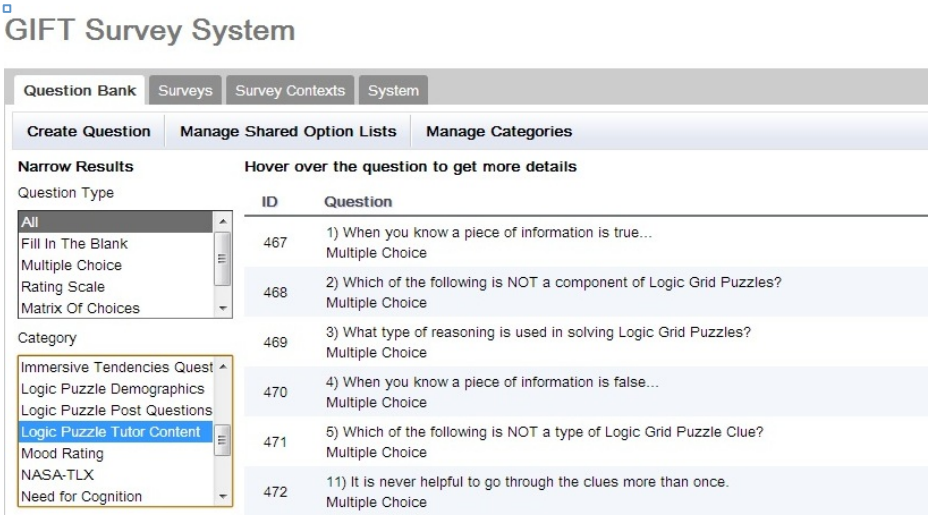



**Fig. 5.** Screenshot of GIFT's Survey Authoring System and a selection of questions associated with the "Logic Puzzle Tutorial" course.

The primary functions of the SAS have remained stable since the design of the original "Logic Puzzle Tutorial" course. However, there are now additional features that would be used. In the design of the original course the outputs of the questions were not automatically scored. Part of the reasoning behind this decision was that many of the scoring features were still in development at the time. Now the scoring features are stable and well documented in GIFT's doc files. Additionally, course examples that use scoring can now be viewed and examined by authors to understand the scoring functionality. Weights can be assigned to the answers in the creation of questions, and surveys can be scored. Additionally, with the development of the Engine for Management of Adaptive Pedagogy (EMAP), question banks can now be created that are associated with specific concepts that the learner can be assessed on. The grading of surveys can now influence remediation that the individual learner is given. The functionality provided by the EMAP may have influenced the design of the logic puzzle tutorial experiment if it was created today, and may have ultimately led to a different experimental design. See Figure 6 for a screenshot of a survey context with a question bank in the SAS that is associated with the functionality of the EMAP. The development of the EMAP has been documented in the literature, which can be referenced for further reading [6,7].

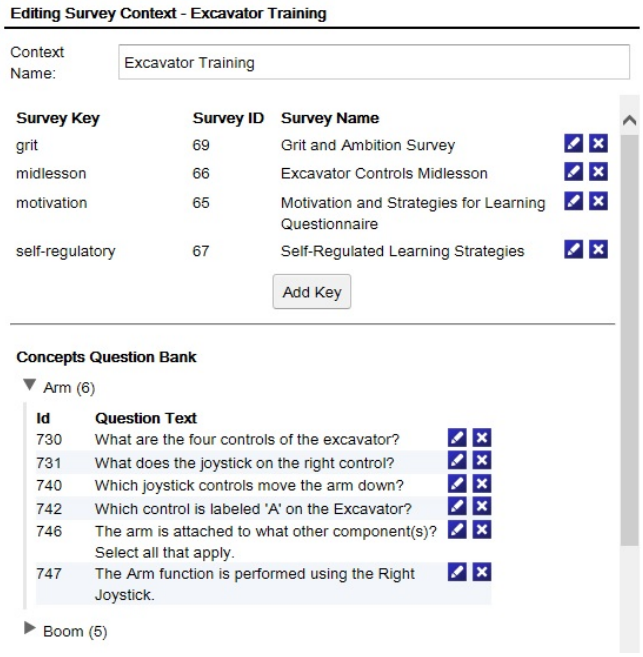



**Fig. 6.** Screenshot of a question bank associated with a Survey Context in GIFT 2014-2's SAS

### 2.3 The Sensor Authoring Tool (SCAT): Then and Now

The SCAT has remained fairly constant since the development of the "Logic Puzzle Tutorial" course. Like the CAT, it is an XML editor based tool. Default configurations for specific sensors are included with GIFT and authors can change the reference for the sensor configuration that will be used when they run GIFT. The sensor configuration is not linked directly to a course, but is used in all instances of the installation of GIFT unless it is adjusted between learners. Future versions of GIFT are expected to move toward making connections between the sensor configuration and the specific course that has been designed and run.

## 3 The Future

GIFT has gone through many iterations through the years, and at each point has added additional functionality and features. More additions and adjustments are expected as GIFT moves forward and in new directions, such as the cloud. One of the current goals of GIFT is to improve usability, which will make current and future features more understandable to all GIFT users. The "Logic Puzzle Tutorial" course which exists in GIFT is an example of using GIFT for an experiment. While it does not include adaptive elements based in GIFT, it offers a demonstration of how GIFT can be used for a traditional psychology experiment. The features of the more recent versions

of GIFT provide more flexibility and options to individuals who will be designing experiments in the future.

## References


1. Sottilare, R. A., Holden, H. K.: Motivations for a Generalized Intelligent Framework for Tutoring (GIFT) for Authoring, Instruction, and Analysis. *AIED 2013 Workshops Proceedings Volume 7* pp. 1 - 8 (2013).
2. Goldberg, B., & Cannon-Bowers, J. (2013, July). Experimentation with the Generalized Intelligent Framework for Tutoring (GIFT): A Testbed Use Case. In *AIED 2013 Workshops Proceedings Volume 7* pp. 27 – 36 (2013).
3. Sinatra, A. M. Using GIFT to Support an Empirical Study on the Impact of the Self-Reference Effect on Learning. In *AIED 2013 Workshops Proceedings Volume 7* pp. 80 – 87 (2013).
4. Sinatra, A. M. The Research Psychologist's Guide to GIFT. In *Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym2)* pp. 86- 93 (2015).
5. Sinatra, A. M., Sims, V. K., Sottilare, R. A. *The Impact of Need for Cognition and Self-Reference on Tutoring a Deductive Reasoning Skill* (No. ARL-TR-6961). Army Research 6. Lab Aberdeen Proving Ground, MD (2014).
6. Wang-Costello, J., Tarr, R.W., Cintron, L.M., Jiang, H., & Goldberg, B. Creating an Advanced Pedagogical Model to Improve Intelligent Tutoring Technologies. In *The Interservice/Industry Training, Simulation & Education Conference Proceedings* (2013).
7. Goldberg, B. What Makes an Effective Pedagogical Model? In ITS 2014 Workshop Proceedings: Pedagogy That Makes A Difference: Exploring Domain-Independent Principles across Instructional Management Research within the ITS Community pp. 1 - 9 (2014).

# Adaptive Course Flow and Sequencing through the Engine for Management of Adaptive Pedagogy (EMAP)

Benjamin Goldberg[1] and Michael Hoffman[2]

[1]U.S. Army Research Laboratory, Orlando, FL
[2]Dignitas Technologies, Inc., Orlando, FL
benjamin.s.goldberg.civ@mail.mil;
mhoffman@dignitastechnologies.com

**Abstract.** The Engine for Management of Adaptive Pedagogy (EMAP) is the Generalized Intelligent Framework for Tutoring's (GIFT) first implementation of a domain-independent pedagogical manager. It establishes a framework within GIFT that adheres to sound instructional system design, while also providing tools and methods to create highly personalized and adaptive learning experiences. In this paper, we present the components of the EMAP, we highlight their utility when authoring an EMAP managed lesson, and we review the limitations associated with its first instantiation.

**Keywords:** Adaptive Instruction; Pedagogical Model; Instructional Management; Engine for Management of Adaptive Pedagogy; Generalized Intelligent Framework for Tutoring

## 1 Introduction

The Generalized Intelligent Framework for Tutoring (GIFT) is being developed as a domain-agnostic solution to authoring, delivering, and evaluating adaptive training solutions across an array of domains and training applications. While GIFT's initial development focused on establishing a standardized architecture for building Intelligent Tutoring System (ITS) functions to support distributed learning events, recent work has centered on extending the adaptive capabilities the framework affords. As a result, the Engine for Management of Adaptive Pedagogy (EMAP) was developed. The EMAP is based on an extensive literature review of instructional strategy focused research within computer-based training [3], and organizes its findings in a domain-independent fashion. At the moment, there are papers that highlight the literature and theory that fed the EMAPs design [3, 4] and that highlight the authoring tools and processes required for implementing its functions [5], but there is nothing that reviews EMAP interactions from the learner's perspective as it relates to event sequencing. In this paper, we present a usecase of a GIFT lesson managed by the EMAP and we review the various architectural components that make it run. We will first highlight the work that went into formalizing the EMAP, the dependencies the EMAP has with

other portions of the GIFT architecture, and we present a usecase of lesson interaction and transitions managed by EMAP logic and configurations.

## 2 Formalizing the EMAP

The EMAP design was the resulting outcome of a collaborative project between the U.S. Army Research Laboratory (ARL) and the Institute for Simulation and Training (IST) at the University of Central Florida. Following an extensive literature review, the team selected David Merrill's Component Display Theory (CDT) as the theoretical framework to structure EMAP requirements around [3,5].

The CDT was conceptually integrated within GIFT as a domain-agnostic framework used for course construction and building guidance/remediation configurations [3]. This requires linking learner relevant information with generalized descriptors of learning content and instructional techniques, strategies and tactics. These relationships were used to establish an initial decision tree that informed real-time adaptations.

It is important to highlight the current attributes represented in a GIFT learner model and their relationship with metadata used to describe learning content. As these variables moderate EMAP configurations that are set and adapted at run-time, it is important to review how each level of data operates and what decisions they inform. For learner model data forms, these include determinations for knowledge states, skill states, affective states, and individualized traits that have been empirically found to impact learning and retention.

### 2.1 Learner Model Dependencies

The EMAP uses pedagogical configurations that are moderated by attributes being tracked in GIFT's learner model. These configurations are coupled to the customized value ranges of available variables supported within the architecture's standardized schema. The configurations implemented in the EMAP are based on both historical and real-time inferences across the various trait and state attribute spaces. As such, the EMAP uses information on prior knowledge along with a set of trait characteristics to personalize lesson materials across the CDT's four quadrants (i.e., Rules, Examples, Recall, and Practice) upfront, and then uses real-time assessment information on knowledge, skill, and affective states to moderate guidance, remediation, and problem selection. The goal is to establish generalized configurations that can translate across different domain spaces and varying training platforms and applications.

For knowledge and skill states, performance is monitored at an objective level. In the latest release, GIFT tracks individual learners across a hierarchy of concepts as they relate to a set of tasks within a specified domain. These concepts are established in the Domain Knowledge File (DKF), where bottom level sub-concepts (i.e., leaf nodes) are assessed against data made available by the training application itself. For each concept and set of sub-concepts, there are currently four possible state determinations: (1) above-expectation, (2) at-expectation, (3) below-expectation, and (4) unknown. Each of these representations can be associated with either a knowledge

state or skill state, where this division is used to differentiate 'knowledge' from 'ability to execute'. This falls in line with the mention of Knowledge/Skills/Abilities (KSAs) defined in most doctrine and helps to make competency badging within a domain more granular. Inference procedures are performed across all concepts to determine a competency level for the domain of instruction, with values being entered as Novice, Journeyman, or Expert.

Variables based on traits found to impact learning are of importance to the EMAP. The individual traits of a learner are believed to be more stable over time and are used to set initial configurations of a lesson based on these associations. Current EMAP logic informed by traits includes motivation, self-regulatory ability, and grit. These items are not inherently tracked in the DKF, but they are used offline to configure lesson materials and sequencing when a lesson is initialized.

In terms of affect represented within GIFT's learner model, these state spaces associate primarily with data made available through sensor technologies that monitor both physiological and behavioral data sources. Affective states of interest include engagement, frustration, boredom, confusion, etc. Regardless of the state space, GIFT is very flexible with respect to affective modeling, as the researcher and/or training developer has the ability to configure what variables to track and what classifiers to apply. These classifiers are used to produce a state determination that is represented in GIFT's DKF across short-term, long-term, and predicted values. For adaptation purposes, much of the affect related information is used to adapt instruction during runtime, as this form of assessment provides insight into a learner's reactive tendencies to an event or interaction.

### 2.2 Metadata Dependencies

Learner model attributes are linked with generic content descriptors that the EMAP is designed to act on. This metadata is used to take domain-independent representations of pedagogical practice and associate it with domain-specific content. The metadata currently in use is based on the Learning Object Metadata (LOM [6]) standard put in place by the Institute for Electrical and Electronics Engineers (IEEE). This provides a set of high level categories (e.g., interactivity type, difficulty, skill level, coverage, etc.) and value ranges (i.e., skill level is broken down into novice, journeyman, and expert) that inform characteristics for a type of interaction. GIFT uses two authoring processes to build the EMAP linkages. First, a lesson developer needs to build metadata files for all associated content and practice materials. Next, the lesson developer must establish what learner model attributes moderate metadata selection, and what value ranges serve as strategy selection thresholds.

### 2.3 EMAP Course Flow Example

The following use case represents the interaction of GIFT transitions across lesson elements and materials. Each event is described in relation to the EMAP and the type of data that informs its application. The usecase is broken down by learner login and course selection; pre-lesson learner model updates and assessments; adaptive lesson

delivery via a Merrill's branching; and After Action Review (AAR) and lesson completion.

**Learner Login and Course Selection.** When a learner interacts with GIFT to initialize a course or lesson, they are first required to login using associated IDs and passwords. Once logged in, the first function GIFT performs is checking for long-term learner model information, such as records of prior training events and any persistent trait variables being stored over time (this latter function is currently being developed). Presently, all prior training events are stored under experience Application Programming Interface (xAPI) specifications within a designated Learner Record Store (LRS) [1]. Out of the box GIFT isn't configured to use an LRS, just the SQL database we have been using for years. However the GIFT in the cloud instance will be configured to use the ADL LRS (but even that clears data out every day or so). No matter if the data is stored in either place, GIFT makes use of that information. Information related to courses taken along with performance outcomes on a concept by concept level are communicated. This information is used to recommend courses based on if any prior training events resulted in below-expectation outcomes. This is the current role xAPI plays in this process. We expect this capability to become more robust over time. Following this update, a learner is then able to select a course from GIFT's Tutor User Interface (TUI). Following this update, a learner has the ability to select their course and progress into the first transitions of a lesson.

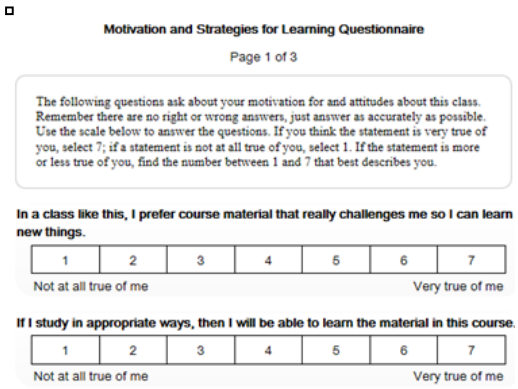



**Fig. 1.** GIFT Survey Interface

**Pre-Lesson Updates and Assessments.** Upon course initialization, GIFT references the EMAPs pedagogical configuration file to determine the trait-based variables that moderates adaptations to the lesson structure. In the current baseline, these variables include motivation, prior knowledge, self-regulatory ability, and grit. Other variables such as skill and goal-orientation can also be applied, which is the current case when a learner enters a practice quadrant of the CDT. A lesson developer has the ability to select which variables to moderate their lesson adaptations around, which impacts the first transitions experienced by a user in a new lesson. GIFT will first check an indi-

vidual's persistent long-term learner model to identify any existing data. If no record is located, GIFT will administer an available survey to collect that information. This interaction is authored in GIFT's Survey Authoring System and is presented directly to the learner on the TUI (see Figure 1). Scoring rules are associated with all administered instruments, which are then used to update learner model attribute values in real-time.

GIFT then establishes learner knowledge and skill states based on associated xAPI data that exists for that domain. If no data is available, then knowledge and skill attributes are set to 'Novice'. Next, if a lesson pre-knowledge assessment is made available by the lesson developer, then the test is presented to the learner through GIFT's TUI. Based on established scoring conditions for that assessment, the learner model is updated accordingly to reflect new predicted competency levels. This information is used to bypass lesson materials on concepts that the learner has exhibited expert understanding of. Bypassing concepts is dependent on the separation of concepts not only in how they are sequenced in the course.xml but also in the content presented. i.e. if there is only 1 piece of content that covers A+B, how can either one be skipped and not the other?

**Adaptive Lesson Delivery via Merrill's Branching.** Once all trait-based information has been established in the learner model and all pre-test assessments have been administered, a learner is then progressed into the adaptive lesson deliver through a set of pre-defined Merrill's Branching points. This entails customized sequencing through the CDT quadrants. This interaction will be outlined through the following collection of bullet points.

- *Rules and Examples Quadrants*: Configure material around defined concepts being instructed and known attributes of the learner that match entries within the EMAP's decision tree
  - Attributes
    - Knowledge; Motivation; Self-Regulatory Ability; Grit
  - Proposed Assessments
    - Affective State: monitor learner to assess emotional and cognitive reactions
    - Behavior: monitor behavior within learning environment to assess gaming behaviors
  - No knowledge/skill updates in learner model will occur within these quadrants
- *Recall Quadrant* (Knowledge Assessment):
  - If a bank of questions for this concept has been authored within the SAS, then deliver randomized recall assessment based on EMAP configuration (configuration is defined within GIFT's Course Authoring Tool; see Goldberg et al., 2015)
    - If established scoring conditions exist, then update learner model based on assessment outcomes
      - Assumption: Only cognitive knowledge is updated based on performance outcomes within a survey delivered assessment within the recall quadrant
  - Guidance Configuration (currently being developed)

- Use known attributes of the learner to configure timing and specificity dimensions
  - Question by Question Feedback vs. Following All Items
    - Attributes that may dictate this decision: Knowledge and Self-Regulatory Ability
  - General to Specific vs. Specific to General Feedback
    - Attributes that may dictate this decision: Knowledge and Grit

— Remediation
- If learner is reported at 'below expectation'/'at expectation' on any items (i.e. concepts), then initiate remediation loop within the defined Merrill Branch
  - Remediation path is dependent on reported cognitive knowledge state based on defined scoring logic in the Course Authoring Tool
    - For each concept:
      - If learner is scored at 'below expectation' based on scoring configuration, select that concept for Rule quadrant remediation
      - If learner is scored at 'at expectation' based on scoring configuration, select that concept for Example quadrant remediation (can be in addition to Rule quadrant remediation)
    - If there is any concept remediation needed, present the Rule remediation for all identified concepts followed by Example remediation.
      - This is where the metadata selection algorithm is used to select different content to deliver to the learner (if available).
- Remediation ends back in Recall Quadrant
  - If items report at 'below expectation' again and there is no new content to present; then allow the learner to select the quadrant they prefer to remediate in (currently being developed).

— If all items in the Recall Assessment are reported at 'above-expectation' then move onto Practice.

• If no questions exist for the concepts within the SAS or the author removed the recall quadrant from the branch, then move onto Practice (not currently supported).

• *Practice Quadrant* (Skill Assessment):

— If no practice has been authored/configured, and the Recall Quadrant has been satisfied, then move onto next transition in the course file

— If a training environment/scenario has been configured, then deliver practice materials through pre-established Gateway and DKF

— Configure material around known attributes of the learner that match entries within the EMAP's decision tree (to be developed)
- Attributes
  - Skill; Motivation; Self-Regulatory Ability; Grit; Goal-Orientation
- Proposed Assessments
  - Affective State: monitor learner to assess emotional and cognitive reaction
  - Behavior: monitor learning environment to assess gaming behaviors
  - Skill: monitor performance in real-time across all identified sub-concepts based on pre-defined assessments authored around Evidence Centered Design (Stealth Assessment; [2])

        - Using established scoring conditions, update learner model based on assessment outcomes
        - Assumption: Only cognitive skill is updated based on performance outcomes within a practice environment
        - A survey authored in the SAS can also be defined as a practice environment (currently being developed).
- Guidance Configuration (currently being developed)
    - Use known attributes of the learner to configure timing and specificity dimensions
        - Number of violations before triggering guidance/feedback
            - Attributes that may dictate this decision: Skill and Self-Regulatory Ability
        - General to Specific vs. Specific to General Feedback
            - Attributes that may dictate this decision: Skill and Grit
        - Static (text or audio alone) vs. interactive (AutoTutor reflection)
- Remediation
    - If learner is reported at 'below expectation'/'at expectation' on any items, then initiate remediation loop within the defined Merrill Branch
        - Remediation path is dependent on a combination of skill and knowledge
            - If learner is novice in skill and expert in knowledge, then re-initialize practice
            - If learner is novice in skill and journeyman in knowledge, then navigate to examples quadrant
    - Remediation ends back in Recall Quadrant (currently being developed)
        - If items report at 'below expectation' again and there is no new content, then allow the learner to select the quadrant they prefer to remediate in
- If all items in the Practice Assessment are reported at 'above-expectation' then move onto next transition in the course file

This sequence of interaction will occur for all identified Merrill's Branching points authored. For instance, in a lesson that instructs across four concepts, an author can decide to break up the material across two branching points. Regardless of the number of Merrill's Branching points, once all exit criteria has been reached, then the lesson transitions into post-test assessments, after-action review and lesson completion.

**Post-Lesson Assessment, After Action Review, and Lesson Completion.** Upon completion of all adaptive lesson transitions across the designated Merrill's Branch points, a course developer will have the ability to administer a post-knowledge and/or post-skill assessment as a means for determining overall competency levels following lesson interventions. These interactions are intended to be void of guidance functions to determine how learners perform on their own. The outcomes are used to establish final score and attribute values for a lesson, with future development offering extended remediation events.

Assessment exercises are followed by a GIFT managed AAR used for reflective and summarization practices. It is during this interaction that a student is directed to

reflect on the experience of the instructional event and their resulting performance outcomes. GIFT's current AAR capability is a web-page that reviews the objectives and concepts of a lesson taken, along with recorded performance measures for all items. A goal is to provide an interactive AAR function that utilizes technology to engage a learner in reflective exercises. Following execution of the AAR transition, the EMAP managed GIFT course is complete. At this instance, GIFT communicated xAPI data for the purposes of updating the LRS with outcomes values of knowledge and skill attributes for all concepts and sub-concepts scored. The learner is then given the option to logout of the system, or to select a new course or lesson to complete.

## 3 Conclusion

In this paper we presented a use case of a conceptual course flow for a GIFT lesson managed by the EMAP. We highlighted architectural dependencies associated with building out an EMAP lesson and we reviewed logic associated with lesson transitions. This paper highlights the EMAP's function at the lesson level, where you can see the various decisions being made and the type of data informing its strategy selection. Enhancements to the EMAP continue, with current developmental plans looking at personalized feedback delivery options. In addition, the authoring process is being converted to web-based interfaces. For an overview of the current authoring process and to see the underlying features of the tools and methods put in place to support a pedagogical model like the EMAP, see [5] for a nice breakdown.

## References


1. Poeppelman, T., Hruska, M., Long, R., Amburn, C. *Interoperable Performance Assessment for Individuals and Teams Using Experience API.* Paper presented at the Second Annual GIFT Users Symposium, Orlando, FL. 2015.
2. Shute, V. J., Kim, Y. J. Formative and stealth assessment *Handbook of research on educational communications and technology* (pp. 311-321): Springer. 2014.
3. Wang-Costello, J., Goldberg, B., Tarr, R. W., Cintron, L. M., Jiang, H. *Creating an Advanced Pedagogical Model to Improve Intelligent Tutoring Technologies.* Paper presented at the The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC). Orlando, FL. 2013.
4. Goldberg, B., Brawner, K. W., Sottilare, R., Tarr, R., Billings, D. R., Malone, N. *Use of Evidence-based Strategies to Enhance the Extensibility of Adaptive Tutoring Technologies.* Paper presented at the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) 2012, Orlando, FL. 2012.
5. Goldberg, B., Hoffman, M., Tarr, R. Authoring Instructional Management Logic in GIFT Using the Engine for Management of Adaptive Pedagogy (EMAP). In R. Sottilare, A. Graesser, X. Hu & K. Brawner (Eds.), *Design Recommendations for Intelligent Tutoring Systems: Authoring Tools (Volume 3)*: U.S. Army Research Laboratory. 2015.
6. Mitchell, J. L., Farha, N. Learning Object Metadata: Use and Discovery. In K. Harman & A. Koohang (Eds.), *Learning Objects: Standards, Metadata, Repositories, and LCMS* (pp. 1-40). Santa Rosa, CA: Informing Science Press. 2007.

# Using Social Media with GIFT to Crowd-source and Enhance Learning Content

Irene T. Boland, PhD[1], Rodney A. Long[2], Ben Farmer, PhD[3], Doug Raum[1], Dan Silverglate[1], Ed Sims, PhD[1]

[1]Vcom3D, Inc., 12124 High Tech Ave., Suite 150, Orlando, FL, USA 32817
[2]Army Research Laboratory (ARL), Human Research and Engineering Directorate (HRED), Simulation and Training Technology Center (STTC), 12423 Research Parkway, Orlando, FL 32826
[3]Defense Equal Opportunity Management Institute (DEOMI). Patrick AFB, Florida
{ireneb, dougr, dans, eds}@vcom3d.com
rodney.a.long3.civ@mail.mil
benjamin.farmer.ctr@us.af.mil

**Abstract.** The US Army recognizes that enterprises that excel at incorporating their latest learning into the mainstream processes of their operations are able to capture and maintain a competitive edge. Among the goals of the Army Learning Concept 2015 is enabling all soldiers to participate in the creation and updating of training without increasing the workload of instructors. In addition to the Generalized Intelligent Framework for Tutoring (GIFT), the Army Research Laboratory (ARL) has funded a Social Media Framework (SMF) that enables an organization to crowd-source and crowd-vet new content and improvements to existing courses. The research questions we seek to answer in our current research include the extent to which the SMF and GIFT can: (a) promote critical thinking, collaboration, adaptability, effective communication, and problem solving; (b) help close the gap between formal training and operational application of the training to missions in the field; (c) reduce the time required to locate and use learning resources; (d) reduce the time required to incorporate feedback from the field into formal instruction; and (e) reduce instructor workload, while maximizing the efficacy of the instructor's time.

**Keywords:** Social media, GIFT, crowd-sourcing, usability, instructional systems design

## 1 Introduction

The US Army trains and educates over a half million individuals per year in a course-based, throughput-oriented system. Much of the Army's web-based instruction is in

the form of static PowerPoint presentations, with little tailoring to individual soldier needs. With the ever-changing landscape of full spectrum operations, today's soldiers are facing ill-structured problems and have little time for the ideal levels of reflection and repetition needed to promote critical thinking, adaptability, and mastery of complex skills. Additionally, the current time frame for updating courses (3 to 5 years) is not supporting the modern Army's fast-paced learning needs.

During the Vcom3D demonstration of GIFT at the 17th International Conference on Artificial Intelligence in Education (AIED), attendees will experience how the breadth and depth of knowledge spread throughout an organization can be harnessed and rapidly incorporated into training for the benefit of those who need to know promptly. In the role of a learner, participants will experience and provide granular feedback on an adaptive course in our web-based GIFT environment. Then participants will discuss and vote on the relevance or accuracy of the content to enable refinement before an instructor reviews it for inclusion in training.

## 2 Background: Social Media Framework

Previously, we investigated a research-based suite of affordances that support the sharing and vetting of information amongst peers. The objectives of the project were to identify lessons learned from: commercial, academic, and US Government applications of social media to knowledge management and learning; and to consider the unique requirements and constraints of the military learning environment and how successful commercial and academic models for learning can be adapted to military applications.

## 3 Current Research

### 3.1 Research Objectives

At a high level, our research aims to investigate the extent to which the integrated SMF and GIFT system can:

- Promote critical thinking, collaboration, adaptability, effective communication, and problem solving,
- Help close the gap between formal training and operational application of the training to missions in the field,
- Reduce the time required to locate and use learning resources,
- Reduce the time required to incorporate feedback from the field into formal instruction,
- Reduce instructor workload, while maximizing the efficacy of the instructor's time.

### 3.2 Experimental Methodology

This research project follows a sequence of overlapping/spiral events, including: Literature Review (ensuring that our proposed research furthers the body of knowledge), Experiential Review (hands-on examination of existing, to ensure that the affordances we test are extending the state of the art), Test Bed Development (creating the suite of affordances to enable testing of our research hypotheses), and Quantitative and Qualitative Research (testing our hypotheses and soliciting feedback from participants).

### 3.3 Test Bed Architecture

Expanding on the existing SMF, a cloud-based 'headless' instance of the GIFT platform has been created, allowing multiple users to connect to GIFT across the internet.

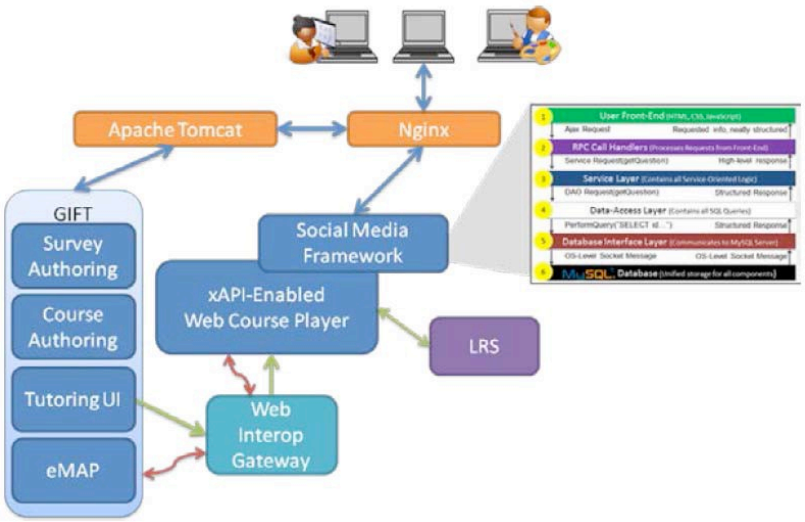



**Fig. 1.** SMF/GIFT Integrated Architecture

The GIFT platform has been extended to include a gateway interoperability module allowing for connection to a web-based course player. The course player, built on a PHP/MySQL platform and using a responsive front-end (suitable for expansion to mobile devices), will play (experience API) xAPI-wrapped course content. Through the gateway interoperability module, the course player will also communicate to the GIFT engine for Management of Adaptive Pedagogy (eMAP), allowing adaptivity within the course driven by GIFT's advanced adaptive capabilities. The course player also generates xAPI statements which are stored in a Learning Record Store (LRS) and usable for learning analytics.

The web-based course player includes the ability for courses to collect social media feedback on granular aspects of the course: paragraphs of text, images, videos, etc.

Using annotation-style commenting, the social feedback is collected and stored within the SMF for crowd-comment and review after the course is completed. In addition, the GIFT tutoring user interface (UI) has been modified to allow other GIFT transitions (surveys, learning materials, after action reviews) to collect social feedback in a similar manner. This feedback, too, will be available within the SMF for crowd-comment and interaction.

### 3.4 Experimental Research

Vcom3D research for the ARL in Social Media-enabled Learning and Knowledge Management has three major phases in 2014-2015, each with a data collection. The recently completed (February 2015) data collection 1 focused on having an Instructional Systems Designer (ISD) and SMEs use a Learning Content Management System (LCMS) to enter content and build a course. The research test bed is a combination of three government-sponsored systems: SMF, GIFT, and an LCMS.

The second data collection (Summer 2015) will involve learners taking the course and providing granular feedback about how they think the course can be improved as well as using social media tools to discuss the feedback of others. Then, in data collection three (Fall 2015), the ISD and SMEs who built the course will review the feedback from learners and decide what improvements they will make to the course. This three-part research demonstrates the speed with which experts in the field and fleet can provide real-world feedback that is then promptly incorporated into the official doctrine course by the schoolhouse. This addresses key goals of the Army Learning Model (ALM) which seeks, among many other goals, to include the ever-evolving knowledge of the field and fleet into the official training as quickly as possible.

**Data Collection 1 Procedure.** Expanding on the existing SMF, a 'headless' instance of ARL's GIFT platform was created, allowing it to run independently of a specific workstation. Utilizing this, we deployed the GIFT Survey Authoring System (SAS) and GIFT (CAT) Course Authoring Tools through our existing Apache Tomcat web application server. Using nginx to serve the existing SMF and act as a proxy to the GIFT instance on the same server gave the participants the experience of a seamless, consolidated system with Single Sign On (SSO) for each subsystem. The experimental test bed was hosted on a dedicated server off site from the research location. Each participant received login credentials and used a separate work station in their lab to access the test bed though the internet from a standard browser.

The researchers guided participants through standard tasks involved in creating learning content. The session was videotaped to allow for detailed analysis afterward. We described the system to our participants as an experimental learning content authoring system the Army has asked us to build and test. We explained that our long-term goal is to grow the system into a powerful tool that is useful to them (and other users) in creating adaptive learning experiences that are easy to update. Having their formative feedback at this early stage will enable us to develop it in the direction that's most useful to users.

We designed their data collection experience to simulate a collaboration to create the course. So, each participant was asked to create a different scenario and then we had them work together to tie it all into a complete course.

**Data Collection 1 Results.** Each of our recommendations has its basis in the time-tested and research-proven principles of UI and User Experience (UX) professions. Our recommendations are meant to help move GIFT closer to its goal of being useful to SMEs who want to author effective courses on their own. The Nielsen/Norman Group of UI/UX professionals defines useful as the result of usability and utility. Utility speaks to the extent that the system has the features the user wants and needs. Usability can be described as having 5 criteria: (1) easy to learn to use, (2) user can complete tasks quickly, (3) user can remember how to use it after being away from it for a while, (4) errors the user makes are few and easily rectified, and (5) the system is enjoyable to use.

*Recommendation 1: Sell the utility, immediately.* Users found that the system contained a large number of steps compared to other systems they had used to build adaptive training or surveys. Some of those steps were unclear in meaning or purpose. The naming conventions used are not consistent with what SMEs would name the features, buttons, and other controls. As a result, they expended a great deal of mental effort (cognitive tolls) to work in the system. Although the researchers explained the long term purpose of the system (to creative adaptive training suited to each individual), the perceived benefits of the system were not sufficient to motivate the users to want to continue using the system in its current state. For all of these reasons, we recommend an early intervention of Selling the utility – making the benefits of the system so clear that new users will be motivated to expend the needed effort to understand and master the system.

We recommend the system provide a short but impactful explainer video that helps users understand the system and what's in it for them. Specific questions that should be answered include: (a) What is Adaptive Learning? (b) Why should I use Adaptive Learning with my learners? (c) What is GIFT? And, why is it better than my other options? (d) How have others similar to me used it (compelling real success stories/visuals)? and (e) How do I use GIFT to create Adaptive Learning?

The military has a long-standing tradition of rigorous ISD which follows a standard ADDIE model (analysis, design, development, implementation, evaluation) of activities. We can reasonably expect a SME to have extensive knowledge of the content being taught. Based on their experience, they may also bring knowledge of the audience (having been a trainer) and the related organizational goals that lead to the SME being asked to share their knowledge. However, there are significant knowledge gaps in ISD for most SMEs. To achieve the long term goal of an independent SME creating effective training, the system must provide the education and support needed by the SME.

*Recommendation 2: Use the process and vocabulary native to the SME.* The current process flow and vocabulary used in the system is not reflective of how most SMEs

think or work. As a result, they are burning significant brain power simply trying to understand the system rather than feeling the reinforcement of accomplishing their goals. To illustrate both of these concepts, we examined a short process – Adding a question to an assessment – as SMEs are accustomed to doing it compared to how SMEs attempt to do it in GIFT.

For this very short sub-process of the larger course creation process, we can compare the expected versus experienced using the scorecard shown in Table 1.

**Table 1.** Cognitive Load Comparison

| ***Measure*** | ***GIFT Experience*** | ***Usual SME Experience*** |
|---|---|---|
| **Steps** | 20 (steps 7-9 repeat 3X) | 9 or less* |
| **Cognitive Load** | High | Low |
| **Time** | Slow | Medium |
| **Other** | Process incomplete. Feedback to be added using additional steps, time and cognitive load in another part of GIFT. | * Ability to upload can make process even shorter. |

*Recommendation 3: Incorporate extensive, yet lean, on-demand contextual support for SME.* We recommend two approaches to providing support to SMEs. First, provide them some fast and simple support when they first arrive. This help should display automatically the first time the user experiences a screen. Afterward, it should be available for the user to display on demand).

Second, offer mouseover-based help for each control, vocabulary term or other element that the SME might not be familiar with. The example in Figure 2 shows that a vocabulary improvement has been made – changing the word Transition to Content, and then providing a mouseover that explains what particular types of content are and alerting the user if they will need to use another part of the system to create that content before trying to use it here.

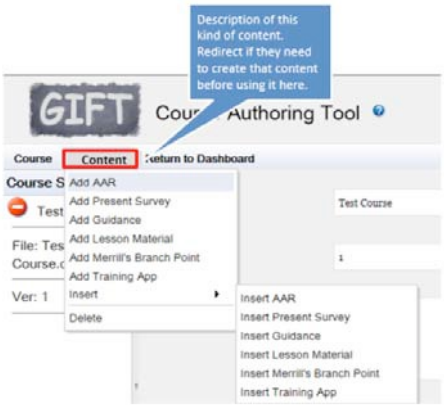



**Fig. 2.** Mouseover Help Example

**Data Collection 2 Procedure.** The SMF will be expanded to include course topics and the actual courses in the training tab. Once launched, courses will be played through the GIFT framework. In GIFT, a course is a series of transitions which might include Surveys, Learning Materials, and Training Applications. To enable a Training Application to play lessons comprised of web-based content, we will implement a new gateway interoperability module. Unlike standard web-based lessons, however, any element of the content can be selected and commented upon. Showing those comments in close proximity to the lesson content could negatively impact the flow of the course for future learners; so instead, the comments will automatically appear as a new conversation thread under the feedback tab of the containing topic page for this course. We will add similar social media commenting capability to other GIFT transitions such as Surveys and Learning Materials. The course material will be furnished by DEOMI ISDs and will be selected for its relevance to the target student participants for specific use in the experimental research. The content will then be prepared for playback by the web-based lesson Training Application and other GIFT transitions.

During the data collection event, multiple sessions of approximately 20 student participants each will access the experimental test bed from work stations in their lab through the internet from a standard browser and using credentials provided by the researchers. Participants will be asked to navigate to a particular topic and take the course associated with that topic. Participants will be encouraged to generate questions or feedback on any content they encounter. After completion of the course, participants can review their comments on the topic page and also see the comments of other participants. They will be able to up vote and down vote the questions, answers, and feedback generated by others as well as contribute to the discussions. Participants in subsequent sessions will the accumulated contributions of all preceding participants. At the end of each session, the participants will complete a survey to provide feedback of their experience.

**Data Collection 3 Procedure.** The third phase of research will explore techniques and algorithms for analyzing the user-created content, surfacing the most relevant comments and activity and connecting them to the most relevant stakeholder. For this data collection with content authors and content owners, the user management section of the SMF will evolve to display a user digest specific to each user and their role in the system. An activity section will highlight the latest contributions by the user. Back-end data analytics will look at factors such as up votes, down votes, and general activity to prioritize the contributions of others relevant to this user. The goal is to highlight trending and actionable issues pertaining to course content owned by this user. Participants will then evaluate the efficacy of the system in surfacing errors, identifying gaps, suggesting content, and reducing ISD work-load.

## 4 Implications for Future Research

At the end of the third phase of the current research, we will have investigated the efficacy of crowd-sourced and crowd-vetted content for applying field knowledge to improve learning content, while reducing instructor workload and turn-around time. However, we believe that social media can provide additional benefits to the learning environment, and to GIFT in particular, by (1) harnessing crowd inputs for the creation and refinement of a Domain Model, or the body of knowledge for a topic and (2) mining social media data to enhance an individual's Learner Profile (or personal history of learning, demographics, and achievements). We have also identified the need to make the user experience more intuitive to its intended end-users (SMEs). At the end of the current research, we will make recommendations for these additional means for applying social media to the integrated learning environment.

Additional areas of research we intend to explore include: (1) harnessing crowd inputs into the creation and refinement of a domain model, or the body of knowledge for a topic, (2) mining social media data to enhance an individual's Learner Profile (or personal history of learning, demographics, and achievements), and (3) developing the user experience to be immediately intuitive to its intended end-users (fielded sub- ject matter experts).

## References


1. Advanced Distributed Learning (ADL). Experience API: Research Summary. Retrieved May 13, 2013 from http://www.adlnet.gov/tla/experience-api
2. Nielsen, J. Usability 101: Introduction to Usability Available. 2012: http://www.nngroup.com/articles/usability-101-introduction-to-usability/
3. US Army TRADOC (2011). The US Army Learning Concept for 2015. Retrieved November 15, 2012 from http://www.tradoc.army.mil/tpubs/pams/tp525-8-2.pdf

# NewtonianTalk: Integration of Physics Playground and AutoTutor using GIFT

Matthew Ventura[1], Xiangen Hu[2,3], Benjamin D. Nye[2], Weinan Zhao[4]

[1]Empirical Games, Tallahassee, FL 32312
[2]University of Memphis, Memphis, TN 38152
[3]Central China Normal University, Wuhan, China. P.R.C.
[4]Florida State University, Tallahassee, FL, 32312
matthewventura@empiricalgames.org; {xhu,bdnye}@memphis.edu;
weinan.zhao@gmail.com

**Abstract.** Despite the popularity of games, there has been limited peerreviewed literature published on game-based learning for science. This paper will describe a project that combined an Intelligent Tutoring System (AutoTutor) with a physics game called Physics Playground. As part of this integration we used the Generalized Intelligent Framework for Tutoring (GIFT) to manage communication between the two technologies. We will also discuss the design of a study comparing two versions of the integration. This study is taking place over Spring of 2015 and will be studying the effects of integrating different levels of tutoring into a gamebased learning system.

**Keywords:** Game-based Learning, Intelligent Tutoring Systems, Physics, Playground, AutoTutor, GIFT

## 1 Introduction

There is growing evidence of video games supporting learning (e.g., Tobias & Fletcher, 2011; Wilson et al., 2009). Such research typically focuses on games explicitly designed for learning. However, games not explicitly designed for learning can also produce significant learning gains. In this research, we look at the potential benefits of adding intelligent tutoring into an existing game. This paper describes the design process for creating an ITS enhanced educational game called NewtonianTalk using the GIFT technology. Before we describe the integration we will briefly review the state of ITS and educational games.

## 2 Background

### 2.1 Intelligent Tutoring Systems

Intelligent tutoring systems (ITS) have proven very effective in improving training outcomes. Meta-analyses show effect sizes on the order of one sigma (Dodds &

Fletcher, 2004; VanLehn, 2011), which is approximately a full letter grade in traditional grading schemes. The long sought-after goal is a 2σ effect size (Bloom, 1984; Corbett, 2001).

Recent advances in natural language processing (NLP), semantic analysis, machine learning, and cognitive modeling have spawned ITSs with the potential to achieve this effect size (Graesser, Conley, & Olney, 2012). Although many of the current computer tutors tend to use heuristics that remain constant as they customize material for individual students, the next generation of tutors will implement more dynamic models that can infer hidden learner characteristics and recognize unanticipated behavior based on learner performance, past experiences, and lessons learned. Aside from these breakthroughs in AI, the next-generation ITSs may include game-like components that further engage the student in the learning experience.

In the research discussed here, the AutoTutor Lite ITS (ATL, Hu et al., 2009) uses an established method of engaging a learner in a natural-language tutorial dialog (Graesser, Olney, Haynes & Chipman, 2005). ATL appears as an animated "talking head" avatar at certain points during the game and engages the learner in conversation about key physics concepts.

### 2.2 Learning Support via Games

Well-designed games can be seen as vehicles for exposing players to intellectual problem solving activities (Gee, 2004). But problem solving can be frustrating, causing some learners to abandon their practice and, hence, learning. This is where the principles of game design come in: Good games can provide an engaging and authentic environment designed to keep practice meaningful and personally relevant. With simulated visualization, authentic problem solving, and instant feedback, computer games can afford a realistic framework for experimentation and situated understanding, and thus act as rich primers for active learning (Shute & Ventura, 2013).

Furthermore, within-game learning support enables learners to do more advanced activities and to engage in more advanced thinking than they could without such help. The complicated part about including learning support in games is providing support that does not disrupt engagement while learners are immersed in gameplay, and reinforcing the emerging concepts and principles that deepen learning and support transfer to other contexts.

### 2.3 Physics Playground

Research into what is called "folk" physics demonstrates that many people hold erroneous views about basic physical principles that govern the motions of objects in the world, a world in which people act and behave quite successfully (Reiner, Proffit, & Salthouse, 2005). Recognition of the problem has led to interest in the mechanisms by which physics students make the transition from folk physics to more formal physics understanding (diSessa, 1982) and to the possibility of using video games to assist in learning (Masson, Bub, & Lalonde, 2011).

The game Physics Playground (PP) was designed to help middle school students understand qualitative physics (Ploetzner, & VanLehn, 1997). We define qualitative physics as a nonverbal understanding of Newton's three laws, balance, mass, conservation of momentum, kinetic energy, and gravity. PP is a 2D sandbox game that requires the player to guide a green ball to a red balloon. The player can nudge the ball to the left and right (if the surface is flat) but the primary way to move the ball is by drawing/creating simple machines on the screen that "come to life" once the object is drawn. Everything obeys the basic rules of physics relating to gravity and Newton's three laws of motion. Using the mouse, players draw colored objects on the screen, which "come to life" as physical objects when the mouse button is released. These objects interact with the game environment according to Newtonian mechanics and can be used to move the ball. When objects interact within the game environment, they act as "agents of force" to move the ball around. The player creates simple levers, pendulums, and springboards to move the ball.

The difficulty of a puzzle was based on a number of factors including: relative location of ball to balloon, number of obstacles present, number of agents required to solve the problem, and novelty of the problem. Difficult problems provide greater weight of evidence to the estimate of a competency level than easy problems. Also, "elegant" solutions (i.e., those using a minimal number of objects) give greater weight to competency level inferences than regular solutions. Preliminary data suggest playing PP for four hours can improve qualitative physics understanding ($t(154) = 2.12$, $p < .05$) with no content instruction or other learning support (Shute, Ventura, & Kim, 2013).

## 3 Methodology: GIFT Management of ATL and PP

As education turns to more game-like ITS learning environments it is important to ensure that their learning pedagogy remain consistent with the learning sciences. To ensure a good balance between the motivating "skin" of the learning experience and the deep "muscle and skeleton" of science-based learning, it is important to adopt a general architecture of ITS learning. The GIFT framework provides such an architecture and allows the integration of independent learning technologies (Graesser, Hu, Nye & Sottilare, In Press). In this work, GIFT manages and controls data communication between ATL and PP.

While the vast majority of the components of an ITS may be made domain independent, there must always be a specific component of the architecture to deal with the problems that the instructor desires to teach. The fundamental problems of domain-dependent components are how to assess student actions, how to respond to instructional changes, how to respond to requests for immediate feedback, and an interface that supports learning (Sottilare, Goldberg, Brawner and Holden, 2012; Goldberg, Sottilare, Brawner, & Holden, 2012). The architecture designed must have built-in support for these types of instructional activities.

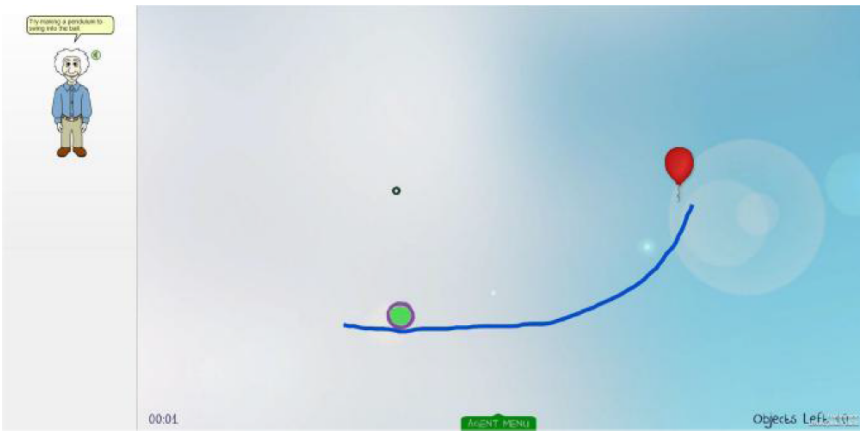



**Fig. 1.** NewtonianTalk Interface

Figure 1 displays the interface of NewtonianTalk. As can be ATL is always displayed on the left next to the PP interface. There are 3 playgrounds in NewtonianTalk. Each playground teaches a physics concept with 3 puzzles (Impulse, Conservation of Momentum, Conservation of Energy). The first design decision that needed to be made was how to most effectively introduce dialogue into PP without disrupting game play. We chose the following pedagogy styles for instruction: information delivery through ATL, scaffolded question and answer selfexplanation in ATL, and PP puzzles with support instruction. The selection of the specific activity is handled by rules specified in the GIFT system that act conditionally on information sent from the PP puzzle as the student interacts with it. Below is the introductory explanation of Impulse to the player:

> *An unbalanced force can cause an object to speed up or slow down. Specifically, an impulse is required to change the speed of an object. Impulse is the product of force times time. To change ball's speed, a springboard exerts a force for an amount of time. Pulling the springboard down further increases the ball's speed even more by applying a greater force for a longer time.*

After the player listens to further explanation as they play three PP puzzles. Figure 2 displays the puzzle for Impulse. As the springboard exerts a force up on the ball for an amount of time, it gives an impulse to the ball that changes the ball's motion. Increasing springboard's force or the time the springboard pushes up on the ball causes it to go even higher.

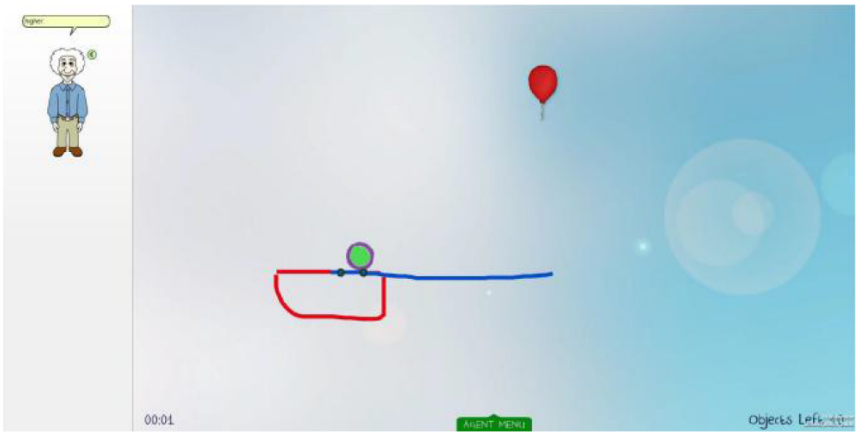



**Fig. 2.** Impulse puzzle with explanatory audio

After the player solves all 3 puzzles in the playground ATL poses a series of questions in natural language. Automated scores are calculated for the learner's performance. Below are questions for impulse:

**Q.** What is impulse? **A.** Impulse is force times time.
**Q.** How does an impulse affect an object? **A.** An impulse can change an object's speed.
**Q.** How could a force make a larger impulse? **A.** Increase the force or increase the amount of time.
**Q.** How can the same impulse be applied if the time of contact is reduced? **A.** To apply the same impulse over a smaller amount of time, the force must increase.

Once the player has answered the questions correctly or has maxed out the attempts (3 per question), the player then moves to the next playground. The player is given feedback in terms of percentages of completing the playgrounds and the ATL questions.

## 4 Discussion and Future Directions

This design process for this integration has identified some of the strengths and challenges for adding intelligent tutoring to an existing game environment that is mainly focused on simulation and experimentation. A strength of adding ITS interactions to such a game is that it allows instruction and discussion of the principles involved as they are encountered in the game (or, alternatively, fill them in when the learner struggles). Prior research on learning through exploring simulations indicates that such help may be important to learn from these activities efficiently (Graesser, Chipman, Haynes and Olney, 2005).

This approach can also be used as a model to enhance noneducational games to make them more effective for learning. For example, the game Portal 2 (despite not being learning-focused) showed significant benefits for certain types of problem solv-

ing skills (Shute, Ventura, and Ke, 2015). The current research integrates ITS into a Unity game, which is a popular engine. Such games may prove powerful learning environments with intelligent tutoring used to highlight and connect the key principles and concepts. However, the primary challenge of this work is to be able to integrate tutoring into an existing interface without being disruptive or introducing too much cognitive load.

We will be collecting data on NewtonianTalk in 2015 on an estimated 100 undergraduate psychology students. In addition to getting valuable usability data we also will test a hypothesis regarding instruction pedagogy. For this study, additional functionality is being specified that will leverage the ability of GIFT to manage and coordinate just-intime feedback based on the learner's activities during a playground. Learners' freedom to explore in a playground may increase transferability of skills, but may also result in unproductive exploration. It is hoped that GIFT support will make exploration more effective.

**Acknowledgments.** This work was supported by the Army Research Lab grant W911NF-12-2-0030 and applies the open-source GIFT system as developed by the ARL Learning in Intelligent Tutoring Environments (ARL-LITE) lab and Dignitas. However, the views of this paper represent only those of the authors.

## References

1. diSessa, A. A. (1982). Unlearning Aristotelian physics: A study of knowledgebased learning. Cognitive Science, 6, 37–75
2. Dodds, P., & Fletcher, J. D. (2004). Opportunities for new" smart" learning environments *enabled by next generation Web capabilities* (No. IDA-D-2952). Institute for Defense Analyses: Alexandria, VA..
3. Gee, J. P. (2004). What Video Games Have to Teach Us about Learning and Literacy. Palgrave Macmillan.
4. Graesser, A. C., Chipman, P., Haynes, B. C., & Olney, A. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *Education, IEEE Transactions on*, *48*(4), 612-618.
5. Graesser, A. C., Conley, M. W., & Olney, A. (2012). Intelligent tutoring systems. In K.R. Harris, S. Graham, T. Urdan, A. G. Bus, S. Major, H. L. Swanson (Eds.) *APA educational psychology handbook, Vol 3: Application to learning and teaching.*( pp. 451-473). Washington, DC, US: APA.
6. Graesser, A.C., Hu, X., Nye, B. & Sottilare, R. (In Press). Intelligent Tutoring Systems, Serious Games and the Generalized Intelligent Framework for Tutoring. In H.F. O'Neil, E.L. Baker & R.S. Perez. (Eds.), *Using games and simulation for teaching and assessment*. Routledge: Abingdon, Oxon, UK.
7. Goldberg, B., Sottilare, R., Brawner, K., & Holden, H. (2012). Adaptive Game- Based Tutoring: Mechanisms for Real-Time Feedback and Adaptation. In *International Defense & Homeland Security Simulation Workshop in Proceedings of the I3M Conference*. Vienna, Austria, September 2012.
8. Hu, X., Cai, Z., Han, L., Craig, S. D., Wang, T., & Graesser, A. C. (2009). AutoTutor Lite. In Artificial Intelligence in Education (*AIED) 2009* (pp. 802-802). IOS.

9. Masson, M. E. J., Bub, D. N., & Lalonde, C. E. (2011). Video-game training and naive reasoning about object motion. Applied Cognitive Psychology, 25, 166–173.
10. Reiner, C., Proffitt, D. R., & Salthouse, T. (2005). A psychometric approach to intuitive physics. *Psychonomic Bulletin and Review, 12*, 740–745.
11. Ploetzner, R., & VanLehn, K. (1997). The acquisition of qualitative physics knowledge during textbook-based physics training. *Cognition and Instruction, 15*(2), 169-205.
12. Shute, V. J., & Ventura, M. (2013). *Measuring and supporting learning in games: Stealth assessment*. Cambridge, MA: The MIT Press
13. 13. Shute, V. J., Ventura, M., & Ke, F. (2015). The power of play: The effects of Portal 2 and Lumosity on cognitive and noncognitive skills. *Computers & Education*, *80*, 58-67.
14. Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and learning of qualitative physics in newton's playground. *The Journal of Educational Research*, *106*(6), 423-430.
15. Sottilare, R. A., Goldberg, B. S., Brawner, K. W., & Holden, H. K. (2012, December). A modular framework to support the authoring and assessment of adaptive computer-based tutoring systems (CBTS). In *Interservice/Industry Training, Simulation, and Education Conference.(I/ITSEC) 2012*. Paper 12017 (pp. 1-13).
16. Tobias, S., & Fletcher, J. D. (Eds.). (2011). *Computer games and instruction*. IAP.
17. VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197-221.
18. Wilson, K. A., Bedwell, W. L., Lazzara, E. H., Salas, E., Burke, C. S., Estock, J. L.,... & Conkey, C. (2009). Relationships between game attributes and learning outcomes review and research proposals. *Simulation & Gaming*, *40*(2), 217-266.

# Rapid Dialogue and Branching Tutors

Keith Brawner[1]

[1]U.S. Army Research Laboratory, Orlando, FL 32826
Keith.w.brawner.civ@mail.mil

**Abstract.** The technology used as part of the Tools for Rapid Automated Development of Expert Models (TRADEM) project has been featured at a number of conferences and publications throughout its creation and development. As a part of these efforts, it has been integrated with the Generalized Intelligent Framework for Tutoring (GIFT) in two fashions: branching, using the Engine for Management of Adaptive Pedagogy (EMAP), and dialogue-based, using open-source chat technology. This technology is nearly ready to be deployed to the public, enabling this workshop to demonstrate its capability, highlight its use, and allow users to make their own tutors centered about their own content.

**Keywords:** intelligent tutoring system, ADDIE process, dialogue based tutoring, branching tutoring

## 1 Introduction

The Tools for Rapid Automated Development of Expert Models (TRADEM) project was first published in 2013 in a simulation venue [1]. The technology was demonstrated last year at the Intelligent Tutoring Systems 2014 conference, as part of a workshop on authoring tools [2], at the Educational Data Mining 2014 conference, as part of an industry session [3], and at the annual GIFT Symposium, as part of general GIFT development [4]. The project has recently come to completion, with the outputs intended to be made publicly available soon, and physically distributed as part of this workshop.

As described by many, including the GIFT foundation paper [5], Intelligent Tutoring Systems (ITSs) contain four components: a domain model, an expert model, a learner (or student) model and a pedagogical model. TRADEM uses a domain model built as a summarization of provided content mixed into a set of topics, as a part of the GIFT Domain Module. The expert model consists of a domain model together with expert-derived information concerning the order of topic learning, information about the content, and a basic manner of assessing learner response. These pieces of information are represented in the GIFT Domain Knowledge File (DKF), and are linked with a series of questions in the Survey Authoring System (SAS). The pedagogical model used as part of TRADEM-produced tutors is simply the GIFT default engine, called the Engine for Management of Adaptive Pedagogy (EMAP), which has been documented in greater detail in other literature [6].

The purpose of the TRADEM project has been to rapidly and mostly-automatically create expert models and sequence domain material from initially provided texts. The traditional teaching model relies upon teachers to select the material for consumption by the learners, where the teacher provides the material. The TRADEM model of development is to condense the material selected for students, where the system provides the learning material created from previously provided learning materials. Naturally, there is some disagreement in the literature as to the nature of an "expert model."; is it the selected materials by the teacher, or the core concepts identified by the system? In the TRADEM formulation, a domain model consists of a set of topics in a domain, while an expert model consists of a domain model together with expert-derived information concerning the order in which topics should be learned and expert-derived data that enables an ITS to present each topic and assess learner knowledge. Expert derived information may take a few different forms. The first of these are the topic names and conventions used as a map of the topics, as shown later in Fig. 3 and Fig. 6. The second part of the expert-derived information is in metadata about the type of information content contains (e.g. Gagne's 9 Events [7] or Merrill's Component Display Theory [8]). The last of the expert-derived information is questions and answers, which are automatically suggested based on the content, and curated by the human expert.

This paper is intended to briefly describe the how the system operates and the technologies which it relies upon, as a short description is helpful to the reader, although not required for practical use. In practice, the purpose of the workshop of this technology is to demonstrate the technology. In short, TRADEM uses automated text analysis techniques to create core groups of "topics" based upon the topics that appear to have been discussed the most. It uses automated summarization techniques to create summary text paragraphs and link it to an exact topic, and uses this text to propose a name for the topic, as content for the topic, and as a basis for creating questions. The technical tasks to perform each of these items are described in other works throughout the literature [1-4].

## 2 Use

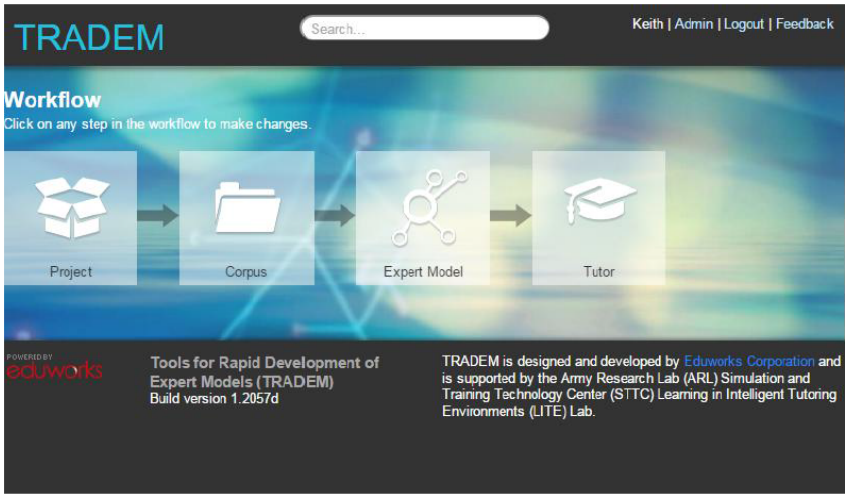



**Fig. 1.** TRADEM User Interface

The basic process of creating a tutor with TRADEM is simple, and relies upon a few basic steps, all of which are shown from the screen following login, as seen in Figure 1. In this section, we will highlight the specific steps required to produce a tutor within the TRADEM authoring workflow.

**Step One:** Create a new project and give it a name.
**Step Two:** Create a corpus, upload documents to it, and save, as seen in Figure 2.

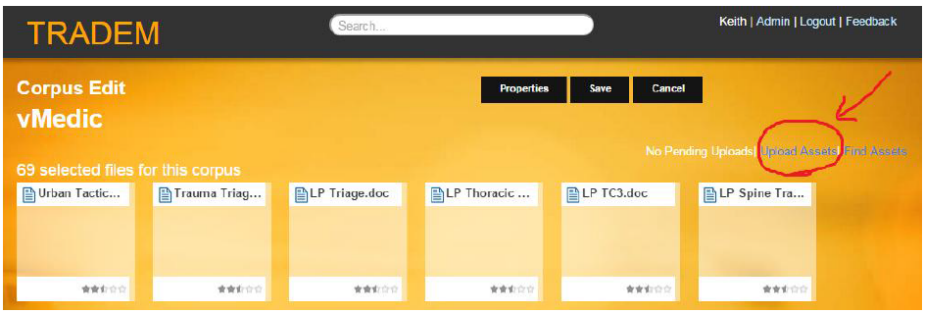



**Fig. 2.** Corpus creation and editing

**Step Three:** Add a new expert model through a selection of features. TRADEM provides an estimation of the number of topics present within your model when using the default settings. If your corpus has a fewer number of documents, or some

of the documents in your corpus are short but contain critical information, you may consider adjusting the expert model parameters to be higher than the default values, shown in Figure 3.

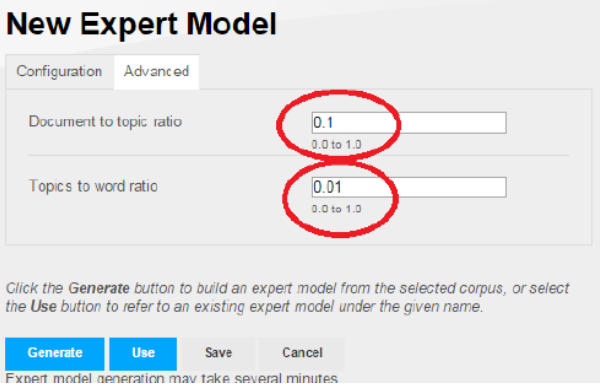



**Fig. 3.** TRADEM Expert Model Parameters

**Step Four:** Edit the expert model and mini-corpus. Be sure to have enough questions on each topic to support the GIFT default exports (3 questions per topic). If TRADEM has not suggested enough questions related to the topic, the user may have to create them manually or generate a new expert model. See the highlighted area in Figure 4 to edit the topic in this manner.

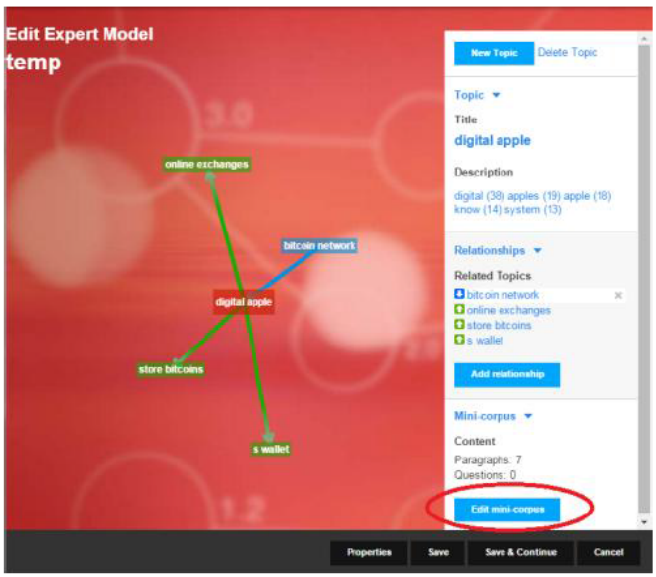



**Fig. 4.** Expert Model Editing

**Step Five:** Export the tutor. At this point you will receive three options to either 1) export as a standard package, 2) export as a GIFT TRADEM-Tutor ("T-Tutor") pack-age, or 3) export as a GIFT PowerPoint (PPT) package. The first of these options exports unadorned slides and questions/answers for presumed import into other Learning Management Systems (LMSs) and traditional training content. The second option exports a dialogue-based "talking head" which can understand basic student inputs and course directions, and can be imported into GIFT. The third of these options exports a series of PowerPoint shows and pre-/post-tests which can be imported into GIFT and managed as a branching course. These options are shown in Figure 5 which shows the "export tutor" option and the "generate export" option after selecting one of the above three choices.

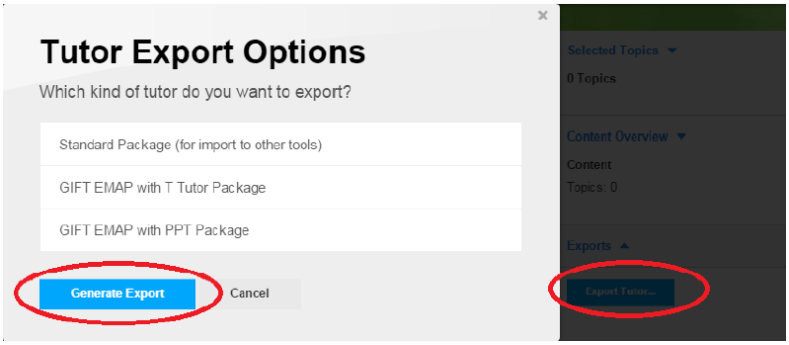



**Fig. 5.** TRADEM Export Tutor Dialogues

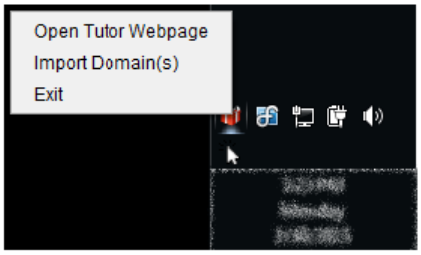



**Fig. 6.** GIFT Import

**Step Six:** Import the package into an existing GIFT installation using the GIFT Import Tool. The GIFT import tool can be found by right-clicking on the GIFT icon as shown in Figure 6, or in the GIFT\scripts\tools\launchControlPanel.bat interface. After import, the EMAP course will be selectable and display as traditional PowerPoint slides, while the "TTutor" export will display with a "talking" head and simplistic dialogue responses, as shown in Figure 7.

**Fig. 7.** TRADEM-Tutor Interface [4]

## 3 Benefits for Use

There are a few benefits to using the TRADEM tool, including aiding in front end analysis of content, automatically summarizing existing documents, or providing the foundation of a GIFT course. This section briefly discusses these three use cases.

One of the manners of TRADEM use is to perform a front end analysis of the content being worked with. The import of content into TRADEM and looking at the structure of the domain can prove valuable to deciding other methods of instruction. As an example, differing domains may represent different manners of instruction, as shown in Figure 8 with a few different domains. This analysis may affect human decisions of how to instruct the material, and can be garnered fairly quickly (minutes).

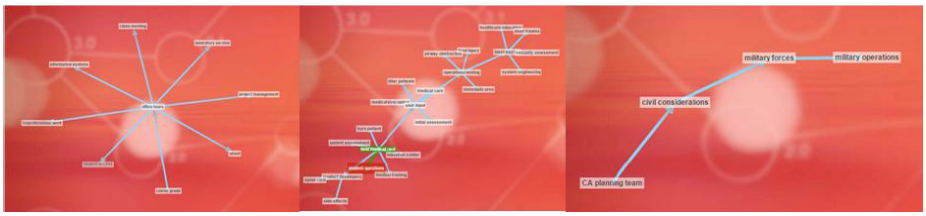



**Fig. 8.** Discovered organizational structures [3], which may be instructed differently

A second manner of technology use is in the automated summary of learning materials. The automated summarization techniques can be used with conference track papers as input, and presented a summary of the things discussed in the individual tracks [1]. Such use may be able to guide conference learners to the sessions of their greatest interest, based on the papers accepted to the tracks.

Further, a GIFT tutor which uses the EMAP can be created with very little effort through the use of TRADEM. Instead of uploading various learning materials, tagging them with metadata, and building a course, the TRADEM tool can be used to integrate checkboxes for metadata, and automatically sequence the content. Given the speed and simplicity of use, such practice may prove standard to the creation of GIFT-EMAP courses. This allows tutor creators to benefit from an extensively researched instructional domain model without significant investment of time, and us-

ing content which can be fine-tuned at a later time with the GIFT authoring tools. Other benefits are more extensively discussed in other works [3].

### 3.1 Licensing

The open-source nature of GIFT means that reproducible code is freely released and updated with each subsequent version. Tutors, the output of GIFT, are free to produce and may be sold or freely provided for community benefit. Developed modules and plug-ins may additionally be sold or donated, while GIFT components may never be sold. While TRADEM is free for both use and modification in Government applications, it is not open source. The close-source encumbrances of TRADEM, however, are not burdensome. The closed-source encumbrances are 1) that the user must agree to a licensing agreement on branding prior to the generation of tutoring materials, and 2) not to remove the branding of the tutoring materials created as part of the TRADEM process. Aside from these issues, the tutors produced using the TRADEM process are free to be used and commercialized as GIFT outputs.

## 4 Future Work

The primary use of TRADEM is for use as an advanced and automated authoring capability [9], but there is a follow-on effort to automate the process of evaluating the weaknesses of the produced courses. The intention is that an instructor, after creating a GIFT or TRADEM course, would be able to analyze the course for the items that produce (or omit) learning gains on the relevant post-test measures. Additional measures are being taken to change the login/logout credentials to match GIFT, to make the Gateway Module plug-in an interoperable and separable service, and to enable web-based learning and software testing. The current architecture and integration is shown in Figure 9, and represents a way for other dialogue tutoring services to integrate into GIFT, as they can either follow this example integration, or the one provided by the AutoTutor webservices.

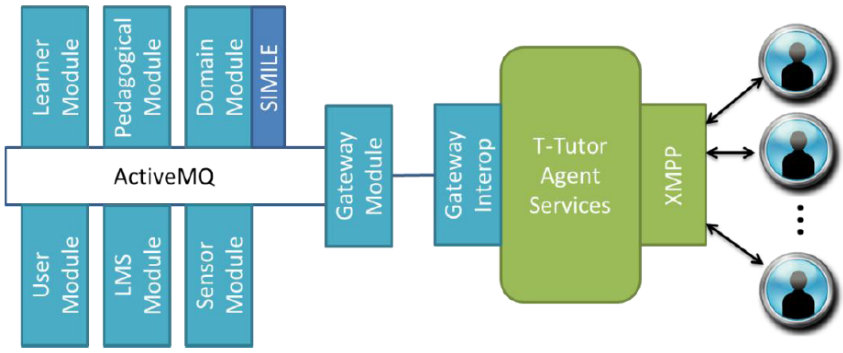



**Fig. 9.** GIFT and TRADEM Combined Architecture

In the above diagram, the agent services for the TRADEM-Tutor are shown as a plugin to the Gateway interoperability section. These interact with Extensible Messaging and Presence Protocol (XMPP) software, for the purpose of interacting with Google Hangouts or other delivery engine. The use of such architecture allows for the combination of traditional GIFT course elements with the newly added TTutor elements. An example of such an integration may be the use of the Student Information Modules for Intelligent Learning Environments (SIMILE) rule assessment engine for digital games [10], as a practice environment for medical training taught by TTutor.

## References


1. Robson, R., Ray, F., Cai, Z.: Transforming Content into Dialogue-Based Intelligent Tutors. The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC), vol. 2013. NTSA, Orlando, FL (2013)
2. Ray, F., Brawner, K., Robson, R.: Automating Addie. Intelligent Tutoring Systems 2014 Conference, Honolulu, Hawaii (2014)
3. Ray, F., Brawner, K., Robson, R.: Using Data Mining to Automate Addie. Educational Data Mining 2014, London, UK (2014)
4. Brown, D., Martin, E., Ray, F., Robson, R.: Using Gift as an Adaptation Engine for a Dialogue-Based Tutor. In: Sottilare, R.A. (ed.) Generalized Intelligent Framework for Tutoring Symposium. www.gifttutoring.org, Pittsburgh, PA (2014)
5. Sottilare, R.A., Brawner, K.W., Goldberg, B.S., Holden, H.A.: The Generalized Intelligent Framework for Tutoring (Gift). (2012)
6. Goldberg, B., Brawner, K., Sottilare, R., Tarr, R., Billings, D.R., Malone, N.: Use of Evidence-Based Strategies to Enhance the Extensibility of Adaptive Tutoring Technologies. In: The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC). NTSA, (2012)
7. Gagne, R.M.: Conditions of Learning and Theory of Instruction. (1985)
8. Merrill, M.D.: Component Display Theory. Instructional-design theories and models: An overview of their current status 1, 282-333 (1983)
9. Olney, A., Brawner, K., Pavlik, P., Keodinger, K.R.: Emerging Trends in Automated Authoring In: Brawner, K. (ed.) Design Recommendations for Intelligent Tutoring Systems: Authoring Tools (Volume 3), vol. 3. U.S. Army Research Laboratory, www.gifttutoring.org. In Press. (2015)
10. Mall, H., Goldberg, B.: Simile: An Authoring and Reasoning System for Gift. In: Sottilare, R.A. (ed.) GIFTSym2, vol. 2, pp. 33-42. Army Research Laboratory, Pittsburgh, PA (2014)

# International Workshop on Affect, Meta-Affect, Data and Learning (AMADL 2015)

*held in conjunction with*

Seventeenth International Conference on Artificial Intelligence in Education (AIED 2015)

Friday, June 26, 2015
Madrid, Spain


**Workshop Co-Chairs:**

Genaro Rebolledo-Mendez[1], Manolis Mavrikis[2], Olga C. Santos[3], Benedict du Boulay[4], Beate Grawemeyer[5] and Rafael Rojano-Cáceres[2]

[1]*Facultad de Estadística e Informática, University of Veracruz*
[2]*London Knowledge Lab, University College London*
[3]*aDeNu Research Group, UNED*
[4]*School of Science and Technology, University of Sussex*
[5]*London Knowledge Lab, Birkbeck*


https://sites.google.com/site/iwamadl2015/

# Table of Contents

# Preface

Emotions and affect play an important role in learning. There are indications that meta-affect (i.e., knowledge about self-affect) also plays a role. There have been various attempts to take them into account both during the design and during the deployment of AIED systems. The evidence for the consequential impact on learning is beginning to strengthen, but the field has been mostly focused on addressing the complexities of affective and emotional recognition and very little on how to intervene. This has largely slowed down progress in this area.

Research is needed to better understand how to respond to what we detect and how to relate that to the learner's cognitive and meta-cognitive skills. One goal might be to design systems capable of recognizing, acknowledging, and responding to learners' states with the aim of promoting those that are conducive to learning by means of tutorial tactics, feedback interventions, and interface adaptations that take advantage of ambient intelligence, among others. Therefore, we need to deepen our knowledge of how changes in learners' affective states and associated emotions relate to issues such as cognition and the learning context.

The papers submitted to the workshop address issues that bridge the existing gap between previous research with the ever-increasing understanding and data availability. In particular, these papers report progress on issues relevant to the broad and interdisciplinary AIED and EDM communities. AMADL 2015 workshop raises the opportunity to bring these two communities together in a lively discussion about the overlap in the two fields. To achieve this, we explicitly address and target both com- munities, as indicated by the workshop's organizers background and the programme committee set up. This workshop builds on the work done in affect related workshops in past AIED conferences, such as Modelling and Scaffolding Affective Experiences to Impact Learning in AIED 2007. The format of the workshop is based on presentations, demonstrations and discussions according to themes addressed by the papers accepted for the workshop.

Genaro Rebolledo-Mendez, Manolis Mavrikis, Olga C. Santos, Benedict du Boulay, Beate Grawemeyer and Rafael Rojano-Cáceres

Workshop Co-Chairs

# The potential of Ambient Intelligence to deliver Interactive Context-Aware Affective Educational support through Recommendations

Olga C. Santos[1], Mar Saneiro[1], M. C. Rodriguez-Sanchez[2], Jesus G. Boticario[1], Raul Uria-Rivas[1], Sergio Salmeron-Majadas[1]

[1] aDeNu Research Group. Artificial Intelligence Dept. Computer Science School, UNED. Calle Juan del Rosal, 16. Madrid 28040. Spain
http://adenu.ia.uned.es
{ocsantos,marsaneiro,jgb,raul.uria,sergio.salmeron}@dia.uned.es
[2] Electronics Department, Universidad Rey Juan Carlos. Calle Tulipán s/n. Móstoles 28933 (Madrid), Spain.
cristina.rodriguez.sanchez@urjc.es

**Abstract.** There is a challenge and opportunity to research if the ambient intelligent support that can be deployed with a recommender system extended with an open hardware infrastructure that can sense and react within the learners' context is of value to supports learners' affectively. In this paper, we summarize the status of our research on eliciting an interactive recommendation for a stressful scenario (i.e., oral examination of a foreign language) that can be delivered through the Ambient Intelligence Context-aware Affective Recommender Platform (AICARP), which is the infrastructure we have designed and implemented with Arduino, an open-source electronic prototyping platform.

## 1 Eliciting Interactive Recommendations with TORMES

We have reported elsewhere [1] our progress on analyzing the potential of Ambient Intelligence to deliver more interactive educationally oriented recommendations that can deal with the affective state of the learner. In particular, following the TORMES methodology [2], we elicited an educational **scenario** focused on helping the learner when preparing for the oral examination in a second language learning course, which is widely considered as a stressful situation.

The **recommendation** identified in this scenario consists in suggesting the learner to breathe slowly (at a rate of 4 breaths/minute) and is aimed to calm her down when she is nervous. The *applicability conditions* that trigger the recommendation take into account physiological (i.e., heart rate, pulse, skin temperature, skin conductance) and behavioral (facial/body movements and speech speed) information that show evidence of restlessness. The recommendation *output* has been coded in a multisensory way by simultaneously modulating light, sound and vibration behavior at aforementioned breath rate, so the learner can perceive the recommended action through alternative sensory channels (i.e., sight, hearing and touch) without interrupting her activity.

## 2 Delivering Interactive Recommendations with AICARP

To deliver the aforementioned recommendation elicited with TORMES, the Ambient Intelligence Context-aware Affective Recommender Platform (AICARP) is being implemented with open source software and open hardware following a modular design controlled by an Arduino board (see [1] for details). In the current version, AICARP receives information from physiological **sensors** regarding changes in the learner affective state through corresponding physiological signals. The sensors are integrated into the e-Health platform [3] and a wireless electrocardiogram system [4]. Taking into account this information, AICARP is able to provide the elicited interactive recommendation to the learner by modulating the output of alternative sensorial **actuators** with the recommended breath rhythm. In particular, the following actuators have already been integrated into AICARP: i) white and red flashlights, ii) an array of blue LEDs, iii) a buzzer that vibrates and sounds, and iv) a speaker reproducing a pure tone at 440 Hz (i.e., "La" musical note).

To get some insight on the users' perception on the recommendation delivery, we have deployed the educational scenario outlined in Section 1 in order to deliver the corresponding recommendation elicited with TORMES. So far, in this context we have carried out **2 pilot studies**, one with 6 university students with various interaction needs -including a blind participant-, and another with 4 participants within the 2014 Madrid Science Week. Since we wanted to test the potential of this approach in detecting not only the physiological information but also the behavioral information, we used the Wizard of Oz method [5]. In this way, the recommendation was triggered by the wizard (in our case, a psycho-educational expert) considering participants' information on both physiological evidences detected with AICARP, as well as body/facial movements and speech speed that the wizard observed while the participants carried out the two tasks defined in the pilots (i.e., talking aloud in English about two specific given topics selected from those usually considered in oral exams).

## 3 Evaluation Outcomes and Open Issues identified

We evaluated AICARP in the 2 pilot studies with the analysis of the participants' responses to the System Usability Scale [6] and to a post-study consisting in a semi structured interview led by the psycho-educational expert. This **evaluation** showed that the implemented infrastructure can actually sense the physiological state of the learner (which seems to be related to some affective state) and deliver ambient intelligent interactive feedback aimed to transform a negative affective (i.e., nervousness) state into a positive one (i.e., relaxation) (see [1] for details on the evaluation results). To the latter, actuators considered aim to provide a natural interaction support not interfering with the participant's task, and consisted of visual, audio and/or tactile feedback.

As discussed in [1], the analysis of the evaluation outcomes has identified several **open issues** to be addressed in future research, as follows:

1. **How to deliver interactive recommendations:** this issue deals with selecting the preferred sensory channels from those available, the format to display the recommendation, the support to understand the purpose of the recommendation and the intrusion level.
2. **When recommendations are to be provided:** in terms of physiological and behavioral changes, while interfering as less as possible with the task. Here, and following TORMES methodology, data mining techniques can be explored to automatically identify the criteria that characterize the appropriate moment to deliver the recommendation [7].
3. **Learners' features of potential relevance in order to design other recommendations:** such as domain dependent attributes (i.e., the English level) and personality traits.
4. **Social aspects involved when collaboration takes place:** in the current scenario, collaboration can occur when learners are asked to perform the oral examination in pairs by dialoging a given situation. The training can be done using a videoconferencing system. In this context, other issues should be considered, such as the intensity of collaboration, the type of collaborative task, the individual acceptance of the technology used to support the collaboration, as well as specific personality traits.

## Acknowledgements

This work was supported by the Spanish Ministry of Economy and Competitiveness (MINECO) under Grant TIN2011-29221-C03-01 (MAMIPEC project).

## References

1. Santos, O.C., Saneiro, M., Rodriguez-Sanchez, M.C. and Boticario, J.G. (2015) Towards Interactive Context-Aware Affective Educational Recommendations in Computer Assisted Language Learning. New Review of Hypermedia and Multimedia, in press.
2. Santos, O.C., Boticario, J.G. (2015) Practical guidelines for designing and evaluating educationally oriented recommendations. In Computers and Education, vol. 81, 354–374.
3. Cooking Hacks. E-Health Platform. Available from: http://www.cooking-hacks.com.
4. Torrado-Carvajal, A., Rodriguez-Sanchez, M.C., Rodriguez-Moreno, A., Borromeo, S., Garro-Gomez, C., Hernandez-Tamames, J. A., and Luaces, M. (2012) Changing communications within hospital and home health care. In 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 6074-6077.
5. Dahlbäck, N., Jönsson, A., and Ahrenberg, L. (1993) Wizard of Oz studies: why and how. In Proceedings of Intelligent User Interfaces, 193–200.
6. Brooke, J. (1996) SUS: a 'quick and dirty' usability scale. In Jordan, P.W., Thomas, B., Weerdmeester, B.A. and McClelland, A.L. Usability Evaluation in Industry. London: Taylor and Francis.
7. Salmeron-Majadas, S., Arevalillo-Herráez, M., Santos, O.C., Saneiro, M., Cabestrero, R., Quirós, P., Arnau, D. and Boticario, J.G. (2015) Filtering of Spontaneous and Low Intensity Emotions in Educational Contexts. 17th Int. Conf. on Artificial Intelligence in Education (AIED 2015). Lecture Notes in Artificial Intelligence, vol. 9112, 429-438.

# The impact of feedback on students' affective states

Beate Grawemeyer[1], Manolis Mavrikis[2], Wayne Holmes[2], Alice Hansen[2], Katharina Loibl[3], and Sergio Gutiérrez-Santos[1]

[1] London Knowledge Lab, Dep of Computer Science and Information Systems, Birkbeck, London, UK
beate@dcs.bbk.ac.uk, sergut@dcs.bbk.ac.uk
[2] London Knowledge Lab, UCL Institute of Education, University College London, London, UK
m.mavrikis@ioe.ac.uk, w.holmes@ioe.ac.uk, a.hansen@ioe.ac.uk
[3] Institute of Educational Research, Ruhr-Universität Bochum, Germany
katharina.loibl@rub.de

**Abstract.** Affective states play a significant role in students' learning behaviour. Positive affective states can enhance learning, while negative affective states can inhibit it. This paper describes a Wizard-of-Oz study that investigates the impact of different types of feedback on students' affective states. Our results indicate the importance of providing feedback matched carefully to the affective state of the students in order to help them transition into more positive states. For example when students were confused affect boosts and specific instructive feedback seem to be effective in helping students to be in flow again. We discuss this and other ways to adapt the feedback, together with implications for the development of our system and the field in general.

## 1 Introduction

This paper reports the results of a set of two Wizard-of-Oz studies which explore the effect of different feedback types on students' affective states.

It is well understood by now that affect interacts with and influences the learning process [9, 6, 2]. While positive affective states such as surprise, satisfaction or curiosity contribute towards constructive learning, negative ones including frustration or disillusionment at realising misconceptions can lead to challenges in learning. The learning process is indeed full of transitions between positive and negative affective states and regulating those is important. For example, a student may seem interested in exploring a particular learning goal, however s/he might have some misconceptions and need to reconsider her/his knowledge. This can evoke frustration and/or disappointment. However, this negative affective state may turn into deep engagement with the task again. D'Mello et al., for example, elaborate on how confusion is likely to promote learning under appropriate conditions [6].

It is important therefore, to deepen our understanding of the role of affective states for learning, and to be able to move students out of states that inhibit learning. Pekrun [13] discusses achievement emotions or affective states, which arise in a learning situation. Achievement emotions are states that are linked to learning, instruction, and achievement. We focus on a subset of affective states identified by Pekrun: flow/enjoyment, surprise, frustration, and boredom. We also add confusion, which has been identified elsewhere as an important affective state during learning [15] for tutor support and for learning in general [6].

As described in Woolf et al. [20] students can become overwhelmed (very confused or frustrated) during learning, which may increase cognitive load [19] for low-ability or novice students. However, appropriate feedback might help to overcome such problems. Carenini et al. [3] describe how effective support or feedback needs to answer three main questions: (i) when the support should be provided during learning; (ii) what the support should contain; and (iii) how it should be presented.

In this paper we focus on the question of *what* the support should contain with respect to affect i.e. the types of feedback that are able to induce a positive affective state.

In related work students' affective states have been used to tailor motivational feedback and learning material in order to enhance the learning experience. For example, Santos et al. [17] show that affect as well as motivation and self-efficacy impact the effectiveness of motivational feedback and recommendations. Additionally, Woolf et al. [20] developed an affective pedagogical agent which is able to mirror a student's affective state, or acknowledge a student's affective state if it is negative. Another example is Conati & MacLaren [5], who developed a pedagogical agent to provide support according to the affective state of the students and the user's personal goal. Also, Shen et al. [18] recommend learning material to the student based on their affective state. D'Mello et al. [7] developed a system that is able to respond to students via a conversation that takes into account the affective state of the student.

In contrast, in this paper, we investigate the impact of different types of feedback on students' affective state and how and whether they can help students regulate their affect and thus improve learning. In what follows we present two sets of Wizard-of-Oz studies where feedback was provided to students interacting with an exploratory learning environment designed to learn fractions. From these studies, the affective states of the students were carefully annotated in order to address our research questions.

## 2 The Wizard-of-Oz studies

### 2.1 Aims

One of our research aims is to develop intelligent support that enhances the learning experience by taking into account the student's affective state. We were specifically interested in identifying how different feedback types modify affective states.

In order to address this question we conducted two sets of ecologically valid Wizard-of-Oz studies (e.g. [11, 8]) which investigated the effect of affective states on different feedback types at different stages of the task.

### 2.2 Participants and Procedure

In total, 26 Year-5 (9 to 10-year old) students took part in the Wizard-of-Oz studies. Each session lasted on average 20 minutes. Each student participated in one Wizard-of-Oz session.

The sessions were run in an ordinary classroom with multiple computers, where additional children were working with the learning platform (not wizarded) in order to support ecological validity. This was important particularly as in early settings we identified that children would not speak that much to the platform if they felt that they were monitored [10]. Figure 1 shows the setup of the studies. Wizards followed a script with pre-canned messages to send mes-

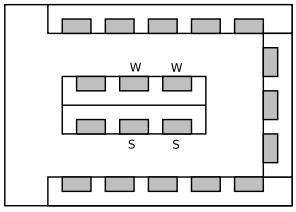



**Fig. 1.** The layout. The Wizard-of-Oz studies took place on the central isle while the rest of the students worked on a version of the system which only sequences tasks and provides minimal support (W=wizard, S=student).

sages to the students through the learning platform and deliberately limited their communication capacity in order to simulate the actual system. To achieve that wizards were only able to see students' screen. An assistant was able to hear students' reactions to reflections or talk-aloud prompts (as prompted by the 'system') and provide recommendations to the wizard with respect to the detected affective state. Any feedback provided was both shown on screen and read aloud by the system to students.

### 2.3 Feedback types

Different types of feedback were presented to students at different stages of their learning task. The feedback provided was based on interaction via keyboard and mouse, as well as speech.

We explore different types of feedback that are known from the literature to support students in their learning and fit our context. The following different feedback types were provided:

– **AFFECT BOOSTS - affect boosts**. As described in [20] affect boosts can help to enhance student's motivation in solving a particular learning

task. These included prompts that acknowledged for example that a task is difficult or that the student may be confused but they should keep trying.
- **INSTRUCTIVE FEEDBACK - instructive task-dependent feedback**. This feedback provided detailed instructions, what subtask or action to perform in order to solve the task.
- **OTHER PROBLEM SOLVING FEEDBACK - task-dependent feedback**. This support was centred on helping students to solve a particular problem that they are facing during their interaction by providing either questions to challenge their thinking or specific hints designed to help them identify the next step themselves.
- **TALK ALOUD PROMPTS - talking aloud**. With respect to learning in particular, the hypothesis that automatic speech recognition (ASR) can facilitate learning is based mostly on educational research that has shown benefits of verbalization for learning (e.g., [1]).
- **REFLECTIVE PROMPTS - reflecting on task performance and learning**. Self-explanation can be viewed as a tool to address students' own misunderstandings [4] and as a 'window' into students' thinking.
- **TALK MATHEMATICS PROMPTS - using particular domain specific mathematics vocabulary**. The aim of this prompt was to encourage students to use mathematical vocabulary in order continually revise their interpretations. In early studies [10] we found that students' reflections were often procedural and pragmatic (e.g. talking about the user interface) rather than mathematical.
- **TASK SEQUENCE PROMPTS - moving to the next task**. This feedback is centred on providing support regarding what action to perform next in order to change the task, such as clicking the 'Next' button.

Table 1 shows examples of the different feedback types.

## 3 Annotation of affective states and feedback reactions

From the Wizard-of-Oz studies we recorded the students' screen display and their voices. From this data, we annotated affective states (e.g. screen interaction and what the students said) before and after feedback was provided.

As described earlier, for the affective state detection we discriminated between five different affective types: enjoyment, surprise, confusion, frustration, and boredom. For the annotation of those affective states we used a similar strategy to that described in [15], where a dialogue between a teacher and a student was annotated retrospectively by categorising utterances in terms of different feedback types. Also, [2] describe how they coded different affective states based on observations of students interacting with a learning environment. Similarly, we annotated student's affective states for each type of feedback provided. In addition to the student's voice we also used the video of the screen capture to support the annotation process. Students' affective states were annotated as follows:

- **FLOW:** Engagement with the learning task. Statements like 'I am enjoying this task' or 'This is fun'. Sustained interaction with the system.

| Feedback type | Example |
|---|---|
| AFFECT BOOSTS | You're working really hard! Keep going! |
| INSTRUCTIVE FEEDBACK | Use the comparison box to compare your fractions. |
| OTHER PROBLEM SOLVING FEEDBACK | If you add fractions, they need to have the same denominators first. |
| REFLECTIVE PROMPTS | What do you notice about the two fractions? |
| TALK ALOUD PROMPTS | Remember to talk aloud, what are you thinking? |
| TALK MATHEMATICS PROMPTS | Can you explain that again using the terms denominator, numerator? |
| TASK SEQUENCE PROMPTS | Well done. When you are ready click 'next' for the next task. |

**Table 1.** Examples of feedback types

- **SURPRISE:** Gasping. Statements like 'Huh?' or 'Oh, no!'.
- **CONFUSION:** Failing to perform a particular task. Statements such as 'I'm confused!' or 'Why didn't it work?'. Uncertain interaction with the system.
- **FRUSTRATION:** Tendency to give up, repeatedly clicking or deleting of objects in the system or repeatedly failing to perform a particular task, sighing, statements such as, 'What's going on?!'.
- **BOREDOM:** Inactivity or statements such as 'Can we do something else?' or 'This is boring'.

## 4 Results

In total 396 messages were sent to 26 students. The video data in combination with the sound files were analysed independently by three researchers (one was independent of the project) who categorised the affective states of students before and after the feedback messages were provided.

The data is combined from two sets of Wizard-of-Oz studies. We use kappa statistics to measure the degree of the agreements of the annotations for reliability. Kappa was .46, $p<.001$. This is generally expected from retrospective annotation of naturalistic affect experiences [14]. We consolidated the annotations based on discussion between the annotators and the rest of the authors of the paper in order to agree upon the annotations that did not match originally. In the second set we had resources to introduce the Baker-Rodrigo Observation Method Protocol (BROMP) and the HART mobile app that facilitates the coding of students affective states in the classroom [12]. Kappa based on the retrospective annotation was still .56, $p<.001$. We first consolidated the data with the same approach as before and then compared against the field annotations. Kappa between the consolidated annotation and the HART data was .71,

$p<.05$ (note that is may appear low but we did not expect the retrospective annotation to get surprise and frustration accurately). We used the HART data to improve the annotation by mapping feedback actions against the observation for 20 seconds prior to the delivery of the feedback to 20 seconds after the student had closed the corresponding feedback window. We marked the changes for an independent annotator to revisit the first set of annotations.

The student's affective states, that occurred before and after the different types of feedback was provided, can be seen in figure 2. Each block shows an affective state *before* feedback was provided. The colour within the bars indicates the type of affective states that occurred *after* the feedback was provided. The number within the bars indicate the number of times the affective state occurred.

In order to investigate whether there was an effect of the feedback on the learning experience, we looked at whether a student's affective state was enhanced, stayed the same or worsened. An affective state was enhanced for example, when it was changed from confusion to flow, or (given the findings about confusion [6]) from frustration to confusion, frustration to flow, boredom to flow etc. An affective state was worsened if it moved for example, from flow to frustration or confusion, or from confusion to frustration.

As the data is categorical [16], we apply chi-square tests to investigate statistical significant differences between the groups. We present them below and discuss in more detail in the next section.

**Flow** When students were in flow, there was no significant difference between the feedback types on whether the affective state stayed in the same flow state ($X^2$(6, N=169) = 4.31, $p>.05$) or worsened ($X^2$(6, N=169) = 4.89, $p>.05$). As flow is the most positive affective state, the affective state in this sub-sample cannot be enhanced.

**Confusion** When students were confused, there was a significant effect of the feedback type on whether students' affective state was enhanced into a flow state ($X^2$(6, N=181) = 13.65, $p<.05$). The most effective feedback types were affect boosts with 68% of the cases, followed by guidance feedback with 67%, and task sequence prompts with 63%. Reflective prompts resulted in a flow state in 48% of the cases, talk aloud prompts 38%, and problem solving support with 34%. Talk maths prompts were the least effective with only 25% of the cases.

There was also a significant effect of the feedback type and whether the affective state stayed the same ($X^2$(6, N=181) = 14.34, $p<.05$). Talk maths prompts were highest associated with a continuing confused state with 75% of the cases. This was followed by problem solving support with 66%, talk aloud prompts with 59%, reflective prompts with 52%, task sequence prompts with 37%, affect boosts with 32%, and the least feedback type that was associated with a continuing confused state were guidance feedback with 29% of the cases.

There was no significant association between the feedback type and whether the affective state worsened ($X^2$(6, N=181) = 4.65, $p>.05$).

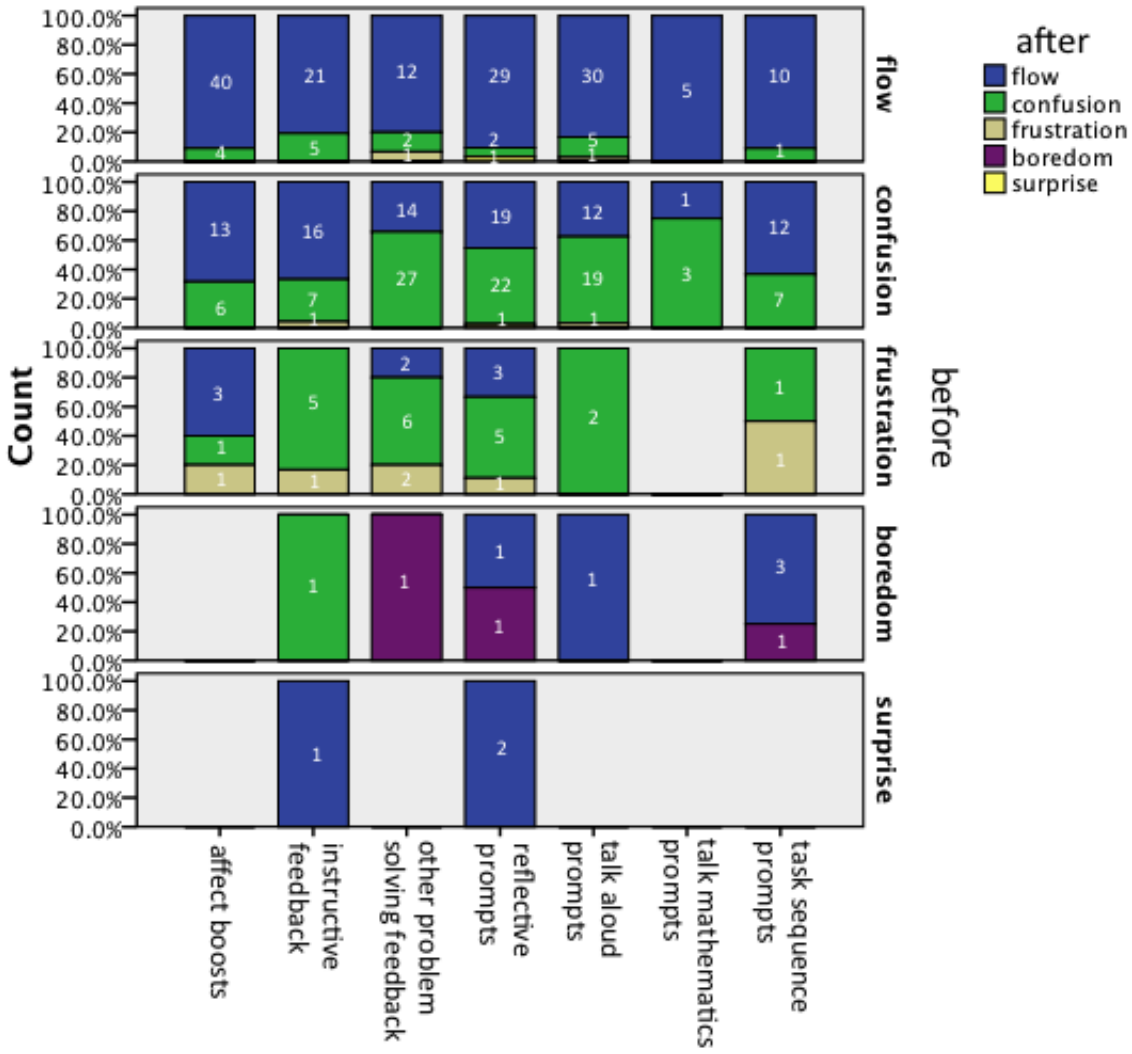



**Fig. 2.** Students' affective states before and after feedback was provided. Each block shows an affective state *before* feedback was provided. The colour within the bars indicates the type of affective states that occurred *after* the feedback was provided. The number within the bars indicate the number of times the affective state occurred.

**Frustration, boredom and surprise** There was not sufficient data available when students were frustrated (36 cases), nor when they were bored (9 cases), or surprised (3 cases) to run a statistical test across the different affective states and feedback types.

However, the data indicates that some of the provided feedback types were better able to change the affective state of the student when they were frustrated, bored or surprised, as can be seen in figure 2. For example, 60% of the affect boosts were able to change frustration into flow, followed by reflective prompts 33% and problem solving support 20%.

## 5 Discussion

The results presented in the previous section show that feedback can enhance students' affective states, and that the impact of the various feedback types mostly depends on the students' affective state before the feedback was provided.

When students were in flow there was no significant difference between the feedback types on whether or not the affective state stayed the same or worsened. This suggests that, when students are in flow, challenging feedback can be provided without negative implications.

However, when students were confused there was a difference between the feedback types on whether the affective state was enhanced, stayed the same or worsened. The feedback types that most effectively moved the student out of a confusion state were affect boosts, instructive, and task sequence prompts. When they were struggling to overcome problems, affect boosts appeared to encourage some students to redouble their efforts without the need for task specific support. We can hypothesise that this enabled students to self-regulate their affect and move forward. As expected, instructive feedback appears to have given the students the next steps that they needed, whereas other problem solving was less successful. Other problem solving feedback seems to have led students to be more confused because of the increased cognitive load caused by them having to understand the hint or the question provided.

While talk aloud prompts and talk maths, encouraged them to vocalize what they are trying to achieve, they appear not to have helped the students address their confusions. Instead, when they were confused, students appeared to have welcomed a new task (the opportunity to abandon the cause of their confusion). While as a strategy this can be pedagogically debatable, there is scope to provide tasks aimed to help them at the same concepts in a different, simpler way or to allow them to practice first some skills in a practice-based rather than exploratory task.

Although there was insufficient data to analyse the impact of the different feedback types on students' affective state when they were frustrated, some tentative observations can be made. For example, it was evident that the affect of students who were frustrated was enhanced whatever the feedback they were provided with. However, it is notable that the frustrated students who were provided affect boosts were most likely to move to a flow. We have other anecdotal evidence in the same scenario with different students that suggest that explicitly

addressing affect and helping students to think of their emotions during learning can help them move to confused or to flow state without need for immediate problem solving support.

It is worth noting that compared to other research we may have been unable to detect more negative states, especially boredom, because of the nature of the environment that the students were using – an exploratory learning environment that encouraged them to speak. The combination of unstructured learning and speech might prevent students from becoming bored.

## 6 Conclusion and future work

The affective state of students can be modified with feedback. There is a difference in the impact of different feedback types according to the affective state the student is in before the feedback was provided. Although there seems not to be too much of a difference when students are in flow, when students were confused different feedback types seem to matter more. While, for example, affect boosts and instructive feedback were able to change confusion into flow, prompting students to use mathematical vocabulary or providing other problem solving support, were associated with the same confused state or even lead to frustration.

In the light of findings like D'Mello et al. [6] for example of the importance of confusion under appropriate conditions in learning, our findings have important implications for learning and teaching in general, and AIED in particular. Problem solving support specifically in exploratory learning environments is difficult to achieve successfully, particularly when students are in a situation that was not previously encountered during a system's design. However, detecting affect may be relatively easier in certain contexts particularly in speech-enabled software like in our case and therefore affective support matters as much, if not more than, problem solving support. In addition, the exact type of support provided when students are frustrated is important. To understand this better we need to investigate more the different types of problem solving support and their combination with affective feedback that can act both as a way to self-regulate affect and take student into a more positive state like confusion or flow.

In our current study we are implying that learning performance is enhanced when students are in a positive affective state. In the future we are planning to evaluate if learning performance will be enhanced when students are moved out of a negative into a positive affective state. Our next step is to train an intelligent system that is able to tailor the type of feedback according to the affective state of the student in order to enhance the learning experience.

## References

1. Askeland, M.: Sound-based strategy training in multiplication. European Journal of Special Needs Education 27(2), 201–217 (2012)
2. Baker, R.S.J.d., DMello, S.K., Rodrigo, M.T., Graesser, A.C.: Better to be frustrated than bored: The incidence, persistence, and impact of learners cognitive-affective states during interactions with three different computer-based learning environments. Int. J. Hum.-Comput. Stud. 68(4), 223–241 (2010)

3. Carenini, G., Conati, C., Hoque, E., Steichen, B., Toker, D., Enns, J.: Highlighting interventions and user differences: Informing adaptive information visualization support. In: Proceedings of CHI 14. pp. 1835–1844 (2014)
4. Chi, M.: Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In: Glaser, R. (ed.) Advances in instructional psychology, pp. 161–238. Mahwah, NJ: Lawrence Erbaum Associates (2000)
5. Conati, C., MacLaren, H.: Empirically building and evaluating a probabilistic model of user affect. User Modeling and User-Adapted Interaction (2009)
6. DMello, S.K., Lehman, B., Pekrun, R., Graesser, A.C.: Confusion can be beneficial for learning. Learning & Instruction 29(1), 153–170 (2014)
7. DMello, S., Craig, S., Gholson, B., Franklin, S., Picard, R., Graesser, A.: Integrating affect sensors in an intelligent tutoring system. In: Affective Interactions: The Computer in the Affective Loop Workshop at IUI 2005. pp. 7–13 (2005)
8. Eynon, R., Davies, C., Holmes, W.: Supporting older adults in using technology for lifelong learning: the methodological and conceptual value of wizard of oz simulations. In: Proceedings of NLC 2012. pp. 66–73 (2012)
9. Kort, B., Reilly, R., Picard, R.: An affective model of the interplay between emotions and learning. In: Proceedings of ICALT 2001. No. 43–46 (2001)
10. Mavrikis, M., Grawemeyer, B., Hansen, A., Gutiérrez-Santos, S.: Exploring the potential of speech recognition to support problem solving and reflection. In: ECTEL 2014. pp. 263–276 (2014)
11. Mavrikis, M., Gutiérrez-Santos, S.: Not all wizards are from Oz: Iterative design of intelligent learning environments by communication capacity tapering. Computers & Education 54(3), 641–651 (2010)
12. Ocumpaugh, J., Baker, R.S.J.d., Rodrigo, M.M.T.: Baker-Rodrigo Observation Method Protocol (BROMP) 1.0. Training Manual version 1.0. Tech. rep., New York, NY: EdLab. Manila, Philippines: Ateneo Laboratory for the Learning Sciences. (2012)
13. Pekrun, R.: The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. J. Edu. Psych. Rev. pp. 315–341 (2006)
14. Porayska-Pomsta, K., Mavrikis, M., DMello, S., Conati, C., de Baker, R.S.J.: Knowledge elicitation methods for affect modelling in education. I. J. Artificial Intelligence in Education 22(3), 107–140 (2013)
15. Porayska-Pomsta, K., Mavrikis, M., Pain, H.: Diagnosing and acting on student affect: the tutors perspective. UMUAI 18(1), 125-173 (2008)
16. Rosenthal, R., Rosnow, R.: Essentials of Behavioral Research: Methods and data analysis. McGraw Hill, 3rd edn. (2008)
17. Santos, O., Saneiro, M., Salmeron-Majadas, S., J.G., B.: A methodological approach to elicit affective educational recommendataions. In: Proceedings of ICALT 2014 (2014)
18. Shen, L., Wang, M., Shen, R.: Affective e-learning: Using emotional data to improve learning in pervasive learning environment. Educational Technology & Society 12(2), 176–189 (2009)
19. Sweller, J., van Merrienboer, J.G., Paas, G.W.: Cognitive Architecture and Instructional Design. Educational Psychology Review 10, 251–296+ (1998)
20. Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., Picard, R.: Affect-aware tutors: recognising and responding to student affect. Int. J. Learning Technology 4(3-4), 129–164 (2009)

# Recognising Perceived Task Difficulty from Speech and Pause Histograms

Ruth Janning, Carlotta Schatten, and Lars Schmidt-Thieme

Information Systems and Machine Learning Lab (ISMLL), University of Hildesheim
{janning,schatten,schmidt-thieme}@ismll.uni-hildesheim.de

**Abstract.** Currently, a lot of research in the field of intelligent tutoring systems is concerned with recognising student's emotions and affects. The recognition is done by extracting features from information sources like speech, typing and mouse clicking behaviour or physiological sensors. In former work we proposed some low-level speech features for perceived task difficulty recognition in intelligent tutoring systems. However, by extracting these features some information hidden in the speech input is loosed. Hence, in this paper we propose and investigate speech and pause histograms as features, which preserve some of the loosed information. The approach of using speech and pause histograms for perceived task difficulty recognition is evaluated by experiments on data collected in a study with German students solving mathematical tasks.

**Keywords:** Intelligent tutoring systems, perceived task difficulty recognition, low-level speech features, speech and pause histograms

## 1 Introduction

Automatic cognition, affect and emotion recognition is a relatively young and very important research field in the area of adaptive intelligent tutoring systems. Some research has been done to identify useful information sources and appropriate features able to describe student's cognition, emotions and affects. Those information sources can be speech input, written input, typing and mouse clicking behaviour or input from physiological sensors. In former work ([5], [6], [7]) we proposed low-level speech features for perceived task difficulty recognition in intelligent tutoring systems. These features are extracted from the amplitudes of speech input of students interacting with the system and contain for instance the maximal and average length of speech phases and pauses. However, by extracting those features some more fine granulated information contained within the sequence of speech and pause segments is loosed and the question arises if there is a way to create features which preserve the loosed information. Histograms contain much more information than only the maximal, minimal and average value. Hence, in this work we propose and investigate speech and pause histograms as features for perceived task difficulty recognition, i.e. for recognising if a student feels *over-challenged* or *appropriately challenged* by a task. Speech and pause histograms share the advantages of low-level speech features

(they do not inherit the error from speech recognition and there is no need that students use words related to emotions or affects, see also sec. 2) and avoid to lose information hidden in the sequences of speech and pause segments.

## 2 Related Work

For the purpose to recognise emotion or affect in speech one can distinct linguistics features, like n-grams and bag-of-words, and low-level features like prosodic features, disfluencies, e.g. speech pauses ([5], [6]), (see e.g. [17]) or articulation features ([7]). If linguistics features are not extracted from written but from spoken input, a transcription or speech recognition process has to be applied to the speech input before emotion or affect recognition can be conducted. Linguistic features for affect and emotion recognition from conversational cues were presented and investigated e.g. in [10] and [11]. Low-level features are used in the literature for instance for expert identification, as in [18], [13] and [8], for emotion and affect recognition as in [12] and [5], [6], [7] or for humour recognition as in [15]. The advantage of using low-level features like disfluencies is that instead of a full transcription or speech recognition approach only for instance a pause identification has to be applied before computing the features. That means that one does not inherit the error of the full speech recognition approach. Furthermore, these features are independent from the need that students use words related to emotions or affects. Another kind of features which is independent from the need that students use words related to emotions or affects are features gained from information about the actions of the students interacting with the system (see e.g. [9]) like features extracted from a log-file (see e.g. [2], [16], [14]). In [9] such kind of features is used to predict whether a student can answer correctly questions in an intelligent learning environment without requesting help and whether a student's interaction is beneficial in terms of learning. Also the keystroke dynamics features used in [4] belong to this kind of features. In [4] emotional states were identified by analysing the rhythm of the typing patterns of persons on a keyboard. A further possibility of gaining features is using the information from physiological sensors as for instance in [1]. However, bringing sensors into classrooms is time consuming and expensive and one has to cope with students' acceptance of the sensors.

## 3 Speech and Pause Histograms

As mentioned above, in this paper we investigate the ability of speech and pause histograms for perceived task difficulty recognition. How these speech and pause histograms are created from students' speech input is described in sec. 3.2 and the data which we used for our experiments is described in the next section.

### 3.1 Data

We conducted a study in which the speech and actions of ten 10 to 12 years old German students were recorded and their perceived task-difficulties were

**Fig. 1.** Graphic of the decibel scale of an example sound file of a student. The two straight horizontal lines indicate the threshold.

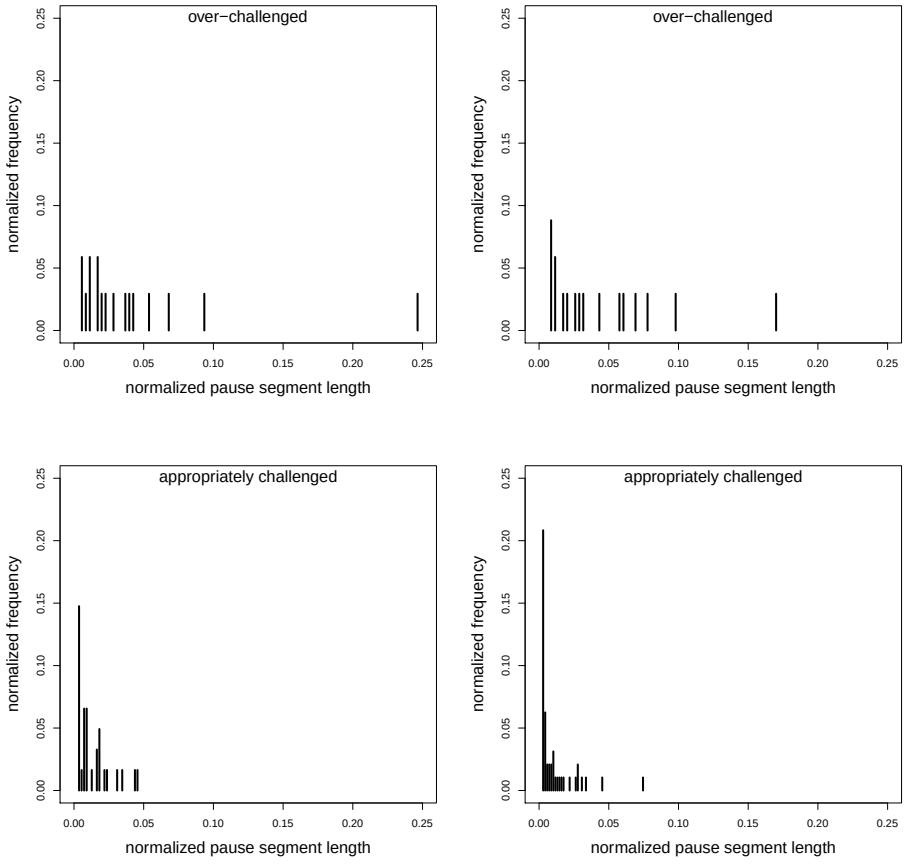



**Fig. 2.** Normalised pause histograms for a task of four different students, where two are labelled as *over-challenged* and the other two as *appropriately challenged*.

reported per task. The labelling of these data was done on the one hand concurrently by a human tutor and on the other hand retrospectively by a second reviewer (with a Cohen's kappa for inter-rater reliability of 0.747, $p < 0.001$). Divergences in the both labellings were clarified later on by discussions between the reviewers. During the study a paper sheet with fraction tasks was shown to the students and they were asked to paint – by means of a software for painting with a computer – their solution and they were prompt to explain aloud their observations and answers. The fraction tasks were subdivided into similar subtasks and covered exercises like assigning fractions to coloured parts of a circle or rectangle, reducing, adding or subtracting fractions and fraction equivalence. Originally, there were 10 tasks with 1 up to 10 subtasks but not each task was seen by each student. We made a screen recording to record the painting of the students and an acoustic recording to record the speech of the students. The screen recordings were used for the retrospective annotation. The acoustic speech recordings, consisting of 10 wav files with a length from 15 up to 20 minutes, were used to gain the speech and pause histograms. The data collection resulted in 36 examples (tasks) labelled with *over-challenged* (12 examples) or *appropriately challenged* (24 examples), respectively 48 examples (24 of class *appropriately challenged*, 24 of class *over-challenged*) after applying oversampling to the smaller set of examples of class *over-challenged* to eliminate the unbalance in the data.

### 3.2 Histograms for Classification

In the above mentioned study we observed that the children often exhibited longer pauses of silence while thinking about the problem when they were *over-challenged* or produced fewer and shorter pauses while communicating when they were *appropriately challenged*. Hence, in this paper we investigate information about pauses and speech segments within the speech input of students in connection with the perceived task difficulty. The first step to gain this information is to segment the acoustic speech recordings for identifying segments containing speech and segments corresponding to pauses. The most easy way to do this is to define a threshold on the decibel scale as done e.g. in [8]. For our study of the data we also used a threshold, which was estimated manually. The manual threshold estimation was done by extracting the amplitudes of the sound files, computing the decibel values and generating a graphic of it like the one in fig. 1. Subsequently, it was investigated which decibel values belong to speech and which ones to pauses to create from this information an appropriate threshold. By means of this threshold the pause and speech segments can be extracted. From the pause segments the pause histogram is generated by counting how often each possible pause length occur. This pause histogram is then normalised, to make the pause histograms of different speech inputs (of different students, different tasks and different lengths) comparable. The normalisation is done by dividing each occurring pause length by the length of the whole speech input as well as dividing the frequency of each occurring pause length by the number of all speech and pause segments, so that the resulting values stem

from the interval between 0 and 1. The same is done with the speech segments for generating the speech histogram. Examples of normalised pause histograms and speech histograms are shown in fig. 2 and fig. 3. The examples stem from the speech input for a task of four different students, where two were labelled as *over-challenged* and the other two as *appropriately challenged*. One can see some

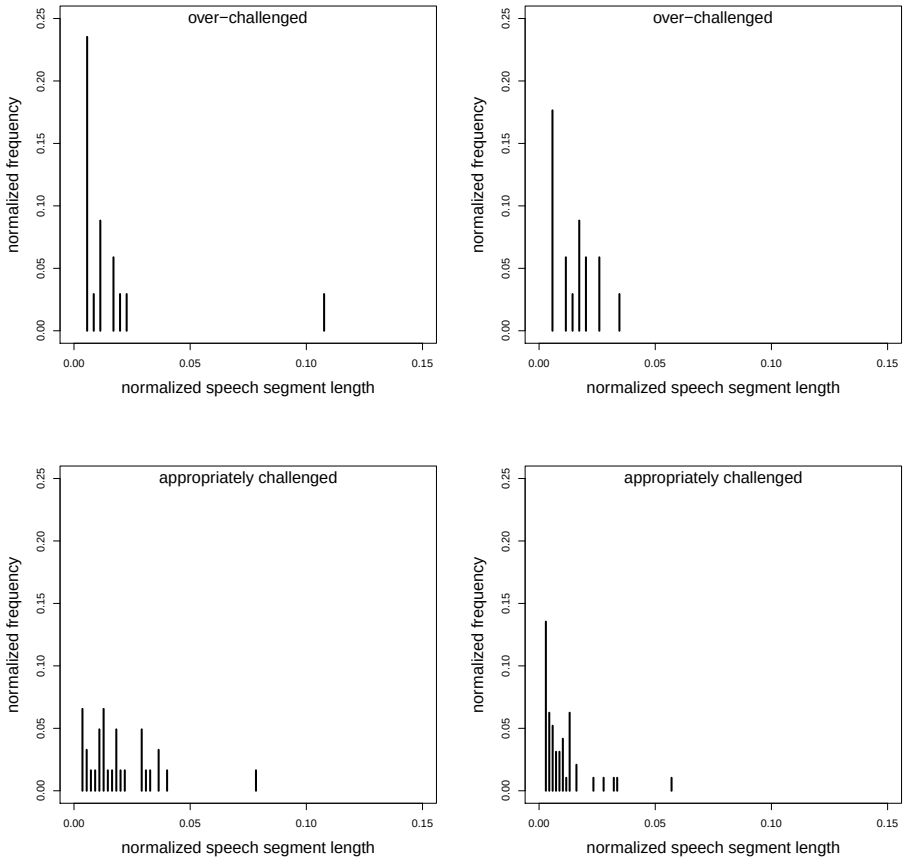



**Fig. 3.** Normalised speech histograms for a task of four different students, where two are labelled as *over-challenged* and the other two as *appropriately challenged*.

differences between the histograms of the *over-challenged* students and the *appropriately challenged* students as well as some similarities of the examples with the same label. The pause histograms of the *appropriately challenged* students show that there are a lot of very small pauses within their speech, but no very large pauses. The pause histograms of the *over-challenged* students in contrast

report long pauses and less smaller pauses than for the *appropriately challenged* students. In the speech histograms one can see that the *over-challenged* students used a lot of very small speech segments of the same length whereas for *appropriately challenged* students there is a large variance in the speech segment length. In the following section we investigate how these histograms can be used for classifying the speech input of a student for a task as either *over-challenged* or *appropriately challenged*.

## 4 Experiments

To investigate if the above described speech and pause histograms are applicable for distinguishing *over-challenged* and *appropriately challenged* students we conducted experiments with the perprocessing and settings described in the following section. The experimental results are reported in sec. 4.2.

### 4.1 Preprocessing and Experimental Settings

To be computationally comparable the normalised histograms still need to be preprocessed, or more explicitly generalised, as the set of possibly occurring segment lengths is infinite (it is a real value between 0 and 1). Hence, we divide the x-axis (the different normalised lengths of pause or speech segments) into a number of equal sized intervals, the *buckets*. Each occurring normalised segment length is then put into the bucket to whose interval it belongs. The number of buckets, or the bucket size respectively, is a hyper parameter and in the experiments we investigated different values for that parameter, i.e. we conducted experiments with 2 up to $1,000,000$ buckets (bucket size 0.5 up to 1.0E-6) where the numbers of buckets are multiples of the numbers by which 100 is divisible without remainder. A comparison of two different histograms can now be done by comparing the content of each bucket in both histograms, that means that for each bucket the normalised frequencies of segments belonging to that bucket are compared. In our experiments we compute the difference between two histograms by computing the differences between the frequencies in all buckets by means of the root mean square error (RMSE):

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{b}(b_i(H_x) - b_i(H_y))^2}{b}}\,, \tag{1}$$

where $H_x$ and $H_y$ are the two histograms to compare, $b_i(H_x)$ and $b_i(H_y)$ are the normalised frequency values belonging to bucket $b_i$ of $H_x$ and $H_y$ and $b$ is the number of buckets. For deciding to which class (*over-challenged* or *appropriately challenged*) a histogram belongs we applied the K-Nearest-Neighbour (KNN) approach. KNN (see e.g. [3]) classifies an example by a majority vote of its neighbours, that is the example is assigned to the class most common among its K nearest neighbours. These K nearest neighbours are the K closest training examples in the feature space. The *closeness* in our case is measured by means

of the RMSE. That is a histogram is assigned to that class to which the majority of the K closest (in terms of RMSE) histograms belongs. K is a further hyper parameter and also for that parameter we tried out different values, i.e. we conducted experiments with a number of 1 up to 35 neighbours where that value is an odd number less than the number of unique examples. For the evaluation we used a Leave-one-out cross-validation in the experiments. The results of our experiments with pause and speech histograms are discussed in the next section.

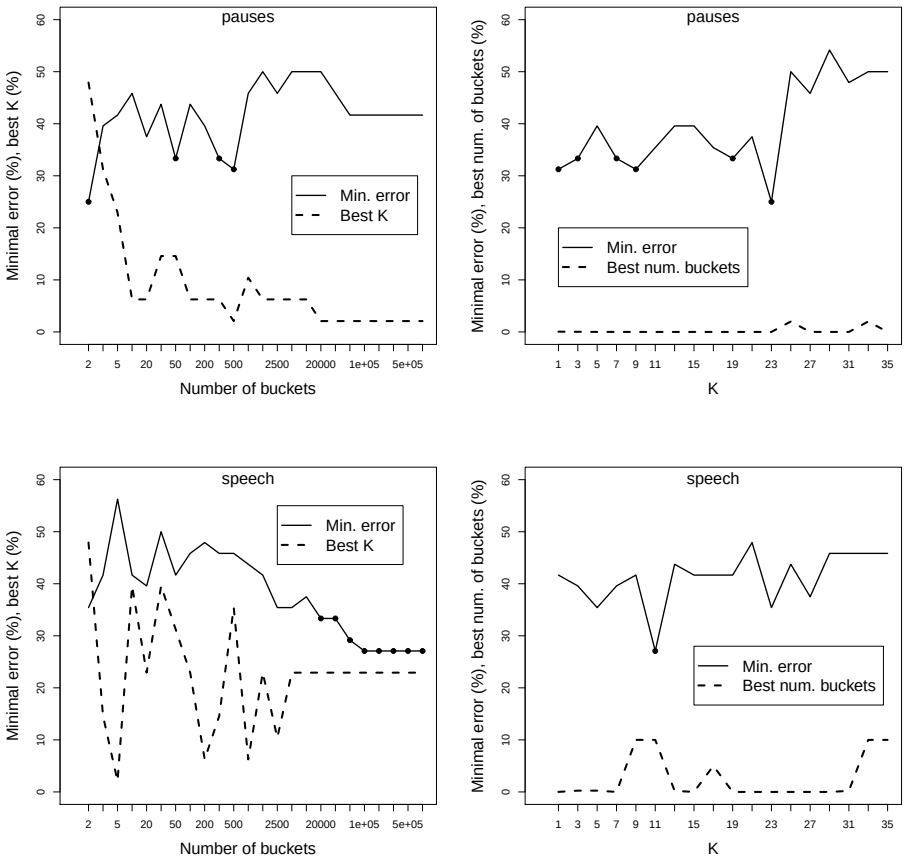



**Fig. 4.** Different numbers of buckets and different numbers K of neighbours mapped to the minimal classification error (%) and the belonging best value for K (% of the number of examples) and for the number of buckets (% of the max. number of buckets) for pause and speech histograms.

### 4.2 Experiments with Speech and Pause Histograms

As mentioned above, we conducted experiments with different numbers of buckets and different values for the K nearest neighbours. In fig. 4 we report the minimal classification error and the belonging best value of K for each bucket number as well as the the minimal classification error and the belonging best number of buckets for each value of K for the pause and the speech histograms. The *classification error* is the number of incorrectly classified histograms divided by the number of all histograms. The black dots in fig. 4 indicate the best results which are also reported in tab. 1 and 2. As one can see in fig. 4 for the

**Table 1.** Number of buckets, bucket size, K, classification error and F-measures of class *over-challenged* & *appropriately challenged* of the experiments with pause histograms with best result (classification error $< 34\%$, black dots in fig. 4).

| Number of buckets | 2 | 2 | 2 | 50 | 250 | 500 |
|---|---|---|---|---|---|---|
| Bucket size | 0.5 | 0.5 | 0.5 | 0.02 | 0.004 | 0.002 |
| K | 9 | 19 | 23 | 7 | 3 | 1 |
| Error (%) | 31.25 | 33.33 | 25.00 | 33.33 | 33.33 | 31.25 |
| F-measure | 0.57, 0.82 | 0.55, 0.80 | 0.67, 0.83 | 0.59, 0.63 | 0.59, 0.57 | 0.60, 0.71 |

**Table 2.** Number of buckets, bucket size, K, classification error and F-measures of class *over-challenged* & *appropriately challenged* of the experiments with speech histograms with best result (classification error $< 34\%$, black dots in fig. 4).

| Number of buckets | 20000 | 25000 | 50000 | 100000 | 200000 | 250000 | 500000 | 1000000 |
|---|---|---|---|---|---|---|---|---|
| Bucket size | 5.0E-5 | 4.0E-5 | 2.0E-5 | 1.0E-5 | 5.0E-6 | 4.0E-6 | 2.0E-6 | 1.0E-6 |
| K | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 |
| Error (%) | 33.33 | 33.33 | 29.17 | 27.08 | 27.08 | 27.08 | 27.08 | 27.08 |
| F-measure | 0.57, 0.73 | 0.57, 0.73 | 0.62, 0.78 | 0.64, 0.77 | 0.64, 0.77 | 0.64, 0.77 | 0.64, 0.77 | 0.64, 0.77 |

pause histograms a smaller number of buckets delivers the best results whereas for the speech histograms the number of buckets has to be large, i.e. a more fine granulated division of the x-axis is needed for good results. The reason might be that the pause histograms of *over-challenged* and *appropriately challenged* students are easier distinguishable as in the pause histogram of an *over-challenged* student there are typically long pause segments which usually do not occur in the speech of *appropriately challenged* students (see also fig. 2). As fig. 3 shows, speech histograms of *over-challenged* and *appropriately challenged* students are not so easy to distinct. Tab. 1 and 2 show the results of the best choices for hyper parameter K and number of buckets and reports the classification error as well as the F-measures of both classes (*over-challenged* and *appropriately challenged*).

The F-measure is a value between 0 and 1 and the closer it is to 1 the better. It is the harmonic mean between the ratio of examples of a class $c$ which are correctly recognised as members of that class (*recall*) and the ratio of examples classified as belonging to class $c$ which actually belong to class $c$ (*precision*). In our experiments the F-measures of class *appropriately challenged* are better than those of class *over-challenged*. The reason could be that originally there were more examples of class *appropriately challenged* and we just oversampled class *over-challenged* to receive a balanced example set. Nevertheless, the best classification errors of 25% and 27.08% and F-measures $0.67, 0.83$ and $0.64, 0.77$ in tab. 1 and 2 indicate that speech and pause histograms are applicable for perceived task difficulty recognition.

## 5 Conclusions and Future Work

We proposed and investigated speech and pause histograms, build from the sequences of speech and pause segments within the speech input of students, as features for perceived task difficulty recognition. To evaluate the approach of using the histograms for distinguishing *over-challenged* and *appropriately challenged* students we applied a K-Nearest-Neighbour classification delivering a classification error of 25% for pause histograms and 27.08% for speech histograms. Next steps will be to try out other classification approaches, for instance from time series classification. Furthermore, the information from the speech histograms and pause histograms could be combined to reach a better classification performance, e.g. by ensemble methods.

**Acknowledgements.** This work is co-funded by the EU project iTalk2Learn (www.italk2learn.eu) under grant agreement no. 318051.

## References

1. Arroyo, I., Woolf, B.P., Burelson, W., Muldner, K. , Rai, D. and Tai, M.: A Multimedia Adaptive Tutoring System for Mathematics that Addresses Cognition, Metacognition and Affect. In International Journal of Artificial Intelligence in Education, Springer, Vol. 24, pp. 387–426 (2014)
2. Baker, R.S.J.D., Gowda, S., Wixon, M., Kalka, J., Wagner, A., Salvi, A., Aleven, V., Kusbit, G., Ocumpaugh, J. and Rossi,L.: Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra. In Proceedings of the 5th International Conference on Educational Data Mining (EDM 2012), pp. 126–133 (2012)
3. Cover, T. and Hart, P.: Nearest neighbor pattern classification. EEE Transactions on Information Theory, Vol. 13(1), pp. 21–27, doi:10.1109/TIT.1967.1053964 (1967)
4. Epp, C., Lippold, M. and Mandryk, R.L.: Identifying Emotional States Using Keystroke Dynamics. In Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems (CHI 2011), pp. 715–724 (2011)
5. Janning, R., Schatten, C., Schmidt-Thieme, L.: Multimodal Affect Recognition for Adaptive Intelligent Tutoring Systems. In Extended Proceedings of the 7th International Conference on Educational Data Mining (EDM 2014), pp. 171–178 (2014)

6. Janning, R., Schatten, C., Schmidt-Thieme, L.: Feature Analysis for Affect Recognition Supporting Task Sequencing in Adaptive Intelligent Tutoring Systems. In Proceedings of the European Conference on Technology Enhanced Learning (EC-TEL 2014), pp. 179–192 (2014)
7. Janning, R., Schatten, C., Schmidt-Thieme, L. and Backfried, G.: An SVM Plait for Improving Affect Recognition in Intelligent Tutoring Systems. In Proceedings of the IEEE International Conference on Tools with Artificial Intelligence (ICTAI) (2014)
8. Luz, S.: Automatic Identification of Experts and Performance Prediction in the Multimodal Math Data Corpus through Analysis of Speech Interaction. Second International Workshop on Multimodal Learning Analytics, Sydney Australia (2013)
9. Mavrikis, M.: Data-driven modelling of students interactions in an ILE. In Proceedings of the International Conference on Educational Data Mining (EDM 2008), pp. 87–96 (2008)
10. DMello, S.K., Craig, S.D., Witherspoon, A., McDaniel, B. and Graesser, A.: Automatic detection of learners affect from conversational cues. User Model User-Adap Inter, DOI 10.1007/s11257-007-9037-6 (2008)
11. D'Mello, S.K. and Graesser, A.: Language and Discourse Are Powerful Signals of Student Emotions during Tutoring. IEEE Transactions on Learning Technologies, Vol. 5(4), pp. 304–317, IEEE Computer Society (2012)
12. Moore, J.D., Tian, L. and Lai, C.: Word-Level Emotion Recognition Using High-Level Features. Computational Linguistics and Intelligent Text Processing (CICLing 2014), pp. 17–31 (2014)
13. Morency, L.P., Oviatt, S., Scherer, S., Weibel, N. and Worsley, M.: ICMI 2013 grand challenge workshop on multimodal learning analytics. In Proceedings of the 15th ACM on International conference on multimodal interaction (ICMI 2013), pp. 373–378 (2013)
14. Pardos, Z.A., Baker, R.S.J.D, San Pedro, M., Gowda, S.M. and Gowda, S.M.: Affective States and State Tests: Investigating How Affect and Engagement during the School Year Predict End-of-Year Learning Outcomes. Journal of Learning Analytics, Vol. 1(1), Inaugural issue, pp. 107–128 (2014)
15. Purandare, A. and Litman, D.: Humor: Prosody Analysis and Automatic Recognition for F * R * I * E * N * D * S *. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pp. 208–215 (2006)
16. San Pedro, M.O.C., Baker, R.S.J.D., Bowers, A. and Heffernan, N.: Predicting College Enrollment from Student Interaction with an Intelligent Tutoring System in Middle School. In Proceedings of the 6th International Conference on Educational Data Mining (EDM 2013), pp. 177–184 (2013)
17. Schuller, B., Batliner, A., Steidl, S. and Seppi, D.: Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication, Elsevier (2011)
18. Worsley, M. and Blikstein, P.: What's an Expert? Using Learning Analytics to Identify Emergent Markers of Expertise through Automated Speech, Sentiment and Sketch Analysis. In Proceedings of the 4th International Conference on Educational Data Mining (EDM '11), pp. 235–240 (2011)

# Analyzing Student Action Sequences and Affect While Playing Physics Playground

Juan Miguel L. Andres[1], Ma. Mercedes T. Rodrigo[1],
Ryan S. Baker[2], Luc Paquette[2], Valerie J. Shute[3], Matthew Ventura[3]

[1] Ateneo de Manila University, Quezon City, Philippines
[2] Teachers College, Columbia University, New York, NY, USA
[3] Florida State University, Tallahassee, FL, USA
{mandres, mrodrigo}@ateneo.edu,
baker2@exchange.tc.columbia.edu, luc.paquette@gmail.com,
{vshute, mventura}@fsu.edu

**Abstract.** Physics Playground is an educational game that supports physics learning. It accepts multiple solutions to most problems and does not impose a stepwise progression through the content. Assessing student performance in an open-ended environment such as this is therefore challenging. This study investigates the relationships between student action sequences and affect among students using Physics Playground. The study identified most frequently traversed student action sequences and investigated whether these sequences were indicative of either boredom or confusion. The study found that boredom relates to poor performance outcomes, and confusion relates to sub-optimal performance, as evidenced by the significant correlations between the respective affective states, and the student action sequences.

**Keywords:** Affect modeling, action sequences, boredom, confusion, Physics Playground

## 1 Introduction

Physics Playground (PP) is an educational game that immerses learners in a choice-rich environment for developing intuitive knowledge about simple machines. As the environment does not impose a stepwise sequence on the learner, and because some problems can have multiple solutions, learners have the freedom to explore, attempt to solve, or abort problems as they wish. The challenge these types of environments impose on educators is that of assessment. Within such an open-ended system, how do educators and researchers assess learning as well as the quality of the learning process?

This study focuses its attention on two main phenomena: student learning and student affect. Student learning within PP refers to how well a player can understand the concepts surrounding four simple machines through their efficient execution in attempting to solve levels, as evidenced by the badges they earn.

Student affect refers to experiences of feelings or emotions. In this study, the affective states of interest are confusion and boredom, as prior studies have shown them to relate significantly with learning [4, 10]. Confusion is uncertainty about what to do next [5]. Confusion is scientifically interesting because it has a positive and negative dimension, wherein it either spurs learners to exert effort deliberately and purposefully to resolve cognitive conflict, or leads learners to become frustrated or bored, and may lead to disengagement from the learning task altogether [7].

Boredom, on the other hand, is an unpleasant, transient affective state in which the individual feels a pervasive lack of interest in and difficulty concentrating on the current activity [8]. Boredom has been a topic of interest because of the negative effects usually associated with it, such as poor long-term learning outcomes when students are not provided any scaffolding [10] and its being characteristic of less successful students [11].

A study conducted by Biswas, Kinnebrew, and Segedy [2] investigated frequently traversed sequences of student actions using bottom-up, data-driven sequence mining, the results of which contributed to the development of performance- and behavior-based learner models. The analyses in this paper seek to perform similar sequence-mining methods in order to find student sequences that inform either of the affective states of interest.

This study conducted data-driven sequence-mining analyses to answer the following research questions:

1. What were the frequently traversed student action sequences among students playing Physics Playground?
2. Are these action sequences indicative of either boredom or confusion?

The analyses in this study are limited to the data collected during gameplay of Physics Playground from six data gathering sessions conducted at a public school in Quezon City in 2013. Data is limited to the interaction logs generated by the game as well as human observation of affect as logged by two coders trained in the Baker-Rodrigo-Ocumpaugh Monitoring Protocol [9].

## 2 Methodology

### 2.1 Participant Profile

Data were gathered from 60 eighth grade public school students in Quezon City, Philippines. Students ranged in age from 13 to 16. Of the participants, 31% were male and 69% were female. As of 2011, the school had 1,976 students, predominantly Filipino, and 66 teachers. Participants had an average grade on assignments of B (on a scale from A to F).

### 2.2 Physics Playground

Physics Playground (PP) is an open-ended learning environment for physics that was designed to help secondary school students understand qualitative physics. Qualitative physics is a nonverbal, conceptual understanding of how the physical world operates [12].

PP has 74 levels that require the player to guide a green ball to a red balloon. An example level is shown in Fig. 1. The player achieves this goal by drawing agents (ramps, pendulums, springboards, or levers) or by nudging the ball to the left or right by clicking on it. The moment the objects are drawn, they behave according to the law of gravity and Newton's 3 laws of motion [12].

**Fig. 1.** Example PP level.

**Performance Metrics.** Gold and silver badges are awarded to students who manage to solve a level. A gold badge is given to a student who is able to solve the level by drawing a number of objects equal to the particular level's par value (i.e., the minimum number of objects needed to be drawn to solve the level). A student who solves a level using more objects will earn a silver badge. A student earns no badge if he was not able to solve the level. Many levels in PP have multiple solutions, meaning a player can solve the level using different agents.

### 2.3 Interaction Logs

During gameplay, PP automatically generates interaction log files. Each level a student plays creates a corresponding log file, which tracks every event that occurs as the student interacts with the game. Per level attempt, PP tracks begin and end times, the agents used, and the badges awarded upon level completion. PP also logs the *Freeform Objects* that player draw, or objects that cannot be classified as any of the four agents. The physics agents within PP are as follows:

- Ramp, any line drawn that helps to guide a ball in motion,
- Lever, an agent that rotates around a fixed point, usually called a fulcrum,
- Pendulum, an agent that directs an impulse tangent to its direction of motion,
- Springboard, an agent that stores elastic potential energy provided by a falling weight.

### 2.4 The Observation Protocol

The Baker-Rodrigo-Ocumpaugh Monitoring Protocol (BROMP) is a protocol for quantitative field observations of student affect and engagement-related behavior, described in detail in [9]. The affective states observed within Physics Playground in this study were engaged concentration, confusion, frustration, boredom, happiness, delight, and curiosity. The affective categories were drawn from [6].

BROMP guides observers in coding affect through different utterances, body language, and interaction with the software specific to each affective state. A total of seven affective states were coded, however, this study focuses on three: concentration, confusion, and boredom. These were identified as follows:

1. Concentration — immersion and focus on the task at hand, leaning toward the computer and attempting to solve the level, a subset of the flow experience described in [5].
2. Confusion — scratching his head, repeatedly attempting to solve the same level, statements such as "I don't understand?" and "Why didn't it work?"
3. Boredom — slouching, sitting back and looking around the classroom for prolonged periods of time, statements such as "Can we do something else?" and "This is boring!"

Following BROMP, two trained observers observed ten students per session, coding students in a round-robin manner, in 20-second intervals throughout the entire observation period of 2 hours. During each 20-second window, both BROMP observers code the current student's affect independently. If the student exhibited two or more distinct states during a 20-second observation window, the observers only coded the first state. The inter-coder reliability for affect for the two observers in the study was acceptably high with a Cohen's Kappa [3] of 0.67. The typical threshold for certifying a coder in the use of BROMP is 0.6, a standard previously used in certifying 71 coders in the use of BROMP (e.g., [9]).

The observers recorded their observations using HART, or the Human Affect Recording Tool. HART is an Android application developed to guide researchers in conducting quantitative field observations according to BROMP, and facilitate synchronization of BROMP data with educational software log data.

### 2.6 Data Collection Process

Before playing PP, students answered a 16-item multiple-choice pretest for 20 minutes. Students then played the game for 2 hours, during which time two trained observers used BROMP to code student affect and behavior on the HART application. A total of 4,320 observations were collected (i.e., 36 observations per participant per each of the two observers). After completing gameplay, participants answered a 16-item multiple-choice posttest for 20 minutes. The pretest and posttest were designed to assess knowledge of physics concepts, and have been used in previous studies involving PP [12].

To investigate how students interacted with PP, the study made use of the interaction logs recorded during gameplay to analyze student performance. Of the 60 participants, data from 11 students were lost because of faulty data capture and

corrupted log files. Only 49 students had complete observations and logs. As a result, the analysis in this paper is limited to these students, and the 3,528 remaining affect observations. Engaged concentration was observed 72% of the time, confusion was observed 8% of the time, and boredom and frustration were observed 7% of the time. Happiness, delight, and curiosity comprise the remaining 6% of the observation time.

## 3 Analyses and Results

### 3.1 Agent Sequences

All PP-generated logs were parsed and filtered to produce a list containing only the events relevant to the study. Sequences were then separated into one of two categories: 1) silver sequences, or the sequences that ultimately led to a silver badge, which comprised 44% of all level attempts, and 2) unsolved sequences, or the sequences that led to the student quitting the level without finding a solution, which comprised 39% of all level attempts. Sequences that ended in gold badges were dropped from the analysis because they only comprised 17% of all level attempts.

Every time a student earns a badge after solving a level, the badge is awarded for one of the four agents (e.g., a player is awarded a silver ramp badge for solving the level using a ramp, and another player is awarded a gold pendulum badge for solving another level using a pendulum). We tracked the agents the badges were awarded for per level, and used this list of badges to relabel the sequences based on correctness. If the level awarded a badge for an agent, that agent was labeled as `correct` for that level; if not, the agent was labeled as `wrong` for the level. For example, on a level that awarded badges for springboards and levers, a sequence of `Lever > Ramp > Springboard > Level End (silver-springboard)` would be relabeled as `correct > wrong > correct > Level End (silver)`.

The relabeling was done because most of the sequences were level-dependent, that is, a majority of some sequences appeared on only one or two levels. By relabeling based on correctness, we were able to ensure level-independence among sequences. Sequences were tabulated and their frequencies calculated (i.e., how many times each of the 49 students traversed each of the sequences). We calculated for distribution of sequence frequencies, and the sequences we found to occur rarely (i.e., less than 30% of the population traversed them) were dropped from the analysis. We found that the gold sequences occurred rarely, which was another reason they were dropped from the analysis. The resulting silver and unsolved sequences can be found in Tables 1 and 2, respectively, along with the frequency means and standard deviations.

Table 1 lists the top 7 silver sequences within PP, which were traversed by more than 30% of the study's population. The Sequences column shows what the respective sequences look like, and the Frequency column shows the average number of times the 49 students traversed them and the standard deviations.

Highlighted sequences showed significant correlations with either boredom or confusion, as discussed further in Section 3.2. Table 2 is presented in the same manner.

**Table 1.** Top 7 silver sequences, their traversal frequency means, and standard deviations.

| | Sequences | Frequency | |
|---|---|---|---|
| | | Mean | SD |
| 1 | correct>Level End (silver) | 3.53 | 2.34 |
| 2 | Level End (silver) | 2.61 | 2.33 |
| 3 | wrong>Level End (silver) | 1.90 | 1.37 |
| 4 | correct>correct>Level End (silver) | 1.61 | 1.15 |
| 5 | wrong>correct>Level End (silver) | 0.90 | 1.01 |
| 6 | correct>correct>correct>Level End (silver) | 0.80 | 1.00 |
| 7 | wrong>correct>correct>Level End (silver) | 0.69 | 0.77 |

The silver sequences in Table 1 show signs of experimentation, with students playing around with the correct and incorrect agents to solve the levels, as seen in sequences 5 and 7. Sequences 1, 4, and 6 show students using the correct agents, but are unable to earn gold badges. This suggests that students, while knowing which agents to use, do not have a full grasp of the physics concepts surrounding the agents' execution. Sequence 3 shows students using wrong objects to solve the levels. While this may suggest that students are still struggling to understand how the agents work and which agent would best solve a level given the ball and the balloon's positions, this may have also been caused by the PP logger labeling the objects they drew as freeform objects, and not one of the correct agents.

Sequence 1 shows the students drawing only the correct agent, but are still unable to earn a gold badge. The sequence-mining algorithm only pulled events related to drawing any of the four main agents, which are enumerated in Section 2.3. Drawing a lever or a springboard, for example, would require drawing more than one component. A lever requires the fulcrum, the board, and the object dropped on the board to project the ball upwards. In order for the agent to work, it has to be executed correctly (i.e., the board must be long enough, with the fulcrum in the right position, and the object dropped on the board must be heavy enough to propel the ball into the air). Sequence 1 may have been caused by students drawing the correct agent, but improperly executing it. For example, the student may not have drawn the right-sized weight to drop on the lever, and thus had to draw another. While drawing another weight to drop on the lever counts towards the level's object count, it was not logged as a separate event by the sequence mining analysis because the player did not draw another agent, only a component of it. Sequence 2, on the other hand, is suspect because despite the student drawing no objects to solve a level, he ends up with only a silver badge. This was most likely caused by the improper logging of the game. The top 7 most frequently traversed silver sequences account for 58% of the total number of silver sequences.

**Table 2.** Top 6 unsolved sequences, their traversal frequency means, and standard deviations.

| | Sequences | Frequency | |
|---|---|---|---|
| | | Mean | SD |
| 1 | `Level End (none)` | 10.69 | 8.17 |
| 2 | `wrong>Level End (none)` | 1.55 | 1.65 |
| 3 | `correct>Level End (none)` | 1.29 | 1.50 |
| 4 | `wrong>wrong>Level End (none)` | 0.45 | 0.65 |
| 5 | `correct>correct>Level End (none)` | 0.41 | 0.73 |
| 6 | `wrong>correct>Level End (none)` | 0.39 | 0.57 |

Table 2, which shows the top 6 unsolved sequences, shows signs of students giving up. Sequence 1 shows students giving up without even drawing a single object, which could have been caused by one of two things: 1) the student saw the level and decided to quit without attempting to solve it, or 2) again, the logger did not log the objects correctly. This sequence is similar to one of the silver sequences in that no objects were drawn. What makes them different, however, is what the sequences ultimately led to. The silver sequences ended in a silver badge, and the unsolved sequences ended in the student earning no badge. The majority of the sequences listed in Table 2 show students experimenting mainly with wrong objects, whether agents or freeform objects. This implies that the students are lacking in the understanding of how to solve the levels. Sequences 3 and 5 are interesting because it is unclear whether or not the students understood the concepts of the agents. That is, students were drawing the correct agents, but could not get the ball to reach the balloon. Despite drawing one or two correct agents, the students decided to give up and quit. The top 6 unsolved sequences account for 81% of the total number of unsolved sequences.

### 3.2 Relationship with Affect

We computed frequencies for each of the 13 sequences that the 49 students traversed. Correlations were then run between each of the 13 arrays and the incidences of confusion and boredom. Because the number of tests introduces the possibility of false discoveries, Storey's adjustment [13] was used as a post-hoc control, which provides a $q$-value, representing the probability that the finding was a false discovery. Tables 3 and 4 show the results. Highlights and asterisks (*) were used on significant findings ($q \leq 0.05$).

Table 3 lists the top 7 most frequently traversed silver sequences, from left to right. The sequences these header numbers represent can be found in Table 1. The table shows the correlation between each of the top 7 silver sequences using a metric that represents the percentage of all attempts that match each of the sequences, the percentage of time the students were observed to be confused (r, con), and the percentage of time the students were observed to be bored (r, bor).

Table 4 is presented in the same manner, with sequence information in Table 2 for the top 6 unsolved sequences.

**Table 3.** Correlations between top 7 silver sequences, confusion, and boredom.

| | Top 7 silver sequences | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| r, con | -0.33 | 0.23 | 0.41* | 0.03 | 0.17 | 0.54* | 0.28 |
| r, bor | -0.20 | -0.17 | -0.19 | -0.05 | 0.14 | -0.19 | -0.20 |

Table 3 shows two significant positive correlations between confusion and the silver sequences. The two sequences showed signs of lesser understanding of the agents. Sequence 3 shows students using only a wrong object to solve a level, which may have been caused either by incorrect object labeling (e.g., PP logged a ramp as a Freeform Object), or the student found a different way of solving the level. Like in most learning environments, players are able to game the system – or systematically misuse the game's features to solve a level [1] – within PP through stacking. Stacking is done when players draw freeform objects to either prop the ball forward or upward, which may have been the case in sequence 3. Sequence 6 shows students drawing only correct agents. These sequences having significant correlations with confusion may imply lesser understanding among confused students as the they are not only dealing with proper agent execution, but also with deciding which agent would best solve the level. Despite the challenges faced by these students, however, they still managed to find a solution to the level. Our findings suggest that the inability to grasp the physics concepts surrounding the agents is a sign of confusion.

**Table 4.** Correlations between top 6 unsolved sequences, confusion, and boredom.

| | Top 6 unsolved learning sequences | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| r, con | -0.17 | 0.00 | -0.12 | -0.01 | -0.06 | 0.04 |
| r, bor | -0.12 | 0.13 | 0.12 | -0.03 | 0.48* | 0.06 |

Table 4 shows that one of the most frequently traversed unsolved sequences has a significant positive correlation with boredom. This sequence shows students using only correct agents, but ultimately deciding to give up. This may have been caused by the inability to execute the agents correctly, which may imply that, unlike confused students, bored students were not likely to exert additional effort to try to solve the level or understand proper agent execution. As mentioned previously, boredom has been found to have significant relationships with negative performance outcomes. In this case, sequences all ultimately led to disengagement: students quitting the level before finding a solution, showing signs of giving up and lack of understanding of any of the four agents.

## 4 Conclusions and Future Work

This study sought to identify the most frequently traversed student action sequences among eighth grade students while interacting with an education game for physics called Physics Playground. Further, the study sought to investigate how these sequences may be indicative of affective states, particularly boredom and confusion, which have been found to significantly affect student learning.

Data-driven sequence mining techniques were conducted to identify most frequently traversed actions sequences in two categories: the sequences that would eventually lead the student to a silver badge, and the paths that would eventually lead the student to not earning a badge.

In the silver sequences, students played around with freeform objects and some of the four agents in attempting to solve the level. The study found confusion to correlate significantly with two of the silver sequences, which supports previous findings regarding the relationship between confusion and in-game achievement, which suggest that because students are unable to grasp the concepts surrounding the agents and their executions, students resort to finding other solutions.

In the unsolved sequences, students would give up and quit without finding a solution, despite already using the correct agents to solve the level. The study found boredom to correlate significantly with one of the unsolved sequences. This finding supports the literature that has shown that boredom relates to poor learning outcomes. This work provides further evidence that boredom and disengagement from learning go hand-in-hand.

This study provides specific sequences of student actions that are indicative of the boredom and confusion, which has implications on the design and further development of Physics Playground. This study also contributes to the literature by providing empirical support that boredom and confusion are affective states that influence performance outcomes within open-ended learning environments, and are thus affective states that learning environments must focus on detecting and providing remediation to. We found that both bored and confused students will tend to continuously use correct agents in attempting to solve levels, but execute them incorrectly. The difference between the two, however, is that confused students tend to end up solving the level, while bored students give up.

The analyses run in this paper were part of a bigger investigation, and as such, there are several interesting ways forward in light of our findings. The paper aims for its findings to contribute to the creation of a tool that can automatically detect affect given a sequence of student interactions, and provide necessary remediation in order to curb student experiences of boredom.

Relationship analyses run between student action sequences and incidences of affect in this paper were done through correlations. However, findings were not able to determine whether boredom or confusion occurred more frequently during specific action sequences. We want to find out whether boredom or confusion occurred before, during, or after the students' execution of the action sequences, and in doing so, see whether or not the affective states were causes or effects of the action sequence executions. We are currently investigating this relationship in a separate study.

**Acknowledgements.** We would like to thank the Ateneo Center for Educational Development, Carmela C. Oracion, Christopher Ryan Adalem, and the officials at Krus Na Ligas High School, Jessica O. Sugay, Dr. Matthew Small, and the Gates Foundation Grant #OP106038 for collaborating with us.

## References

1. Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task behavior in the cognitive tutor classroom: when students game the system. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 383-390). ACM.
2. Biswas, G., Kinnebrew, J. S., & Segedy, J. R. (2011). Using a Cognitive/Metacognitive Task Model to analyze Students Learning Behaviors.
3. Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 20 (1960), 37-46.
4. Craig, S., Graesser, A., Sullins, J., & Gholson, B. (2004). Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. Journal of Educational Media, 29(3), 241-250.
5. Csikszentmihalyi, M. (1990). Flow: The psychology of optimal experience. New Y ork: Harper Perennial.
6. D'Mello, S. K., Craig, S. D., Witherspoon, A., McDaniel, B., & Graesser, A. (2005). Integrating affect sensors in an intelligent tutoring system. In Proceedings of the Workshop on Affective Interactions: The computer in the affective loop workshop, International conference on intelligent user interfaces (pp. 7- 13). New York: Association for Computing Machinery.
7. D'Mello, S., Graesser, A. (2012). Dynamics of affective states during complex learning. Learning and Instruction, 22(2): 145-157.
8. Fisherl, C. D. (1993). Boredom at work: A neglected concept. Human Relations, 46(3), 395-417.
9. Ocumpaugh, J., Baker, R.S., Rodrigo, M.M.T. (2015) Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual. Technical Report. New York, NY: Teachers College, Columbia University. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.
10. Pardos, Z. A., Baker, R. S., San Pedro, M., Gowda, S. M., & Gowda, S. M. (2014). Affective States and State Tests: Investigating How Affect and Engagement during the School Year Predict End-of-Year Learning Outcomes. Journal of Learning Analytics, 1(1), 107-128.
11. San Pedro, M. O. Z., d Baker, R. S., Gowda, S. M., & Heffernan, N. T. (2013, January). Towards an understanding of affect and knowledge from student interaction with an Intelligent Tutoring System. In Artificial Intelligence in Education (pp. 41-50). Springer Berlin Heidelberg.
12. Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and Learning of Qualitative Physics in Newton's Playground. The Journal of Educational Research, 106(6), 423-430.
13. Storey, J.D. (2002). A direct approach to false discovery rates. Journal of the Royal Statistical Society, Series B, 64: 479-498.

# La Mort du Chercheur: How well do students' subjective understandings of affective representations used in self-report align with one another's, and researchers'?

Wixon[1], Danielle Allessio[2], Jaclyn Ocumpaugh[3], Beverly Woolf[2], Winslow Burleson[4] and Ivon Arroyo[1]

[1]Worcester Polytechnic Institute, Worcester Massachusetts
{mwixon, iarroyo}@wpi.edu
[2]University of Massachusetts, Amherst Massachusetts
{allessio@educ, bev@cs}.umass.edu
[3]Teachers College, Columbia University, New York, New York
jocumpaugh@wpi.edu
[4]New York University, New York, New York
Wb50@nyu.edu

**Abstract.** We address empirical methods to assess the reliability and design of affective self-reports. Previous research has shown that students may have subjectively different understandings of the affective state they are reporting [18], particularly among younger students[10]. For example, what one student describes as "extremely frustrating" another might see as only "mildly frustrating." Further, what students describe as "frustration" may differ between individuals in terms of valence, and activation. In an effort to address these issues, we use an established visual representation of educationally relevant emotional differences [3, 8, 25]. Students were asked to rate various affective terms and facial expressions on a coordinate axis in terms of valence and activation. In so doing, we hope to begin to measure the variability of affective representations as a measurement tool. Quantifying the extent to which representations of affect may vary provides a measure of measurement error to improve reliability.

**Keywords:** Affective States; Intelligent Tutoring Systems; Reasons for Affect

## 1 Introduction

The evaluation of students' affective states remains an incredibly difficult challenge. While recognized as a key indicator of student engagement [14, 17, 26], there remains no clear gold-standard for identifying an affective state, leading to researchers such as Graesser & D'Mello [13] to call for greater attention to the theoretical stances that certain research methods entail. A full theoretical review is beyond the scope of this paper; instead, the current work presents a pilot study designed to empirically evaluate the reliability of two different types of affective self-reports in an educational

context. Reliability is measured both in terms of inter-rater reliability (the degree of agreement between students), and "inter-method" reliability (i.e. given words or facial expressions as representations of affective states, which representation produces more consistent results).

A considerable body of research has been devoted to affect computing, and in particular to affect detection in educational software [9]. Progress has been made with methods that include self-report [8, 10], physiological sensors [1, 24], video-based retrospective reports [5, 15], text-based [11, 19], and field observation [16, 23] data. However, much of this research evaluates success based on the ability of a model to predict when a training label is present or absent, without giving deeper consideration to questions about the appropriateness of the training label itself.

Even within limited to the body of research that relies on self-report research, there are serious concerns about how methodological decisions might impact student responses. In addition to issues about the frequency and timing of surveys, one primary area of concern is that students may have subjectively different understandings of the state they are reporting [19], an effect that is likely to be even greater among younger students [10]. For example, Graesser and D'Mello [13] have suggested that a students' tolerance of cognitive disequilibrium (e.g., confusion or frustration) is probably conditioned by their knowledge and prior success with the topic they are interacting with. Further, what students describe as "frustration" in itself may differ between individuals in terms of dimensional component measures of affect: valence, activation, and dominance. The former two dimensions are typically used to differentiate affective states [4], and the latter used in some cases [7].

In this study, we explore these interpretative issues using three different types of representations that have been employed in previous self-report studies: words, facial expressions, and dimensional measures. In particular, we are interested in verifying that students' understanding of the meaning of these representations aligns with interpretations of these labels that are present in the literature (as constructed by experts). To this end, we use dimensional measures (valence & activation) to compare how students respond to both linguistic representations and pictorial representations, further testing hypotheses that the latter might be more appropriate for surveying students [19, 21, 22]. Our goal is to determine the extent to which this student population shows variance in the interpretation of these two different types of representations, since substantial variation in student perception should be accounted for in subsequent research. Last, while we might achieve researcher agreement in terms of methods and terminology for self-reported affects, that will do little good if there is a large degree of variance in terms of our subject pool's agreement on the meaning of these constructs.

### 1.1 Methods

Students surveyed included eighty one 7th graders from two Californian middle schools in a major city (among the 30 most populous cities in California), where a majority of census respondents identified as Hispanic or Latino and median household

income was within one standard deviation of California's overall median household income. They were surveyed at the end of the academic year.


**Fig. 1.** Blank Valence & Activation Sheet given to Students

Students were asked to place both textual and facial representations of affect on an XY axis of Activation=X Valence=Y. Textual representations of affect were selected based on the affective states that have been used in the past [2, 12], that corresponded to quite different levels of activation x valence according to us researchers, so that words would theoretically cover all quadrants. These terms and their researcher-hypothesized valence x arousal placements included: Angry (low valence x high activation), Anxious (low valence x high activation), Bored (low valence x low activation), Confident (high valence x low activation), Confused (med-low valence x med-high activation), Enjoying (high valence x medium activation), Excited (high valence x high activation), Frustrated (low valence x high activation), Interested (high valence x medium activation) and Relieved (high valence x med-low activation). In general, it was clear to the researchers which word corresponded to which face, with a few exceptions, such as the level of activation that should be associated to enjoying and interest. An established set of emoticons were chosen from previous affective research [8] that corresponded to extreme emoticon states of activation x valence x dominance. While the emoticons possessed these three attributes, our participants were asked only to orient them based on activation and valence.

Each student was presented with a sheet of paper depicting a coordinate axis with activation from "sleepy" to "hyper" on the x-axis and "bad" to "great" on the y-axis. These terms were used to express what valence and activation mean experientially, using language that children are familiar with and could relate to. Activation is then expressed more as a physical experience of arousal, while Valence is expressed not as much as a physical experience but as a judgment of the positive or negative nature of

the experience. Later, during coding, these axes were mapped discretized into a seven point scale of -3 to 0 to +3 at either extreme of each axis, defining a grid of 7 x 7.

Students were also given stickers for the 10 separate affective terms: Angry, Anxious, Bored, Confident, Confused, Enjoying, Excited, Frustrated, Interested, & Relieved, see Figure 2; as well as 8 stickers to depict each extreme emoticon expression from the ends of each of the 3 axis coordinate systems including: pleasure, activation, and dominance [8]. Students placed each of these stickers on their coordinate axes according to where they felt each term or emoticon should be placed with respect to valence and activation.

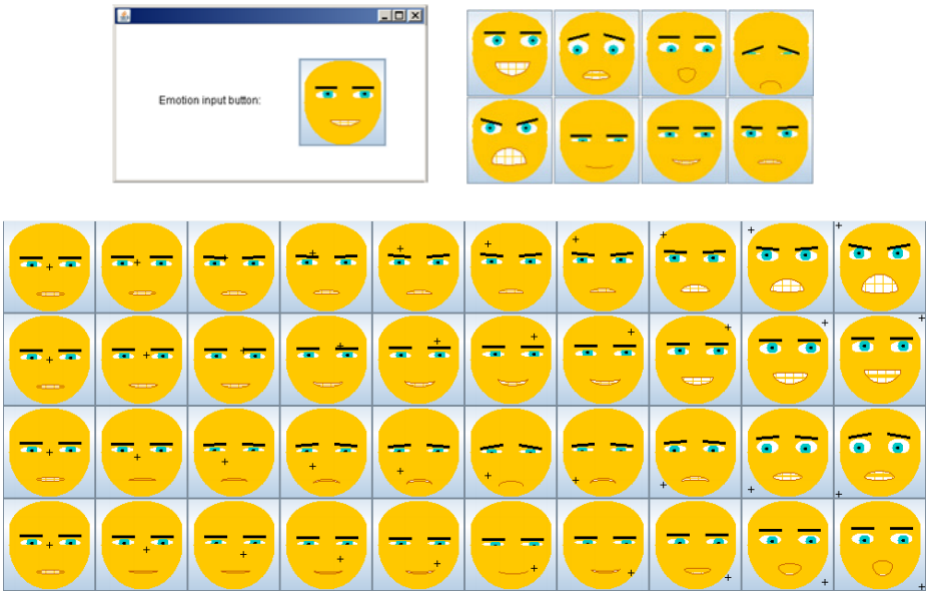



**Fig. 2.** Directly from Broekens, & Brinkman, 2013 [8]. Top left displays the affect button interface. Students use the cursor to change the expression in the inter-face. Depending on their actions, one of 40 affective expressions may be displayed; these expressions, shown across the bottom of this figure, are designed to vary based on pleasure (valence), activation, and dominance (PAD for brevity). From left to right first row: elated (PAD=1,1,1), afraid (-1,1,-1), surprised (1,1,-1), sad (-1,-1,-1). From left to right second row: angry (-1,1,1), relaxed (1,-1,-1), content(1,-1,1), frustrated (-1,-1,1). Top right displays PAD extremes, which serve as the basis for this research.

## 2 Results

Mean positioning results are displayed visually in figure 3, corresponding to the position that each word or emoticon sticker was placed averaged across all respondents. Missing data occurred in which students may not have placed every sticker. On average any given term or emoticon was missing 16.6 reports, with a maximum of 23 students of 81 missing reports for boredom, frustration, and relief. The average stu-

dent was only missing 3.7 out of 18 terms and emoticons from their sheet, and there were 5 students who turned in completely blank sheets.

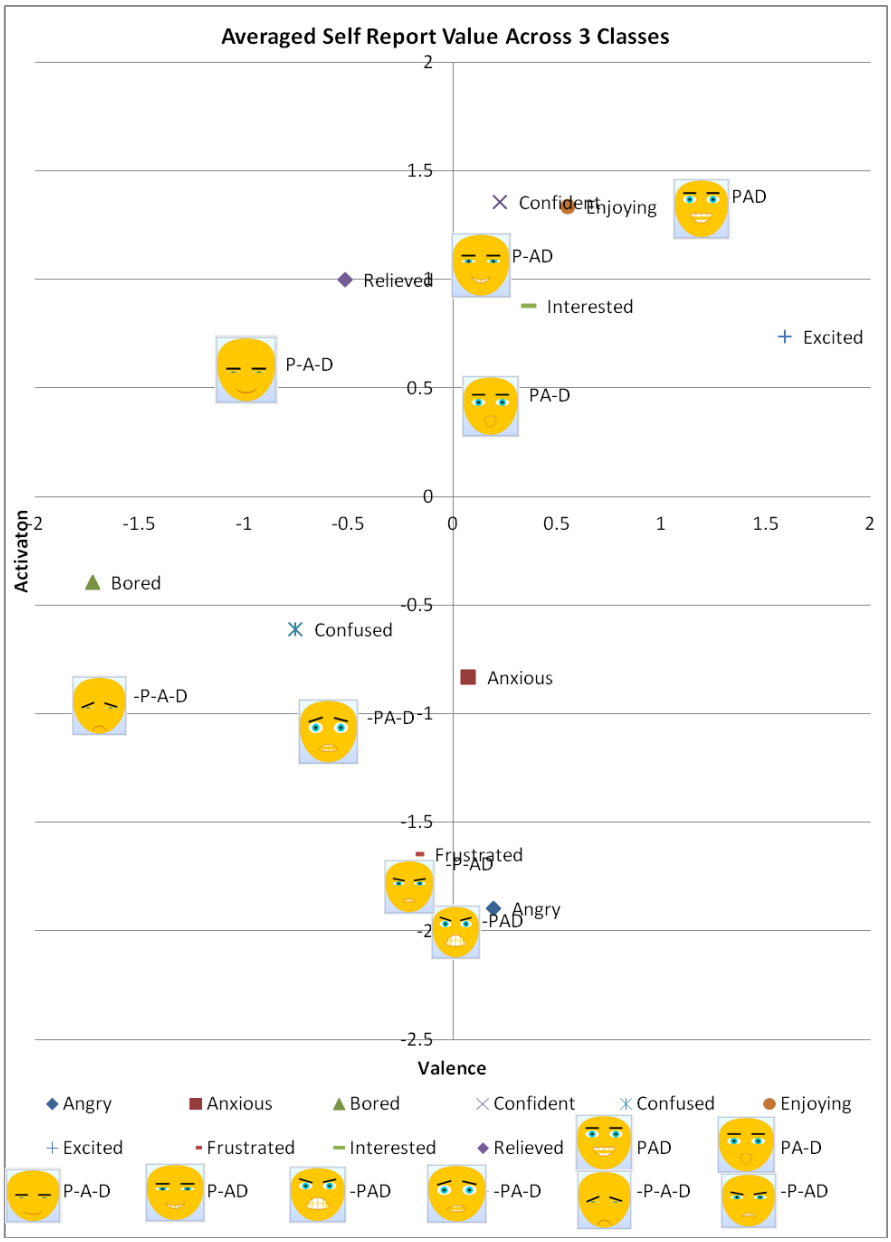



**Fig. 3.** Averaged Placement of Text and Emoticon Stickers

Interestingly, the placement of -PAD and -P-AD (negative sign indicating most extreme negative activation, pleasure, dominance, lack of a negative indicating most extreme positive, see figure 2 caption) match up with their respective terms "Angry" and "Frustrated" very closely. However, while both seem to be at the extreme end of negative valence, on average both seem to be viewed as fairly neutral in terms of activation by students. Although all emoticons and terms fall under the expected half of the coordinate axes in terms of valence (i.e. those we would expect to be pleasurable are categorized as above the origin, those displeasurable below it), activation does not follow this trend. For example anxiety is rated as neutral activation. One possible explanation, consistent with the results, is that students may be grouping activation and dominance together as a single measure. Emoticons with both negative activation and dominance were rated most negatively in terms of activation, those with either negative activation or dominance tended to fall in the middle, and the rating with all positive PAD was the emoticon with the highest rated activation.

| Text or Emoticon | Activation Mean (StdDev) | Valence Mean (StdDev) |
|---|---|---|
| Angry | 0.19 (1.09) | -1.9 (0.99) |
| Anxious | 0.07 (1.78) | -0.87 (1.19) |
| Bored | -1.72 (1.28) | -0.4 (1.02) |
| Confident | 0.23 (1.22) | 1.35 (0.99) |
| Confused | -0.75 (1.36) | -0.61 (1.12) |
| Enjoying | 0.55 (1.18) | 1.34 (1.14) |
| Excited | 1.59 (1.04) | 0.74 (1.26) |
| Frustrated | -0.17 (1.33) | -1.65 (1.05) |
| Interested | 0.36 (1.34) | 0.88 (0.98) |
| Relieved | -0.52 (1.43) | 1 (1.12) |
| Face_PAD | 1.25 (1.3) | 1.38 (1.13) |
| Face_PA-D | 0.28 (1.86) | 0.47 (0.93) |
| Face_P-A-D | -0.89 (1.57) | 0.61 (0.91) |
| Face_P-AD | 0.2 (1.26) | 1.11 (1.08) |
| Face_-PAD | 0.05 (0.95) | -1.95 (0.93) |
| Face_-PA-D | -0.5 (1.39) | -1.01 (1.01) |
| Face_-P-A-D | -1.61 (1.41) | -0.91 (1.11) |
| Face_-P-AD | -0.12 (1.15) | -1.69 (0.89) |
| Average | -0.08 (1.33) | -0.12 (1.05) |

**Table 1.** Means and Standard Deviations of Students' placement of stickers.

One key goal of this work was to determine the degree of variance between students in terms of where they placed each term or emoticon. Given any affective term, there was little difference between the standard deviation for terms (mean S.D for terms = 1.20) and faces (mean S.D. for faces = 1.18). However, there was a larger

difference between the standard deviation in activation (mean S.D for activation of terms or faces = 1.33) and valence (mean S.D for valence of terms or faces = 1.05), suggesting that students may have a greater degree of agreement in regarding rating the valence of affective representations than the activation it produces in them, which is consistent with the finding that affective representations fall on the division between positive and negative valence as we would categorize them, but not necessarily in terms of activation.

## 3 Discussion

The results presented in this article highlight a few different conclusions: a) students did not necessarily match emoticons or affective terms to the quadrants where researchers would have placed them, mostly in relation to activation; b) there is a large variation across these middle-school students in terms of where they placed a specific emotion within the axes of valence x arousal.

Characterizing researcher common expectations for arousal or activation is difficult, as many researchers only tentatively suggest how emotional states may be characterized in terms of activation. Pekrun found data to support boredom being somewhat deactivating, [18]. Russell [25] explores the components of affect and offers a few hypotheses which are summarized in figure 1 of Baker et al 2010 [3] wherein boredom is characterized as deactivating, while frustration, surprise, and delight are characterized as activating. Broekens' [8] emoticons follow the scheme outlined in the figure 2 caption: elation, fear, surprise, and anger are seen as activating, while sadness, relaxation, contentment, and frustration are seen as deactivating.

Students seem to agree that delight or elation is highly activating along with excitement, and boredom is deactivating along with sadness and relaxation. However, we found that students viewed an emoticon of fear as deactivating, and other affective states placed relatively close to neutral in terms of activation.

There are a few points of methodological concern. Firstly, the order that the students' place their stickers may be important: beyond a simple priming effect of considering one term/emoticon before another, by placing one item first students are changing the affordance of the coordinate axis itself by adding a milestone in the form of a term or emoticon. In future research, we could consider including fewer stimuli for placement or giving students a clean chart for each stimuli.

A second point of concern is one of validity. The terms, emoticons, and even the coordinate axis itself are abstract descriptors of affective states, which in this experiment are divorced from the actual experiences students may be having.

By placing our study outside the experimental environment we are likely reducing the validity of this work in exchange for simplicity of study design (i.e. not requiring students to respond with faces and words on the axis at various points in their experience).

The work of Bieg et al. [6] tells a much larger story than recommending against self-reports out of context. Out of context self-reports were found to bias in a consistent direction as compared to in context self-reports. However we maintain this

method is “less valid” rather than “invalid”. Further, if we take into consideration the savings in class time an out of context self-report may actually be a better study design choice in some cases. It is our position that establishing more quantitative comparisons of reliability will yield better relative comparisons of validity and allow for improved study design.

This argument can be extended to affective research in general in the distinction between emotional experience and appraisal. We conceptualize the experience itself as the construct, and the cognitive appraisal process as a means of communicating that experience. The appraisal may be performed to send communication (e.g. having an experience and generating a representation of that experience for others), or receive communication (e.g. identify a representation as signifying an emotional state).

From this standpoint we suggest that the fewer steps of appraisal exist, the greater the face validity of an appraisal is in terms of reflecting an emotional experience. This is consistent with the findings of [6] wherein aggregate appraisal may differ from immediate contextual appraisal and we tend to view immediate appraisal as having greater face validity. This hypothesis also lends credence to the belief that external appraisal of an unconsciously generated representation (which may still be unconsciously meant to communicate an experience), in the form of facial expressions may be more valid than self-report measures wherein experiences are appraised by both subject and researcher. However, while passing through multiple appraisals may risk loss of information, the quality and richness of the appraisal may also play a role.

While validity remains very difficult to establish with regard to affect by testing “inter-method” or “representational” reliability perhaps we can building convergent and discriminant validity: multiple representations indicating the same construct across multiple participants. We maintain that reliability and validity are continuous rather than discrete traits of models. Therefore, we wish to reach consensus on methods of determining reliability and validity and then begin applying them to methods of inferring the experience of emotion. This work is a means of determining reliability between appraisals of representations of emotion rather than reliability of appraisals of emotions themselves. This is to say that matching particular facial expression to their personal lexicon of categorical affective terms, a high degree of agreement may validate the relationship between depictions of affect textually and facially, but not between either of those representations and the experience of an emotion.

A potential way towards greater validity and reliability could be to cognitively induce an emotional experience by asking students to respond to how they would feel given a particular situation (e.g. “Report on how you’d feel if you failed a math test.”). Of course there may be a distinction between induced affect and “organic” affect, further there will be a broad degree of subjectivity based on how individual students might feel about any given situation. Therefore the variance in responses could be attributed at least to two types of factors: those pertaining to both how students’ believe they would feel in a given context, and those pertaining to students’ ability to report that subjective experience through self-report measures. While there isn’t a clear way to disambiguate between which type of factor is responsible for the variance here, such an approach might be able to establish a conservative maximum of error in self-report measurements, because two students might have very different

feelings about failing a math exam. In essence, we have measured variance in reliability here, not validity.

Finally, while reliability of self-report measures should inform their design, there may be cases of diminishing returns where a slight improvement in reliability has heavy costs for implementation workload, response time, or other practical concerns. We need not pick the measure with the highest available reliability; however it would be good to have some empirical handle on the relative reliabilities of different types of self-report measures. Perhaps the greatest thing to come out of this work would be future collaborations which might better address these concerns.

## 4 References


1. AlZoubi, O., D'Mello, S. K., & Calvo, R. A. (2012). Detecting naturalistic expressions of nonbasic affect using physiological signals. Affective Computing, IEEE Transactions on, 3(3), 298-310.
2. Arroyo, I., Woolf, B.P., Royer, J.M. and Tai, M. (2009b) 'Affective gendered learning companion', Proceedings of the International Conference on Artificial Intelligence and Education, IOS Press, pp.41–48.
3. Baker, R.S.J.d., D'Mello, S.K., Rodrigo, M.M.T., Graesser, A.C. (2010) Better to Be Frustrated than Bored: The Incidence, Persistence, and Impact of Learners' Cognitive-Affective States during Interactions with Three Different Computer-Based Learning Environments. *International Journal of Human-Computer Studies*, 68 (4), 223-241.
4. Barrett, L. F. (2004). Feelings or Words? Understanding the Content in Self-Report Ratings of Experienced Emotion. Journal of Personality and Social Psychology, 87(2), 266–281.
5. Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., Shute, V., Ventura, M., & Zhao, W. (2015). Automatic Detection of Learning-Centered Affective States in the Wild. In Proceedings of the 2015 International Conference on Intelligent User Interfaces (IUI 2015). ACM, New York, NY, USA.
6. Bieg, M., Goetz, T., & Lipnevich, A.A. (2014). What Students Think They Feel Differs from What They Really Feel – Academic Self-Concept Moderates the Discrepancy between Students' Trait and State Emotional Self-Reports. PLoS ONE 9(3): e92563.
7. Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the Self-Assessment Manikin and the Semantic Differential. Journal of Behav Ther Exp Psychiatry, 25, 49-59.
8. Broekens, J., & Brinkman, W.-P. (2013). AffectButton: a method for reliable and valid affective support. International Journal of Human-Computer Studies, 71(6), 641-667.
9. Calvo, R. A., D'Mello, S., Gratch, J., & Kappas, A. (Eds.) (2015). The Oxford Handbook of Affective Computing. Oxford University Press: New York, NY.
10. Conati, C., & Maclaren, H. (2009). Empirically building and evaluating a probabilistic model of user affect. User Modeling and User-Adapted Interaction, 19(3), 267-303.
11. D'Mello, S. , Craig, S. D., Sullins, J., & Graesser, A. C. (2006). Predicting Affective States expressed through an Emote-Aloud Procedure from AutoTutor's Mixed-Initiative Dialogue. International Journal of Artificial Intelligence in Education, 16(1), 3-28.
12. D'Mello, S., & Graesser, A. C. (2012). Language and Discourse Are Powerful Signals of Student Emotions during Tutoring. IEEE Transactions on Learning Technologies, 5(4): 304–317.

13. Graesser, A., & D'Mello, S. (2011). Theoretical perspectives on affect and deep learning. In New perspectives on affect and learning technologies (pp. 11-21). Springer New York.
14. Linnenbrink-Garcia, L., & Pekrun, R. (2011). Students' emotions and academic engagement. Introduction to the special issue. Contemporary Educational Psychology, 36, 1–3.
15. McDaniel, B. T., D'Mello, S., King, B. G., Chipman, P., Tapp, K., & Graesser, A. C. (2007). Facial features for affective state detection in learning environments. In Proceedings of the 29th Annual Cognitive Science Society (pp. 467-472).
16. Ocumpaugh, J., Baker, R.S., Rodrigo, M.M.T. (2015) Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual.. Technical Report. New York, NY: Teachers College, Columbia University. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.
17. Pardos, Z. A., Baker, R. S., San Pedro, M. O., Gowda, S. M., & Gowda, S. M. (2013). Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. Proc. 3rd Int.Conf. Learning Analytics & Knowledge, 117-124.
18. Pekrun, R., Goetz, T., Daniels, L. M., Stupnisky, R. H., & Perry, R. P. (2010). Boredom in achievement settings: Exploring control–value antecedents and performance outcomes of a neglected emotion. *Journal of Educational Psychology, 102*(3), 531-549.
19. Porayska-Pomsta, K., Mavrikis, M., D'Mello, S., Conati, C., Baker, R.S.J.d. (2013) Knowledge Elicitation Methods for Affect Modeling in Education. International Journal of Artificial Intelligence in Education, 22 (3), 107-140.
20. Porayska-Pomsta, K., Mavrikis, M., & Pain, H. (2008). Diagnosing and acting on student affect: the tutor's perspective. User Modeling and User-Adapted Interaction, 18(1-2), 125-173.
21. Read, J., McFarlane, S., and Cassey, C. (2002). Endurability, engagement and expectations: Measuring children's fun. In Proceedings of International Conference for Interaction Design and Children.
22. Read J. C. and MacFarlane, S.(2006). Using the fun toolkit and other survey methods to gather opinions in child computer interaction. In Proceedings of the 2006 conference on Interaction design and children (IDC '06). ACM, New York, NY, USA, 81-88.
23. Rodrigo, M. M. T., Baker, R. S. J. d., Lagud, M. C. V., Lim, S. A. L., Macapanpan, A. F., Pascua, S. A. M. S., et al. (2007). Affect and Usage Choices in Simulation ProblemSolving Environments. In R. Luckin, K. R. Koedinger & J. Greer (Eds.), Proceeding of the 2007 conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts that Work (Vol. Frontiers in Artificial Intelligence and Applications 158). Amsterdam: IOS Press.
24. Rowe, J. P., Mott, B. W., & Lester, J. C. (2014) It's All About the Process: Building Sensor-Driven Emotion Detectors with GIFT. In Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym2) (p. 135).
25. Russell J.A, Barrett L,F. (1999) Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. J. Pers. Soc. Psychol. 76(5):805–19.
26. San Pedro, M.O.Z., Baker, R.S.J.d., Bowers, A.J., Heffernan, N.T. (2013) Predicting College Enrollment from Student Interaction with an Intelligent Tutoring System in Middle School. Proceedings of the 6th International Conference on Educational Data Mining, 177-184.

# Cultural aspects related to motivation to learn in a Mexican context

Erika-Annabel Martínez-Mirón[1,*], Genaro Rebolledo-Méndez[2]
[1]Universidad Politécnica de Puebla, Puebla, México
*Corresponding Author: erika.martinez@uppuebla.edu.mx
[2]Universidad Veracruzana, Xalapa, México
g.rebolledo@gmail.com

**Abstract.** The development of motivationally intelligent tutoring systems has been based on a variety of motivational models from the psychology field. These models mainly consider characteristics from de areas of values, expectancies and feelings [1]. However, this paper proposes to take into account some cultural aspects when operationalizing such models. The basis of this proposal is presented from the perspective of some cultural aspects that effect career choice, in particular for a Mexican context.

**Keywords:** Motivation, career choice, Mexican cultural context

## 1 Introduction

Research in motivation to learn when using educational technology has operationalized different motivational models found in the psychological literature in order to develop motivationally intelligent tutoring systems. According to these models, motivationally aware tutoring systems should combine expertise and knowledge about user's cognitive, affective, meta-cognitive and meta-affective levels in order to appropriately react and be able to favor user's learning [2, 3]. That is, these models should mainly consider characteristics from the areas of values, expectancies and feelings [1].

However, this paper argues also for the inclusion of other aspects that have been seldom taken into account so far. We refer to cultural aspects inherent to each group of individuals from a certain background. Since there is evidence that students from different cultural origin react to the same motivational strategy in a different way [4, 5, 6] or have different attitudes for online assessment [7], the cultural aspect of learning with technology becomes an important issue. For instance, if a female student from a highly gender-stereotyped cultural background is asked to attend a course considered to be strongly oriented to men, then she might perceived to be in the wrong course and probably will not exert her maximum effort. Or even she might believe that her role in society is to be protected by someone, and she attends courses just to be in the possibility to meet that expectation. It will not matter what motivational strategy the teacher uses, since the female student's cultural belief is in an apparently superior level and she will only be concerned to learn at the minimum, just to continue studying until meeting her protector [8].

In order to develop the arguments to support the inclusion of cultural aspects in the design of motivationally-aware tutoring systems, the following sections describe some of these elements within a Mexican context from the perspective of career choice, based on the findings that instrumental motivation is an important predictor for course selection, career choice, and performance [9, 10]. That is, students may pursue to perform well in some tasks because they are important for future goals, even if the student is not interested on the task.

## 2 Motivation, career guidance and cultural context

Motivation is related to the student's desire to participate in the learning process. Current research findings suggest that motivational constructs do change over time [11, 12, 13] and/or contexts [14, 15, 16]. In particular, it is well documented that cultural differences affect achievement motivation [4, 5, 6].

We believe that if teachers truly want to promote the success of all students, they must recognize how achievement motivation varies culturally within the population it serves.
Similarly, career counseling must incorporate different variables and different processes to be effective for students from different cultural contexts. Career counseling is defined as "the process of assisting individuals in the development of a life-career with focus on the definition of the worker role and how that role interacts with other life roles" [17].
According to Rivera [18], there are characteristics that prevail among Hispanic/Latino American children and adolescents, such as: A) Restraint of feelings, particularly anger and frustration; B) Limited verbal expressions toward authority figures; C) Preference for closer personal space; avoidance of eye contact when listening or speaking to authority figures; D) Relaxation about time and punctuality; and immediate short-term goals; E) Collective, group identity; interdependence; cooperative rather than competitive; emphasis on interpersonal relations. To certain extent, these characteristics can be considered part of one of the four sources of information, social persuasion, included in the model of the Socio Cognitive Career Theory [19], (see Table 1). This framework conceptualizes career choice as a process with multiple stages and different sources of information. We propose that cultural aspects of the Mexican context might have an impact not just the process of choosing a career, but on the way students undertake their learning activities as described in the following paragraphs.

Table 1. Sources of information proposed in the model of social cognitive influences on career choice behavior [19]

| **Source of information** | **Description** |
|---|---|
| Performance accomplishment | Success in performing the target task or behavior |
| Vicarious learning or modeling | To watch others who could perform the target behavior successfully. |
| Emotional arousal | Anxiety when performing the target behavior |
| Social persuasion | Support and encouragement from others in the process of performing the target behavior. |

### 2.1 Machismo

There is growing research supporting that achievement differences between genders are smaller during early years of school or being reduced [20]. The succession of career behaviors for women is far more complex than for men. In particular, in Mexican students, the complexities might lay in the cultural aspect of machismo. In Mendoza's review [21], machismo is defined as a strong sense of masculine pride, and it is suggested that machismo should be considered in any Latino study, but it is often forgotten. The social behavior pattern associated to machismo includes the expectation of men being caring, responsible, decisive, strong of character, and the protector of probably extended family. At the same time, negative aspects of machismo include aggressiveness, physical strength, emotional insensitivity, and a womanizing attitude towards the opposite sex.
Galanti [22], cited in [21], surveyed a group of Latino students who reported that the relationship between male and female would be of protector and protected. More specifically, according to them, the role of the traditional Hispanic woman is to look after the family; her job is to cook, clean, and care for the children. Other characteristics of a good wife include submission and obedience to her husband's orders without questioning him but rather standing behind whatever he decides, even if she disagrees. She must also be tolerant of his behavior. Taking into account these views it is understandable that women´s career choice might be influenced by the fulfillment of this profile rather than freely choosing a career that may imply a great amount of dedication. In some Mexican contexts, women may prefer to undertake studies that are less demanding. Women also must strive to overcome obstacles such as gender discrimination and sex stereotyping. For instance, Gallardo-Hernández *et. al.* reported the results of a questionnaire applied to 637 first-year medical nutrition, dentistry and nursing students

[23]. The findings suggest that among women of low socioeconomic strata, more traditional gender stereotypes prevail which lead them to seek career choices considered feminine. Among men, there is a clear relationship between career choice, socioeconomic level and internalization of gender stereotypes.

### 2.2 Social orientation

Cooperative learning is very important for Mexicans [24]. They do not seem to openly want to show what they know for fear of embarrassing those who do not know [25]. It is not common in a Hispanic family to encourage children to excel over siblings or peers but rather, it is considered bad manners. It is worth noting that most of the studies reported have taken into account the Mexican context around Mexican American students but no studies so far focus on comparison between this population and a Mexican population living in Mexico. Nevertheless, their findings can, to some extent, be considered valid for Mexican population. For instance, Ojeda and Flores [26] considered the educational aspirations of 186 Mexican American high school students to test a portion of social-cognitive career theory [19]. Their results indicated that perceived educational barriers significantly predicted students' educational aspirations above and beyond the influence of gender, generation level, and parents' education level. Similarly, Flores, Romero and Arbona [27] found that Mexican American men and women with high measures of ethnic loyalty might be at risk for perceiving social costs of pursuing a higher education.

### 2.3 Perception of time and career guidance

Mexicans are oriented toward present time; they are focused on “right now” rather than on the past or on future events or outcomes. They often live the phrase “Dios dirá” or "God will tell," that is, time is relative. To arrive late for an engagement is called in the southwest "Mexican time." This perception permeates career-counseling programs in the Mexican context, since its interventions start in the educational level just behind the university program [28]. Therefore, students have to decide in a relatively short period of time which career suits them best. Sometimes the students might have a great amount of career information, making it difficult to make a good analysis of each of the options. But it also might occur that there is little availability of information and students might end up making an inadequate career choice.

## 3 Discussion

Increasingly, researchers are calling for studies of change in motivation, rather than treating motivation as a static trait-like factor [1], [4]. However, those studies mainly consider motivation to be influenced by characteristics from the areas of values, expectancies and feelings [1], without taking into account that some cultural aspects like machismo, social orientation or perception of time might also be influencing how students approach to a learning activity. For instance, women could be avoiding pursuing a career that would not allow them to easily integrate their expected roles as mother and spouse with their future professional activity. Also, the perception of educational barriers, such as gender and ethnicity, nurtured by the social context could reinforce the idea of choosing a career according to the students’ sex, which in turn might influence students’ motivation to learn a particular area of study. Although there is little research evidence that establishes a direct connection between career choice and motivation to learn a particular topic, this paper reviewed some cultural aspects in the Mexican context that have an impact on students’ learning behavior. Based on this, we consider plausible to do some research that consider these aspects when designing a motivationally tutoring system. For example, in a Mexican context, a tutoring system for Mathematics could emphasize women’s capacity to solve problems regardless of their gender, like providing feedback including mentions to important contributions from female scientists, or listing the advantages of achieving personal professional success as a woman, or maybe using a very strong female character showing high IQ as the main avatar.

## 4 REFERENCES


1. du Boulay, B. Towards a Motivationally-Intelligent Pedagogy: How should an intelligent tutor respond to the unmotivated or the demotivated? In R. A. Calvo & S. D'Mello (Eds.), New Perspectives on Affect and Learning Technologies (pp. 41-54). New York: Springer (2011)
2. Avramides, K. and du Boulay, B. Motivational Diagnosis in ITSs: Collaborative, Reflective Self-Report. In V. Dimitrova, R. Mizoguchi, B. du Boulay & A. Graesser (Eds.), Artificial Intelligence in Education. Building Learning Systems that Care: from Knowledge Representation to Affective Modelling AIED2009 14th International Conference on Artificial Intelligence in Education (Frontiers in Artificial Intelligence and Applications No. 200 pp. 587-589). Amsterdam: IOS Press (2009)
3. du Boulay, B., Rebolledo Mendez, G., Luckin, R. & Martinez Miron, E. (2007). Motivationally Intelligent Systems: Diagnosis and Feedback. In R. Luckin, K. Koedinger & J. Greer (Eds.), Artificial Intelligence in Education: Building Technology Rich Learning Contexts. Proceedings of AIED2007, Los Angeles (Frontiers in Artificial Intelligence and Applications No. 158 pp. 563-565). Amsterdam: IOS (2007)
4. Henderlong, J., and Lepper, M. R. The effects of praise on children's intrinsic motivation: A review and synthesis. *Psychological Bulletin, 128,* 774–795 (2002)
5. Kaplan, A., Karabenick, S., & De Groot, E. Introduction: Culture, self, and motiva- tion: The contribution of Martin L. Maehr to the fields of achievement motivation and educational psychology. In A. Kaplan, S. Karabenick, & E. De Groot (Eds.), *Culture, self, and motivation: Essays in honor of Martin L. Maehr* (pp. vii-xxi). Charlotte, NC: Informa- tion Age Publishing (2009)
6. Trumbull, Elise; Rothstein-Fisch, The Intersection of Culture and Achievement Motivation. Carrie School Community Journal, v21 n2 p25-53 (2011)
7. Terzis V., Moridis C., Economides A.A., Rebolledo-Mendez G. Computer Based Assessment Acceptance: A Cross-Cultural Study in Greece and Mexico. *Journal of Educational Technology and Society. 16(3),* 411-424 (2013)
8. Schmitz, K. and Diefentahler, S. An examination of traditional gender roles among men and women in Mexico and the United States. Retrieved, Vol. 12, p. 2008. (1998)
9. Wigfield, A., and Eccles, J.S. (1992) The development of achievement task values: A theoretical analysis. Developmental Review 12: 265 – 310.
10. Wigfield, A., Eccles, J.S., and Rodriguez, D. (1998) The development of children's motivation in school context. Review of Research in Education 23: 73-118.
11. Bong, M., and Skaalvik, E. M. Academic Self-Concept and Self-Efficacy: How Different Are They Really?. Educational Psychology Review, Vol. 15, No. 1, 1-40 (2003)
12. Chouinard, R. and Roy, N. Changes in high-school students' competence beliefs, utility value and achievement goals in mathematics. British Journal of Educational Psychology, Vol. 78, No. 1, 31-50 (2008)
13. Corpus, J., McClintic-Gilbert, M. S., and Hayenga, A. O. Within-year changes in children's intrinsic and extrinsic motivational orientations: Contextual predictors and academic outcomes. Contemporary Educational Psychology, Vol. 34, No. 2, 154-166 (2009)
14. Otis, N., Grouzet, F. E. and Pelletier, L. G. Latent Motivational Change in an Academic Setting: A 3-Year Longitudinal Study. Journal Of Educational Psychology, Vol. 97, No. 2, 170-183 (2005)
15. Caprara, G. V., Fida, R., Vecchione, M., Del Bove, G., Vecchio, G. M., Babaranelli, C. and Bandura, A. Longitudinal analysis of the role of perceived self-efficacy for self-regulated learning in academic continuance and achievement. Journal of Educational Psychology, Vol. 100, 525-534 (2008)
16. Wang, Q. and Pomerantz, E. The motivational landscape of early adolescence in the United States and China: a longitudinal investigation. Child Development, Vol. 80, No. 4, 1272-1287 (2009)
17. National Career Development Association. http://www.ncda.org/aws/NCDA/pt/sd/news_article/37798/_self/layout_ccmsearch/true
18. Rivera, B. D., and Rogers-Adkinson. D. Culturally sensitive interventions: Social skills training with children and parents from culturally and linguistically diverse backgrounds. Intervention in School and Clinic. 33(2), 75-80 (1997)
19. Lent, R. W., Brown, S. D., and Hackett, G. Toward a Unifying Social Cognitive Theory of Career and Academic Interest, Choice, and Performance, Journal of Vocational Behavior, 45, p. 93 (1994)
20. Hyde, J., Lindberg, S., Linn, M., Ellis, A., & Williams, C. Gender similarities characterize math performance. Science, 321(5888), 494 – 495 (2008)
21. Mendoza, E. Machismo Literature Review. Center for Public Safety Initiatives. Rochester Institute of Technology (2009)

22. Galanti, G. The Hispanic Family and Male-Female Relationships: An overview. Journal of Transcultural Nursing, 14(3), 180-185 (2003)
23. Gallardo-Hernández, G., Ortiz-Hernández, L., Compeán-Dardón, S., Verde-Flota, E., Delgado-Sánchez, G., Tamez-González, S. Intersection between gender and socioeconomic status in medical sciences career choice. Gaceta Médica Mex. 2006 Nov-Dec; 142(6):467-76 (2006)
24. Gorodnichenko, Y. and Roland, G. Culture, Institutions and the Wealth of Nations, CEPR Discussion Paper No 8013 (2010). http://eml.berkeley.edu/~ygorodni/gorrol_culture.pdf
25. Losey, K. M. Mexican American students and classroom interaction: An overview and critique. Review of Educational Research, 65, 283-318 (1995)
26. Ojeda, L. and Flores, L. The Influence of Gender, Generation Level, Parents' Education Level, and Perceived Barriers on the Educational Aspirations of Mexican American High School Students. The Career Development Quarterly 57, 1, 84–95. (2008)
27. Flores, Y.N., Romero A., and Arbona, C. Effects of Cultural Orientation on the Perception of Conflict Between Relationship and Educational Goals for Mexican American College Students. Hispanic Journal of Behavioral Sciences 22, 1: 46-63 (2000)
28. POV (2011). Programa de Orientación Vocacional para el bachillerato general, tecnológico y profesional técnico. www.dgb.sep.gob.mx/04-m2/.../Programa_Orientacion_Vocacional.pdf