Workshop on Non-Cognitive Factors & Personalization for Adaptive Learning (NCFPAL)

Many computer-based learning environments adapt to individual learners based on cognitive factors like skill mastery, but recently research has been increasingly directed at improving personalization and adaptation in such systems by harnessing non-cognitive factors such as learner affect, motivation, preferences, self-efficacy, self-regulation, and grit. This workshop brings together researchers studying non-cognitive factors in a variety of environments and contexts, using various experimental, measurement, and/or data mining and statistical methods. In addition to presenting on-going research on specific non-cognitive factors and their impact of learning outcomes, speakers at the workshop will present various creative approaches to address methodological issues endemic to research on non-cognitive factors.

Of one invited paper and five accepted papers, three papers explore non-cognitive factors in intelligent tutoring systems (ITSs) used in K-12 schools. Walkington and collaborators, in an invited paper, provide an account of various text-based features of mathematics word problems that are associated with learner performance in ITSs (specifically, Carnegie Learning's Cognitive Tutor). While explanations that point to both cognitive and non-cognitive factors may account for this association, Bernacki and Walkington follow up this observational study by exploring an intervention in the same ITS wherein word problems are personalized based on learners' out-of-school interests in areas like sports and music and find that personalization has benefits for both learner interest and measures of learning. A third study by Ostrow and colleagues considers an intervention in the ASSISTments system in which learners were presented with different types of "growth mindset" motivational messages (e.g., animations, audio, etc.). The impact of these messages on measures like persistence and learning are considered.

The next three papers consider data from college-level courses and learners. Ezen-Can and Boyer present an unsupervised method for classifying dialogue acts (e.g., ask a question, give a command) when learners interact with (human) tutors in a text-based dialogue environment; their method leverages gender and learner self-efficacy as noncognitive factors along which sub-populations of learners can be identified so that dialogue acts can be better classified. Next, Moretti and colleagues mine data about university computer science courses that are publicly available on the web to determine factors (e.g., choice of programming language and grading criteria) that are associated with learner feedback and other aspects of instruction. Finally, Gray and colleagues provide an analysis, using both classification and regression methods, of various psychometric measures of non-cognitive factors as predictors of whether students are "at risk" or likely to fail in their university courses.

The papers that comprise these proceedings represent a diverse set of measurement and analytical approaches and of student populations and learning platforms to which they are applied. We take this as a sign of developments to come, especially as researchers and developers in the learning sciences, educational data mining, and learning analytics increasingly turn to non-cognitive factors as possible "levers" to adapt and personalize learning experiences in more and more sophisticated technology-enhanced learning platforms and environments.

We gratefully acknowledge the following members of the workshop program committee:

Vincent Aleven, Carnegie Mellon University
Ryan S.J.d. Baker, Columbia University
Matt Bernacki, University of Nevada, Las Vegas
Alan Drimmer, Apollo Group, Inc.
Andrew Krumm, SRI International
Timothy Nokes-Malach, University of Pittsburgh
John Stamper, Carnegie Mellon University
Candace Walkington, Southern Methodist University
Michael Yudelson, Carnegie Learning, Inc.

The NCFPAL workshop organizers
Steven Ritter
Stephen E. Fancsali

Table of Contents NCFPAL

The Impact of Cognitive and Non-Cognitive Text-Based Factors on Solving Mathematics Story Problems

Candace Walkington
Southern Methodist University
3011 University Blvd. Ste. 345
Dallas, TX, 75205
1-214-768-3072
cwalkington@smu.edu

Virginia Clinton
University of Wisconsin - Madison
1025 West Johnson Street
Madison, WI 53706
1-608-890-4259
vclinton@wisc.edu

Steven Ritter
Carnegie Learning, Inc.
437 Grant Street, Suite 918
Pittsburgh, PA 15219
1-888-851-7094 x122
sritter@carnegielearning.com

Mitchell Nathan
University of Wisconsin - Madison
1025 West Johnson Street
Madison, WI 53706
1-608-262-0831
mnathan@wisc.edu

Stephen E. Fancsali
Carnegie Learning, Inc.
437 Grant Street, Suite 918
Pittsburgh, PA 15219
1-888-851-7094 x219
sfancsali@carnegielearning.com

ABSTRACT

Intelligent tutoring systems (ITSs) that personalize instruction to individual learner background and preferences have emerged in K-16 classroom settings all over the world. In mathematics instruction, ITSs may be especially important for tracking mathematical skill development over time. However, recent research has pointed to the importance of text-based measures when solving mathematics word problems, suggesting that in order to accurately model the student it is important to understand how they respond to text characteristics. We investigate the impact of text-based factors (readability and problem topic) on the solving of mathematics story problems using a corpus of N = 3394 students working through an ITS for algebra, Cognitive Tutor Algebra. We leverage recent advances in computerized text-mining to automate fine-grained text analyses of many different word problems. We find that several elements of the text of mathematics word problems matter for performance – including the concreteness of the problem's topic, the length and conciseness of the story's text, and the words and phrases used.

Keywords

Intelligent tutoring system, readability, mathematics, word problems, personalization

1. INTRODUCTION

Since the 1980s, Intelligent Tutoring Systems (ITSs) have risen as an important instructional tool to support student learning in classrooms, especially in middle and high school. ITSs typically consist of at least three components: (1) the *domain model* of the appropriate steps needed to correctly solve each problem, (2) the *student model*, which captures the evolution of an individual student's cognitive states as they relate to the domain model, and (3) the *tutoring model* which selects tutor actions based on the domain model and student model [1]. It is through the construction of the student model and its contribution to the tutoring model that ITSs can enact *personalization* where they adapt to the needs and backgrounds of individual learners. Here we explore cognitive and non-cognitive factors related to how students react to and understand the text of mathematics story problems. We argue that these non-mathematical factors may be an important element to consider for an ITS in secondary mathematics. In particular, we provide evidence suggesting that both the students' reading level (a cognitive factor) and the students' interests, preferences, and motivational outlooks (non-cognitive factors) have the potential to influence how they respond to text-based mathematics problems situated in "real world" contexts.

Cognitive Tutor Algebra (CTA; [2]) is a prominent mathematics ITS used in many schools across the United States. CTA uses *model-tracing approaches* to relate student actions to the domain model and provides individualized error feedback. CTA also uses *knowledge-tracing approaches* to track students' learning from one problem to the next, using this information to identify the students' strengths and weakness in terms of production rules (i.e., knowledge components or skills). The software then uses this analysis to individualize the selection of problem tasks. However, missing from this tutoring model is a consideration of other non-mathematical characteristics of the story problem texts – including the *reading difficulty* of the text respective to students' reading ability and preferences, and the *real-world topic* of the text respective to students' interests and preferences.

For example, a learner presented with a mathematics word problem that is difficult to read – with high-level vocabulary, complex sentence structure, etc. - may lack the reading ability to appropriately comprehend that problem. This cognitive element of the problem's difficulty is not typically monitored by ITSs for mathematics learning. In

Published in CEUR-WS:
NCFPAL workshop (Ritter and Fancsali)
In EDM 2014 Extended Proceedings (Gutierrez-Santos and Santos)

addition, such a problem may inhibit the students' motivation – a non-cognitive factor. In particular, even if the learner is technically able to read the problem, they may be intimidated by the problem text, and request a hint instead of putting forth the effort of understanding the text of the problem. ITSs also do not typically monitor the learner motivation for reading and understanding text-based problems.

Another non-mathematical element of the text of mathematics story problems is the real world topic – whether the story is about working at a part-time job or harvesting a field of grain. The way in which students react to the topic of the story problem is also based on both cognitive and non-cognitive factors. Students may be unfamiliar with elements of the context that are important for fully comprehending the problem – for example, in a banking context, they may not know what "break even" means. In this way, they may lack the prior knowledge needed to interpret the story. Similarly, different real world topics may differ in the motivation they elicit from students – students may experience greater motivation when solving a problem about a familiar, interesting context than about a context they find boring or unfamiliar.

We next provide a theoretical framework that provides an explanation of how students comprehend story problems and how cognitive and non-cognitive factors may interact as they solve story problems.

2. THEORETICAL FRAMEWORK

2.1 Cognitive Factors

Nathan and colleagues [3] proposed a model of mathematics story problem solving where students navigate three levels of representation as they comprehend and solve story texts: (1) a *textbase* containing the propositional statements made in the story problem, (2) a *situation model*, a qualitative representation of the actions and events in the story, and (3) a *problem model*, containing the formal mathematical equations, variables, and operands. Because mathematics word problems are stated in verbal language (rather than mathematics notation), we hypothesize that the reading difficulty and topic of the problem matters for the construction of the situation model and its successful coordination with the problem model.

Various aspects of the reading difficulty, including *readability measures*, may be important in situation model construction. Readability measures often include the kinds of words used, the length of the story, and the structure of the sentences. These elements of the text's structure may make it more difficult to comprehend, especially for students with weaker reading skills.

Another aspect of reading difficulty is the *topic* of the problem – whether it is about, for example, farming or banking. Walkington and colleagues [4] proposed that story contexts that are related to topics that are familiar and accessible to students are easier for them to solve because these contexts can facilitate situation model construction because of their relatedness to learner prior knowledge. In related work [5], they also identified the prevalence of issues with verbal interpretation of mathematics story problems, finding that even high school students struggle to understand difficult vocabulary words and construct an accurate propositional textbase and situation model from a story problem's text.

2.2 Non-Cognitive Factors

An important precursor to students' motivation is their level of *interest* – defined as the state of engaging and the predisposition to re-engage with particular topics, ideas, or activities [6]. Two types of interest have been described in the literature. First, *situational interest* is an immediate, temporary state of heightened attention and affective engagement that stems from elements of a learning environment that are surprising, salient, evocative, challenging, personally relevant, etc. Situational interest can be *triggered* in response to a stimuli within a learning environment, and then may or may not become *maintained* over time [6]. A second type of interest is *individual interest* – learners' enduring predispositions to engage with certain activities or topics over time.

Elements of a story problem's text have the potential to both trigger and maintain situational interest. In particular, story problems that are accessible, easy to read, and situated within the topics and contexts that a particular learner finds relevant and interesting may trigger and maintain interest. In the other hand, difficult reading passages disconnected from a learner's experiences and interests may not trigger interest and may cause disengagement if interest has previously been triggered.

2.3 Research Purpose

If text-based measures like readability and problem topic matter for student performance, these might be important elements to add to future systems for personalized learning in mathematics. For example, an ITS might present weak readers with problems with simplified verbal language as these learners are initially mastering a new mathematical skill. As the student gains expertise with the mathematics by mastering skills, additional levels of verbal difficulty could be layered on by the ITS. Similarly, learners that lack motivation may be presented with story problems that are less intimidating to read and situated within their interests, with this support faded out over time. By neglecting to model this aspect of the user's experience in the ITS, the system may be generating inferences about learner knowledge states that are inaccurate.

3. LITERATURE REVIEW

3.1 The Impact of Reading Difficulty on Solving Mathematics Story Problems

Recent research has found that reading ability is especially important as students solve mathematics word problems [7]. Studies examining the association of reading difficulty of mathematics word problems and U.S. student

performance on large-scale assessments has found that problems that use words with multiple meanings, complex verbs, and mathematics vocabulary words are more difficult [8]; the effect is especially pronounced for students who speak English as a second language [9]. A small study of students working in CTA found that extraneous text that provided a real world context for the problem, as well as references to concrete people, places, and things, were associated with less concentration and more confusion in the tutor [10]. However, a similar study found that the extraneous text was also associated with fewer unproductive "gaming the system" behaviors in the tutor [11]. Converging evidence suggests text characteristics relating to reading difficulty are important when solving mathematics word problems, but studies are needed that address which elements of reading difficulty are most important.

3.2 The Impact of Problem Topic on Solving Mathematics Story Problems

The topic of mathematics story problems also has an important relationship to students' prior knowledge and motivation. A study of high school students solving either standard story problems or story problems personalized to topics they were interested in (e.g., sports, video games, social networking) within one unit of CTA found that personalized stories were associated with higher performance. This performance gain was present in two tasks – labeling independent and dependent quantities given in algebra story problems, and writing algebraic expressions from the story scenarios [12]. It was hypothesized that during these two tasks, students are working closely with the problem text, constructing their situation model and coordinating it with a problem model. This study also found that students receiving problems in the context of their out-of-school interests were less likely to game the system – to exploit regularities in hints and feedback provided by CTA in order to avoid productive learning behaviors. Further, students who received personalization had stronger performance in future units where the problems were no longer personalized.

In a recent follow-up study [13], story problems in four units of CTA were personalized to topics students were interested in, and students solving personalized problems were compared to a control group solving normal problems. Results showed that personalized problems both triggered students' situational interest and enhanced students' individual interest for learning algebra. Personalization was associated with greater learning gains than a control condition only when the personalization was matched to deep features of the students' interest area. This was contrasted with personalization that was only matched surface features of the learners' interests – i.e., modifications to the problems that simply involved inserting familiar pop-culture words rather than considering how learners might actually use relationships between quantities in their everyday activities. Thus converging evidence points to the importance of considering the real world topic of mathematics story problems and its relationship to students' interests and experiences. However, more research is needed to determine which topics may be more or less likely to trigger and maintain students' interest.

3.3 Research Questions

In the present study, we investigate the relationship between readability and topic measures and student performance on mathematics story problems. We examine these issues within an ITS for Algebra I, Cognitive Tutor Algebra (CTA), that tracks student hint requests in addition to whether they get problems correct or incorrect. We investigate two research questions: (1) How are readability and topic measures associated with correct answers and hint requests when students label independent and dependent quantities in stories in CTA? (2) How are readability and topic measures associated with correct answers and hint requests when students write algebraic expressions from stories in CTA? Answers to these questions could inform the design of future ITSs for personalized instruction.

4. METHOD

Data from N = 3394 students with active CTA accounts were collected from 9 high schools and 1 middle school that were diverse in terms of their socio-economic, racial, and achievement background (Table 1). Data were collected for students solving 151 distinct word problems accross the first 8 units of CTA; later units were not included because many students did not advance beyond these units. We collapsed for all analyses (i.e., treat as identical) problems containing an identical story but using slightly different numbers. On average, each problem had been solved by 742 students (SD = 495). Each problem included a story scenario that outlined one or more linear functions within a real world situation (Figure 1). The student was asked to complete steps in which they identified the independent and dependent quantities in the story, wrote a linear algebraic expression for the story, and solved their expression for different x and y values; we consider only the first two skills.

CTA log data from students in the selected schools were uploaded to DataShop (pslcdatashop.web.cmu.edu), an online repository of detailed student interaction data. These logs contained information on whether the student got each problem correct, incorrect, or requested a hint on their first attempt; because requesting a hint is a distinct outcome, correct and incorrect are not completely repetitive measures. Thus, for each problem, we compiled the percentage of students who had gotten the problem correct on the first attempt, incorrect, or requested a hint. This percentage was our dependent measure in three distinct regression models.We analyzed the text of the introduction to each story problem (i.e., the initial text that gives the linear rate of change and intercept; see Figure 1) with the *Coh-Metrix* and *LIWC* text-mining programs. Coh-Metrix

[14] measures a large number of aspects of text readability, including the amount semantic overlap between sentences, the number of verbs, use of concrete versus abstract words, the average sentence length, and others.

Table 1. Demographic characteristics of schools in study

ID	Math Prof %	State Prof %	School Enrollment	School Type
1	88%	70%	797	Middle
2	81%	47%	1,482	High
3	95%	84%	2,163	High
4	55%	46%	708	High
5	27%	NA	1,875	High
6	68%	59%	986	High
7	2%	31%	602	High
8	76%	84%	1,333	High
9	19%	39%	397	High
10	68%	79%	800	High

ID	White	Black	Hispanic	F/R Lunch
1	72%	7%	15%	21%
2	90%	4%	2%	4%
3	84%	10%	3%	6%
4	99%	1%	1%	41%
5	20%	4%	72%	77%
6	9%	2%	88%	41%
7	1%	99%	1%	82%
8	36%	60%	2%	48%
9	100%	0%	0%	45%
10	38%	51%	11%	62%

Because some of our story introductions had only one sentence, measures that pre-supposed multiple sentences were ommitted. LIWC [15] was used to determine the topic of the story problems – this program counts how many words in the story fall into various word categories, including social processes (family, friends, people), affective processes (positive emotions and negative emotions), biological processes (body, health, ingestion), cognitive processes (insight, causation, discrepancy, tentativeness, certainty, inhibition, inclusive/exclusiveness), perceptual processes (see, hear, feel), relativity processes (motion, space, time), and personal concerns (work, achievement, leisure, home, money, religion). If a story contained any words that fell into one of these topic categories, that story was coded as a 1 for that category; otherwise it was coded as a 0.

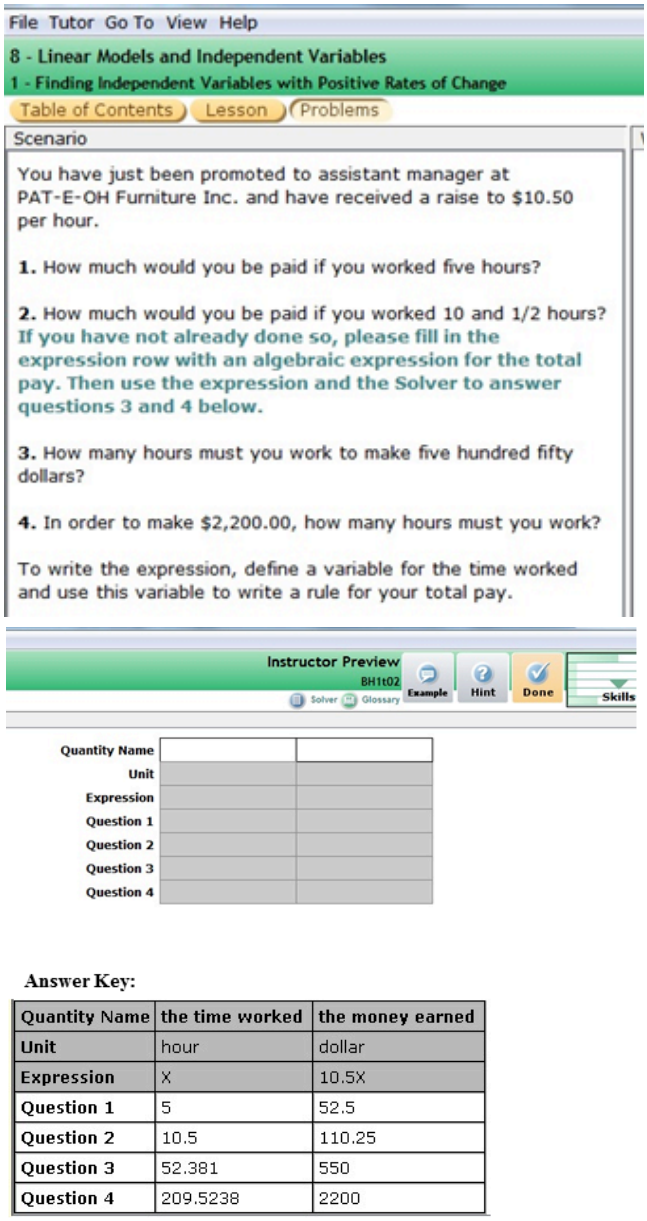


Figure 1. Screenshot of algebra story problem in CTA with answer key superimposed

For each category in Coh-Metrix and LIWC, the correlation was computed between the list of each problem's score on that category, and the percentage of students who got each problem correct, incorrect, or requested a hint. Correlations that were significantly different from 0 were tested for inclusion as fixed effects in regression models predicting the performance measures (hints, corrects, incorrects). These models included random effects that described various aspects of the problem's mathematical structure, including the unit and section it came from in CTA, and the numbers it used. Models were initially fit using the *lmer()* command in *R* including all potential fixed and random effects. Then we used the *step()* command in *R* to perform backwards elimination on fixed

and random effects, leaving a model with only the effects that significantly improved the fit of the model. These analyses were carried out separately for a dataset that included only instances of students labeling independent and dependent quantities, and a dataset that included only instances of students writing algebraic expressions.

5. RESULTS

5.1 Labeling Independent and Dependent Variables

Regression results showing the relationship between performance measures (% incorrect, hint, and correct) and readability and topic measures for labeling quantities in story problems are provided in Table 2. Table 2 shows that problems that use adverbial phrases (*DRAP*) were associated with fewer incorrect answers. Adverbial phrases are phrases that add on to verbs, answering the questions where, when, or how? In the present data set, adverbial phrases mostly answered when the action occured, and often included words like *currently*, *already*, *next*, *first*, *every day/week*, and *not yet*. However, some of these adverbs also answered the how question, relying information about quantities that might be useful to cue students to the constraints of the problem – examples of words used in this manner included *only*, *completely*, and *evenly*. These words may have given important details about how the quantities involved in the story were changing as the action in the story proceeded.

Table 2. Regression tables relating performance measures on labeling quantities to readability/topic categories

	Estimate	Std. Err	t value	Pr(>|t|)	
% Incorrect					
(Intercept)	0.182	0.032	5.63	0.00018	***
DRAP	-0.0008	0.0003	-2.34	0.02104	*
motion	0.036	0.0137	2.65	0.00899	**
% Hint					
(Intercept)	0.045	0.014	3.21	0.01407	*
inhibition	0.023	0.008	2.83	0.00543	**
% Correct					
(Intercept)	0.784	0.044	17.87	0.00000	***
motion	-0.042	0.0182	-2.29	0.02370	*

Stories that involve *motion words* (e.g., *go*, *move*, *ran*, *arrive*, *come*, *enter*, *threw*) are associated with more incorrect answers and fewer correct answers. These stories often incldued contexts where people were walking, biking, hot-air-balooning, driving, or actively constructing something. In terms of the quantities used, there was often a rate of change (e.g., per hour, per minute, a day) that involved this motion, and students had to identify the two quantities that made up this rate of change. Using more abstract physics quantities – like distance and speed – may have been more difficult for students than using quantities relating to specific concrete objects (e.g., accumulating cards, toys, or money). Finally, *inhibition words* were associated with more hint requests. Inhibition words were often included in story problems that discussed safety issues or saving money. Students may have persieved these less concrete, finance- or safety-oriented contexts as less accessible, making them more likely to request a hint rather than attempt to write the labels. These problems often involved money as the dependent variable, but the label for this variable may have been complex because the actor in the story might have already saved or spent some money when the story started. Thus a label of simply *money* may not be appropriate, and the student would have to enter a label that captured that it was *total money* or *net money* saved or spent.

5.2 Writing the Algebraic Expression

Regression results showing the relationship between performance measures and readability and topic measures for writing the expression are shown in Table 3. We again see that *inhibition words* – often associated with financial contexts – are more difficult for students – they are associated with more incorrect answers, more hint requests, and fewer correct answers. The conceptual difficulty of this topic area might become especially important as students move from formulating their situation model to coordinating their situation model with a problem model.

Table 3. Regression tables relating performance measures on writing expressions to readability/topic categories

	Estimate	Std. Err	t value	Pr(>|t|)	
% Incorrect					
(Intercept)	0.195	0.060	3.26	0.00167	**
WRDPOLc	0.0494	0.013	3.91	0.00014	***
inhibition	0.086	0.034	2.52	0.01286	*
% Hint					
(Intercept)	0.055	0.014	3.95	0.00050	***
One sentence	(ref.)				
Two sentences	-0.045	0.016	-2.82	0.00548	**
Three Sentences	-0.057	0.017	-3.48	0.00067	***
4 + Sentences	-0.033	0.019	-1.77	0.07868	
RDL2	0.002	0.001	3.51	0.00061	***
family	0.030	0.015	2.05	0.04282	*
inhibition	0.052	0.011	4.74	0.00001	***
motion	0.025	0.009	2.77	0.00637	**
% Correct					
(Intercept)	0.334	0.17478	1.91	0.05778	
LDTTRc	0.428	0.169	2.53	0.01242	*
WRDPOLc	-0.041	0.01469	-2.78	0.00609	**

inhibition	-0.128	0.03909	-3.28	0.00132	**

Another factor that stands out in the regression results is word polysemy (*WRDPOLc*) – or the number of different meanings that a word has (for example, in English, *mine* can be something you own or an explosive device). The results show that stories that contain words with more potential meanings are associated with more incorrect answers and fewer correct answers. Polysemous words have been found to make mathematics word problems more difficult to interpret accross other studies [8-9].

Results also showed that higher type-token ratios (*LDTTRc*) are associated with more correct answers. As type-token ratio increases, more unqiue words are being used in the story problem, and fewer words are being repeated. These results suggest that students have an easier time writing the expression in a story that is relatively concise with little reptition of ideas. While it makes sense that this type of story may be more amenable to translation into mathematics notation, this result contrasts with research in text comprehension in reading tasks [14] which generally finds that repitition and lower type-token ratios facilitate reading comprehension. However, the story problems with high levels of word repetition frequently discuss complex topics of which students may lack familiarity, including operating capital, business inventory, and wholesale prices. In this way, a high type-token ratio may be indicative of a complex topic rather than increased readability in these story problems.

Students‘ tendency to seek hints when writing the algebraic expression is associated with a number of different readability factors. First, we see an effect for the length of the story text; students are more likely to seek hints for *one sentence story problems*, compared to problems that have two or more sentences. Having only one single sentence in a story problem might not be enough to ground or fully describe a linear rate of change as it arises in a real-world situation, and these overly-sparse stories might consequently inhibit performance.

In addition to greater difficulty of inhibition words, stories with *family words* and *motion words* were associated with greater hint-seeking. Only 13 of the problems involved family words, and these were often complex scenarios where multiple actors (e.g., a main character and his brother) were each contributing to the algebraic rate of change in their own way (e.g., saving/earning/splitting money together). Motion words often involved physics contexts (e.g., traveling in a car or plane) in which students had to track distance, rate, and time. This suggests that keeping track of multiple individuals engaging in mathematical actions and solving problems with physical distances and rates may be significant difficulty factors when writing expressions.

Finally, the regression results showed that scoring higher on Coh-Metrix’s second language readability measure (*RDL2*) was associated with greater hint-seeking when writing expressions. This measure is calculated through measures of word frequency (with words that occur more frequently in the English language yielding higher scores), sentence syntax similarity (with sentences that have similar grammatical structures yielding higher scores), and word overlap (with words that share semantic meaning yielding higher scores; [16]). Given that a higher second language readability score is typically associated with greater ease in comprehending the text [17], it is suprising that stories that score higher on this measure would be associated with students seeking more hints. The explanation of this finding may be similar to that for our finding with type-token ratio; story problems that use similar words and sentence structures often use a lot of reptition as a way to present complex ideas. Stories that are simple and concise may be easier for students to solve.

6. DISCUSSION

Results indicate that readability and topic measures have important associations with students‘ performance when solving mathematics word problems in an ITS. In particular, it was more difficult for students to name the independent and dependent quanitities in problems relating to motion (physics) and inhibition (saving and safety), while adverbial cues facilitated this skill. When writing algebraic expressions, we again see that motion and inhibition topics are difficult, but also find other important readability measures that matter. Words with multiple meanings make story problems more difficult, which corresponds to previous findings in both mathematics and reading education.

However, mathematics stories that use concise language with little repitition, which in terms of their readability level makes them technically *less* readable, are actually easier for students to solve. Thus measures of readability that stem from research on reading comprehension may need to be considered differently when working with mathematics problems. Results also suggest that while a story problem that includes only a single sentence is concise, it might present difficulty for students by not providing necessary context and information for them to feel they can respond without needing a hint.

Overall, our results suggest that mathematics story problems that have story texts that are more accessible to students have several characteristics: (1) they are concise with little repetition, but not a single sentence only, (2) they use only a single actor performing actions, (3) they use simple words with clear meanings, (4) they avoid more abstract physics or financial contexts, instead focusing on familiar contexts involving accumulation or loss of concrete physical objects, and (5) they make use of adverbial cues. Story problems with these characteristics may allow students to more easily construct a situation model from a propositional textbase. They may promote situation-model construction by both increasing students‘

ability to comprehend the semantics of the problem, and by increasing students' interest in working on the problem.

7. CONCLUSION

Future adaptive ITSs will be designed to model student characteristics at an extremely fine-grained level, as technology for personalized learning continues to advance. Here we argue that an important element of these future adaptive systems will be a consideration of the non-mathematical text-based characteristics of the problem tasks they present to students. Making inferences about students' current level of mathematical knowledge or motivation without considering these characteristics may lead to misspecifications.

Readability and topic measures may be an important consideration for ITSs to model in a variety of domains, including when considering tasks from history, social studies, and science. Future research should focus on the readability and topic measures that are most important for students of different age groups in different subject domains, and narrow down which characteristics are most critical to include in student and domain models as we build future ITSs. In current work, we are analyzing the mathematics problems on the National Assessment of Educational Progress (NAEP) and Trends in International Mathematics and Science Study (TIMSS) to examine how readability and topic measures impact the performance of 4th and 8th graders in the United States, and how these factors interact with cognitive and non-cognitive student background characteristics.

8. REFERENCES

[1] Padayachee, I. 2002. Intelligent tutoring systems: Architecture and characteristics. University of Natal, Durban, Information Systems & Technology, School of Accounting & Finance.

[2] 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.

[3] Nathan, M. J., Kintsch, W., Young, E.: A theory of algebra-word-problem comprehension and its implications for the design of learning environments. *Cognition and Instruction*, 9(4), 329-389 (1992)

[4] Walkington, C., Petrosino, A., Sherman, M. 2013. Supporting algebraic reasoning through personalized story scenarios: How situational understanding mediates performance and strategies. *Mathematical Thinking and Learning*, 15(2), 89-120.

[5] Walkington, C., Sherman, M., & Petrosino, A. 2012. 'Playing the game' of story problems: Coordinating situation-based reasoning with algebraic representation. *Journal of Mathematical Behavior*, 31(2), 174-195.

[6] Hidi, S., & Renninger, K. 2006. The four-phase model of interest development. *Educational Psychologist*, *41*(2), 111-127.

[7] Vilenius-Tuohimaa, P. M., Aunola, K., Nurmi, J. E. 2008. The association between mathematical word problems and reading comprehension. *Educational Psychology*, 28(4), 409-426.

[8] Shaftel, J., Belton-Kocher, E., Glasnapp, D., Poggio, J. 2006. The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment*, 11(2), 105-126.

[9] Wolf, M. K., Leon, S. 2009. An investigation of the language demands in content assessments for English language learners. *Educational Assessment*, 14(3-4), 139-159.

[10] Doddannara, L. S., Gowda, S. M., Baker, R. S., Gowda, S. M., De Carvalho, A. M. 2011. Exploring the relationships between design, students' affective states, and disengaged behaviors within an ITS. In: *Proceedings of the 16th International Conference on Artificial Intelligence and Education*, pp. 31-40.

[11] Baker, R. S., de Carvalho, A. M. J. A., Raspat, J., Aleven, V., Corbett, A. T., Koedinger, K. R. 2009. Educational software features that encourage and discourage "gaming the system." In: *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, pp. 475-482.

[12] Walkington, C. 2013. Using learning technologies to personalize instruction to student interests: The impact of relevant contexts on performance and learning outcomes. *Journal of Educational Psychology*, 105(4), 932-945.

[13] Bernacki, M. & Walkington, C. 2014. The Impact of a Personalization Intervention for Mathematics on Learning and Non-Cognitive Factors. Submitted to the *2014 International Conference of Educational Data Mining*, London.

[14] Graesser, A. C., McNamara, D. S., Louwerse, M. M., Cai, Z.: Coh-Metrix: Analysis of text on cohesion and language. Behavior Research Methods, Instruments, & Computers, 36(2), 193-202 (2004)

[15] Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., Booth, R. J.: The development and psychometric properties of LIWC2007. Austin, TX, LIWC. Net. (2007)

[16] Crossley, S., Allen, D., McNamara, D. 2011. Text readability and intuitive simplification: A comparison of readability formulas. Reading in a Foreign Language, 23(1), 84-101.

[17] Crossley, S. A., Greenfield, J., McNamara, D. S. 2008. Assessing text readability using cognitively based indices. TESOL Quarterly, 42(3), 475-493.

The Impact of a Personalization Intervention for Mathematics on Learning and Non-Cognitive Factors

Matthew Bernacki
University of Nevada, Las Vegas
4505 S. Maryland Parkway
Las Vegas, NV 89012, USA
1-702-895-4013
matt.bernacki@unlv.edu

Candace Walkington
Southern Methodist University
3011 University Blvd. Ste. 345
Dallas, TX, 75205, USA
1-214-768-3072
cwalkington@smu.edu

ABSTRACT
Personalization of learning environments to the background characteristics of learners, including non-cognitive factors, has become increasingly popular with the rise of advanced technology systems. We discuss an intervention within the Cognitive Tutor ITS where mathematics problems were personalized to the out-of-school interests of students in topic areas such as sports, music, and movies. We found that relative to a control group receiving normal problems, personalization had benefits for interest and learning measures. However, personalization that included deeper connections to students' interests seemed to be more effective than surface-level personalization.

Keywords
Personalization; interest; mathematics; intelligent tutoring systems

1. INTRODUCTION
The question of how to enhance the interest and motivation of adolescents has gained increasing prominence [1] especially in secondary mathematics [2]. Students often find mathematics, especially the math in middle and high school, to be disconnected from their interests, everyday lives, and typical ways of thinking about relationships and quantities [3]. At the same time, young people are using increasingly sophisticated and technology-driven ways to pursue and learn about their non-academic interests, and have become accustomed to a high level of customization, interaction, and control when seeking knowledge [4].

As a result, the idea of designing and advancing highly *personalized* systems for student learning has become a central focus for educational stakeholders [5]. Technology systems that enact personalized learning in the classroom have the potential to intelligently adapt to students' prior knowledge, interests, preferences, and goals [4]. In mathematics, these systems can make explicit connections between the interests students pursue outside of school – like sports, video games, or social networking – and the academic concepts they are learning. Algebra in particular is a rich space for such connections to be made [6] – students experience mathematical concepts like rate of change as they gain points in their favorite video game, track their pace in cross country, or accumulate followers on Instagram. As Algebra is often considered to be a gatekeeper to higher-level mathematics [7], and a subject that adolescents struggle to see as relevant [3], it may be a particularly important area for the development of interventions for personalized learning. We posit that 1) using a technology-based system for personalization that grounds algebra problems in students' out-of-school interests has the potential to elicit students' interest in the mathematics content to be learned, and 2) that personalization to well-developed individual interests can have a long-term effect on students' learning of algebraic concepts and their motivation to learn mathematics.

2. THEORETICAL FRAMEWORK
Interest has been defined as being both the state of engaging and the predisposition to re-engage with particular activities, events, and ideas over time [8]. Researchers have defined two types of interest. *Situational interest* is a state of heightened attention and increased engagement elicited by elements of an environment that are surprising, salient, evocative, or personally relevant. Situational interest can be *triggered* in response to stimuli, and becomes *maintained* over time as a learner engages further with the stimuli [8]. *Individual interest* is an enduring preference for certain objects or activities that persists over time and involves knowledge, value, and enjoyment; individual interest can be *emerging* or *well-developed*.

Situational interest can also be subdivided into interest based on *enjoyment* of the activity and interest based on *valuing* of the activity with respect to other things the learner values. Value-based situational interest has also been referred to as utility value – a learner's awareness of the usefulness of a topic to their life and goals [9]. Interventions that are intended to trigger students' situational interest are sometimes called "catch" interventions – the idea is to immediately grab students' attention through salient, evocative, relevant, or surprising characteristics of the instructional materials. Interventions that are designed to promote maintained situational interested as sometimes called "hold" interventions – they often reveal the value of the content to students' lives and goals, seeking to empower students [10-12]. For example, Mitchell [4] proposed that activities involving group work, computers, and puzzles function as "catch" mechanisms in the secondary mathematics classroom, while meaningfulness and involvement "hold" situational interest. Research has shown that when individuals are interested in a task or activities, they engage in more productive learning behaviors and have improved learning outcomes [e.g., 13].

An important question, then, is how to elicit and develop learners' interests for academic content areas. *Personalization* is a particular kind of intervention that can be used in learning environments to accomplish this goal. Personalization interventions identify topics for which learners have emerging or well-developed individual interest, and then connect these topics to academic content topics they are learning about in school (like algebra), for which they may have a lower level of interest. For example, consider a student who has a well-developed individual interest in music, but is not interested in Algebra. In their Algebra I class, they may engage with a variety of problems and projects that explore the mathematics behind musical pieces. Over time, the connection between these two areas might support her in developing situational interest based on her enjoyment of the incorporation of music as a context and the value perceived for music-themed problems, ultimately leading to the development of individual interest in Algebra [14]. By making explicit

Published in CEUR-WS:
NCFPAL workshop (Ritter and Fancsali)
In EDM 2014 Extended Proceedings (Gutierrez-Santos and Santos)

connections to students' interests, personalization interventions are hypothesized to trigger situational interest in the academic content being learned, which can be maintained over time and eventually develop into individual interest in that content area. Personalization can increase students' engagement in the math task, improve their performance on personalized math tasks and future math tasks that are not personalized [15], and may even increase students' interest in the math they now see as relevant to their personal interests. However, little research has investigated the mechanisms by which personalization promotes these learning outcomes. In this study, we test this *situational interest hypothesis* by monitoring students' interest in math units via embedded self-report surveys and examining whether personalization induces higher levels of situational interest, and whether this situational interest transforms into individual interest. Thus we test whether increased situational interest is an important mechanism through which personalization may gain its effect.

In addition to possessing enjoyment and value components, Renninger, Ewen, and Lasher [16] accentuate that interest also involves knowledge. Learners tend to possess useful prior knowledge related to their areas of interest, but this knowledge may be intuitive and informal with respect to underlying principles, making connections to concepts being learned in school (like algebra) difficult to acknowledge or articulate. In addition to possessing the potential to spur enjoyment and value-driven reactions to an academic content area, personalization is advantageously positioned to formalize students' intuitive prior knowledge about their interests by explicitly connecting it to a concept learned in school. For example, a learner with substantial knowledge of musical composition may have implicit understandings of the mathematical or numerical underpinnings of music, and this knowledge can potentially act as a support when they are learning formal algebra. In mathematics education, this follows a "funds of knowledge" perspective [17], which accentuates that students bring with them to the classroom powerful quantitative ways of reasoning from their home and community lives. These informal, interest-based funds of knowledge are potential strengths that can be leveraged through thoughtful instructional approaches like personalization to develop students' algebraic knowledge. In this study, we test the *funds of knowledge hypothesis* by examining whether solving personalized problems that incorporate deeper features of one's interest (e.g., mechanics of a popular video game) elicit stronger effects on learning than problems personalized based on shallower features of a learner's interest (e.g. passing reference to a game title in a problem about snacking) or non-personalized problems. Thus we test whether increased activation of prior knowledge is an important mechanism through which personalization gains its effect.

Whereas outside interests can be leveraged by personalization, initial interest in mathematics may moderate the effectiveness of personalization interventions. Durik and Harackiewicz [10] found that an intervention designed to "catch" (i.e., trigger [8]) student interest (adding colorful, vivid decorations to instructional materials) was most effective for learners with low individual interest in mathematics (IIM), but hampered learners with high IIM. Conversely, they found that an intervention designed to "hold" (i.e., maintain based on value [8]) student interest (informing students of the value of the content being learned) was beneficial for high IIM students, and detrimental for low IIM students.

In order for personalized instructional materials to successfully activate knowledge, trigger interest, and enhance perceptions of value, Walkington and Bernacki [14] identified three key features designers must consider. First is the *depth* of the intervention – whether the personalization draws upon surface level aspects of a learners' interest (e.g., simply inserting familiar objects or names into an already-designed task), or whether the personalization involves deep, authentic connections to actual experiences the learner has pursuing an interest like music. Second is the *grain size* of the intervention – whether the personalization is targeted to the specific experiences of an individual, or to the generic experiences of an entire group. When considering grain size, it is important to remember that some topics will tend to tap into the interests of larger groups of students more than others – for example, a problem about the specifics of football may match the fine-grained interests of more ninth graders than a problem about field hockey. Use of these topics that relate to many students' experiences may be a productive way to allow materials to be personalized at a finer grain size. Third is the *ownership* of the personalization – whether the students themselves take a role in generating the connections between the academic content area and their interests, or if teachers or curriculum developers control the personalization. In this study, we examined students' interest in mathematics and algebra learning when exposed to a personalization intervention of medium grain size (i.e., personalized for local users based on interest interviews conducted at the same school in a prior year) versus a standard set of problems (i.e., broad grain size written by curriculum developers for all Algebra I students who use the curriculum). In the fourth unit of the intervention, we also varied the depth of problems by personalizing on surface or deep features of the problem to examine the effects of depth on interest and learning (i.e. the funds of knowledge hypothesis). No manipulation of problem ownership was conducted.

In the present study, we pursue the following research questions by implementing a personalization intervention for Algebra I:

1) What is the immediate impact of a personalization intervention on students' situational interest in algebra instructional units?
2) What long-term effect does personalization have on students' individual interest in algebra?
3) What is the impact of a personalization intervention on students' learning of algebra concepts?
4) How does depth influence the impact of personalization on interest and learning?

Based on prior work examining the effects of personalization on learning [15] and theoretical assumptions about the development of interest [8] including the situational interest hypothesis, we hypothesize that 1) Personalized problems should trigger greater situational interest in algebra units than standard problems; 2) Students completing personalized problems that incorporate out of school interests will report greater individual interest in algebra; and 3) Students who complete personalized problem solving units will achieve greater increases in their algebra performance than students completing standard problem solving units. In accordance with the funds of knowledge hypothesis, we expect 4) that students who complete problems that are personalized based on deeper features of their interest area should outperform those completing problems personalized on surface features of the problems and standard problems.

3. METHODS

3.1 Participants and Environment

Total participants included $N = 152$ ninth grade Algebra I students in the classes of two Algebra I teachers. Students attended a rural

Northeastern school that was 96% Caucasian with 21% of students eligible for free or reduced price lunch. In 2012, 71% of students passed the state standardized test in Mathematics, which is administered in the 11th grade. The sample was 51% female. Because one teacher at the school site did not administer the pretest before students began using the Cognitive Tutor, eighty-three students completed pretest, posttest and all questionnaires delivered in the CTA software and compose the primary sample for this study.

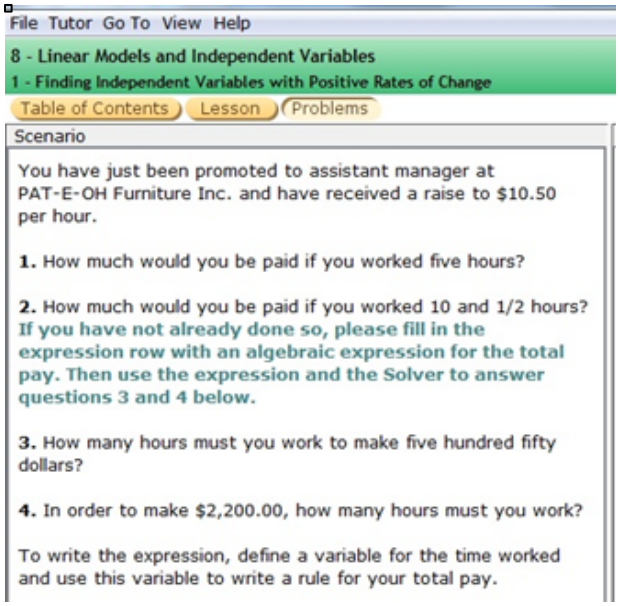
The school at which the study took place used the Cognitive Tutor Algebra (CTA) curriculum [18]. CTA is an intelligent tutoring system for Algebra I that uses *model-tracing approaches* to relate the students' actions back to the domain model to provide individualized error feedback. CTA also uses *knowledge-tracing approaches* to track learning from one problem to the next, using this information to identify strengths and weakness in terms of production rules. CTA presents learners with algebra story problems where they must navigate tabular, graphical, and symbolic representations of functions (Figure 1). Students in schools that use CTA typically use the software 2 days per week.

4. Personalization Intervention

Before entering the first unit in CTA (Unit 1), all participants were given an interests survey where they would rate their level of interest in 10 topic areas – music, art, cell phones, food, computers, games, stores, TV, movies, and sports. Participants were then assigned to one of two main conditions: (1) a Control Condition that received the standard algebra story problems in all units in CTA including Units 1, 3, 7, and 9 covering linear equations, (2) an Experimental Condition that received versions of these same problems with the same underlying structure that were matched to the interests they indicated on the interests survey for Units 1, 3, 7, and 9 (i.e. Personalization Condition). In unit 9, we tested the funds of knowledge hypothesis by further subdividing learners in the Personalization condition to (A) a Deep Personalization condition where they received personalized problems with greater depth – i.e., the personalized problems the Deep Personalization group received in Unit 9 were written to better correspond to ways that adolescents might actually use linear functions when pursuing their interests, and were intended to draw upon "funds of knowledge" more explicitly. The remaining students were assigned to (B) a Surface Personalization Condition where they received problems that contained stories with only superficial references to their identified interests. These problems should elicit situational interest, but not draw upon knowledge about one's interests.

In the first sample Control problem in Table 1, students must identify the relationship between dosage and weight. This relationship is grounded in a story that provides a context that likely to be of limited relevance to the student. In the Surface Personalization problem the structure of the problem remains consistent, but a topic that corresponds to the learners' personal interests has been applied. In the Deep Personalization version, the personal interest is applied more intentionally. Like the surface-level personalization problem, The Clash of Clans problem matches students' reported interest in games. However it is also intended to draw upon the learner's knowledge of the game's architecture to frame the underlying algebraic relationship to be learned in a deeply relevant context (i.e. it is actually useful to keep track of the relationship between elapsed time and how goals are accomplished, and this quantity is explicitly tracked and displayed for the player within the game interface). We consider this to be a deeper level of personalization compared to the Surface Personalization condition, as it seems less likely that despite an interest in games, a teen would care about or track exactly how frequently they consume snacks during play. Personalized problems were written based on surveys (N = 45) and interviews (N = 23) with Algebra I students at the school where they discussed their out-of-school interests.

Deep Personalization problems were written to more closely correspond to quantitative information given by students in the interviews and open-ended surveys about their out-of-school interests, including interviews with Algebra I students at the school where the study was conducted. In these interviews, students discussed how they consider rate of change as they play video games, participate in sports, track their rate of texting and battery usage on their cell phone, engage in cooking, work at part-time jobs, activities, and so on. (see [6] for a full analysis of student interviews).



Answer Key:

Quantity Name	the time worked	the money earned
Unit	hour	dollar
Expression	X	10.5X
Question 1	5	52.5
Question 2	10.5	110.25
Question 3	52.381	550
Question 4	209.5238	2200

Figure 1. Screenshot of Cognitive Tutor Algebra environment with answer key superimposed

Table 1. Study Conditions

	Control	Surface Personalization	Deep Personalization
GAMES	The correct dosage of a certain medicine is two milligrams per 25 pounds of body weight.	While playing cards a person typically eats two snacks for every 25 minutes of playing time in a card game.	When playing Clash of Clans a player can build two barracks for every 25 minutes of playing time.
SPORTS	Three out of every five people in a recent survey supported the President's Health Plan.	Three out of five people have attended a Pittsburgh Steelers game in their lifetime.	Three out of five free throws are successful for NBA players.
FOOD	Directions for a swimming pool chemical that controls the growth of algae state that you should use six fluid ounces of chemical for every 500 gallons of water.	Looking through a collection of online recipes, there are six recipes that require powdered sugar for every 500 recipes that you find online.	In a family recipe you use six drops of hot pepper oil for every 500 ounces of chili that is being cooked.

Problems across the 3 conditions were written to hold constant factors like order of information given, numbers, sentence structure and length, mathematical vocabulary, readability, pronoun use, and distractor information. The personalized problems did *not* require that students have additional knowledge of specific numerical mathematical information in their interest area (e.g., knowing how many points a field goal is worth) – all information given was matched across problem types.

All instructional units involved in the study involved linear functions. Of the core sample comprising most of our analyses, 31 participants were assigned to the Control, 34 were assigned to Surface Personalization, and 27 were assigned to Deep Personalization.

4.1 Measures

We collected the following measures from all participants:

4.1.1 Paper-Based Pre/Post Assessments

At the beginning of the school year, prior to entering the tutor, all students completed a paper-based pre-test on linear functions. The test contained 4 story problems where a linear function was described that either had a slope and intercept (2 problems) or had only a slope (2 problems). Participants first were given an x value in the linear function and asked to solve for y, then they were given a y value in the linear function and asked to solve for x. Finally, they were asked to write the linear function using algebra symbols. A post-test was administered to all students around the midterm of their ninth grade year (i.e., four months later). The post-test contained 4 matched items containing slightly different wording and numbers. Students' responses to each part of each problem were scored as correct or incorrect.

4.1.2 Domain-Level Motivational Surveys

Prior to entering Unit 1 (pre-) and Unit 10 (post-) in CTA, the software presented students with a survey asking them to rate their attitudes about algebra. Specifically, they rated their individual interest in mathematics (IIM), as well as their maintained situational interest–enjoyment and maintained situational interest-value for mathematics. Subscales were adopted from a larger set of scales from Linnenbrink-Garcia et al. [19]. Sample items for each scale appear in Table 2.

4.1.3 Unit-Level Motivational Surveys

After each unit impacted by the personalization intervention (Figure 2; Units 1, 3, 7, and 9), participants were also given a unit-level motivational survey that assessed the degree to which that unit triggered their situational interest and maintained their situational interest in the CTA unit. These scales were adapted based on measures from Linnenbrink-Garcia et al. [19] with the math unit as the referent. Sample items for each scale appear in Table 2, as do Cronbach's alphas for the initial administration of each survey. An overview of the survey measures and CTA units completed by participants in this study is provided in Figure 2.

Table 2. Interest Measures

Interest Measure	Sample item	α
Individual Interest in Mathematics	Thinking mathematically is an important part of who I am.	.92
Maintained Situational Interest in Math- Value	What we are studying in math class is useful for me to know.	.92
Maintained Situational Interest in Math- Enjoyment	I really enjoy the math we do in this class.	.89
Triggered Situational Interest in Math	The topics in this unit grabbed my attention.	.84
Maintained Situational Interest in Unit - Value	The math in this unit is useful for me to know.	.90
Maintained Situational Interest in Unit - Enjoyment	In this unit, I really enjoyed the math.	.84

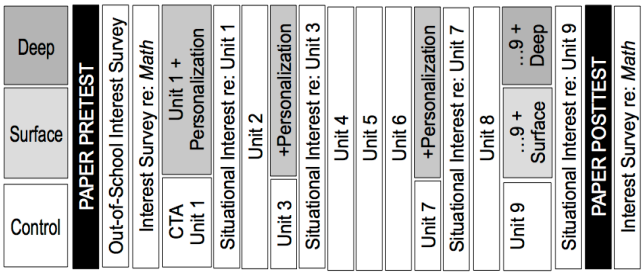


Figure 2. Measures

5. RESULTS

We report results as they address the first three research questions in section 2. We do not provide a separate section for research question 4 (impact of depth of personalization), and instead discuss the results for depth of personalization within each of the other three sections.

5.1 What is the impact of personalization on students' situational interest in algebra units?

To assess the effect of the personalization interventions on students' situational interest, we conducted a series of analyses of covariance examining students' reported triggered and maintained interest in CTA units. All students were given unit-level surveys assessing their level of interest in the instructional unit after each of the units impacted by the personalization treatment (Units 1, 3, 7, and 9). We controlled for initial individual interest in mathematics (IIM) as indicated on the domain survey before Unit 1 (Figure 2).

Students in the two Personalization conditions (i.e., Surface Personalization and Deep Personalization are identical in Units 1, 3, and 7) consistently reported significantly higher levels of triggered situational interest than students assigned to the Control condition (Table 3; Unit 1 $F(1,80) = 5.19$, MSe = .96, $p = .03$, Unit 3 $F(1,80) = 5.31$, MSe = .98, $p = .02$; Unit 7 $F(1,80) = 3.82$, MSe = .91, $p = .05$).

Significant differences between any of the 3 groups in triggered situational interest were not obtained in Unit 9. The level of triggered situational interest reported by the Deep Personalization was consistent with prior units with the triggered interest for the Surface Personalization group was slightly lower. The Control group, however, reported greater triggered situational interest, and the inclusion of three groups (two with smaller Ns) further diminished the statistical power available to detect effects.

No significant differences in maintained situational interest were found between groups on any of the four units observed, Fs < 3.73, ps = ns. Directionally, measures of maintained situational interest generally favored the personalization groups.

5.2 What effect does personalization have on students' individual interest in algebra?

All students were given domain-level surveys assessing their interest towards learning algebra prior to the intervention and after the final personalized unit (i.e., Unit 9). A repeated measures analysis of variance examining change in Individual Interest in Mathematics (i.e., Post-Pre) between the two Personalization conditions (i.e., Deep & Surface) versus Control was conducted to examine the main effect of Time and Interaction between Time X Condition. Results indicated a significant main effect of Time, F (1, 81) = 5.39, MSe = 1.75, p = .023. Overall, students' individual interest in mathematics declined from pretest to posttest. Analyses also indicated a marginally significant interaction between Time and Condition, F (1, 81) = 3.73, $p = .057$. Students in the control group significantly *reduced* their rating of individual interest in algebra an average of 0.37 points over the 10-unit span (Table 3; $t(29) = 3.21$, $p < .01$), while students in the Deep and Surface Personalization groups maintained their individual interest in algebra ($M = 0.04$ decline). Thus personalization had a positive effect in that it preserved students' individual interest in algebra. Within the Personalization condition, no differences were found between students who received Surface versus Deep Personalization.

5.3 What is the impact of personalization on students' learning of Algebra I concepts?

The pre- and post- test scores on the algebra learning measures for each of the three conditions is shown in Table 4. A linear regression model predicting amount of absolute gain from pre- to post-test (i.e., post-test score minus pre-test score) was fit to the data, with students' class period as a random effect. Adding a predictor for Condition significantly improved the fit of the model ($\chi^2(2) = 6.39$, $p = 0.04$), as did a control variable for students' initial level of individual interest in mathematics (IIM) prior to the intervention ($\chi^2(1) = 4.07$, $p = 0.04$). The interaction of Condition and IIM also significantly improved the fit of the model ($\chi^2(2) = 14.43$, $p < .001$).

Table 3. Estimated Marginal Means Controlling for Individual Interest in Math

Variable			Personalization[a]		Control[b]		
	Unit		EMM	SE	EMM	SE	
Triggered Situational Interest	1		2.86	0.13	2.33	0.19	*
	3		2.82	0.13	2.27	0.19	*
	7		2.69	0.13	2.25	0.18	*
	9	D[c]	2.82	0.18	2.55	0.19	
		S[d]	2.56	0.20			
Maintained Situational Interest - Value	1		2.95	0.13	2.77	0.19	
	3		3.07	0.13	2.74	0.18	
	7		2.76	0.13	2.76	0.18	
	9	D	2.84	0.19	2.82	0.18	
		S	2.70	0.17			
Maintained Situational Interest - Enjoyment	1		2.76	0.12	2.46	0.17	
	3		2.81	0.13	2.40	0.18	
	7		2.66	0.12	2.35	0.17	
	9	D	2.62	0.19	2.50	0.18	
		S	2.33	0.17			
Individual Interest in Math	Pre		2.87	.14	3.34	.20	
	Post		2.83	.16	2.94	.22	

Notes. *- $p < .05$, EMM = Estimated Marginal Mean, SE = Standard Error, D = Deep personalization, S = Surface Personalization, [a] - N = 55, [b] - N = 28, [c] - N = 24, [d] - N = 31

Table 4. Scores on Knowledge tests by Condition

		Pretest		Posttest	
Condition	N	M	SD	M	SD
Control	32	0.68	0.2	0.83	0.12
Surface Personalization	29	0.73	0.15	0.82	0.15
Deep personalization	32	0.63	0.22	0.84	0.18

The regression output is shown in Table 5. The reference category is the Control Group, and we interpret all significant simple effects regardless of whether they are displayed in the table. The IIM control measure was dichotomized to separate students with high IIM (average rating of 3 or more) from low IIM (average rating less than 3) to aid interpretability and to be consistent with prior work [e.g., 14]. As can be seen from Table 5, for students with low individual interest in math, Deep Personalization was significantly more effective than Control ($p < 0.05$). Additional contrasts not shown in the table compared Surface Personalization to Deep Personalization, and found that for students with low IIM,

Deep Personalization was significantly more effective than Surface Personalization (B = 0.24, *SE* (B) = 0.07, $p < 0.001$). Finally, within the Deep Personalization condition, students with high IIM gained significantly less than students with low IIM (B = .17, *SE*(B) = .07, $p = .01$).

Table 5. Regression Output for Pre/Post Learning Gains

	B	SE (B)	*t*	*p*
(Intercept)	.13	.07	1.81	.07
Control	(ref.)			
Surface Personalization	-.10	.08	-1.33	.18
Deep Personalization	.14	.07	1.97	.05
Low IIM	(ref.)			
High IIM	.00	.07	-.07	.94
Surface Personalization × High Initial Individual Interest	.08	.10	.82	.41
Deep Personalization × High Initial Individual Interest	-.17	.10	-1.71	.09

6. DISCUSSION & CONCLUSION

This study examined whether personalizing algebra problems to students' out-of-school interests would increase their situational interest in CTA algebra problems, increase their interest in mathematics, and improve their acquisition of algebra knowledge (i.e., the situational interest hypothesis). It additionally tested whether solving problems that incorporated deep features of an interest into problems would produce greater benefits that solving problems that incorporated interests superficially or standard problems (i.e. the funds of knowledge hypothesis). Students who received problems personalized to their out-of school interests reported significantly higher triggered situational interest for CTA math units. Compared to a Control group that experienced a drop in their individual interest in mathematics, Personalization also had a preserving effect on students' interest in mathematics. After accounting for students' initial individual interest in mathematics, significant differences in learning gains were found between groups of students in the Deep Personalization, Surface Personalization and Control Conditions. These findings are next discussed in light of prior theory and research.

6.1 Personalization and Situational Interest

Students who completed algebra problems personalized to their interests reported greater triggered situational interest compared to students who completed standard CTA problems, however students who solved personalized problems did not report significantly greater maintained interest resulting from enjoyment or perceptions of value. The finding that personalization was effective in triggering situational interest is encouraging as we consider the Control condition to be a considerably strong control. That is, the standard problems included in tutor units might be considered to be personalized to student interests at a very broad grain size [11] – they were generally written by teachers and curriculum writers with this student population in mind (i.e., adolescent algebra learners). The personalized problems in the intervention, on the other hand, had a medium grain size – they were written for and provided to subsets of the student population that had particular topic interests (e.g., sports, video games). The change from a large to a medium grain size was sufficient to elicit changes in triggered situational interest, though additional effort may be necessary to elicit sufficient enjoyment or perception of value to maintain students' situational interest. Indeed, in another personalization study [20], we found that a personalization intervention with a much smaller grain size where students wrote and solved problems that incorporated features of their personal interests produced increases in students' maintained situational interest associated with perceived value. This intervention also involved a higher level of ownership of the personalization on the part of the students [14], which suggests that personalization at a medium grain size may successfully trigger situational interest, but a personalization at a smaller grain size with some level of ownership may be necessary to achieve more enduring situational interest in math units. This type of intervention may be especially important given that it takes the burden of generating fine-grained instructional materials away from teachers and curriculum developers and places it on students.

6.2 Personalization and Individual Interest

Despite a failure to elicit maintained situational interest, the Personalization intervention did have a significant effect on students' individual interest in mathematics. Importantly, the individual interest items assessed how students felt about the domain of mathematics as a whole, rather than how they felt about the particular math class they were enrolled in or the particular units they were working on. This preservation of individual interest in algebra over half a year of high school coursework is a desirable outcome, given research that documents declines in interest in math over adolescence [21, 22]. In sum, the findings from the first two research questions support the situational interest hypothesis. We consider this finding in light of theory on interest development in section 6.4.

6.3 Deep Personalization and Algebra Learning

Walkington [12] found that a one-unit personalization intervention improved students' long-term learning of algebra concepts within the CTA environment, relative to a control condition. This study extends that work and indicates that, when personalization incorporates deep features of students' out-of-school interests, it can also induce learning gains that transfer outside of an intelligent tutoring environment (i.e. to delayed, paper-based tests). However, these effects are moderated by students' initial level of individual interest in mathematics, with Deep Personalization being beneficial mainly for low IIM students. Walkington [15] did not collect such interest measures in her study, but did find that personalization was most effective for students who were making slower progress through CTA– a variable known to track closely with interest in math [23]. We consider these findings in light of proposed hypotheses that personalization may obtain effects on learning by activating students' funds of knowledge in their out-of-school interest, and that personalization may trigger greater situational interest in math tasks. The current study showed that Deep Personalization was significantly less effective for learners with high IIM, compared to learners with low IIM. This, along with the results that personalization triggers but does not maintain situational interest, suggests that even Deep Personalization may achieve its effects on learning as a "catch" intervention, immediately eliciting triggered situational interest. That is, solving personalized problems triggered students' interests, but did not maintain them. This provides some promise as prior research has shown catch interventions that trigger interest to be beneficial primarily for learners with low IIM [10]. This is contrasted with a "hold" intervention that maintains situational interest, often by communicating the value of the content being learned. In this study personalization did not increase students' perceptions that

algebra problems had value, but additional interventions aimed at boosting perceived value and relevance [11, 12] could potentially be incorporated to ITSs to also obtain this effect and its benefits for learning.

Although we termed our Condition "Deep" Personalization, the connections made to learners' actual experiences may not have been uniformly deep depending on students more specific interests within a topic area, and thus may not have elicited value-based reactions from some students. This stems from issues with the grain size of the intervention – students merely indicated their level of interest in a broad topic (e.g., "sports"), and were then given problems that could cover the entire space of activities that fell within that topic (e.g., basketball, hockey, football), without considering students more specific interest in a subtopic (e.g., just hockey). Although attempts were made to use the "high-leverage" interest sub-topics that many students would have specific knowledge of (i.e., football rather than field hockey) this approach likely allowed for the personalization to have highly variable level of correspondence to students' exact interests. The level of correspondence depended on the overlap between a student's interest and the commonly reported interests by peers in surveys and interviews prior to problem development. Walkington and Bernacki [20] found significant increases in maintained situational interest (value) for students who authored problems about their specific interests, suggesting that the smaller grain size and increased ownership of the personalization intervention in that study allowed it to function more as a "hold" intervention.

Finally, the current study showed that Deep Personalization was significantly more effective than Surface Personalization for students with low IIM. This suggested that personalization may need to have at least a moderate level of depth for it to be effective at all for supporting learning outcomes for any subgroup of students. Indeed, a number of recent personalization interventions that employed relatively surface-level personalization have reported null findings [24, 25]. Thus we conclude from all of these analyses that a personalization intervention with a moderate depth and grain size can potentially have long-term effects on student learning for students who begin with limited interest in mathematics. However, increasing depth and personalizing at an even smaller grain size may have more powerful effects, especially for students with higher IIM for whom value-based connections may be most critical.

Although learning gains were produced for low IIM students who received Deep Personalization (rather than Surface Personalization), these students did not show differences in situational or individual interest measures within Unit 9 compared to the Surface Personalization group. There were also no differences between Surface and Deep in individual interest over the course of the entire intervention. This suggests that Deep Personalization may gain its effectiveness over Surface Personalization by connecting to students' prior knowledge (funds of knowledge hypothesis) rather than triggering and maintaining differing levels of situational interest (situational interest hypothesis). However, ultimately comparisons between these two groups are of limited usefulness given the relatively small sample sizes. Thus we find limited but promising support for the funds of knowledge hypothesis.

6.4 Theoretical Implications

When viewed through the lens of interest development theory [8], the findings regarding personalization and interest development are somewhat puzzling. Per Hidi and Renninger's [8] theory, interest is 1) triggered by environmental stimuli and 2) maintained when engagement in the environment is enjoyable or confers value through consistent or repeated situational interest. This supports 3) the emergence of an individual interest, which 4) becomes well developed over time. In this study, analyses reveal a triggering of situational interest among students in the Surface and Deep Personalization conditions, no reported maintenance of situational interest via enjoyment or value, but a significant effect of Personalization on individual interest. Thus individual interest developed without being maintained during learning; this requires that we consider alternate explanations by which such effects on individual interest may have been obtained.

One potential explanation is that the way instructors used Cognitive Tutor in the math classes may have reproduced some of the behaviors expected when students' situational interest is maintained. In their model, Hidi and Renninger [8] describe that those who maintain interest in a topic tend to repeatedly engage with content involving the topic (e.g., a student who is interest in dolphins may seek more opportunities to learn about them by reading books about them in school or choose "dolphins" as a topic for school assignments). While students' did not report that personalized Cognitive Tutor Algebra units maintained their interest to a degree that we would expect them to voluntarily seek out opportunities to learn using Cognitive Tutor, the compulsory use of the Cognitive Tutor in math class twice a week for many months effectively ensured repeated engagement in (personalized) problem solving via CTA use. Thus we could conclude that the continued exposure to math content personalized to one's out-of-school interests approximated behavioral outcomes of maintained situational interest and created an alternate pathway by which individual interest was preserved in Personalization conditions (i.e., no drop in interest), but not in the Control condition where there was no initially triggered interest. Much like the typical adolescent whose interest in math declines over time, students in the Control condition were required to complete math units that did not trigger situational interest and subsequently reported declines in their interest in mathematics.

6.5 Conclusion

The results obtained in this study provide important insight about the ways depth and grain size of personalization may impact the development of students' interests in their math course, the domain of mathematics, and ultimately their long-term learning of algebra concepts. In future analyses, we will analyze additional data from students participating in this study, and look for difference in in behavior and performance within intervention and subsequent CTA units, including analyses of learning behaviors using log-files and automated detectors.

7. ACKNOWLEDGMENTS

Both authors contributed equally to this manuscript. The authors thank Steve Ritter, Susan Berman, Tristan Nixon and Steve Fancsali (Carnegie Learning), Gail Kusbit (Carnegie Mellon University & LearnLab) and participating teachers. Funding for the study was provided by a subgrant of National Science Foundation Award # SBE-0354420. Additional funding as provided by IES Award # R305B100007.

8. REFERENCES

[1] Hidi, S., & Harackiewicz, J. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. Review of Educational Research, 70, 151–179.

[2] Mitchell, M. (1993). Situational interest: Its multifaceted structure in the secondary school mathematics classroom. *Journal of Educational Psychology*, *85*, 424–436.

[3] McCoy L. P. (2005). Effect of demographic and personal variables on achievement in eighth-grade algebra. *Journal of Educational Research*, *98*(3), 131–135.

[4] Collins, A., & Halverson, R. (2009). *Rethinking Education in the Age of Technology: The Digital Revolution and Schooling in America*. New York: Teachers College Press.

[5] U.S. Department of Education, Office of Educational Technology, Transforming American Education: Learning Powered by Technology, Washington, D.C., 2010. http://www.ed.gov/sites/default/files/netp2010.pdf

[6] Walkington, C., Sherman, M., & Howell, E. (in press). Connecting Algebra to sports, video games, and social networking: How personalized learning makes ideas "stick." *Mathematics Teacher*.

[7] Moses, R., & Cobb, C. (2001). *Radical Equations: Math Literacy and Civil Rights*. Boston: Beacon Press.

[8] Hidi, S., & Renninger, K. (2006). The four-phase model of interest development. *Educational Psychologist*, *41*(2), 111-127.

[9] Eccles, J. (1983). Expectancies, values and academic behaviors. In R. C. Atkinson, G. Lindzey, & R. F. Thompson (Eds.), *Achievement and achievement motives: Psychological and sociological approaches* (pp. 75-146). San Francisco: W.H. Freeman & Co

[10] Durik, A., & Harackiewicz, J. (2007). Different strokes for different folks: How individual interest moderates effects of situational factors on task interest. *Journal of Educational Psychology*, *99*(3), 597-610.

[11] Hulleman, C. S., Godes, O., Hendricks, B. L., & Harackiewicz, J. M. (2010). Enhancing interest and performance with a utility value intervention. J*ournal of Educational Psychology*, *102*, 880-895.

[12] Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science*, *326*(5958), 1410–1412.

[13] Harackiewicz, J., Durik, A., Barron, K. Linnenbrink, E., & Tauer, J. (2008). The role of achievement goals in the development of interest: Reciprocal relations between achievement goals, interest, and performance. *Journal of Educational Psychology*, *100*(1), 105-122.

[14] Walkington, C., & Bernacki, M. (in press). Motivating students by "personalizing" learning around individual interests: A consideration of theory, design, and implementation issues. In S. Karabenick & T. Urdan (eds.) *Advances in Motivation and Achievement*, Emerald Group Publishing.

[15] Walkington, C. (2013). Using learning technologies to personalize instruction to student interests: The impact of relevant contexts on performance and learning outcomes. *Journal of Educational Psychology*, *105*(4), 932-945.

[16] Renninger, K., Ewen, L., & Lasher, A. (2002). Individual interest as context is expository text and mathematical word problems. *Learning and Instruction*, *12*, 467-491.

[17] Civil, M. (2007). Building on community knowledge: An avenue to equity in mathematics education. In N. Nassir. and P. Cobb (Eds.) *Improving access to mathematics: Diversity and equity in the classroom* (pp. 105-117). Teachers College Press.

[18] Carnegie Learning (2013). Cognitive Tutor Algebra [software]. Carnegie Learning, Inc. Pittsburgh, PA, USA.

[19] Linnenbrink-Garcia, L., Durik, A., Conley, A., Barron, K., Tauer, J., Karabenick, S., & Harackiewicz, J. (2010). Measuring situational interest in academic domains. *Educational Psychological Measurement*, *70*, 647-671.

[20] Walkington, C., & Bernacki, M. (2014). Students authoring personalized "algebra stories": Problem-posing in the context of out-of-school interests. Presentation at the 2014 Annual Meeting of American Educational Research Association.

[21] Fredricks, J. A., & Eccles, J. (2002). Children's competence and value beliefs from childhood through adolescence: Growth trajectories in two male-sex-typed domains. *Developmental Psychology*, *38*, 519–533.

[22] Frenzel, A. C., Goetz, T., Pekrun, R., & Watt, H. M. G. (2010). Development of mathematics interest in adolescence: Influences of gender, family, and school context. *Journal of Research on Adolescence*, *20*, 507–537.

[23] Bernacki, M. L., Nokes-Malach, T.J., Aleven, V., & Glick, J. (2014). Intelligent tutoring systems promote achievement in middle school mathematics, especially for students with low interest. Presentation at the 2014 Annual Meeting of the American Educational Research Association.

[24] Bates, E., & Wiest, L. (2004). The impact of personalization of mathematical word problems on student performance. *The Mathematics Educator*, *14*(2), 17-26.

[25] Caker, O., & Simsek, N. (2010). A comparative analysis of computer and paper-based personalization on student achievement. *Computers & Education*, *55*, 1524-1531.

Promoting Growth Mindset Within Intelligent Tutoring Systems

Korinn S. Ostrow
Worcester Polytechnic Institute
100 Institute Road
Worcester, MA 01609
ksostrow@wpi.edu

Sarah E. Schultz
Worcester Polytechnic Institute
100 Institute Road
Worcester, MA 01609
seschultz@wpi.edu

Ivon Arroyo
Worcester Polytechnic Institute
100 Institute Road
Worcester, MA 01609
iarroyo@wpi.edu

ABSTRACT

When designing adaptive tutoring systems, a myriad of psychological theories must be taken into account. Popular notion follows cognitive theory in supporting multi-channel processing, while working under assumptions that pedagogical agents and affect detection are of the utmost significance. However, motivation and affect are complex human characteristics that can muddle human-computer interactions. The following study considers the promotion of the growth mindset, as defined by Carol Dweck, within middle school students using an intelligent tutoring system. A randomized controlled trial comprised of six conditions is used to assess various delivery mediums of growth mindset oriented motivational messages. Student persistence and mastery speed are examined across multiple math domains, and self-response items are used to gauge student mindset, enjoyment, and perception of system helpfulness upon completion of the assignment. Findings, design limitation, and suggestions for future analysis are discussed.

Keywords

Motivational messages, growth mindset, pedagogical agents, multi-media learning principles, e-learning design.

1. INTRODUCTION

The optimal design of adaptive tutoring systems is a continuous debate for researchers in the Learning Sciences. Decisions when authoring content can be immense, including not only the user interface and tutor material, but also the presence of adaptive feedback strategies such as hints or scaffolding, the use of affect detectors, and in growing popularity, the use of pedagogical agents. While many adaptive tutors share designs rooted in cognitive theory, creators should also incorporate elements that improve student motivation, engagement, persistence, metacognition, and self-regulation skills. These elements aid in the promotion of active learning, an experience that has been shown to heighten the creation of mental connections [10]. However, successful adaptive tutoring systems are not just a random conglomeration of these learning goals. All too often, adaptive tutors are designed under the assumption that students are ideal learners, driven and motivated, ready to employ a full range of self-regulation skills coupled with technological prowess [1]. Thus, researchers have recently undertaken a more thorough examination of how to universally encourage and motivate students while still promoting self-regulated learning skills and optimizing system design [3, 8].

Human motivation has historically been explained and argued by an array of theories, as intrinsic or as extrinsic, as static or as the constant flow of needs, emotions, and cognitions [13]. In a somewhat similar sense, recent research promoting affect detection within educational technology suggests that affect plays a primary role in learning success [2]. How can researchers incorporate deeply rooted human characteristics like motivation and affect into the design of an adaptive tutoring system? A renowned leader in the field of psychology, Carol Dweck has helped to establish theories of intelligence that marry these complex constructs within the confines of learning studies [5]. Her research has shown that students approach learning tasks largely with one of two 'mindsets.' The *fixed mindset* is characterized by the notion that intelligence is somehow innate or immutable. Students who live within this fixed realm generally emit lower learning and performance outcomes as well as higher attrition rates based in the notion that effort will not lead to intellectual advancement [6]. Much of American society is rooted in this view; strong emphasis is placed on standardized testing and zero sum competition, with the goal of comparing student intelligence rather than promoting learning. Alternatively, students with a *growth mindset* believe that intelligence is malleable and that effort and persistence can lead to success. While Dweck [7] argues that neither mindset is necessarily 'correct,' she promotes the notion that mindset can be altered, and explains the growth mindset as offering a healthier mental lifestyle. Altering mindset is best achieved by varying the type of praise students receive and by realigning their definition of successful learning. By highlighting the learning process rather than the student's intelligence or performance, 'process praise' and the promotion of malleable intelligence has led to positive, long-term learning gains [5]. Students trained in the growth mindset show increased enjoyment in difficult learning tasks as well as higher overall achievement and performance [6].

An expert in his own right, Richard Mayer has devoted much of his career to promoting a series of multi-media learning principles that enhance e-learning design. These principles call for learning environments to be driven by active learning processes while considering the cognitive load and working memory of users [4]. As such, those authoring adaptive tutors

Published in CEUR-WS:
NCFPAL workshop (Ritter and Fancsali)
In EDM 2014 Extended Proceedings (Gutierrez-Santos and Santos)

should utilize audio, animation, graphics, video, and other hypermedia elements to appease multiple sensory channels and thereby reduce the user's overall cognitive load. It is important to note that powerful design requires a fine balance of these resources, as exorbitance may serve to distract or disrupt learners. The evolution of pedagogical agents and learning companions within adaptive tutoring systems has served as a primary way to incorporate both multi-media elements and non-cognitive support. As guidelines for the design of human-computer interaction have followed those set forth by human-human interaction, the art of appropriating the cognitive and affective responses of pedagogical agents has been of major concern [9]. Agents are typically designed with the premise that they should respond happily to student successes and with a shared disappointment upon failures [9].

Considering the optimal design of adaptive tutoring systems and the incorporation of hypermedia and pedagogical agents to engage students in active learning, the current study seeks to analyze the promotion of Dweck's growth mindset theory within ASSISTments, an adaptive mathematics tutor. The following research questions were derived from themes relevant to Dweck's [6] work, in combination with adaptive tutoring structures unique to ASSISTments:

1. Does the addition of motivational messaging within the tutoring system affect the likelihood of student persistence or attrition?
2. Does the presence of motivational messaging within the tutoring system affect mastery speed as defined by how many items, on average, it takes for students to complete the problem set?
3. Can specific elements within message delivery be pinpointed as significantly powerful? That is, can researchers isolate an element (e.g., the presence of a pedagogical agent, the audio component, static images, or a combination of these elements) that is responsible for the majority of variance in persistence and learning efficiency?

It is hypothesized that students randomly assigned to a messaging condition will be more likely to show continued, persistent effort than those in the control condition. Similarly, regardless of the delivery medium, researchers expect students who receive mindset messages to show improved mastery speed, with fewer items, on average, required to complete a problem set. In the assessment of message delivery, it is hypothesized that motivational messages delivered using an animated version of Jane, a learning companion that originates from partnering tutor Wayang Outpost, will have a stronger effect on student persistence and learning efficiency than alternative message mediums.

2. METHODS

To determine appropriate math content for this study, the tutor's database was queried to compile a historical record of usage data for a variety of problem sets that fit within Common Core State Standards across various grade levels. All observed problem sets were of a style unique to the ASSISTments tutor, requiring students to answer three consecutive questions correctly in the same day in order to complete the assignment. If the student were to reach a preset 'daily limit' (i.e., ten problems) while attempting to solve three consecutive questions, they are prompted to consult with their teacher and try again tomorrow.

Five problem sets were chosen based on high usage, with math content spanning grades four through seven. The skill topics assessed by these problem sets included finding missing values using percent on a circle graph, equivalent fractions, multiplying decimals, rounding, and order of operations. The goal in designing multiple problem sets was three-fold: to increase data collection, to determine any significant effect for student skill level, and to determine if content was linked to student motivation, perhaps due to difficulty level. Six conditions were then established for each problem set, as defined in Table 1. These conditions were designed following the principles set forth by Mayer [4], to test matched content messages across a variety of processing channels.

Table 1. Motivational messaging conditions.

Control	ASSISTments as usual; no messages added
Animation	Jane, a female pedagogical agent, delivers messages with motion and sound
Static Image with Text	The agent is presented as a static image, with a speech bubble to deliver motivational text messages
Static Image with Audio	The agent is presented as a static image, supplemented by audio files to deliver motivational messages
Word Art	A speech cloud shows motivational text messages, with no agent involvement
Audio	The agent's voice delivers motivational messages with no graphical changes to tutor content

The student experience for each problem set was formatted in the same manner. An introductory 'question' explained the format of the problem set and alerted the student to turn on their computer volume and to use headphones if necessary. The second 'question' tested whether or not the student was able to see and hear the pedagogical agent Jane as she introduced herself as a problem-solving partner. This question was included to test the compatibility of the HTML files that supported the pedagogical agent's animation and sound conditions, thus serving as confirmation of fair random assignment. Researchers then relied on a randomization feature unique to ASSISTments that randomly assigned students to one of the six conditions depicted in Table 1. Math content was isomorphic across conditions, and was thus considered comparable in difficulty. A test drive of the student experience for each problem set can be found at [12].

Motivational message content, as depicted in Table 2, was matched across conditions to reduce confounding. These messages were validated in and derived from [1]. Each problem set was designed to randomly select questions from a pool of approximately 100 problems, containing two types of motivational message delivery: *general attributions*, in which the motivational message was presented with the primary question, and *incorrect attributions*, in which the motivational message was presented alongside content feedback if the student responded incorrectly or employed a tutoring strategy. Following this design structure, students saw general attributions on approximately half of the questions, with the remaining half displaying incorrect attributions only to students who answered a problem correctly. Therefore, each student's experience of motivational messaging may have differed slightly, even within each condition. This design was established to reduce persistent message delivery and to avoid inundating students with messages on each question, with the goal of optimizing the effects of motivational messages while retaining a primary focus on math content. All visual motivational messages appeared within the tutor and remained until the student completed the problem; audio messages were played once upon loading the problem or tutoring strategy.

Table 2. Motivational message item content.

General Attributions
1. Did you know that when we learn something new our brain actually changes? It forms new connections inside that help us solve problems in the future. Pretty amazing, huh?
2. Did you know that when we practice to learn new math skills our brain grows and gets stronger? That is so cool!
3. Hey, I found out that people have myths about math… like that only some people are "good" at math. The truth is we can all be successful in math if we give it a try.
4. I think the most important thing is to have an open mind and believe that one can actually do math!
5. I think that more important than getting the problem right is putting in the effort and keeping in mind the fact that we can all be good at math if we try.
Incorrect Attributions
1. Making a mistake is not a bad thing. It's what learning is all about!
2. When we realize we don't know why that was not the right answer, it helps us understand better what we need to practice.
3. We may need to practice a lot, but our brains will develop with what we learn.

At the end of each problem set, students were asked to partake in a series of four survey questions developed based on previously validated content from [11], to assess student mindset, goal orientation, and perceptions of enjoyment and system helpfulness. All students received these questions regardless of condition. All survey content can be accessed at [12].

3. PROCEDURE

Teachers in the state of Massachusetts who frequently use ASSISTments with their students were approached with a brief presentation explaining the study and providing examples of the conditions, motivational messages, and math content. Teachers assigned one or more of the problem sets to their students in accordance with the teachers' usual use of the tutoring system (i.e., as either classwork or homework). Material was assigned as current content and/or review, for a total of 765 student assignments. Log data was compiled for each student's performance. Prior to analysis of persistence and mastery speed, students were removed if they had noted experiencing technical difficulties or if they failed to log enough progress to enter one of the six conditions. Additional students were removed prior to survey analysis due to incompletion. Students remaining after each step are examined across problem sets in Table 3.

Table 3. Explanation of Students Remaining After Removals.

Problem Set	*A*[1]	*MA**	*SA***
Percent on a Circle Graph	87	69	62
Equivalent Fractions	255	208	205
Multiplying Decimals	62	48	47
Rounding	253	208	205
Order of Operations	108	88	86
REMAINING	**765**	**621**	**605**

A[1] = Assigned. MA = Math Analysis. SA = Survey Analysis.
*Students were removed prior to math analysis due to technical difficulties or failure to initiate a condition.
**Additional students were removed prior to survey analysis due to incompletion.

An ex post facto judgment of student gender was determined for 570 students within the sample remaining for math content analysis. Due to incompletion rates within this subset of students, gender was determined for 554 students within the sample remaining for survey content analysis.

4. RESULTS

Analyses of student persistence and mastery speed were performed at the condition level for each problem set, as well as for an aggregate of the five sets to serve as a composite analysis of the conditions across math content. To determine if an effect existed within a particular processing channel, similar conditions were compiled based on delivery elements. For example, all conditions utilizing audio were compiled to assess the effect of audio (i.e., audio, animation, static image with audio). Similar analyses were performed to determine the effect of textual messages and the effect of the pedagogical agent's presence. Researchers also compared a compilation of all conditions containing motivational messages to the control condition in order to determine the effectiveness of motivational messages in general. Initial findings suggested that in general, the sample was too advanced for the math content as students were found to be at ceiling across many of the problem sets. Thus, secondary analyses examined gender differences and assessed the aforementioned variables for a subset of students operationally defined as "strugglers," or those requiring more than three questions to complete their assignment.

When considering student persistence, as defined by continuing until reaching completion, ANOVA results suggested null results ($p > .05$) across all problem sets except for multiplying decimals $F(5, 42) = 2.57$, $p < .05$, $\eta^2 = 0.23$. No significant results were observed when the problem sets were compiled or when specific delivery elements were isolated, and there was no significant difference between messaging conditions and the control. For the full sample, gender was found to differ significantly on persistence, $F(1, 568) = 3.84$, $p = 0.051$, $\eta^2 = 0.01$, with girls showing significantly more persistence ($M = 0.99$, $SD = 0.12$) across conditions than boys ($M = 0.96$, $SD = 0.20$). While girls were found to be approaching completion in all conditions ($p < .05$), boys showed lower completion overall, with the lowest performance apparent in the control condition.

When considering mastery speed, as defined by the number of questions required for problem set completion, ANOVA results suggested null results ($p > .05$) across all problem sets analyzed

individually. Further, no significant results were observed when problem sets were compiled or when specific delivery elements were isolated, and there was no significant difference between messaging conditions and control. Although there was no significant difference in mastery speed across genders, trends suggested that girls had faster mastery speed in general, requiring consistently fewer questions to complete problem sets regardless of condition (M = 4.25, SD = 2.65) than boys (M = 4.43, SD = 2.86). Means and standard deviations for the full sample are presented in Table 4.

ANOVA comparisons of the survey measures of mindset, enjoyment, and system helpfulness similarly conveyed null results within the full sample. The "mindset" variable was established from an average of two binary survey questions, with a composite score scaled from 0-2 representing the spectrum from fixed mindset (0) to growth mindset (2). The "enjoyment" variable was based on one question with Likert scale scores from 0-3, representing how much the student enjoyed their assignment. The "helpfulness" variable is represented in the same manner, based on the student's perception of how helpful the tutoring system was in completing their assignment. Null results were found for all three measures across problem sets when analyzed individually, and no significant differences were observed between conditions when problem sets were compiled or when specific delivery elements were isolated. Further, there was no significant difference between all messaging conditions and the control group. Gender was found to have a significant effect on enjoyment, regardless of condition $F(1, 552) = 19.50$, $p < .001$, $\eta^2 = 0.03$, with girls measuring more enjoyment on average (M = 1.84, SD = 0.81) than boys (M = 1.52, SD = 0.90). As shown by Table 4, the Control was found to be the most enjoyable condition, while WordArt was enjoyed significantly less ($p < .10$). Gender was also approaching significance on the mindset measure, $F(1, 552) = 3.31$, $p = 0.069$, $\eta^2 = 0.01$, with boys exhibiting a lower mindset in general (M = 0.93, SD = 0.78) than girls (M = 1.05, SD = 0.77). Gender was not found to have a significant effect on student's perception of tutor helpfulness.

In an attempt to answer our third research question, elements within message delivery were collapsed based on similarity to better understand if a certain processing channel (i.e., audio) was providing the main effect for messaging results. As noted briefly in results for persistence, mastery speed, and survey measures, researchers were not able to isolate any significant differences among delivery elements ($p > .05$).

While few significant findings were observed in the full sample, it became clear that many students were at ceiling in the math content and therefore showing high persistence (completion) in minimum mastery speed (three consecutive correct questions). When we reassessed the sample for students operationally defined as 'struggling,' or those who required more than three questions to complete their assignments, our analysis became a bit more informative. Among 253 student assignments, no significant differences were found among conditions in persistence or mastery speed ($p > .05$). However, findings suggested that it took struggling students less questions on average to reach mastery when in the audio condition (M = 5.59, SD = 2.00) compared to all other conditions, as shown in Table 5.

When considering gender, struggling boys exhibited lower mastery in conditions including audio ($p < .05$) yet were found to persevere more when an image of Jane was present, while girls persevered less with the female presence ($p < .05$). Survey results for struggling students suggested that boys exhibited the lowest mindset measures after experiencing the control condition ($p < .05$), and trends suggested that regardless of condition, girls exhibited the growth mindset more consistently (M = 1.00, SD = 0.79) than boys (M = 0.91, SD = 0.75). As with the primary analysis, trends suggested that boys exhibited the growth mindset after experiencing the animation condition ($p < .10$). It was also found that regardless of condition, girls enjoyed their assignments (M = 1.72, SD = 0.87) significantly more than boys (M = 1.42, SD = 0.92), $p < .05$, and that girls consistently found the tutoring system more helpful in completing their assignment (M = 2.10, SD = 0.83) than did boys (M = 1.92, SD = 0.90).

Table 4. Means and Standard Deviations for Persistence, Mastery Speed, and Survey Measures Across Control and Messaging Conditions for All Students.

	Control (104^a, 99^b)		All Messaging (517^a, 506^b)		Animation (106^a, 103^b)		Static Image with Text (116^a, 113^b)		Static Image with Audio (117^a, 115^b)		Word Art ($90^{a,b}$)		Audio (88^a, 85^b)	
Analysis	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Persistence	0.95	0.21	0.98	0.14	0.97	0.17	0.97	0.16	0.98	0.13	**1.00**	**0.00**	0.97	0.18
Mastery Speed	**4.74**	**3.35**	4.32	2.67	4.24	2.69	4.62	2.83	4.32	2.42	4.28	3.33	4.09	1.91
Mindset	**1.06**	**0.81**	0.96	0.78	1.01	0.80	0.96	0.77	1.02	0.77	1.00	0.79	0.78	0.75
Enjoyment	**1.83**	**0.80**	1.67	0.89	1.74	0.87	1.66	0.90	1.77	0.82	1.49	0.91	1.67	0.96
Helpfulness	1.99	0.85	1.94	0.86	1.86	0.89	**2.01**	**0.89**	**2.01**	**0.77**	1.82	0.79	1.95	0.95

[a]Sample size for Persistence and Mastery Speed.
[b]Sample size for Mindset, Enjoyment, and Helpfulness.
Note. "Mindset" is measured by two questions (0 = Fixed Mindset, 1 = Growth Mindset) and scores are compiled. "Enjoyment" is measured by one question (Likert Scale, 0-3). "Helpfulness" is measured by one question (Likert Scale, 0-3).

Table 5. Means and Standard Deviations for Persistence, Mastery Speed, and Survey Measures Across Control and Messaging Conditions for Struggling Students.

	Control (46^a, 45^b)		All Messaging (207^a, 204^b)		Animation (42^a, 41^b)		Static Image with Text (49^a, 47^b)		Static Image with Audio ($49^{a,b}$)		Word Art ($28^{a,b}$)		Audio ($39^{a,b}$)	
Analysis	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Persistence	0.98	0.15	0.99	0.12	0.98	0.15	0.96	0.20	**1.00**	**0.00**	**1.00**	**0.00**	**1.00**	**0.00**
Mastery Speed	7.07	3.95	6.34	3.32	6.17	3.48	6.84	3.24	6.14	2.88	**7.11**	**4.95**	5.59	2.00
Mindset	0.93	0.75	0.95	0.78	1.00	0.81	0.89	0.73	**1.04**	**0.82**	0.82	0.86	0.92	0.70
Enjoyment	1.60	0.86	1.58	0.94	**1.76**	**0.92**	1.45	1.00	1.71	0.79	1.43	1.07	1.51	0.97
Helpfulness	1.98	0.92	2.01	0.87	1.98	0.94	1.98	0.82	2.04	0.87	2.00	0.82	**2.05**	**0.94**

[a]Sample size for Persistence and Mastery Speed.
[b]Sample size for Mindset, Enjoyment, and Helpfulness.
Note. "Mindset" is measured by two questions (0 = Fixed Mindset, 1 = Growth Mindset) and scores are compiled. "Enjoyment" is measured by one question (Likert Scale, 0-3). "Helpfulness" is measured by one question (Likert Scale, 0-3).

Approximately 60% of students in the full sample exhibited the growth mindset in their survey responses, regardless of condition. Noting Table 4, students in the control condition actually reported the highest levels of growth mindset (M = 1.06, SD = 0.81), with those in the audio condition reporting the lowest levels (M = 0.78, SD = 0.75). Among struggling students, the highest levels of growth mindset were reported by students in the static image with audio condition (M = 1.04, SD = 0.82), while those in the word art condition reported the lowest levels (M = 0.82, SD = 0.86). Responses to measures of enjoyment and helpfulness followed normal distributions, with approximately 60% finding the assignments at least "somewhat" enjoyable, and approximately 78% finding the tutoring system at least "somewhat" helpful.

5. DISCUSSION

Within the current study, the addition of motivational messaging to the ASSISTments tutor did not significantly affect the likelihood of student persistence or mastery speed. Further, there was little evidence that the motivational messages had the intended effect on mindset within the full sample. Trends suggested that those in messaging conditions experienced a slight increase in persistence and a decrease in mastery speed in comparison to those in the control condition. However, students in the messaging conditions also exhibited consistently lower levels for measures of mindset, enjoyment of the assignment, and perception of system helpfulness. A larger student population would be required to discern a truly significant effect within these trends.

Interestingly, struggling students appeared to benefit from the presence of messages, showing an increase in persistence, a decrease in mastery speed, and slightly increased measures of the growth mindset. It can be argued that struggling students, or those facing a challenge, are most in need of motivational interventions, and that they are more likely to respond to messaging, regardless of condition. Motivational messages produced distinctly higher adoption of the growth mindset in struggling students who experienced the static image with audio condition. Thus when designing motivational content for struggling students, current findings promote the addition of audio as an alternative processing channel to assist students. Researchers were not able to pinpoint an optimal processing channel for the delivery of growth mindset messages when targeting the general population.

One participating teacher requested that her students use a feature within the tutoring system to comment on their experience while completing their assignment. Feedback was predominantly negative, with students citing the messages as distracting or confusing. One student specifically questioned why the animated learning companion simply repeated messages rather than helping to solve the problems. This suggests that students are familiar with systems that utilize pedagogical agents, and that they have developed expectations for characters that are associated with learning. This echoes the argument set forth by Kapoor, et al. [9] regarding the necessity for tutors to provide appropriate cognitive and affective responses, and aids in the design of tutoring systems hoping to incorporate learning companions.

This study had a variety of limitations. The ASSISTments math content chosen due to popular usage lead to a high percentage of ceiling effects within the sample. Teachers assigned multiple problem sets to their students, often as review. Thus, many students easily mastered the content intended for lower grades and thereby skewed rates of persistence and mastery speed. Further, the null effects found in the full sample raise important questions regarding the generalizability of mindset interventions outside of struggling student populations. Within the context of an adaptive mathematics tutor, students who appear to be at ceiling in math content may not require motivational messaging, and it may become detrimental to the learning process.

We also note that approximately 18.8% of students reported having technical difficulties and were removed prior to analysis. The incompatibility of simple HTML files serves as a reminder that many classrooms struggle to maintain up-to-date technological resources. Students are often required to share computers or iPads that come equipped with outdated software and generally slow internet connections. Future research should incorporate allowance for these issues within the experimental design, as incompatibilities may lead to selection bias.

It is also difficult to justify whether or not students consistently attended to the motivational messages. As students were simply presented the messages and were not asked to respond in any manner, the levels of message internalization may be broad. We also note that the duration of the intervention may have been too short to observe reliable differences among messaging conditions. In much of her work, Dweck has provided longer interventions upfront, coupled with 'reminders' such as the messages used in the current study [7]. Further, her studies often run longitudinally across the course of a school year or more. Still, regardless of condition, the majority of students in our sample exhibited the growth mindset. Future research should include a pretest mindset survey to determine if these results can be credited solely to the motivational messages provided throughout the learning experience.

Finally, it should be noted that researchers relied on the tutoring system to perform random assignment. While prior research has suggested that this practice is sound, assignment for this study appears to have favored the static image with audio condition. Future research using ASSISTments should take this bias into consideration.

Future iterations of this study should focus on struggling students, or those undertaking challenging academic tasks. Future research should also seek to assess these conditions in an even more adaptive environment. It seems as though students were not reaping the benefits of the "persona effect" found in prior research [1], due to a lack of bonding with the agent. A truly adaptive agent, one consistently present and building rapport, may be more effective in message delivery. Rather than repeating the same select set of general and incorrect attributions, struggling students may require motivational messages linked with the tutor content and their progress. Perhaps just as a pedagogical agent, these messages must be fine-tuned to a student's cognitive and affective states. Alternative message delivery methods, including video feedback with human tutors used as hints, scaffolding, and misconception messages, should also be considered in future research.

6. ACKNOWLEDGEMENTS

We acknowledge funding from NSF (#1316736, 1252297, 1109483, 1031398, 0742503), ONR's "STEM Grand Challenges," and IES (#R305A120125, R305C100024). Thanks to S. & L.O.

7. REFERENCES

[1] Arroyo, I., Burleson, W., Tai, M., Muldner, K., Woolf, B.P. 2013. Gender differences in the use and benefit of advanced learning technologies for mathematics. Journal of Educational Psychology. 105, 4, 957-969.

[2] Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A. & Koedinger, K. 2008. Why students engage in "Gaming the System" behavior in interactive learning environments. *Journal of Interactive Learning Research*. 19, 2, 185-224.

[3] Bernacki, M. L., Nokes-Malach, T. J., & Aleven, V. 2013. Fine-grained assessment of motivation over long periods of learning with an intelligent tutoring system: Methodology, advantages, and preliminary results. In *International handbook of metacognition and learning technologies*. Springer New York. 629-644.

[4] Clark, R.C. & Mayer, R. E. 2003. e-Learning and the science of instruction: proven guidelines for consumers and designers of multimedia learning. San Francisco, CA: Pfeiffer.

[5] Dweck, C.S. 2002. Messages that motivate: how praise molds students' beliefs, motivation, and performance (in surprising ways). Improving Academic Achievement: Impact of Psychological Factors in Education. Ed. Joshua Aronson. New York.

[6] Dweck, C.S. 2006. Mindset: The new psychology of success. Random House.

[7] Dweck, C.S. 2013. Mindsets: Helping Students Fulfill Their Potential. Smith College Lecture Series, North Hampton, MA. September 19.

[8] Graesser, A., Chipman, P., King, B., McDaniel, B., & D'Mello, S. 2007. Emotions and learning with autotutor. *Frontiers in Artificial Intelligence Applications*, *158*, 569.

[9] Kapoor, A., Burleson, W., & Picard, R. W. 2007. Automatic prediction of frustration. International Journal of Human-Computer Studies. 65, 724-736.

[10] Mayer, R.E. 2014. Incorporating motivation into multimedia learning. Learning and Instruction. Volume 29, 171-173.

[11] Mueller, C. & Dweck, C. 1998. Praise for Intelligence Can Undermine Children's Motivation and Performance. Journal of Personality and Social Psychology, Vol. 75, No. 1, 33-52.

[12] Ostrow, K.S. 2013. Motivational Message Study. Accessed 12/12/2013. Student Experience, RCT & All Data: https://sites.google.com/site/korinnostrow/research

[13] Reeve, J. 2009. Understanding motivation and emotion. (5th ed.). Hoboken, NJ: Wiley.

Toward Adaptive Unsupervised Dialogue Act Classification in Tutoring by Gender and Self-Efficacy

Aysu Ezen-Can
Department of Computer Science
North Carolina State University
aezen@ncsu.edu

Kristy Elizabeth Boyer
Department of Computer Science
North Carolina State University
keboyer@ncsu.edu

ABSTRACT

For tutorial dialogue systems, classifying the dialogue act (such as questions, requests for feedback, and statements) of student natural language utterances is a central challenge. Recently, unsupervised machine learning approaches are showing great promise; however, these models still have much room for improvement in terms of accuracy. To address this challenge, this paper presents a new unsupervised dialogue act modeling approach that leverages non-cognitive factors of gender and self-efficacy to better model students' utterances during tutorial dialogue. The experimental findings show that for females, leveraging learner characteristics within dialogue act classification significantly improves performance of the models, producing better accuracy. This line of investigation will inform the design of next-generation tutorial dialogue systems, which leverage machine-learned models to adapt to their users with the help of non-cognitive factors.

Keywords

Tutorial dialogue, learner characteristics, dialogue act classification, unsupervised machine learning, adaptive learning.

1. INTRODUCTION

Tutorial dialogue is a highly effective form of instruction, and much of its benefit is thought to be gained from the rich natural language dialogue exchanged between tutor and student [7, 17, 36]. In order to model tutorial dialogue for the purposes of building tutorial systems or for studying human tutoring, *dialogue acts*, which capture both cognitive and non-cognitive aspects of dialogue utterances, provide a valuable level of representation. Dialogue acts represent the underlying intention of utterances (for example, to ask a question, agree or disagree, or to give a command) [3, 32]. Within the computational linguistics and dialogue systems literature, automatically classifying dialogue acts has been a focus of research for several decades [6, 14, 35]. For tutorial dialogue systems, dialogue act classification is crucial to understanding students' utterances and developing tutorial strategies [8, 24].

Today's tutorial dialogue systems utilize a variety of dialogue act classification strategies, some rule-based and some statistical [13]. Historically when machine learning has been used to devise tutorial dialogue classifiers, these have been *supervised* classifiers, which require training on a manually labeled corpus. The same is true within the broader dialogue systems research community: dialogue act classifiers have historically either been handcrafted and rule-based, or learned with supervised machine learning techniques [11, 14, 22, 29]. However, supervised techniques face substantial limitations in that they are labor-intensive due to the manual annotation and handcrafted dialogue act taxonomies that are usually domain-specific. To overcome these challenges, unsupervised dialogue act modeling techniques including hidden Markov models [20, 21, 30], Dirichlet Process clustering [12, 23], *k*-means clustering [31], and query-likelihood clustering [15] have been investigated in recent years.

Despite this growing focus on developing unsupervised dialogue act classifiers, these models still underperform compared to supervised approaches in their accuracy for classifying according to manual tags. However, while unsupervised models to date have considered such things as lexical features (the words found in the utterance) and syntactic features (the structure of the sentence), they have not considered non-cognitive factors, such as gender and self-efficacy, which are believed to influence the structure of tutorial dialogue [10]. Cognitive factors such as skill mastery has been widely studied in learning environments. However, there is a smaller body of work on adaptive learning environments using non-cognitive factors. A variety of learner characteristics, including non-cognitive factors, play an influential role in learning, not only in tutoring but in classroom settings [1], and in web-based courses [19]. Prior work on learner characteristics has focused on building adaptive systems based on different user groups [16], tutorial feedback selection [9] and identifying students that need remedial support [27]. Identifying clusters of student characteristics is also an active area of research [4, 25–27].

This paper investigates whether the performance of an unsupervised dialogue act classifier can be improved by taking these factors into account. Because non-cognitive factors are shown to affect language, we believe that training dialogue act classifiers tailored to specific learner characteristics can help tutorial dialogue systems to understand students better. We utilize two learner characteristics: gender, as self-reported by students on a survey and domain-specific self-efficacy, as measured by a validated instrument for determining a student's confidence in her own abilities. Specifically, we train unsupervised dialogue act models that are tailored to students of specific gender and self-efficacy level, and we compare those models to corresponding ones trained without restricting by that learner characteristic. This unsupervised training is conducted entirely without the use of manual tags. We then test all of the models on held-out test sets within leave-one-student-out cross validation, and compare the resulting classification accuracy according to their previously applied manual tags. The results show that for female students, utilizing learner characteristics statistically significantly improves dialogue act classification models. For self-efficacy groups, improvement is observed but not at a statistically reliable level. This paper constitutes the first research toward incorporating non-cognitive factors into unsupervised dialogue act classifiers for

Published in CEUR-WS:
NCFPAL workshop (Ritter and Fancsali)
In EDM 2014 Extended Proceedings (Gutierrez-Santos and Santos)

tutorial dialogue with the overarching goal of providing personalized learning for students. We first administered a survey to collect these characteristics via self-report, and then learned a dialogue act classifier tailored to those characteristics. These results can inform the way that next-generation tutorial dialogue systems conduct their real-time dialogue act classification and language adaptation.

2. RELATED WORK

Dialogue act modeling is an important level of representation within dialogue systems. Following theories proposed several decades ago within philosophy and linguistics [3, 32], dialogue act classification aims to capture the intention of an utterance; for example, in tutoring some dialogue acts involve asking questions or giving or requesting feedback. While a long-standing line of investigation has focused on handcrafted or supervised machine learning techniques for dialogue act classification [11, 14, 22, 29], only recently is a body of work emerging on unsupervised approaches to this problem. Most of this work has been done outside of educational domains, with a proposed hidden Markov model in the domains of Twitter posts [30] and emails [21], Dirichlet Process Mixture Models for a train fare dialogue domain [12] and for navigating buildings [23], and a Chinese Restaurant Process approach for spoken Japanese [20].

Another important difference between the current work and prior research is in the features used, namely the non-cognitive characteristics of gender and self-efficacy. Prior work has used a variety of features for performing supervised dialogue act classification, including prosodic and acoustic features which involve the profile of the sound signal itself [35], lexical features such as words and sequences of words [34], syntactic features including part-of-speech tags [6, 24], dialogue structure features such as taking the initiative and the previous dialogue act [33] as well as task/subtask features in tutorial dialogue [8, 18]. Within unsupervised dialogue act classification a subset of these features have also been used such as words [12], state transition probabilities in Markov models [23], topic words [30], function words [15], a smaller subset of words containing beginning portions of utterances [31], part-of-speech tags and dependency trees [21]. While a variety of experiments have demonstrated the utility of these features in several domains, no prior work has reported on an attempt to include the factors considered here, in order to improve the performance of an unsupervised dialogue act classifier. To investigate this, we build dialogue act classifiers that learn from utterances of specific learner groups and predict dialogue acts of students according to their learner characteristics.

3. CORPUS

The corpus used in this study consists of student-tutor interactions in an introductory computer science programming task [18]. Throughout the data collection, freshman engineering students and tutors communicated through a textual dialogue-based learning environment while working on Java programming. The ethnicity of students participated in this study is distributed as follows: 26 white, 9 Asian, 3 Latino, 2 African American, 1 Middle Eastern and 1 Asian American. An excerpt from the corpus is shown in Table 1.

Students were given a pre-survey that included survey items on computer science self-efficacy, such as 'I am sure I can learn programming'. This self-efficacy scale was adapted directly from the Domain-specific Self-Efficacy Scale [5], with five items measured on a Likert scale from 1-5 (1 being lowest self-efficacy, 5 being highest). Students also completed a demographic questionnaire from which gender was obtained. For self-efficacy, students were divided into classes based on the median score across all students on that scale. Along with gender, this produces two partitions of the 42 students: females (12) and males (30), low (24) and high self-efficacy students (18).

Table 1: Excerpt of dialogue with a *male* student in the *low self-efficacy* group

Role	Utterance	Dialogue Act
Tutor	You'll need to end every Java statement with a semi colon	*S*
Student	Got it!	*ACK*
Tutor	This is to let Java know where each statement ends	*S*
Tutor	Ah no prompt!	*S*
Tutor	Why do you think that is?	*Q*
Student	I wish I knew...	*A*
Student	I don't think I spelled anything wrong	*S*
Tutor	Ah it's actually pretty easy	*S*
Tutor	The order of the lines matters	*S*

The corpus containing 1640 student utterances was manually annotated with dialogue act tags in previous work [18] (Table 2). These dialogue act tags are not available during model training, but we use them for evaluation purposes to calculate accuracy on a held-out testing set.

Table 2: Student dialogue acts and distributions

Student Dialogue Act	Example	Distribution
A (answer)	*yeah I'm ready!*	39.95%
ACK (acknowledgement)	*Alright*	21.31%
S (statement)	*i am taking basic fortran right now never seen literal before*	21.20%
Q (question)	*what does that mean?*	15.15%
RF (request feedback)	*better?*	0.98%
C (clarification)	**html messing*	0.79%
O (other)	*haha*	0.61%

4. DIALOGUE ACT MODELING BASED ON LEARNER CHARACTERISTICS

We hypothesize that dialogue act models built using unsupervised machine learning will perform substantially better when customized to specific learner groups. Specifically, we investigate whether by training a model only on students of a particular learner characteristic, that model would perform significantly better at predicting the dialogue acts of unseen students with the same learner characteristic compared to a model that was trained on students of all learner characteristics.

We note that because the same corpus is being partitioned in two different ways, the same student will occur in one of the gender groups and in one of the self-efficacy groups. This choice to partition in 2-way splits rather than *2n*-way splits where *n* is the number of learner characteristics is because of issues that arise with sparsity. This interdependence between partitions is a limitation to note; however, as discussed in Section 5, this

interdependence can be taken into account for making decisions within a tutorial dialogue system by employing a suite of classifiers within a voting scheme.

4.1 Experimental Design

For gender and self-efficacy, we will test whether an unsupervised dialogue act classifier trained only on students with that characteristic outperforms a classifier that is not specialized by this characteristic. In order to gather accuracy data across these characteristics, we conduct leave-one-student-out training and testing folds. The testing set for each of the *n* folds (where *n* varies depending on which learner group is being considered) consists of all of a single student's dialogue utterances and the model is trained on the remaining *n-1* students. The average number of utterances per student in the corpus is 36.8 (σ=12.07; min=16; max=64). These are therefore the average, minimum, and maximum number of utterances across the leave-one-student-out test sets.

We compute the average test set performance of the model across all folds for each non-cognitive characteristic partition. The performance metric utilized in this study is *accuracy* compared to the manually labeled dialogue acts described in the previous section, where accuracy is computed as the number of utterances in the test set that were classified according to their manual label, divided by the number of utterances total in the test set. As described in 4.2, the process of labeling via unsupervised classification involves taking the majority vote within each cluster.

For constructing the folds, we take an approach to balance the sample size available to model training. This balancing approach is needed to ensure that each model is trained on a similar size of data. Consider, for example, the partition of gender. Without a balanced sampling approach the leave-one-student-out testing folds for the un-specialized classifier for female students would include n_{female}=12 test folds but the available data for each training fold would be n_{total}-1 = 41. In contrast, the specialized classifier trained only on female students would still include n_{female}=12 test points but the available data for each training fold would be n_{female}-1 = 11. Therefore, each un-specialized classifier was trained on a randomly selected subset of the corpus. In the case of females, each of the 12 testing folds will utilize a model trained on 11 data points. The specialized classifier will use 11 female data points, and the un-specialized classifier will use 11 randomly selected data points. In this way, we investigate how well a model predicts dialogue acts of a student with and without utilizing learner characteristic information.

4.2 Unsupervised Dialogue Act Models

Our unsupervised dialogue act classification approach leverages the *k*-medoids clustering technique [28]. This approach groups similar utterances together, and is similar to the more familiar *k*-means algorithm except that in *k*-medoids, the centroid of each cluster must be an actual data point within the corpus rather than a potentially artificial data point computed as the mean of distances. Our experiments with *k*-medoids have demonstrated that it outperforms a variety of other unsupervised machine learning approaches for the task of dialogue act classification in tutorial dialogue, although the results of such experiments are beyond the scope of this paper since our goal is to investigate the *differential benefit* of adding learner characteristic features to the model, not to compare different unsupervised approaches.

The *k*-medoids algorithm requires seeding clusters at the beginning of each training fold and then proceeds by distributing data points to clusters according to their closest centroids until convergence upon the model. In the standard *k*-medoids algorithm, the seeds are randomly selected. However, we employ a greedy seed selection approach intended to mitigate the effects of the unbalanced distribution of dialogue acts in the corpus [2]. Within this greedy seed selection, an initial seed is randomly selected and then each of the subsequent seeds are selected by choosing the point that maximizes its distance from the already-selected seeds. The goal in using this approach is to select the seeds from diverse utterances so the algorithm produces better clusters, and our initial experiments indicated that it substantially improves the model.

In addition to its seeding approach, the *k*-medoids approach requires the number of clusters *k* to be set prior to model training. To discover the number of clusters, we experimented with *X*-Means and Expectation Maximization clustering, both of which attempt to identify the optimal number of clusters. Both of these algorithms converged at four clusters as the optimal choice, so we proceed with *k*=4. However, perhaps in part due to the benefit of the greedy seed selection made possible by *k*-medoids, these models performed with substantially worse overall accuracy than *k*-medoids.

The utterances were represented as vectors with each column matching a token (punctuation and words) in the corpus and each row matching an utterance. There were a total of 877 distinct tokens.

With these parameters in place, first the clusters were formed using each training set, and then for each utterance of the student held out within the leave-one-student-out fold, we computed the closest cluster to that utterance as indicated by average cosine distance to each point in the cluster. The closest cluster was selected as the cluster to which the test utterance belongs, and the majority vote of the cluster was assigned to the test utterance as its dialogue act label. For each leave-one-student-out testing fold, the accuracy was computed by comparing these cluster-assigned labels to the manual dialogue act tags.

4.3 Experimental Results

This section presents experimental results for unsupervised dialogue act classification based on learner characteristics. We compare each model built separately by gender and self-efficacy level to the models that are built using utterances from randomly selected students, *i.e.* not utilizing learner characteristic information. Each comparison in this section is conducted with a one-tailed *t*-test with a post-hoc Bonferroni correction. The threshold for statistical reliability after the correction has been taken as α=0.05.

Gender. As shown in Figure 1, the average leave-one-student-out cross-validation accuracy for the model built using female students' utterances (n_{female}=12) is higher than the model built on randomly selected students. In each test run, all of one female's utterances were left out to be used as the test set, and the dialogue act model was built on the remaining eleven female students' utterances. This process was repeated for each female student. Note that for each of the eleven students, all utterances from that student were considered. Average test set accuracy for the model with randomly selected students was 0.41 (σ=0.2), whereas the average test set accuracy for the dialogue act classification model that was built utilizing female students' utterances only was 0.56 (σ=0.19). After a Bonferroni correction this difference was statistically significant (p_{Bonf}<0.05).

For male students (n_{male}=30), the average accuracy is only slightly higher with the models tailored to males 0.43 (σ =0.13) than the models learned for randomly selected students 0.40 (σ =0.12), and this difference is not statistically significant (Figure 1). Looking more closely at the results, we find that for eight of the thirty males within the corpus, a tailored model outperformed the random model (with five of these seeing more than 10% increase in accuracy), while twenty-two of the cases saw no difference in accuracy between the random and tailored conditions. Two of the males saw a decrease in accuracy for the tailored condition.

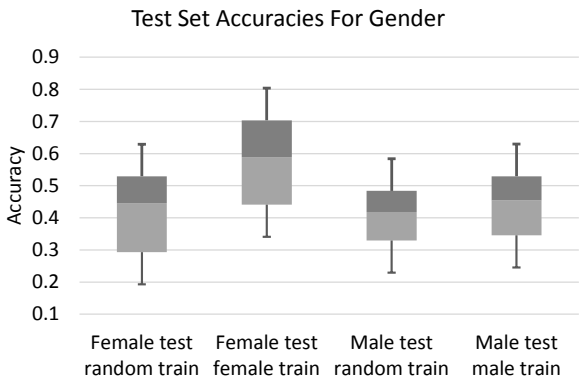


Figure 1: Leave-one-student-out test set accuracies for models by gender

Self-efficacy. Models built using the self-efficacy learner characteristic predict the unseen utterances' dialogue acts marginally more successfully than models that do not use this information, though these differences are not statistically reliable. For students with low self-efficacy (n_{lowEff}=24) the average test set accuracy for dialogue act models that selected students randomly is 0.38 (σ=0.16) and it increases to 0.43 (σ=0.17) with dialogue act models that learn only from low-self-efficacy students' utterances (Figure 2). In fifteen out of twenty-four cases the dialogue act models tailored to low self-efficacy groups outperform models that are trained on randomly selected students (eight of the cases with more than a 10% increase), while in seven of the cases the performance is decreased by utilizing the learner characteristic (five of them by more than a 5%) and in two of the cases the accuracy remains the same.

The improvement obtained by utilizing learner characteristics in dialogue act classification task is also marginal for high-self-efficacy students, where $n_{highEff}$=18. The average performance for the random model is 0.41 (σ=0.14) whereas the model achieves 0.47 (σ=0.11) accuracy when trained only on utterances of high-self-efficacy students. This improvement was statistically significant before Bonferroni correction but not afterward. In seven out of eighteen cases, models trained on utterances of high self-efficacy students improved test set accuracy (five of them above 15% improvement) and in two of the cases the learner characteristic decreases the performance (both of them below 5% decrease). Nine of the cases remained unaffected in their dialogue act classification accuracy.

The average accuracies over the leave-one-student-out cross-validation folds can be found in Table 3. Models tailored to learner groups uniformly outperform their counterpart, and the improvement is statistically significant for females.

Figure 2: Leave-one-student-out test set accuracies for models by self-efficacy

Table 3: Average test set accuracies for each learner characteristic (p<0.05 after Bonferroni correction)**

Learner characteristic group	Model restricted by learner characteristic	Model built on randomly selected students
Females	**0.56****	0.41
Males	**0.43**	0.40
Low self-efficacy	**0.43**	0.38
High self-efficacy	**0.47**	0.41

5. DISCUSSION

Dialogue act classification is a central task for tutorial dialogue systems. Without accurate dialogue act classification, systems cannot adapt and respond appropriately. Unsupervised machine learning approaches to dialogue act classification are a highly promising new area of study, and we have presented the first unsupervised dialogue act classifier tailored to learner characteristics. The experimental results demonstrated that dialogue act classifiers that leverage the non-cognitive factors of gender and self-efficacy outperform those that do not, and in the case of female students the improvement was statistically significant. This section presents some examples of the learned dialogue act clusters and discusses the implications of this work for tutorial dialogue systems.

First, we examine clusters from the gender-tailored unsupervised dialogue act classifier. Table 4 displays a selection of utterances that were clustered together during the unsupervised training of the model, and afterward the clusters were labeled for testing purposes using the manual tags that comprise the majority of each cluster. For those in Table 4 the clusters were labeled as Acknowledgments and Questions. By examining the structure of these clusters we gain some intuition as to the types of regularities that help the tailored models to perform significantly better. We see females in this study tended to use acknowledgment phrases such as, "oh I see" and "makes sense," while males tended to use the phrasing, "got it" more frequently. Within the cluster labeled as questions, we observe that females tended to request more feedback, an observation that also emerged in prior work within a different corpus in the same domain collected approximately six years earlier [10]. On the other hand, male students tended to ask more general questions.

In addition, we observe some example clusters from the models based on self-efficacy in Table 5. Students with high self-efficacy tend to use more confident utterances such as "absolutely" compared to "ok" used by low-self efficacy students. We note that questions in the low self-efficacy group often make an implicit

request for reassurance within their task-based questions, such as, "and that is it?". In contrast, students in the high self-efficacy group more often ask contentful questions.

Table 4: Selected utterances from clusters tailored to gender

	Females	Males
Acknowledgements	- oh I see - make sense - yup - aha! -hahaha its ok	- got it - ok i got it - alright i got it - gotcha alrigth - cool - sure thing
Questions	-is this right? -does that work? -should I run it? -was i supposed to put that before something? -so for line number could i have typed system out println monopoly instead of println x if i wanted to?	-so will testing always be related to running the program -so it is kinda like saying x number or something in algebra? -why does not it stop on the next line in this case

Table 5: Selected utterances from clusters tailored to self-efficacy

	Low Self-Efficacy	High Self-Efficacy
Acknowledgements	- ok - yes there were a lot of things i felt like i had to switch around - that makes sense now	-cool! -oh ok that works - yep got that - absolutely
Questions	-so what exactly am i supposed to be doing? - is there something specific i need to call my game - i finished reading should i click compile again? -and that is it?	-what is the best way to do that? - ok so tell me if this makes sense string declares the variable and then line number tells me what that variable is value is?

Limitations. The present work has several notable limitations. First, as mentioned previously, the partitions of the corpus are not independent; that is, the same student, and associated utterances, are present within one gender group and one self-efficacy group. Because these partitions are not independent, care must be taken when interpreting the findings. Furthermore, it is possible that the self-efficacy of students can change in the course of tutoring, which would not be handled by a classifier built using a one-time self-report. However, we believe that the current approach holds great promise for real-time tutorial dialogue classification. By building separate classifiers by learner characteristic, a suite of classifiers (each smaller and faster than one built on the entire corpus) can be run in parallel and can vote for the classification of a given students' utterance. However, as is the case with the work presented here, splitting the corpus results in a substantially reduced sample size on which to train, which partially explains the lack of statistically reliable results observed here. Our work has begun to explore the use of intrinsic metrics for accuracy (rather than relying on manual tags), which has the potential to dramatically increase the available data to any dialogue act classifier and mitigate issues of sparsity that arise when splitting by learner characteristics.

6. CONCLUSION AND FUTURE WORK

More accurately understanding student natural language within intelligent tutoring systems is a critical line of investigation for tutorial dialogue systems researchers. The field has only begun to explore unsupervised approaches and to investigate the range of features that are beneficial within this paradigm. We have presented a first attempt to leverage non-cognitive factors within such a dialogue act classification model, achieving statistically significant improvements in dialogue act modeling for female students, and increasing the models' performance by small margins for the self-efficacy groups.

Building upon these first steps, there are several promising future directions. First, while sample size prohibited exploring some other learner characteristics here, other characteristics are likely highly influential and should be investigated. These may include ethnicity, personality, and other non-cognitive factors. Additionally, while the current work focused on analyzing dialogue, another aspect of the tutorial interaction that presents challenges in understanding is the task model. Models that aim to understand students' problem-solving activities and infer their goals or plans may benefit substantially from leveraging learner characteristics. It is hoped that the research community can continue to build richer models of natural language understanding for students of all learner characteristics in order to improve the student experience and enhance learning by adaptation.

ACKNOWLEDGMENTS

The authors wish to thank the members of the Center for Educational Informatics at North Carolina State University for their helpful input. This work is supported in part by the National Science Foundation through Grant DRL-1007962 and the STARS Alliance, CNS-1042468. Any opinions, findings, conclusions, or recommendations expressed in this report are those of the participants, and do not necessarily represent the official views, opinions, or policy of the National Science Foundation.

REFERENCES

[1] Ames, C. and Archer, J. 1988. Achievement Goals in the Classroom: Students' Learning Strategies and Motivation Processes. *Journal of Educational Psychology*. 80, 3, 260–267.

[2] Arthur, D. and Vassilvitskii, S. 2007. k-means++: The Advantages of Careful Seeding. In *Proceedings Of The Eighteenth Annual ACM-SIAM Symposium On Discrete Algorithms*. 1027–1035.

[3] Austin, J.L. 1962. *How To Do Things With Words*. Oxford University Press.

[4] Azarnoush, B., Bekki, J.M. and Bernstein, B.L. 2013. Toward a Framework for Learner Segmentation. *JEDM*. 5, 2, 102–126.

[5] Bandura, A. 2006. Guide for Constructing Self-Efficacy Scales. *Self-efficacy Beliefs Of Adolescents*. 5, 307–337.

[6] Bangalore, S., Di Fabbrizio, G. and Stent, A. 2008. Learning the Structure of Task-Driven Human-Human Dialogs. *IEEE Transactions on Audio, Speech and Language Processing*. 16, 7, 1249–1259.

[7] Bloom, B.S. 1984. Sigma of Problem: The Methods Instruction One-to-One Tutoring. *Educational Researcher*. 4–16.

[8] Boyer, K.E., Ha, E.Y., Phillips, R., Wallis, M.D., Vouk, M.A. and Lester, J.C. 2010. Dialogue Act Modeling in a Complex Task-Oriented Domain. In *Proceedings of SIGDIAL*. 297–305.

[9] Boyer, K.E., Phillips, R., Wallis, M., Vouk, M. and Lester, J. 2008. Balancing Cognitive and Motivational Scaffolding in Tutorial Dialogue. In *Proceedings of ITS*, 239–249.

[10] Boyer, K.E., Vouk, M.A. and Lester, J.C. 2007. The Influence of Learner Characteristics on Task-Oriented Tutorial Dialogue. In *Proceedings of AIED,* 365–372.

[11] Buckley, M. and Wolska, M. 2008. A Classification of Dialogue Actions in Tutorial Dialogue. In *Proceedings of the 22nd International Conference on Computational Linguistics*. 1, 73–80.

[12] Crook, N., Granell, R. and Pulman, S. 2009. Unsupervised Classification of Dialogue Acts Using a Dirichlet Process Mixture Model. In *Proceedings of SIGDIAL*. 341–348.

[13] Dzikovska, M.O., Farrow, E. and Moore, J.D. 2013. Combining Semantic Interpretation and Statistical Classification for Improved Explanation Processing in a Tutorial Dialogue System. In *Proceedings of AIED*. 279–288.

[14] Eugenio, B. Di, Xie, Z. and Serafin, R. 2010. Dialogue Act Classification , Higher Order Dialogue Structure, and Instance-Based Learning. *Dialogue & Discourse*. 1, 2, 1–24.

[15] Ezen-Can, A. and Boyer, K.E. 2013. Unsupervised Classification of Student Dialogue Acts With Query-likelihood Clustering. In *Proceedings of EDM*, 20–27.

[16] Forbes-Riley, K. and Litman, D.J. 2009. A User Modeling-Based Performance Analysis Of A Wizarded Uncertainty-Adaptive Dialogue System Corpus. In *Proceedings of INTERSPEECH*, 2467–2470.

[17] Graesser, A.C., Wiemer-Hastings, K., Wiemer-Hastings, P. and Kreuz, R. 1999. AutoTutor: A Simulation Of A Human Tutor. *Cognitive Systems Research*. 1, 1, 35–51.

[18] Ha, E.Y., Grafsgaard, J.F., Mitchell, C.M., Boyer, K.E. and Lester, J.C. 2012. Combining Verbal and Nonverbal Features to Overcome the 'Information Gap' in Task-Oriented Dialogue. In *Proceedings of SIGDIAL*, 247–256.

[19] Hershkovitz, A. and Nachmias, R. 2011. Online Persistence In Higher Education Web-Supported Courses. *The Internet and Higher Education*. 14, 2, 98–106.

[20] Higashinaka, R., Kawamae, N., Sadamitsu, K., Minami, Y., Meguro, T., Dohsaka, K. and Inagaki, H. 2011. Unsupervised Clustering of Utterances Using Non-Parametric Bayesian Methods. In *Proceedings of INTERSPEECH*, 2081–2084.

[21] Joty, S., Carenini, G. and Lin, C.-Y. 2011. Unsupervised Modeling Of Dialog Acts In Asynchronous Conversations. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*. 1807–1813.

[22] Keizer, S., Akker, R. and Nijholt, A. 2002. Dialogue Act Recognition with Bayesian Networks for Dutch Dialogues. In *Proceedings of the SIGDIAL Workshop*, 88–94.

[23] Lee, D., Jeong, M., Kim, K., Ryu, S. and Geunbae, G. 2013. Unsupervised Spoken Language Understanding for a Multi-Domain Dialog System. *IEEE Transactions On Audio, Speech, and Language Processing*. 21, 11, 2451–2464.

[24] Marineau, J., Wiemer-Hastings, P., Harter, D., Olde, B., Chipman, P., Karnavat, A., Pomeroy, V., Rajan, S. and Graesser, A. 2000. Classification of Speech Acts in Tutorial Dialog. In *Proceedings of the Workshop On Modeling Human Teaching Tactics And Strategies at ITS*. 65–71.

[25] Meece, J.L. and Holt, K. 1993. A Pattern Analysis Of Students' Achievement Goals. *Journal Of Educational Psychology*. 85, 4, 582–590.

[26] Merceron, A. and Yacef, K. 2003. A Web-Based Tutoring Tool With Mining Facilities to Improve Learning and Teaching. In *Proceedings of AIED*, 201–208.

[27] Merceron, A. and Yacef, K. 2005. Clustering Students To Help Evaluate Learning. *Technology Enhanced Learning*. 171, 31–42.

[28] Ng, R.T. and Han, J. 1994. Efficient and Effective Clustering Methods for Spatial Data Mining. In *Proceedings of the 20th International Conference on Very Large Data Bases*, 144–155.

[29] Reithinger, N. and Klesen 1997. Dialogue Act Classification Using Language Models. In *Proceedings of EuroSpeech*, 2235–2238.

[30] Ritter, A., Cherry, C. and Dolan, B. 2010. Unsupervised Modeling of Twitter Conversations. In *Proceedings of the Association for Computational Linguistics*, 172–180.

[31] Rus, V., Moldovan, C., Niraula, N. and Graesser, A.C. 2012. Automated Discovery of Speech Act Categories in Educational Games. In *Proceedings of EDM*, 25–32.

[32] Searle, J.R. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

[33] Serafin, R. and Di Eugenio, B. 2004. FLSA: Extending Latent Semantic Analysis With Features For Dialogue Act Classification. In *Proceedings of the Association for Computational Linguistics*, 692–699.

[34] Sridhar, V.K.R., Bangalore, S. and Narayanan, S.S. 2009. Combining Lexical, Syntactic and Prosodic Cues For Improved Online Dialog Act Tagging. *Computer Speech & Language*. 23, 4, 407–422.

[35] Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C. Van and Meteer, M. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*. 26, 3, 339–373.

[36] VanLehn, K., Jordan, P.W., Rosé, C.P., Bhembe, D., Bottner, M., Gaydos, A., Makatchev, M., Pappuswamy, U., Ringenberg, M., Roque, A., Siler, S. and Srivastava, R. 2002. The Architecture Of Why2-Atlas: A Coach For Qualitative Physics Essay Writing. In *Proceedings of ITS*, 158–167.

Mining the Web to Leverage Collective Intelligence and Learn Student Preferences

Antonio Moretti†, José P. González-Brenes*, Katherine McKnight†
†Center for Educator Learning & Effectiveness
*Center for Digital Data, Analytics & Adaptive Learning
Research & Innovation Network, Pearson
{antonio.moretti, jose.gonzalez-brenes, kathy.mcknight}@pearson.com

ABSTRACT

University professors of conventional offline classes are often experts in their research fields, but have little training on educational sciences. Current educational data mining techniques offer little support to them. In this paper we propose a novel algorithm, Analyzing CurrIculum Decisions (ACID), that leverages collective intelligence to model student opinions to help instructors of traditional classes. ACID mines publicly available educational websites, such as student ratings of professors and course information, and learns student opinions within a statistical framework. We demonstrate ACID to discover patterns in learner feedback and factors that affect Computer Science instruction. Specifically, we investigate the choice of a programming language for introductory courses, the grading criteria and the posting of a publicly available online syllabus.

Keywords

offline teacher support, collective intelligence, web mining

1. INTRODUCTION

There are thousands of undergraduates in computer science programs throughout the US, roughly 24% of whom will switch majors to non-computing fields [7]. An essential component of retaining students is the quality of instruction that students receive in introductory courses [7]. While clear instruction and good pedagogy are widely acknowledged as fundamental to retention, supports for instructors to improve their educational practice are often based on old data; the languages used in computer science courses quickly evolve and old surveys are not useful. In this paper, we develop a data mining technique that will help provide insight into learner feedback which can be translated into changes that affect course quality. In general, our approach is similar to large scale surveys that attempt to be representative of student populations. The benefits of our approach are that it is rapid and inexpensive due to its use of publicly available information on the Web.

The field of educational data mining has been cultivating a strong interest in creating technologies to mine data collected from sophisticated online systems such as intelligent tutoring systems, virtual learning environments, and recently from Massive Open Online Courses (MOOC). The merits of these complex online systems have been demonstrated empirically [2, 8] with controlled studies. MOOCs are a powerful resource that allow educators to study student behavior and social learning in a controlled environment, however the scope of the impact of such technologies is limited. For example, a recent survey of active MOOC users in 200 countries and territories revealed that an overwhelmingly majority of students on these courses correspond to the most educated elite of their respective countries [3]. It is clear that improving basic education worldwide is necessary before MOOCs can deliver their promise. Moreover, because most education still happens offline, it is important to provide educational technologies that can utilize the power of internet to understand student behavior and to deliver these technologies to traditional offline classes. It is not clear how existing educational data mining technologies can help bridge this divide.

We discuss the *Analyzing CurrIculum Decisions*(ACID) [11] methodology, which has been presented and applied briefly. In this paper we elaborate on both our methodology and statistical model and expand upon our results. ACID is an algorithm that leverages collective intelligence within a statistical framework. ACID supports the decisions of instructors of traditional offline courses by extracting from the web teaching syllabi data, and using crowd-sourcing to pair it up with students' course ratings, comments and sentiment to analyze the relationship between the two.

This paper reports a case study of using the ACID methodology to explore three questions that instructors of computer science courses face when designing their courses. In addition we discuss ACID's heuristic value within a larger educational framework. We address the following questions:

1. **What course activities and grading rubric correlate with clear instruction?** The question of how to design a grading rubric and weight course activities determines what students focus on within a course. It is important for instructors to optimize course activities and grading criteria with respect to the student experience.

Published in CEUR-WS:
NCFPAL workshop (Ritter and Fancsali)
In EDM 2014 Extended Proceedings (Gutierrez-Santos and Santos)

Algorithm 1 ACID pseucode

n universities to analyze, z reviews to analyze

procedure ACID

  while $|R| < z$ **do**

    $s \leftarrow$ sample of n universities

    $s \leftarrow$ Remove non-English speaking universities

    $R \leftarrow$ Search_The_Web_For_Reviews(s)

    $R \leftarrow$ ratings rated by more than ϵ students

  $Q \leftarrow$ CrowdSource_Questionnaire(R)

  Analyze_Data(Q)

2. **For introductory classes, which programming language(s) correlate with clear instruction?** Academics and industry professionals disagree as to the programming language that is best suited for beginners [16]. For example, some argue that introductory courses should use interpreted languages that allow for a faster understanding of the applications of programming rather than compiled languages that rely heavily on language-specific syntax. Others believe that developing skill with compiled languages is necessary for future work in computer science. The choice of a first programming language likely affects students' decision to continue education within the field of computer science.

3. **Are students more interested in courses with publicly available online syllabi?** The choice to make a syllabus publicly available adds to information available to prospective students on the Web. We hypothesize that the posting of an online syllabus can be used as a proxy for factors including instructor organization and motivation, and that students will both be more interested in and prefer these courses.

The rest of this paper is organized as follows. § 2 explains the ACID methodology; § 3 describes three case studies of evaluating teaching decisions using ACID; § 4 relates to prior work; § 5 concludes.

2. ANALYZING CURRICULUM DECISIONS

Pseudocode for the ACID methodology is presented in Algorithm 1. For a given number of reviews, we sample n universities, remove the non-English speaking universities, scrape and parse the relevant reviews from a ratings website and retain ratings rated by more than a given number of students. We then extract information from these courses using crowd-sourcing, and analyze the data. We describe the process in detail below.

To evaluate the relative impact of different course features, we mine the web for data that reflect:

- **Curriculum decisions** University professors often upload information about their classes. This information is targeted towards prospective or enrolled students. This information includes syllabi with detailed descriptions of course material such as textbooks, projects, home-works and exams. We make use of this data to infer teaching strategies.

- **Student perceptions of the course.** We make use of self-selected student evaluations collected from a third-party website. The validity and usefulness of self-selected online rating systems, have been assessed in the literature [1, 12]. For example, evidence suggests that online ratings do not lead to substantially more biased ratings than those done in a traditional classroom setting [1] and that online ratings are a proxy to measure student learning [12]: student learning can often be modeled as a latent variable that causes patterns of observed faculty ratings. Researchers hypothesize a non-linear or concave relationship between student learning and the perceived difficulty level of a course [12]; students learn most when a course is not too difficult or too easy. Our work relies on self-selected ratings as a metric to study learner opinion.

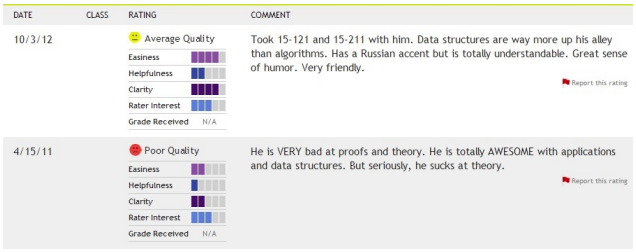


Figure 1: Two Examples from the Ratings Sample

Table 1: Statistics for the Ratings Sample

	Easiness	Helpfulness	Clarity	Interest
Mean	2.84	3.30	3.24	3.35
Std. Dev.	1.33	1.62	1.59	4.00
Median	3.00	4.00	4.00	1.38

We use publicly available self-selected ratings of professors from a third-party website, *Rate My Professor*[1] (RMP). This site allows students to rate the professors of the courses they have taken. The database contains data from over 13 million ratings for 1.5 million professors. They collect ratings on a 1—5 scale (being 1 the lowest possible score, and 5 the highest) under the categories of "easiness", "helpfulness" and "clarity." Additionally students may fill out an "interest" field in which they indicate how appealing the class was before enrolling, and a 350 character summary of their class experience. We focus on perceived clarity because of the direct link between clarity and quality of instruction.

For the purposes of this paper, we focus on Computer Science courses due to our familiarity with the content. Since we do not have access to the ratings database, we develop a process to sample data from the website. For this, we first select a random sample of 50 international universities that teach Computer Science from the Academic Ranking of

[1] `ratemyprofessor.com`

World Universities[2] [14]. From this sample we only consider the 41 universities are English speaking.

We find, scrape and parse the reviews of the ratings data-set for all professors within the computer science departments of the universities in our sample. We remove the ratings from faculty that were rated by fewer than 30 students. More than one professor can teach the same course. For our analysis, we describe one course listing taught by two different professors as two separate courses. Table 1 shows the mean, standard deviation and median of the ratings in our sample. Figure 1 shows two sample ratings for one professor from our sample. The professor name and course names are removed for privacy.

We use Amazon Mechanical Turk, a crowdsourcing platform, to find course features for each of the courses in our ratings sample. We do this by asking respondents to fill out a survey. The survey requests to provide the URL for the online syllabus that corresponds to the course and professor from which we have ratings that is closest to the date of the student review online. Then, using the syllabus, respondents are asked to to provide the programming language(s) used, the textbook(s) used, and the percentage of the grade that was determined by homework, projects, quizzes, exams and whether the course was taught online or in a blended format (both face-to-face and online). However, when we reviewed the responses to the blended format question, it appeared that most syllabi did not provide enough information by which to make an accurate response.

From our original sample of 1,112 courses taught by a unique professor, respondents find an online syllabus matching the professor for 342 courses (∼31%). We hypothesize three explanations for the missing syllabi: (i) the syllabi may be accessed only with a password through a course management system, such as blackboard, (ii) the syllabi may not be available only, or (iii) the respondents are not able to find the syllabi.

3. DATA ANALYSIS: WHAT MAKES A BETTER CLASS?

We report our results of applying the ACID methodology to evaluate teaching decisions. In § 3.1 we assess the quality of the data collected by the crowd sourcing platform. In § 3.2 we discuss the statistical model we use. In § 3.3 we report the results of using ACID.

3.1 Data Quality

We now report the how we attempt to collect high-quality data through the use of crowd-sourcing and how we assess the quality of our data.

Mechanical Turk provides a "master" qualification level to respondents that are more reliable. Masters-level respondents require higher compensation for crowd-sourcing tasks than non-masters level respondents although their "acceptance rate," or proportion of approved tasks is much higher. We ran a preliminary experiment, to decide whether respondents on master level qualification provide better quality data for our purposes. We ask respondents to find the syllabus corresponding to a random sample of 30 courses and to answer a set of questions. Table 2 shows the accuracy and interrater agreement of Masters and non-Masters level respondents.

Table 2: Respondent Validation

	Accuracy	Interrater Agreement
Masters	100%	96.67%
non-Masters	85.56%	6.07%

In the pretest we used a screening question to evaluate the accuracy of respondents' data on each task. We asked respondents to find the URL of the website of a randomly selected faculty member at Carnegie Mellon University from a set of 8, from which we knew the answer. We compared the URL they provided with the correct URL to assess accuracy. Of the 13 responses of non-masters workers that did not provide an exact URL match, five responses left the validation question blank. We found that respondents with master level qualification were significantly more accurate (i.e. answered the validation item correctly) than the non-Masters level respondents (p-value = 0.0002).

Additionally, we tested interrater agreement by asking 3 respondents to carry out the same task, i.e. finding the same URL (for a total of 3x30 or 90 tasks). We used a dummy variable to code whether the three respondents provided the same URL for the course syllabus. Our measure of agreement is calculated by taking the proportion of total responses in which all three respondents provide the same URL. Masters-level respondents agreed (i.e. all three provided the same URL) 100% of the time, whereas the non-Masters level respondents performed much worse – only 6% agreed. As a result of these comparisons, we decided to hire only Masters-level respondents to complete the crowdsourcing experiment.

After collecting the data using Masters level respondents, we performed a post-hoc analysis by examining the responses to the screening question. From the final group of 342 responses that provided a link to an online syllabus, 325 responses (95.03%) provided the correct URL for the faculty website. It should be noted that 13 of the 17 responses that did not provide an exact URL match provided the website for a different faculty member from the set of 8, suggesting that they copied and pasted their previous response without checking to see that the prompt had changed for the new response. Two of the 17 responses provided a link to the directory website for the faculty member rather than the faculty member's personal website. One response provided the correct faculty member's website within the department of Statistics rather than the department of Computer Science (the faculty member is in both departments).

3.2 Model

We describe our general linear mixed model. We provide descriptive statistics and model selection criteria.

[2]Academic Ranking of World Universities is also known as Shanghai Ranking `shanghairanking.com`

Table 3: VPC and ICC Statistics

	University	Professor	Course
VPC	0.0646	0.3365	0.2355
ICC	0.0728	0.3425	0.1982

We explore the relationship between student reviews and features collected from online syllabus data using general linear mixed modeling. Student reviews are organized at three levels: by university, professor and course. It is important to note the non-independence of the student reviews due to the hierarchical or clustered nature of the data. We suspect that student ratings within each course, professor and perhaps university are correlated. We begin by estimating the amount of variance attributed to each of these three levels. The simplest multilevel model does not yet include explanatory variables:

$$y_{i,j} = \beta_0 + u_{0,j} + \epsilon_{i,j} \tag{1}$$

The dependent variable $y_{i,j}$ is the clarity rating that student i gave to level j. The term β_0 represents the intercept or mean student clarity rating across all observations. The term $u_{0,j}$ represents the mean clarity rating for level j. The term $\epsilon_{i,j}$ represents the error attributed to student rating i at level j. For comparison we fit a null or single-level model:

$$y_{i,j} = \beta_0 + \epsilon_{i,j} \tag{2}$$

We calculate the percentage of variation in the data set that is separately attributed to each of the three levels of the data. Conventionally the variance partition coefficient (VPC) and intraclass correlation coefficient (ICC) can be interpreted similarly to an R-squared term and are reported in Table 3.

$$\rho = 1 - \frac{\sigma_e^2}{\sigma_e^2 + \sigma_u^2} \tag{3}$$

The VPC and ICC are denoted by ρ, the residual variance is denoted by σ_e^2 and the variance of the effect is denoted by σ_u^2. The ICC is a statistic that is similar to the VPC. However, since the parameter values of the within and between level variance are estimated using sample data, there may be bias due to sampling variation, particularly when there are fewer observations within a given level. The ICC as described by Bartko [1] corrects for this bias by making a small computational adjustment.[3] Observe that the ICC term appears to give slightly less weight to the course effect. It is clear from both statistics that the main effect is the professor effect.

We examine the professor level-residuals and their associated standard errors to look for variation in clarity ratings across professors. The caterpillar plot displays the professor residuals in rank order together with 95% confidence intervals. Wider intervals occur for professors with more student reviews. Observe that the majority of the intervals do not overlap and thus there are significant differences between professors. The blue circles on the far left represent professors who are rated two standard deviations below the mean clarity rating, whereas those on the far right are 1.5 standard deviations higher than the mean clarity rating. The red horizontal line refers to the "average" professor.

[3]For a description of the computation of the ICC, see the documentation and source code for the R library *lme.*

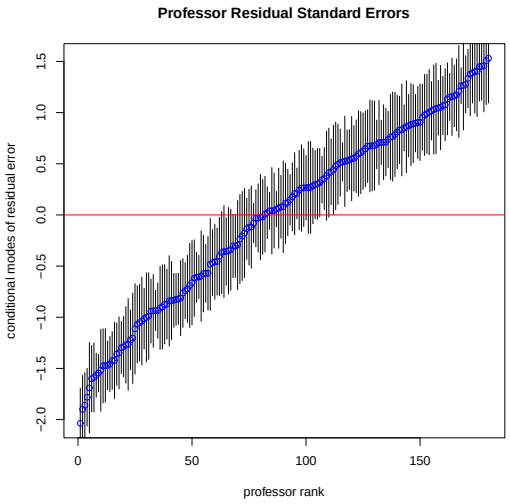


Figure 2: 95% CI for Professor Residual Error

We calculate a Chi-squared likelihood ratio statistic by taking the difference between log likelihood values of two successive models. We begin by comparing the null model and the course level model to compare the significance of including the course effect. We continue by adding each of the additional effects. We do not report the values of the test statistic although all additional levels of complexity are statistically significant. We consider the Bayesian information criterion (BIC) and Akaike information criterion (AIC) as model selection tools to avoid over-fitting the data. The BIC and AIC penalize the log-likelihood of a model for the inclusion of extra parameters. The parameters are estimated using restricted maximum likelihood estimation (REML).

We choose the model with the minimum BIC. A two-level mixed model including course effect and professor effect provides the optimal Bayesian information criterion value. Two and three way interaction effects were considered although they did not decrease the AIC or BIC of any of the models. While the log likelihood value is maximized by including the university effect, a simpler model is preferable because it involves fewer parameter estimates and is more likely to generalize. The model can be written in matrix form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\nu} + \boldsymbol{\epsilon} \tag{4}$$

$\mathbf{Y}$ denotes the response variable observations (student ratings). The matrix $\boldsymbol{\beta}$ represents a vector of fixed-effects parameters with a design matrix $\mathbf{X}$. $\mathbf{Z}$ is a design matrix of indicator variables denoting group membership across random-effect levels and $\boldsymbol{\nu}$ is a vector containing random-effect parameters. $\boldsymbol{\epsilon}$ is a vector of error terms.

3.3 Case Studies

We show the results of using the ACID methodology to answer three course design questions.

Table 4: Programming Language Statistics

	Value	Std.Err	t-value	Pr<|t|	n
C	3.38	0.32	10.58	0.0000	109
C++	3.30	0.31	10.65	0.0000	214
Java	3.62	0.19	19.33	0.0000	353
Python	3.70	0.26	14.50	0.0000	133
Scheme	4.06	0.47	8.61	0.0000	32
Scratch	3.91	0.84	4.67	0.0000	49

3.3.1 *For introductory classes, which programming language do students associate with clear instruction?*

Professors teaching introductory level courses in computer science choose between a number of programming languages and textbooks. We make use of the data collected to provide insights into which programming languages beginning students associate with clear instruction. We filter the data to only include introductory level courses (one which does not require any prerequisite coursework in computer science). Our restricted sample includes 1,024 reviews; 34.58% of all reviews with syllabus data are of introductory courses. We explore the relationship between clarity ratings and programming language with random professor and course effects. Programming languages with less than 30 student reviews are not reported[4]. Table 4 gives the estimates for student ratings of clarity by programming language and their associated p-values. An intercept is not modeled in order to make the results easily interpretable. The mean clarity rating for introductory courses is 3.599.

We found C and C++ had the lowest coefficients (i.e. compiled languages had the lowest perceived clarity ratings). Scheme and Scratch have the highest clarity ratings followed by Python and Java. We note that the standard errors are largest for Scheme and Scratch and smallest for Java and Python. This suggests that results for Java and Python are stronger. Students in our sample associate clearer instruction with interpreted languages rather than compiled languages. Also, both Python and Java are associated with clearer instruction than C or C++.

3.3.2 *What mix of course activities – exams, quizzes, homework and projects – do students associate with clear instruction?*

To assess students' course ratings of clarity based on the percentage of the grade due to exams, quizzes, homework and projects, we created a factor made up of four clusters representing four ways of weighting homework, projects, exams, quizzes and miscellaneous (such as extra credit) for the students' grade. We begin by sorting the data to only include observations in which the grading criteria (percentage of the grade determined by homework, projects, exams, quizzes and miscellaneous) is available and sums to 100. Of the 2,935 observations with syllabus data, there are 2,225 observations with full grading criteria. The difference in these numbers represents 710 ratings for which the respondents were not able to find a complete grade breakdown from the online syllabus.

[4]SQL is a special purpose programming language used only for relational databases and is not reported.

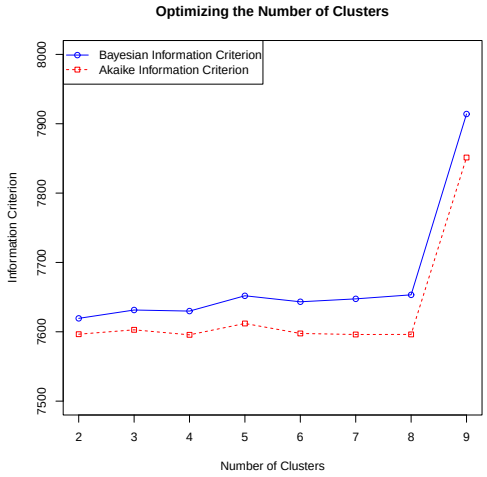


Figure 3: Information Criterion

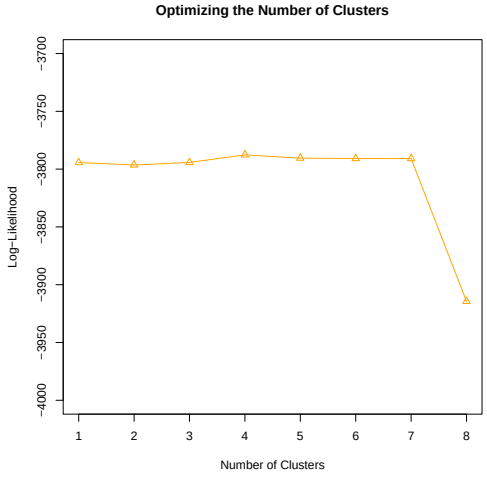


Figure 4: Log Likelihood

Table 5: Cluster Statistics

	HW	Projects	Exams	Quizzes	Other
Cluster1	18.11	2.36	76.66	0.61	2.25
Cluster2	20.59	7.90	48.90	12.46	10.15
Cluster3	7.00	40.18	46.23	3.51	3.08
Cluster4	42.93	0.76	54.61	0.70	2.00

Table 6: Grading Criteria Statistics

	Clarity	Std.Err	t-value	Pr<|t|	n
Exam Heavy	3.23	0.12	26.91	0	726
Equal Mix	3.52	0.14	26.04	0	484
Exam Proj	3.65	0.13	27.76	0	610
Exam HW	3.12	0.13	23.53	0	415

We use k-means clustering to partition the 2,225 observations with complete grading criteria information based on the five aforementioned variables. We optimize k, our number of clusters, by examining how the BIC and AIC of the mixture model change based on the number of clusters selected. Figure 3 displays the information criterion and Figure 4 displays the log-likelihood values for each number of clusters respectively. A solution involving two clusters minimizes the BIC of the model, whereas a four cluster solution minimizes the AIC. The log likelihood is optimized with the four cluster solution. We consider both two and four cluster models as optimal and we find that they lend themselves to similar interpretation. The cluster means for the four cluster solution are presented in table 5.

The first cluster represents courses that are heavily weighted towards exams with a smaller weight towards homework. The second cluster represents a more even weighting of exams, homework, projects and quizzes. The third cluster represents an equal weighting towards exams and projects. The fourth cluster represents courses that are heavily weighted towards exams and homework. The cluster membership is treated as a predictor variable and modeled using equation 4. Table 6 displays the estimated clarity ratings within each group for the four cluster solution.

The exams and projects cluster has the highest estimate of clarity. We find that weighting projects equally with exams is associated with a clearer course experience. The equal mix cluster also is associated with higher clarity estimates. The exam heavy cluster and the exam and homework heavy clusters are associated with lower student clarity ratings. We find that a rubric that weights exams and projects evenly has higher perceived clarity ratings to a rubric which is weighted heavily towards exams and homework. This result extends to both two and four cluster solutions.

3.3.3 *Does the posting of a syllabus online translate into higher ratings?*

We hypothesize the posting of the syllabus online is a proxy for organization, perhaps motivation or drive of the professor. We make use of all of the data collected to compare student reviews of professors who have a publicly available syllabus and of those who do not. Many professors may choose to only post a syllabus through course management systems that require a password. Potential students of these courses are unable to access the syllabus to determine whether the course would be a good fit. We treat the posting of an online syllabus as a factor and test for differences in clarity ratings between the two groups using our model.

We find statistically significant differences between clarity, helpfulness and interest ratings and report the clarity estimates for the two groups in Table 7. We note that the difference in easiness ratings is not statistically significant. We find evidence that students are more interested in professors and courses in which the syllabus is made publicly available. We note that the parameter estimates for the two groups are within one standard error of one another which suggests that the conclusions are modest.

Table 7: Online Syllabi

	Clarity	Std. Err	t-value	Pr<|t|	n
Available	3.33	0.07	44.48	0	2953
Not Found	3.26	0.07	46.03	0	7702

4. RELATION TO PRIOR WORK

Research has recently focused on online faculty ratings with mixed conclusions. Felton et al. [4] found that online instructor ratings were associated with perceived easiness, and that a "halo effect" existed in which raters gave high scores to instructors perhaps because their courses were easier. We find that student ratings of clarity and easiness are correlated (ρ=0.45) although not as strongly associated as clarity and helpfulness. We do find that student ratings of clarity and helpfulness are highly correlated (ρ=0.84). We chose to focus on clarity ratings as we assumed these were less susceptible to a "halo effect" and other bias relative to the overall ratings of a course or professor. Otto et al [13] found issues related to bias in online ratings stating that online ratings are characterized by selection bias as anyone can enter faculty ratings at any time. Carini et al [1], Hardy [5], McGhee and Lowell [6] had contradictory results finding that an online format did not lead to more biased ratings. Otto et al. [12] hypothesized that instructor clarity and helpfulness as captured by Rate My Professor are more positively associated with student learning than easiness.

Several approaches have been proposed to synthesize responses using crowd sourcing systems such as Amazon's Mechanical Turk. Majority voting is perhaps the simplest way to combine crowd responses using equal weights irrespective of respondent experience. The results of our preliminary analysis in accessing the accuracy of non-Masters level respondents correspond to the steep drop in respondent accuracy noted by Karger [9] when low-quality respondents are present. Whitehill et al [15] proposed a probabilistic model for combining crowd responses called Generative model of Labels, Abilities and Difficulties (GLAD). The GLAD methodology makes use of the EM algorithm to calculate parameter estimates of unobserved variables including an approximation of the expertise of the rater. Khattak and Salleb-Aouissi compared the accuracy and percentage of bad responses using majority voting, probabilistic models, and their novel approach entitled Expert Label Injected Crowd Estimation (ELICE) [10]. ELICE makes use of a few "ground truth" responses and incorporates expertise of the labeler, difficulty of the instance and an aggregation of labels. Khattak and Salleb-Aouissi found that their approach was robust and outperformed GLAD and iterative methods even when bad labelers were present. Our simple approach was to use Masters level respondents from Mechanical Turk although GLAD and ELICE are alternative methods to reduce the number of expert level respondents required while also obtaining high quality data.

5. CONCLUSIONS, LIMITATIONS AND FUTURE WORK

We demonstrate how the Analyzing CurrIculum Decisions (ACID) methodology can be used to leverage collective intelligence and learn student preferences. In introductory

computer science courses, we find that students that are taught interpreted languages find their classes clearer. We also that find students who are given an even weighting of exams and projects find their classes clearer; and that interest in a course corresponds to the availability of an online syllabus. Our study does not necessarily suggest that teachers should change their programming language. Further research is needed before drawing causal inferences. We argue that ACID is a beneficial tool to discover patterns in student behavior. Syllabus data and course ratings data are becoming increasingly available on the Web. This data is used by millions of students and worthy of further research.

This study can be expanded in several ways. Student evaluations often include free form text where students can describe their experience in the course. Sentiment analysis is a probabilistic approach for categorizing student comments as being either positive or negative. One extension is to regress text sentiment on course features. There is arguably a strong association between comment sentiment and student preference. Another way ACID can be applied is to disciplines other than computer science, or to discover patterns in syllabi across disciplines that can provide insight into learner experiences.

6. REFERENCES

[1] R. Carini, J. Hayek, G. Kuh, J. Kennedy, and J. Ouimet. College student responses to web and paper surveys: does mode matter? *Research in Higher Education*, 44(1):1–19, 2003.

[2] A. Corbett. Cognitive computer tutors: Solving the two-sigma problem. In M. Bauer, P. Gmytrasiewicz, and J. Vassileva, editors, *User Modeling 2001*, volume 2109 of *Lecture Notes in Computer Science*, pages 137–147. Springer Berlin Heidelberg, 2001.

[3] E. J. Emanuel. Online education: Moocs taken by educated few. *Nature*, 503(7476):342–342, 2013.

[4] J. Felton and J. Mitchell. Web based student evaluations of professors: the relations between perceived quality, easiness and sexiness. *Assessment and Evaluation in Higher Education*, 29(1):91–108, 2004.

[5] N. Hardy. Online ratings: fact and fiction. *New Directions for Teaching and Learning*, (96):31–38, 2003.

[6] N. Hardy. Psychometric properties of student ratings of instruction in online and on-campus courses. *New Directions for Teaching and Learning*, 2003(96):39–48, 2003.

[7] M. Haungs, C. Clark, J. Clements, and D. Janzen. Improving first-year success and retention through internet-based cs0 courses. *ACM SIGCSE*, pages 549–594, 2012.

[8] S. Jaggars and T. Bailey. Effectiveness of fully online courses for college students: Response to a department of education meta-analysis. *Teachers College: Community College Research Center*, 2010.

[9] S. Karger, D. Oh and D. Shah. Budget–optimal task allocation for reliable crowdsourcing systems. *CoRR*, arXiv:1110.3564, 2011.

[10] F. Khattak and A. Salleb-Aouissi. Robust crowd labeling using little experience. *Discovery Science*, 8140:94–109, 2013.

[11] A. Moretti, J. Gonzalez-Brenes, and K. McKnight. Towards data–driven curriculum design: Mining the web to make better teaching decisions. *EDM*, 2014.

[12] J. Otto, D. A. Sanford Jr, and D. N. Ross. Does ratemyprofessor. com really rate my professor? *Assessment & Evaluation in Higher Education*, 33(4):355–368, 2008.

[13] J. Otto, D. A. Sanford Jr, and W. Wagner. Analysis of online student ratings of university faculty. *Journal of College Teaching & Learning*, 2(7):25–30, 2005.

[14] Shanghai. Academic ranking of world universities. *Retrieved from http://www.shanghairanking.com/*, Accessed at 2013 12 01.

[15] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Neural Information Processing Systems*, pages 2035–2043, 2009.

[16] J. Zelle. Python as a first language. *Retrieved from http://mcsp.wartburg.edu/zelle/python/python-first.html/*, Accessed at 2014 02 23.

APPENDIX

A. SAMPLE OF UNIVERSITIES SELECTED

	Country	n Professors	n Courses	n Reviews
Colorado State	USA	1	9	32
Carnegie Mellon University	USA	3	21	102
North Carolina State	USA	2	10	63
Pennsylvania State	USA	12	74	938
Rensselaer Polytechnic Institute	USA	3	22	131
Rutgers	USA	8	30	468
Simon Fraser	Canada	27	98	1873
SUNY Stony Brook	USA	8	55	505
UC Davis	USA	10	44	589
UNC Chapel Hill	USA	1	4	49
University of Alberta	Canada	2	6	69
University of Arizona	USA	3	13	158
University of Delaware	USA	15	56	806
University of Florida Gainsville	USA	5	36	321
University of Illinois at Urbana	USA	5	14	339
University of Massachusetts	USA	6	39	405
University of Montreal	USA	1	6	59
University of Toronto	Canada	14	66	775
University of Utah	USA	2	17	66
University of Virginia	USA	3	19	131
University of Waterloo	Canada	46	125	2700
Vanderbilt University	USA	2	10	76

Non-cognitive factors of learning as predictors of academic performance in tertiary education

Geraldine Gray, Colm McGuinness, Philip Owende
Institute of Technology Blanchardstown
Blanchardstown Road North
Dublin 15, Ireland
geraldine.gray@itb.ie

ABSTRACT

This paper reports on an application of classification and regression models to identify college students at risk of failing in first year of study. Data was gathered from three student cohorts in the academic years 2010 through 2012 (n=1207). Students were sampled from fourteen academic courses in five disciplines, and were diverse in their academic backgrounds and abilities. Metrics used included non-cognitive psychometric indicators that can be assessed in the early stages after enrolment, specifically factors of personality, motivation, self regulation and approaches to learning. Models were trained on students from the 2010 and 2011 cohorts, and tested on students from the 2012 cohort. Is was found that classification models identifying students at risk of failing had good predictive accuracy ($> 79\%$) on courses that had a significant proportion of high risk students (over 30%).

Keywords

Educational data mining, learning analytics, academic performance, non cognitive factors of learning, personality, motivation, learning style, learning approach, self-regulation

1. INTRODUCTION AND LITERATURE REVIEW

Learning is a latent variable, typically measured as academic performance in continuous assessment and end of term examinations [33]. Identifying predictors of academic performance has been the focus of research for many years [20, 34], and continues as an active research topic [6, 8], indicating the inherent difficulty in generating models of learning [29, 46]. More recently, the application of data mining to educational settings is emerging as an evolving and growing research discipline [40, 43]. Educational Data Mining (EDM) aims to better understand students and how they learn through the use of data analytics on educational data [42, 10]. Much of the published work to date is based on ever-increasing volumes of data systematically gathered by education providers, particularly log data from Virtual Learning Environments and Intelligent tutoring systems [16, 2]. Further work is needed to determine if gathering additional predictors of academic performance can add value to existing models of learning.

Research from educational psychology has identified a range of non-cognitive psychometric factors that are directly or indirectly related to academic performance in tertiary education, particularly factors of personality, motivation, self regulation and approaches to learning [8, 9, 35, 39, 44, 25]. Personality based studies have focused on the Big-5 personality dimensions of conscientiousness, openness, extroversion, stability and agreeableness [9, 22, 27]. There is broad agreement that conscientiousness is the best personality based predictor of academic performance [44]. For example, Chamorro et al. [9] reported a correlation of r=0.37 ($p<0.01$, n=158) between conscientiousness and academic performance. Correlations between academic performance and openness to new ideas, feelings and imagination are weaker. Chamorro et al. [9] reported a correlation of r=0.21 ($p<0.01$, n=158) but lower correlations were reported in other studies (see Table 1) which may be explained by variations in assessment type. Open personalities tend to do better when assessment methods are unconstrained by submission rules and deadlines [27]. Studies are inconclusive on the predictive validity of other personality factors [44].

A meta-analysis of 109 studies analysing psychosocial and study skill factors found two factors of motivation, namely self-efficacy (90% CI [0.444,0.548]) and achievement motivation (90% CI [0.353, 0.424]), had the highest correlations with academic performance [39]. Distinguishing between learning (intrinsic) achievement and performance (extrinsic) achievement goals, Eppler and Harju [19] found learning goals (r=0.3, $p<0.001$, n=212) were more strongly correlated with academic performance than performance goals (r=0.13, $p> 0.05$, n=212). Covington [13] however argues that setting goals in itself is not enough, as ability to self-regulate learning can be the difference between achieving, or not achieving, goals set. Self-regulated learning is recognised as a complex concept to define as it overlaps with a number of other concepts including personality, self-efficacy and goal setting [4]. Ning and Downing [35] reported high correlations between self regulation and academic performance, specifically self-testing (r=0.48, $p<0.001$) and monitoring understanding (r= 0.42, $p<0.001$). On the other hand, Komarraju and Nadler [31] found effort management, includ-

Published in CEUR-WS:
NCFPAL workshop (Ritter and Fancsali)
In EDM 2014 Extended Proceedings (Gutierrez-Santos and Santos)

ing persistence, had higher correlation with academic performance (r=0.39, p<0.01) than other factors of self-regulation and found that self-regulation (monitoring and evaluating learning) did not account for any additional variance in academic performance over and above self-efficacy, but study effort and study time did account for additional variance.

Research into approaches to learning has its foundations in the work of Marton & Säljö [32] who classified learners as shallow or deep. Deep learners aim to understand content, while shallow learners aim to memorise content regardless of their level of understanding. Later studies added strategic learners [18, pg. 19], whose priority is to do well, and will adopt either a shallow or deep learning approach depending on the requisites for academic success. Comparing the influence of approaches to learning on academic performance, Chamorro et al [9] reported a deep learning approach (r=0.33, p<0.01) had higher correlations with academic performance than a strategic learning approach (r=0.18, p<0.05). Cassidy [8] on the other hand found correlations with a deep learning approach (r=0.31, p<0.01) were marginally lower than with a strategic learning approach (r=0.32, p<0.01). Differences found have been explained, in part, by assessment type [49], highlighting the importance of assessment design in encouraging appropriate learning strategies.

Knight, Buckingham Shum and Littleton argued learning measurement should go beyond measures of academic performance [29], promoting greater focus on learning environment and encouragement of malleable, effective learning dispositions. Disposition relates to a tendency to behave in a certain way [6]. An effective learning disposition describes attributes and behaviour characteristic of a good learner [6]. A range of non-cognitive psychometric factors have been associated with an effective learning disposition such as a deep learning approach, ability to self-regulate, setting learning goals, persistence, conscientiousness and subfactors of openness, namely intellectual curiosity, creativity and open-mindednesss [6, 29, 47]. A lack of correlation between such non-cognitive factors and academic performance is in itself insightful, suggesting assessment design that fails to reward important learning dispositions. It has been argued that effective learning dispositions are as important as discipline specific knowledge [6, 29].

Statistical models have dominated data analysis in educational psychology [15], particularly correlation and regression [25]. Relatively high levels of accuracy were reported in regression models of academic performance that included cognitive and non-cognitive factors. For example, Chamorro-Premuzic et al [9] reported a coefficient of determination (R^2) of 0.4 when predicting 2^{nd} year GPA (based on essay type examinations) in a regression model that included prior academic ability, personality factors and a deep learning approach. Robbins [39] reported similar results (R^2=0.34) in a meta-analysis of models of cognitive ability, motivation factors and socio-economic status. Models of non-standard students were less accurate, for example Swanberg & Martinsen [44] reported R^2=0.21 in models of older students (age: m=24.8) based on prior academic performance, personality, learning strategy, age and gender. Lower accuracies were also reported in studies not including cognitive ability. Robbins [39] reported R^2=0.27 in a meta-analysis of models of factors of motivation. Komarraju et al. [30] predicted GPA (R^2=0.15) from variables of personality and learning approach, while Bidjerano & Dai [4] had similar results (R^2=0.11) with factors of personality and self-regulation.

Linear regression assumes constant variance and linearity between independent and dependent attributes. There is evidence to suggest variance is not constant for some non-cognitive factors. For example, De Feyter et al. [14] found low levels of self-efficacy had a positive, direct effect on academic performance for neurotic students, and for stable students, average or higher levels of self-efficacy only had a direct effect on academic performance. In addition, Vancouver & Kendall [48] found evidence that high levels of self-efficacy can lead to overconfidence regarding exam preparedness, which in turn can have a negative impact on academic performance. Similarly, Poropat [38] cites evidence of non-linear relationships between factors of personality and academic performance, including conscientiousness and openness. It is therefore pertinent to ask if data mining's empirical modelling approach is more appropriate for models based on non-cognitive factors of learning.

A growing number of educational data mining studies have investigated the role of non-cognitive factors in models of learning [6, 41, 36]. Bergin [3] cited an accuracy of 82% using an ensemble model based on prior academic achievement, self-efficacy and study hours, but due to the small sample size (n=58) could not draw reliable conclusions from the findings. The class label distinguished strong (grade>55%) versus weak (grade<55%) academic performance based on end of term results in a single module. Gray et al. [23] cited similar accuracies (81%, n=350) with a Support Vector Machine model using cognitive and non cognitive attributes to distinguish high risk (GPA<2.0) from low risk (GPA≥2.5) students based on first year GPA. Model accuracy was contingent on modelling younger students (under 21) and older students (over 21) separately.

The focus of this study was to investigate if non-cognitive factors of learning, measured during first year student induction, were predictive of academic performance at the end of first year of study. We evaluated both regression models of GPA and classification models that predicted first year students at risk of failing. Participants were from a diverse student population that included mature students, students with disabilities, and students from disadvantaged socio-economic backgrounds.

2. METHODOLOGY

The following sections report on study participants and the study dataset. Data analysis was conducted following the CRoss Industry Standard for Data Mining (CRISP-DM) using RapidMiner V5.3 and R V3.0.2.

2.1 Description of the study participants

The participants were first year students at the Institute of Technology Blanchardstown (ITB), Ireland. The admission policy at ITB supports the integration of a diverse student population in terms of age, disability and socio-economic background. Each September 2010 to 2012, all full-time, first-year students at ITB were invited to participate in the study by completing an online questionnaire administered

Table 1: Correlations with Academic Performance in Tertiary Education

Study	N	age	AP	Temperament		Motivation			Learning Approach			Learning Strategy		
				Conscientious	Open	Self Efficacy	Intrinsic Goal	Extrinsic Goal	Deep	Shallow	Strategic	Self Regulation	Study Time	Study Effort
[4]	217	m=22	self reported GPA										0.33**	0.0.23**
[8]	97	m=23.5	GPA			0.397***			0.398**	-0.013	0.316**			
[9]	158	18-21	GPA	0.37**	0.21**				0.398*	-0.15	0.18*			
[17]	146	17-52	GPA	0.21	0.06				0.097	-0.054	0.153			
[19]	212	m=19.2	GPA				0.3***	0.13						
[27]	133	18-22	GPA	0.46**	-0.08									
[30]	308	18-24	self reported GPA	0.29**	0.13*									
[31]	257	m=20.5	GPA			0.3**						0.14*	0.31**	0.39**
[35]	581	20.48	GPA											0.0.24**
[39]	meta	analysis, 18+	GPA			0.496		0.179						
[44]	687	m=24.5	single exam						0.16	-0.25				

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < 0.001$

during first year student induction. A total of 1,376 (52%) full-time, first year students completed the online questionnaire. Eliminating students who did not give permission to be included in the study (35) and invalid data (134) resulted in 45% of first year full time students participating in the study (n=1207).

Participants ranged in age from 18 to 60, with an average age of 23.27; of which, 355 (29%) were mature students (over 23), 713 (59%) were male and 494 (41%) were female. There were 32 (3%) participants registered with a disability. Students were enrolled on fourteen courses across five academic disciplines, Business (n=402, 33%), Humanities (n=353, 29%), Computing (n=239, 20%), Engineering (n=172, 14%) and Horticulture (n=41, 3%).

Academic performance was measured as GPA, an aggregate score of between 10 and 12 first year modules, range 0 to 4, and was calculated on first exam sitting only. The GPA distribution (profiled sample) was compared with the GPA distribution of the full cohort of students for that year (reference sample) using a Kolmogorov-Smirnov non-parametric test. The recorded differences in the distribution for 2010 (D=0.032, p=0.93), 2011 (D=0.036, p=0.90) and 2012 (D=0.042, p=0.69) were not statistically significant. The distribution of GPA was also similar across the three years of study. The largest difference was between the 2010 and 2012 profiled samples (D=0.063, p=0.37) and was not significant. To pass overall, a student must achieve a GPA $\geq$ 2.0 and pass each first year module. 89% of students with GPA $>$ 2.5 passed all modules indication a low risk group that can progress to year two. 84% of students with a GPA $<$ 2 failed three or more modules, indicating a high risk group falling well short of progression requirements. Of the students in GPA range [2.0, 2.49], 39% passed all modules, 36% failed one module, 18% failed two modules, and 7 % failed more than two modules. This is a less homogenous group in terms of academic profile, but could be generally regarded as borderline, either progressing on low grades or required to repeat one or two modules in the repeat exam sittings. Figure 1 and Table 2 illustrate GPA distribution by course.

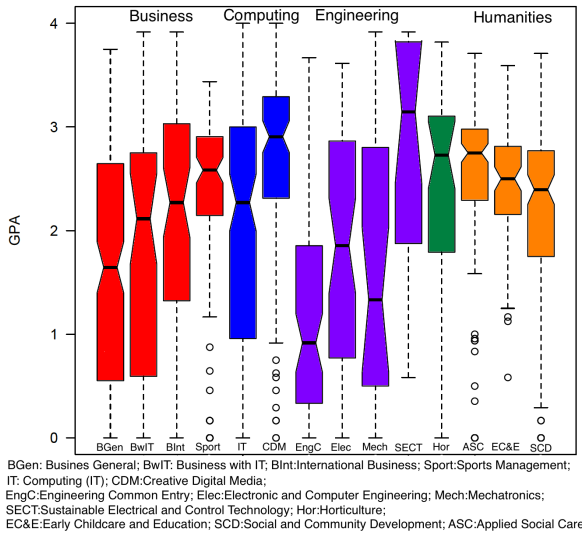


Figure 1: Notched box plots for GPA by course

2.2 The Study Dataset

Table 3 lists the psychometric factors included in the dataset, collected using an online questionnaire developed for the study (www.howilearn.ie). With the exception of learning modality, questions were taken from openly available, validated instruments, with some changes to wording to suit the context. Where two questions were similar on the published instrument, only one was included. This choice was made to reduce the overall size of the questionnaire, despite the likely negative impact on internal reliability statistics. Questionnaire validity and internal reliability were assessed using a paper-based questionnaire that included both the revised wording of questions used on the online questionnaire (reduced scale), and the original questions from the published instruments (original scale). The paper questionnaire was administered during scheduled first year lectures across all academic disciplines. Pearson correlations between scores calculated from the reduced scale, and scores calculated from the original scale, were high for all factors (>=0.9) except intrinsic goal orientation and study time and environment, confirming the validity of the study instrument for those factors. Internal reliability was assessed using Cronbach's alpha. All factors had acceptable reliability (>0.7)[1] given the small number of questions per scale (between 3 and 6), with the exception again of intrinsic goal orientation and study time and environment. Learner modality data (Visual, Auditory, Kinaesthetic (VAK) [21]) was based an instrument developed by the National Leaning Network Assessment Services (NLN) (www.nln.ie).

[1]While generally a Cronbach alpha of $>$ 0.8 indicates good internal consistency, Cronbach alpha closer to 0.7 can be regarded as acceptable for scales with fewer items [12, 45].

Table 2: Academic profile by course

Course Name	n	GPA*	high risk	borderline	low risk
all participants	1207	2.1±1.1	28%	16%	46%
Computing (IT)	137	2.0±1.2	47%	11%	42%
Creative Digital Media	102	2.6±1.0	20%	8%	72%
Engineering common	73	1.1±0.9	79%	8%	13%
Electronic & computer eng.	52	1.8±1.2	52%	10%	38%
Mechatronics	27	1.6±1.2	63%	7%	30%
Sustainable Electrical & Control Technology	20	2.8±1.1	30%	5%	65%
Horticulture	41	2.4±1.1	27%	2%	71%
Business General	183	1.7±1.1	56%	15%	29%
Business with IT	60	1.8±1.2	46%	22%	32%
Business International	64	2.2±1.1	41%	14%	45%
Sports Management	95	2.3±0.9	22%	24%	54%
Applied Social Care	146	2.5±0.7	15%	16%	69%
Early Childcare	80	2.4±0.6	20%	28%	52%
Social & Community Development	127	2.2±0.9	30%	27%	43%

*GPA mean and standard deviation.

Table 3: Study factors, mean and standard deviation

Category & Instrument	Study Factor
Personality: IPIP scales (ipip.ori.org) [22]	Conscientiousness (5.9±1.5) Openness (6.1±1.3)
Motivation: MSLQ [37]	Intrinsic Goal Orientation (7.1±1.4) Self Efficacy (6.9±1.4) Extrinsic Goal Orientation (7.8±1.4)
Learning approach: R-SPQ-2F [5]	Deep Learner (5.4±2.9) Shallow Learner (1.3±1.9) Strategic Learner (3.4±2.5)
Self-regulation: MSLQ [37]	Self Regulation (5.9±1.4) Study Effort (5.9±1.8) Study Time & Environment (6.2±2.3)
Learner modality: NLN profiler	Visual (7.2±2.1) Auditory(3.3±2.2) Kinaesthetic(4.5±2.4)
Other factors:	Preference for group work (6.5±3.4) Age (23.27±7.3) Male=713 (59%), Female=494 (41%)

Note: All ranges are 0 to 10 apart from age.

Prior knowledge of the student available to the college at registration, namely age, gender and prior academic performance, was also available to the study. Access to full time college courses in Ireland is based on academic achievement in the Leaving Certificate, a set of state exams at the end of secondary school. College places are offered based on CAO[2] points, an aggregate score of grades achieved in a student's top six leaving certificate subjects, range 0 to 600. Table 4 summarises participant profile by course.

Table 4: Participant profile based on prior knowledge, means and standard deviation

Course Name	n	CAO points	age	%age male	Z*
Computing (IT)	137	232±67	24±8	91%	9
Creative Digital Media	102	305±79	23±7	68%	7
Engineering common	73	220±61	20±3	92%	8
Electronic & computer eng	52	232±53	22±7	92%	3
Mechatronics	27	238±46	21±3	85%	1
Sustainable Electrical & Control Technology	20	199±97	27±7	95%	0
Horticulture	41	273±66	28±11	8%	4
Business General	183	256±57	21±5	54%	10
Business with IT	60	229±75	22±5	60%	6
Business International	64	248±51	21±5	24%	6
Sports Management	95	306±86	23±6	84%	8
Applied Social Care	146	259±84	28±9	32%	10
Early Childcare	80	308±78	22±5	6%	7
Social & Community Development	127	266±78	25±8	29%	9

*Number of study factors differing significantly from a normal distribution ($p<<0.001$).

3. RESULTS

Correlation and regression were used to analyse relationships between study factors and GPA. Subsequent analysis used classification techniques to identify students at risk of failing. Unless otherwise stated, models are based age, gender and non-cognitive factors of learning as listed in Table 3.

All non-cognitive factors of learning failed the Shapiro−Wilk normality test which is common in data relating to education and psychology [26]. However factors of personality were normally distributed within each discipline except for business. Intrinsic motivation and study effort were also normally distributed for engineering and computing students. There were further improvements when analysing subgroups by academic course. Factors of personality, self regulation and intrinsic motivation were normally distributed for all courses. With the exception of approaches to learning, learner modality, preference for group work and GPA, other factors were normally distributed for most courses. Table 4 illustrates the number of attributes that differed significantly from a normal distribution by course. Larger groups were more likely to fail tests of normality.

3.1 Correlations with Academic Performance

Correlations between study factors and GPA were assessed using Pearson's product-moment correlation coefficient (PPMCC). As some attributes violated the assumption of normal distribution, significance was verified with bootstrapped 95% confidence intervals using the bias corrected and accelerated method [7] on 1999 bootstrap iterations.

Bootstrap correlation coefficients are given in Table 5. With the exception of learning modality, all non-cognitive factors were significantly correlated with GPA. The highest correlations with GPA were found for approaches to learning, specifically deep learning approach (r=0.23, bootstrap 95% CI[0.18, 0.29]), and study effort (r=0.19, bootstrap 95% CI [0.13, 0.24]). Age also had a relatively high correlation with GPA (r=0.25, bootstrap 95% CI [0.19, 0.3]). A shallow learning approach (r=-0.15, bootstrap 95% CI[-0.21, -0.09]) and preference for group work (r=-0.076, bootstrap 95% CI [-0.14, -0.02]) were negatively correlated with GPA. Openness had one of the weakest significant correlations with GPA (r=0.08, bootstrap 95% CI [0.03, 0.14]). Correlations were comparable with other studies that included a diverse student population [4, 9, 28] with the exception of self efficacy (r=0.12, bootstrap 95% CI [0.06, 0.17])) which was lower than expected. This may be reflective of the low entry requirements for some courses.

[2]CAO refers to the Central Applications Office with responsibility for processing applications for undergraduate courses in the Higher Education Institutes in Ireland.

3.2 Regression models

Regression models predicting GPA from non-cognitive variables were run for the full dataset and for subgroups by disciplines and by course. The coefficient of determination (R^2) is reported to facilitate comparison with other studies. However R^2 is influenced by the variability of the underlying independent variables. Consequently Achen [1, pg 58-61] argued that prediction error is a more appropriate fitness measure for psychometric data. Therefore absolute error mean and standard deviation is also reported.

A regression model for all participants ($R^2 = 0.14$) was comparable with other reported models of non-cognitive factors [4, 30]. However when modelling students by discipline and by course, there were significant differences in model performance. A chow test [11] comparing the residual error in a regression model of all participants (full model) with the residual errors of models by discipline (restricted models) showed significant differences between the full and restricted models (F(17,1098)=22.02, p=0). There was also significant differences between models based on a particular discipline (full model) and models of courses within that discipline (restricted models). In computing, significant differences of F(17,205)=2.22 (p=0.005) were found between the full model and the two restricted models. Within engineering, a model combining mechatronics with electronic & computing engineering was not significantly different from a model of those two courses individually (F(17,79)=0.58, p=0.89), but including either common entry students and/or sustainable electrical & control technology resulted in significant differences between the full and restricted models. Sustainable electrical & control technology was therefore excluded from further consideration because of the small sample size (n=20). Significant differences were also found in models of each of the three humanities courses compared with those courses combined (F(17,302)=2.22, p=0.004). The least significant differences were found in models of business students provided sport management was excluded (F(17, 307)=1.95, p=0.015). Adding sports management further increased the difference in model residual errors (F(17,334)=8.36, p=0). Table 6 gives model details by course and factors used in each model. Electronic & computer engineering students and mechatronic students were combined.

In general, models based on technical courses had a higher R^2 than models for non technical courses. For example, engineering courses, computing (IT) and business with IT all had $R^2 > 0.3$. Absolute error for these courses was in the range [0.63,0.8]. The difference between the highest absolute error (m=0.8, s=0.56[3]) and the lowest absolute error (m=0.63, s=0.54) was not significant (t(15)=1.74, p=0.1). Regression results for International Business was also relatively good (R^2=0.27). For the remaining non-technical disciplines R^2 was lower (range [0.12,0.17]) but the absolute error was more varied. Early childcare had the lowest absolute error (m=0.37, s=0.34) while general business had the highest absolute error (m=0.9, s=0.53). The difference was significant (t(15)=10.3, p<0.001) and may be explained by the greater distribution of GPA scores in general business.

There was little agreement across models on which study factors were most predictive of GPA. Approaches to learning and age were significant for models of all participants, computing students and engineering students, but motivation and learning strategy were more significant for Business with IT. Factors of motivation, learning strategy and approaches to learning were also relevant to models in the humanities courses. All regression models improved when prior academic performance was included in the model. The most significant increase was for sports management, R^2 increased from 0.16 to 0.30. Business with IT and applied social care also increased by more than 0.1. For all other regression models, R^2 increased by between 0.05 and 0.09

3.3 Classification models

Classification models were generated using four classification algorithms, namely Naïve Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM), and k-Nearest Neighbour (k-NN). A binary class label was used based on end of year GPA score, range [0-4]. The two classes were: high risk students (GPA<2, n=459); and low risk students (GPA≥2.5, n=558) giving a dataset of n=1017. Borderline students (2.0 ≤ GPA ≤ 2.49) have not been considered to date. Gray et al. [24] found that cross validation over-estimated model accuracy compared to models applied to a different student cohort. Therefore models were trained on participants from 2010 and 2011 and tested on participants from 2012. All datasets were balanced by over sampling the minority class, and attributes were scaled to have a mean of 0 and standard deviation of 1. Significant attributes were identified by finding the optimal threshold for selecting attributes by weight. Attributes were weighted based on uncertainty[4] for DT, k-NN and Naïve Bayes models, and based on SVM weights for SVM models. Table 6 shows the accuracies achieved and factors used in each model.

k-NN had the highest accuracy for models of all students (66%). Accuracies for DT (61%), SVM (62%) and Naïve Bayes (62%) were similar. The most significance attributes by weight were age, deep learning approach and study effort. Including factors of prior academic performance improved model accuracy marginally to 72%.

Model accuracy improved when modelling each course separately. In general, k-NN had either the highest accuracy, or close to the highest accuracy, for all groups with the exception of two courses, international business and early childcare & education. Naïve Bayes had the highest accuracy for both those courses and their attributes of significance were normally distributed. Five courses had accuracies marginally higher than the model for all students, social & community development (70%), applied social care (68%), early childcare & education (69%), creative digital media (67%) and sports management (70%). As illustrated in Table 1, these courses were distinguished by a high average GPA and a low failure rate. Consequently, patterns identifying high risk students may be under represented in these groups. Accuracies for other courses were significantly higher (≥ 79%). For example the difference between sports management (70%) and the next highest accuracy (Engineering other, 79%) was significant (Z=5.86, p<0.001)[5].

[3] m=mean, s=standard deviation

[4] Symmetrical uncertainty with respect to the class label.

[5] Accuracy comparisons were based on the mean accuracy of

Table 5: Bootstrap correlations of non-cognitive factors with GPA

Study Factors:	Temperament		Motivation			Learning Approach			Learning Strategy			Other			Modality		
	C	O	SE	IM	EM	De	Sh	St	SR	ST	StE	Group	Age	Gen	V	A	K
Correlation with GPA (n=1207):	0.15 ***	0.08 **	0.12 ***	0.15 ***	0.12 ***	0.23 ***	-0.15 ***	-0.16 ***	0.13 ***	0.1 **	0.19 ***	-0.08 **	0.25 ***	0.09 **	0.06	0.02	0.06

*$p < .05$, **$p < .01$, ***$p < 0.001$; C:Conscientiousness; O:Openness; SE:Self Efficacy; IM:Intrinsic Goal Orientation; EM:Extrinsic Goal Orientation; De:Deep Learner; Sh: Shallow Learner; St: Strategic Learner; SR: Self Regulation; ST:Study Time; StE: Study Effort; Group:Likes to work in groups; Gen=Gender; V:Visual Learner; A:Auditory Learner; K:Kinaesthetic Learner.

Table 6: Regression and classification models by discipline, using non-cognitive factors only

Regression models:				Temperament		Motivation			Approach			Strategy			Other			Modality		
Course	N	Absolute error	R^2	C	O	SE	IM	EM	De	Sh	St	SR	ST	StE	G	age	In	V	A	K
All	1207	0.83±0.56	0.125	+	+	+	+	***	****	***	***	**		***	***	****	*	+		
Computing	137	0.8 ±0.56	0.34	+	+		+			**	+				*	****		*		
Creative Dig Media	103	0.68±0.58	0.11			+		+	****	****	****	+	+	+	+		***			
Eng Common Entry	73	0.67±0.53	0.34		*			+	+	+	+		+			***	***	+		+
Engineering other	99	0.72±0.5	0.43		+	***	+	+	+	**	**	**	*	+	*	****			+	
Horticulture	41	0.63±0.54	0.34	+	+	+	+	+	****	****	****	+	+	+	+	*	****		**	
General Business	183	0.9±0.53	0.13	+		+		+	+	+	+			+	+	+	**			+
Business With IT	60	0.67±0.52	0.48	+			**	**	*			+		***	**	**	**			**
International Business	64	0.78±0.5	0.27		***	+	+		*						+	*	****			+
Sports Management	95	0.64±0.53	0.16	+	+		+			**		+	***							
Applied Social Care	146	0.5±0.5	0.08	+	+	+	+	+	+	*	*	+	+	+	+		****			
Early childcare	80	0.37±0.34	0.17				+	+	+	*	**	+	+	+		+		+		
Social & Comm Dev	127	0.63±0.5	0.12				+	+	+	+				**	+					+
Classification models:				Temperament		Motivation			Approach			Strategy			Other			Modality		
Course	N	Learner	Accuracy	Kappa	C	O	SE	IM	EM	De	Sh	St	SR	ST	StE	G	age	gen	V	A	K
All	1017	11-NN	66%	0.33	✓	✓	✓	✓		✓	✓	✓	✓		✓		✓	✓			
Computing	122	SVM	81%	0.62		✓		✓	✓		✓				✓	✓			✓		✓
Creative Dig Media	94	2-NN	67%	0.35		✓		✓	✓	✓		✓			✓					✓	
Eng Common Entry	73	SVM	94%	0.88	✓	✓				✓	✓	✓			✓		✓		✓		✓
Engineering other	72	DT	79%	0.58						✓					✓		✓				
Horticulture	40	7-NN	86%	0.71	✓	✓	✓	✓	✓		✓	✓			✓					✓	✓
Business General	156	5-NN	85%	0.69										✓	✓						
Business With IT	47	7-NN	83%	0.67		✓	✓	✓			✓	✓	✓		✓		✓				✓
International Business	55	NB	80%	0.6		✓	✓														
Sports Mgmt	72	SVM	70%	0.39	✓				✓		✓		✓	✓							
Applied Social Care	122	4-NN	68%	0.37	✓		✓	✓		✓			✓	✓							✓
Early childcare	58	NB	69%	0.38	✓				✓		✓						✓				
Community dev	93	2-NN	70%	0.39	✓						✓						✓				

Significant model coefficients: +$p > .05$, *$p < .05$, **$p < .01$, ***$p < 0.001$, ****$p << 0.001$; ✓: factors included in the classification model
C:Conscientiousness; O:Openness; SE:Self Efficacy; IM:Intrinsic Goal Orientation; EM:Extrinsic Goal Orientation; De:Deep Learner; Sh: Shallow Learner; St: Strategic Learner; SR: Self Regulation; ST:Study Time; StE: Study Effort; G:Likes to work in groups; IN:Regression model intercept; gen=Gender; V:Visual Learner; A:Auditory Learner; K:Kinaesthetic Learner; Engineering others: Mechatronics and Electrical & Computer Engineering.

It could be argued that the smaller sample size of course groups over estimated model accuracy as smaller samples may under represent the complexity of patterns predictive of academic achievement. Therefore 30 samples randomly generated from the full dataset (n=100) were also modelled. Model accuracy for the random samples was normally distributed, with mean=63.12% (s=11%), which was marginally lower than the model of all students (Z=2.68, p=0.017).

There was little agreement across models on which study factors were most predictive of high risk and low risk students. Conscientiousness, study effort and a shallow learning approach were used most frequently, followed by openness, intrinsic motivation and age. There was no significant improvement in model accuracy when prior academic performance was included in each model. For example, the largest increase in accuracy was from 79% to 82% in a model of Engineering students.

4. CONCLUSIONS

Results from this study suggest that models of academic performance, based on non-cognitive psychometric factors measured during first year student induction, can achieve good predictive accuracy, particularly when individual courses are modelled separately. A deep learning approach, study effort and age had the highest correlations with GPA across all disciplines. These factors were also significant in both the regression model and classification model of all students. Extrinsic motivation, preference for working alone and self regulation were also significant in the regression model, while all factors except extrinsic motivation, preference for working alone and study time were significant in a classification model of all students. Models of individual courses also differed in the range of factors used. The lack of consensus in identification of significant factors may be explained by an overlap in the constructs measured by each [24]. Openness appeared frequently in both classification and regression models despite its relatively low correlation with GPA.

In general, regression models for students in technical disciplines, such as engineering, computing and business with IT, had a higher coefficient of determination (R^2) than models of non technical disciplines. However the coefficient of determination did not reflect prediction error, highlighting the underlying variability in independent variables. For example, early childcare (R^2=0.17) and sports management (R^2=0.16) had the same R^2, but sports management had a higher absolute error (0.64±0.53) than early childcare (0.37 ± 0.34). The difference was significant (t(15)=3.996, p=0.001). Prediction error was reflective of the GPA distribution for each course regardless of discipline.

Classification models that distinguished between high and low risk students based on GPA had good accuracy for both technical and non technical disciplines, particularly for courses with a significant proportion (>30%) of high risk students. As with regression, models of individual courses outper-

100 bootstrap samples from each group.

formed both models of the full dataset and models of random samples taken from the full dataset. This would suggest models trained for specific courses can outperform models generalising patterns for all students. k-NN, a non-linear classification algorithm, gave optimal or near optimal accuracies for most course groups. This may be reflective of non-linear patterns in the dataset.

Including a cognitive factor of prior academic performance did not improve the accuracy of classification models significantly. On the other hand, Gray et al. [23] reported that predictive accuracy of models based on cognitive factors only (prior academic performance) increased marginally when non-cognitive factors were included in the model. This would suggest a high overlap in constructs captured by both cognitive and non-cognitive factors of learning.

Model accuracies are based on a heuristic search of attribute subsets. A more exhaustive search is needed to verify optimal attribute subsets. Further work is also required to investigate principal components amongst non-cognitive factors. In addition, results are based on full time students in a traditional classroom setting at one college. Further work is needed to determine if these results generalise to students in other colleges, and other delivery modes.

5. ACKNOWLEDGMENTS

The authors would like to thank Institute of Technology Blanchardstown for their support in facilitating this research, and staff at the National Learning Network for assistance administering questionnaires during student induction.

6. REFERENCES

[1] Achen, C. *Intrepreting and Using Regression*. Number 07-029 in Quantitative Applications in the Social Sciences. Sage Publications, Inc, 1982.

[2] Baker, R. S. J. D. and Yacef, K. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1):3–17, 2010.

[3] Bergin, S. *Statistical and machine learning models to predict programming performance*. PhD thesis, Computer Science, NUI Maynooth, 2006.

[4] Bidjerano, T. and Dai, D. Y. The relationship between the big-five model of personality and self-regulated learning strategies. *Learning and Individual Differences*, 17:69 – 81, 2007.

[5] Biggs, J., Kember, D., and Leung, D. The revised two-factor study process questionnaire: R-SPQ-2F. *British Journal of Education Psychology*, 71:133–149, 2001.

[6] Buckingham Shum, S. and Deakin Crick, R. Learning dispositions and transferable competencies. pedagogy, modelling and learning analytics. In *2nd International Conference on Learning Analytics and Knowledge*, pages 92–101, Vancouver, BC, Canada, 2012.

[7] Carpenter, J. and Bithell, J. Bootstrap confidence intervals - when, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, 19:1141–1164, 2000.

[8] Cassidy, S. Exploring individual differences as determining factors in student academic achievement in higher education. *Studies in Higher Education*, 37(7):1–18, 2011.

[9] Chamorro-Premuzic, T. and Furnham, A. Personality, intelligence and approaches to learning as predictors of academic performance. *Personality and Individual Differences*, 44:1596–1603, 2008.

[10] Chatti, M. A., Dychhoff, A. L., Schroeder, U., and Thüs, H. A reference model for learning analytics. *International Journal of Technology Enhanced Learning. Special Issue on State of the Art in TEL*, pages 318–331, 2012.

[11] Chow, G. C. Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, 28(3):591–605, 1960.

[12] Cooper, A. J., Smillie, L. D., and Corr, P. J. A confirmatory factor analysis of the mini-IPIP five-factor model personality scale. *Personality and Individual Differences*, 48(5):688–691, 2010.

[13] Covington, M. V. Goal theory, motivation, and school achievement: An integrative review. *Annual Review of Psychology*, 51:171–200, 2000.

[14] De Feyter, T., Caers, R., Vigna, C., and Berings, D. Unraveling the impact of the big five personality traits on academic performance. The moderating and mediating effects of self-efficacy and academic motivation. *Learning and Individual Differences*, 22:439–448, 2012.

[15] Dekker, G., Pechenizkiy, M., and Vleeshouwers, J. Predicting students drop out: a case study. In Barnes, T., Desmarais, M. C., Romero, C., and Ventura, S., editors, *Proceedings of the 2nd International Conference on Educational Data Mining*, pages 41–50, Cordoba, Spain, 2009.

[16] Drachsler, H. and Greller, W. The pulse of learning analytics. Understandings and expectations from the stakeholders. In *2nd International Conference on Learning Analytics and Knowledge*, pages 120–129, Vancouver, BC, Canada, 29 April- 2 May 2012. ACM.

[17] Duff, A., Boyle, E., Dunleavy, K., and Ferguson, J. The relationship between personality, approach to learning and academic performance. *Personality and Individual Differences*, 36:1907–1920, 2004.

[18] Entwhistle, N. Contrasting perspectives in learning. In Marton, F., Hounsell, D., and Entwhistle, N., editors, *The Experience of Learning*, pages 3–22. Edinburgh: University of Edinburgh, Centre for Teaching, Learning and Assessment, 2005.

[19] Eppler, M. A. and Harju, B. L. Achievement motivation goals in relation to academic performance in traditional and nontraditional college students. *Research in Higher Education*, 38 (5):557–573, 1997.

[20] Farsides, T. and Woodfield, R. Individual differences and undergraduate academic success: The roles of personality, intelligence, and application. *Personality and Individual Differences*, 34:1225–1243, 2003.

[21] Fleming, N. D. I'm different, not dumb. Modes of presentation (VARK) in the tertiary classroom. *Research and Development in Higher Education, Proceedings of the 1995 Annual Conference of the Higher Education and Research Development Society of Australasia*, 18:308–313, 1995.

[22] Goldberg, L. R. The development of markers for the

big-five factor structure. *Psychological Assessment*, 4 (1):26–42, 1992.

[23] Gray, G., McGuinness, C., and Owende, P. An investigation of psychometric measures for modelling academic performance in tertiary education. In D'Mello, S. K., Calvo, R. A., and Olney, A., editors, *Sixth International Conference on Educational Data Mining*, pages 240–243, Memphis, Tennessee, July 6-9 2013.

[24] Gray, G., McGuinness, C., and Owende, P. An application of classification models to predict learner progression in tertiary education. *4th IEEE International Advanced Computing Conference*, pages 549–554, February 2014.

[25] Gray, G., McGuinness, C., Owende, P., and Carthy, A. A review of psychometric data analysis and applications in modelling of academic achievement in tertiary education. *Journal of Learning Analytics*, 1(1):75–106, 2014.

[26] Kang, Y. and Harring, J. R. Reexamining the impact of non-normality in two-group comparison procedures. *Journal of Experimental Education*, in press.

[27] Kappe, R. and van der Flier, H. Using multiple and specific criteria to assess the predictive validity of the big five personality factors on academic performance. *Journal of Research in Personality*, 44:142–145, 2010.

[28] Kaufman, J. C., Agars, M. D., and Lopez-Wagner, M. C. The role of personality and motivation in predicting early college academic success in non-traditional students at a hispanic-serving institution. *Learning and Individual Differences*, 18:492 – 496, 2008.

[29] Knight, S., Buckingham Shum, S., and Littleton, K. Epistemology, pedagogy, assessment and learning analytics. In *Third Conference on Learning Analytics and Knowledge (LAK 2013)*, pages 75–84, Leuven, Belgium, April 2013.

[30] Komarraju, M., Karau, S. J., Schmeck, R. R., and Avdic, A. The big five personality traits, learning styles, and academic achievement. *Personality and Individual Differences*, 51:472–477, 2011.

[31] Komarraju, M. and Nadler, D. Self-efficacy and academic achievement. Why do implicit beliefs, goals, and effort regulation matter? *Learning and Individual Differences*, 25:67–72, 2013.

[32] Marton, F. and Säljö, R. Approaches to learning. In Marton, F., Hounsell, D., and Entwhistle, N., editors, *The Experience of Learning*, pages 36–58. Edinburgh: University of Edinburgh, Centre for Teaching, Learning and Assessment, 2005.

[33] Mislevy, R. J., Behrens, J. T., and Dicerbo, K. E. Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *Journal of Educational Data Mining*, 4 (1):11–48, 2012.

[34] Moran, M. A. and Crowley, M. J. The leaving certificate and first year university performance. *Journal of Statistical and Social Enquiry in Ireland*, XXIV, part 1:231–266, 1979.

[35] Ning, H. K. and Downing, K. The reciprocal relationship between motivation and self-regulation: A longitudinal study on academic performance. *Learning and Individual Differences*, 20:682–686, 2010.

[36] Pardos, Z. A., Baker, R. S. J. D., San Pedro, M. O. C. A., Gowda, S. M., and Gowda, S. M. Affective states and state test. Investigating how affect throughout the school year predicts end of year learning. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge (LAK '13)*, pages 117–124, Leuven, Belgium, April 2013. ACM.

[37] Pintrich, P., Smith, D., Garcia, T., and McKeachie, W. A manual for the use of the motivated strategies for learning questionnaire. Technical Report 91-B-004, The Regents of the University of Michigan, 1991.

[38] Poropat, A. E. A meta-analysis of the five-factor model or personality and academic performance. *Psychological Bulletin*, 135(2):322–338, 2009.

[39] Robbins, S. B., Lauver, K., Le, H., Davis, D., and Langley, R. Do psychosocial and study skill factors predict college outcomes? A meta analysis. *Psychological Bulletin*, 130 (2):261–288, 2004.

[40] Sachin, B. R. and Vijay, S. M. A survey and future vision of data mining in educational field. In *Advanced Computing Communication Technologies (ACCT), 2012 Second International Conference on*, pages 96–100, Jan 2012.

[41] Shute, V. and Ventura, M. *Stealth Assessment. Measuring and Supporting Learning in Video Games*. The John D. and Catherine T. MacArthur Foundation Reports on Digital Media and Learning. MIT Press, 2013.

[42] Siemens, G. Learning analytics. Envisioning a research discipline and a domain of practice. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pages 4–8, 2012.

[43] Siemens, G. and Baker, R. S. J. D. Learning analytics and educational data mining. Towards communication and collaboration. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pages 252–254, 2012.

[44] Swanberg, A. B. and Martinsen, Ø. L. Personality, approaches to learning and achievement. *Educational Psychology*, 30(1):75–88, 2010.

[45] Tavakol, M. and Dennick, R. Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2:53–55, 2011.

[46] Tempelaar, D. T., Cuypers, H., van de Vrie, E., Heck, A., and van der Kooij, H. Formative assessment and learning analytics. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge (LAK '13)*, pages 205–209, New York, NY, USA, 2013. ACM.

[47] Tishman, S., Jay, E., and Perkins, D. N. Teaching thinking disposition: From transmission to enculturation. *Theory into Practice*, 32:147–153, 1993.

[48] Vancouver, J. B. and Kendall, L. N. When self-efficacy negatively relates to motivation and performance in a learning context. *Journal of Applied Psychology*, 91(5):1146–53, 2006.

[49] Volet, S. E. Cognitive and affective variables in academic learning: the significance of direction and effort in students' goals. *Learning and Instruction*, 7(3):235–254, 1996.