Preface

Emotions and affect play an important role in learning. There are indications that meta-affect (i.e., knowledge about self-affect) also plays a role. There have been various attempts to take them into account both during the design and during the deployment of AIED systems. The evidence for the consequential impact on learning is beginning to strengthen, but the field has been mostly focused on addressing the complexities of affective and emotional recognition and very little on how to intervene. This has largely slowed down progress in this area.

Research is needed to better understand how to respond to what we detect and how to relate that to the learner's cognitive and meta-cognitive skills. One goal might be to design systems capable of recognizing, acknowledging, and responding to learners' states with the aim of promoting those that are conducive to learning by means of tutorial tactics, feedback interventions, and interface adaptations that take advantage of ambient intelligence, among others. Therefore, we need to deepen our knowledge of how changes in learners' affective states and associated emotions relate to issues such as cognition and the learning context.

The papers submitted to the workshop address issues that bridge the existing gap between previous research with the ever-increasing understanding and data availability. In particular, these papers report progress on issues relevant to the broad and interdisciplinary AIED and EDM communities. AMADL 2015 workshop raises the opportunity to bring these two communities together in a lively discussion about the overlap in the two fields. To achieve this, we explicitly address and target both com- munities, as indicated by the workshop's organizers background and the programme committee set up. This workshop builds on the work done in affect related workshops in past AIED conferences, such as Modelling and Scaffolding Affective Experiences to Impact Learning in AIED 2007. The format of the workshop is based on presentations, demonstrations and discussions according to themes addressed by the papers accepted for the workshop.

Genaro Rebolledo-Mendez, Manolis Mavrikis, Olga C. Santos, Benedict du Boulay, Beate Grawemeyer and Rafael Rojano-Cáceres

Workshop Co-Chairs

Cultural aspects related to motivation to learn in a Mexican context

Erika-Annabel Martínez-Mirón[1,*], Genaro Rebolledo-Méndez[2]
[1]Universidad Politécnica de Puebla, Puebla, México
*Corresponding Author: erika.martinez@uppuebla.edu.mx
[2]Universidad Veracruzana, Xalapa, México
g.rebolledo@gmail.com

Abstract. The development of motivationally intelligent tutoring systems has been based on a variety of motivational models from the psychology field. These models mainly consider characteristics from de areas of values, expectancies and feelings [1]. However, this paper proposes to take into account some cultural aspects when operationalizing such models. The basis of this proposal is presented from the perspective of some cultural aspects that effect career choice, in particular for a Mexican context.

Keywords: Motivation, career choice, Mexican cultural context

1 Introduction

Research in motivation to learn when using educational technology has operationalized different motivational models found in the psychological literature in order to develop motivationally intelligent tutoring systems. According to these models, motivationally aware tutoring systems should combine expertise and knowledge about user's cognitive, affective, meta-cognitive and meta-affective levels in order to appropriately react and be able to favor user's learning [2, 3]. That is, these models should mainly consider characteristics from the areas of values, expectancies and feelings [1].

However, this paper argues also for the inclusion of other aspects that have been seldom taken into account so far. We refer to cultural aspects inherent to each group of individuals from a certain background. Since there is evidence that students from different cultural origin react to the same motivational strategy in a different way [4, 5, 6] or have different attitudes for online assessment [7], the cultural aspect of learning with technology becomes an important issue. For instance, if a female student from a highly gender-stereotyped cultural background is asked to attend a course considered to be strongly oriented to men, then she might perceived to be in the wrong course and probably will not exert her maximum effort. Or even she might believe that her role in society is to be protected by someone, and she attends courses just to be in the possibility to meet that expectation. It will not matter what motivational strategy the teacher uses, since the female student's cultural belief is in an apparently superior level and she will only be concerned to learn at the minimum, just to continue studying until meeting her protector [8].

In order to develop the arguments to support the inclusion of cultural aspects in the design of motivationally-aware tutoring systems, the following sections describe some of these elements within a Mexican context from the perspective of career choice, based on the findings that instrumental motivation is an important predictor for course selection, career choice, and performance [9, 10]. That is, students may pursue to perform well in some tasks because they are important for future goals, even if the student is not interested on the task.

2 Motivation, career guidance and cultural context

Motivation is related to the student's desire to participate in the learning process. Current research findings suggest that motivational constructs do change over time [11, 12, 13] and/or contexts [14, 15, 16]. In particular, it is well documented that cultural differences affect achievement motivation [4, 5, 6].

We believe that if teachers truly want to promote the success of all students, they must recognize how achievement motivation varies culturally within the population it serves.
Similarly, career counseling must incorporate different variables and different processes to be effective for students from different cultural contexts. Career counseling is defined as "the process of assisting individuals in the development of a life-career with focus on the definition of the worker role and how that role interacts with other life roles" [17].
According to Rivera [18], there are characteristics that prevail among Hispanic/Latino American children and adolescents, such as: A) Restraint of feelings, particularly anger and frustration; B) Limited verbal expressions toward authority figures; C) Preference for closer personal space; avoidance of eye contact when listening or speaking to authority figures; D) Relaxation about time and punctuality; and immediate short-term goals; E) Collective, group identity; interdependence; cooperative rather than competitive; emphasis on interpersonal relations. To certain extent, these characteristics can be considered part of one of the four sources of information, social persuasion, included in the model of the Socio Cognitive Career Theory [19], (see Table 1). This framework conceptualizes career choice as a process with multiple stages and different sources of information. We propose that cultural aspects of the Mexican context might have an impact not just the process of choosing a career, but on the way students undertake their learning activities as described in the following paragraphs.

Table 1. Sources of information proposed in the model of social cognitive influences on career choice behavior [19]

Source of information	**Description**
Performance accomplishment	Success in performing the target task or behavior
Vicarious learning or modeling	To watch others who could perform the target behavior successfully.
Emotional arousal	Anxiety when performing the target behavior
Social persuasion	Support and encouragement from others in the process of performing the target behavior.

2.1 Machismo

There is growing research supporting that achievement differences between genders are smaller during early years of school or being reduced [20]. The succession of career behaviors for women is far more complex than for men. In particular, in Mexican students, the complexities might lay in the cultural aspect of machismo. In Mendoza's review [21], machismo is defined as a strong sense of masculine pride, and it is suggested that machismo should be considered in any Latino study, but it is often forgotten. The social behavior pattern associated to machismo includes the expectation of men being caring, responsible, decisive, strong of character, and the protector of probably extended family. At the same time, negative aspects of machismo include aggressiveness, physical strength, emotional insensitivity, and a womanizing attitude towards the opposite sex.
Galanti [22], cited in [21], surveyed a group of Latino students who reported that the relationship between male and female would be of protector and protected. More specifically, according to them, the role of the traditional Hispanic woman is to look after the family; her job is to cook, clean, and care for the children. Other characteristics of a good wife include submission and obedience to her husband's orders without questioning him but rather standing behind whatever he decides, even if she disagrees. She must also be tolerant of his behavior. Taking into account these views it is understandable that women´s career choice might be influenced by the fulfillment of this profile rather than freely choosing a career that may imply a great amount of dedication. In some Mexican contexts, women may prefer to undertake studies that are less demanding. Women also must strive to overcome obstacles such as gender discrimination and sex stereotyping. For instance, Gallardo-Hernández *et. al.* reported the results of a questionnaire applied to 637 first-year medical nutrition, dentistry and nursing students

[23]. The findings suggest that among women of low socioeconomic strata, more traditional gender stereotypes prevail which lead them to seek career choices considered feminine. Among men, there is a clear relationship between career choice, socioeconomic level and internalization of gender stereotypes.

2.2 Social orientation

Cooperative learning is very important for Mexicans [24]. They do not seem to openly want to show what they know for fear of embarrassing those who do not know [25]. It is not common in a Hispanic family to encourage children to excel over siblings or peers but rather, it is considered bad manners. It is worth noting that most of the studies reported have taken into account the Mexican context around Mexican American students but no studies so far focus on comparison between this population and a Mexican population living in Mexico. Nevertheless, their findings can, to some extent, be considered valid for Mexican population. For instance, Ojeda and Flores [26] considered the educational aspirations of 186 Mexican American high school students to test a portion of social-cognitive career theory [19]. Their results indicated that perceived educational barriers significantly predicted students' educational aspirations above and beyond the influence of gender, generation level, and parents' education level. Similarly, Flores, Romero and Arbona [27] found that Mexican American men and women with high measures of ethnic loyalty might be at risk for perceiving social costs of pursuing a higher education.

2.3 Perception of time and career guidance

Mexicans are oriented toward present time; they are focused on “right now” rather than on the past or on future events or outcomes. They often live the phrase “Dios dirá” or "God will tell," that is, time is relative. To arrive late for an engagement is called in the southwest "Mexican time." This perception permeates career-counseling programs in the Mexican context, since its interventions start in the educational level just behind the university program [28]. Therefore, students have to decide in a relatively short period of time which career suits them best. Sometimes the students might have a great amount of career information, making it difficult to make a good analysis of each of the options. But it also might occur that there is little availability of information and students might end up making an inadequate career choice.

3 Discussion

Increasingly, researchers are calling for studies of change in motivation, rather than treating motivation as a static trait-like factor [1], [4]. However, those studies mainly consider motivation to be influenced by characteristics from the areas of values, expectancies and feelings [1], without taking into account that some cultural aspects like machismo, social orientation or perception of time might also be influencing how students approach to a learning activity. For instance, women could be avoiding pursuing a career that would not allow them to easily integrate their expected roles as mother and spouse with their future professional activity. Also, the perception of educational barriers, such as gender and ethnicity, nurtured by the social context could reinforce the idea of choosing a career according to the students’ sex, which in turn might influence students’ motivation to learn a particular area of study. Although there is little research evidence that establishes a direct connection between career choice and motivation to learn a particular topic, this paper reviewed some cultural aspects in the Mexican context that have an impact on students’ learning behavior. Based on this, we consider plausible to do some research that consider these aspects when designing a motivationally tutoring system. For example, in a Mexican context, a tutoring system for Mathematics could emphasize women’s capacity to solve problems regardless of their gender, like providing feedback including mentions to important contributions from female scientists, or listing the advantages of achieving personal professional success as a woman, or maybe using a very strong female character showing high IQ as the main avatar.

4 REFERENCES

1. du Boulay, B. Towards a Motivationally-Intelligent Pedagogy: How should an intelligent tutor respond to the unmotivated or the demotivated? In R. A. Calvo & S. D'Mello (Eds.), New Perspectives on Affect and Learning Technologies (pp. 41-54). New York: Springer (2011)
2. Avramides, K. and du Boulay, B. Motivational Diagnosis in ITSs: Collaborative, Reflective Self-Report. In V. Dimitrova, R. Mizoguchi, B. du Boulay & A. Graesser (Eds.), Artificial Intelligence in Education. Building Learning Systems that Care: from Knowledge Representation to Affective Modelling AIED2009 14th International Conference on Artificial Intelligence in Education (Frontiers in Artificial Intelligence and Applications No. 200 pp. 587-589). Amsterdam: IOS Press (2009)
3. du Boulay, B., Rebolledo Mendez, G., Luckin, R. & Martinez Miron, E. (2007). Motivationally Intelligent Systems: Diagnosis and Feedback. In R. Luckin, K. Koedinger & J. Greer (Eds.), Artificial Intelligence in Education: Building Technology Rich Learning Contexts. Proceedings of AIED2007, Los Angeles (Frontiers in Artificial Intelligence and Applications No. 158 pp. 563-565). Amsterdam: IOS (2007)
4. Henderlong, J., and Lepper, M. R. The effects of praise on children's intrinsic motivation: A review and synthesis. *Psychological Bulletin, 128,* 774–795 (2002)
5. Kaplan, A., Karabenick, S., & De Groot, E. Introduction: Culture, self, and motiva- tion: The contribution of Martin L. Maehr to the fields of achievement motivation and educational psychology. In A. Kaplan, S. Karabenick, & E. De Groot (Eds.), *Culture, self, and motivation: Essays in honor of Martin L. Maehr* (pp. vii-xxi). Charlotte, NC: Informa- tion Age Publishing (2009)
6. Trumbull, Elise; Rothstein-Fisch, The Intersection of Culture and Achievement Motivation. Carrie School Community Journal, v21 n2 p25-53 (2011)
7. Terzis V., Moridis C., Economides A.A., Rebolledo-Mendez G. Computer Based Assessment Acceptance: A Cross-Cultural Study in Greece and Mexico. *Journal of Educational Technology and Society. 16(3),* 411-424 (2013)
8. Schmitz, K. and Diefentahler, S. An examination of traditional gender roles among men and women in Mexico and the United States. Retrieved, Vol. 12, p. 2008. (1998)
9. Wigfield, A., and Eccles, J.S. (1992) The development of achievement task values: A theoretical analysis. Developmental Review 12: 265 – 310.
10. Wigfield, A., Eccles, J.S., and Rodriguez, D. (1998) The development of children's motivation in school context. Review of Research in Education 23: 73-118.
11. Bong, M., and Skaalvik, E. M. Academic Self-Concept and Self-Efficacy: How Different Are They Really?. Educational Psychology Review, Vol. 15, No. 1, 1-40 (2003)
12. Chouinard, R. and Roy, N. Changes in high-school students' competence beliefs, utility value and achievement goals in mathematics. British Journal of Educational Psychology, Vol. 78, No. 1, 31-50 (2008)
13. Corpus, J., McClintic-Gilbert, M. S., and Hayenga, A. O. Within-year changes in children's intrinsic and extrinsic motivational orientations: Contextual predictors and academic outcomes. Contemporary Educational Psychology, Vol. 34, No. 2, 154-166 (2009)
14. Otis, N., Grouzet, F. E. and Pelletier, L. G. Latent Motivational Change in an Academic Setting: A 3-Year Longitudinal Study. Journal Of Educational Psychology, Vol. 97, No. 2, 170-183 (2005)
15. Caprara, G. V., Fida, R., Vecchione, M., Del Bove, G., Vecchio, G. M., Babaranelli, C. and Bandura, A. Longitudinal analysis of the role of perceived self-efficacy for self-regulated learning in academic continuance and achievement. Journal of Educational Psychology, Vol. 100, 525-534 (2008)
16. Wang, Q. and Pomerantz, E. The motivational landscape of early adolescence in the United States and China: a longitudinal investigation. Child Development, Vol. 80, No. 4, 1272-1287 (2009)
17. National Career Development Association. http://www.ncda.org/aws/NCDA/pt/sd/news_article/37798/_self/layout_ccmsearch/true
18. Rivera, B. D., and Rogers-Adkinson. D. Culturally sensitive interventions: Social skills training with children and parents from culturally and linguistically diverse backgrounds. Intervention in School and Clinic. 33(2), 75-80 (1997)
19. Lent, R. W., Brown, S. D., and Hackett, G. Toward a Unifying Social Cognitive Theory of Career and Academic Interest, Choice, and Performance, Journal of Vocational Behavior, 45, p. 93 (1994)
20. Hyde, J., Lindberg, S., Linn, M., Ellis, A., & Williams, C. Gender similarities characterize math performance. Science, 321(5888), 494 – 495 (2008)
21. Mendoza, E. Machismo Literature Review. Center for Public Safety Initiatives. Rochester Institute of Technology (2009)

22. Galanti, G. The Hispanic Family and Male-Female Relationships: An overview. Journal of Transcultural Nursing, 14(3), 180-185 (2003)
23. Gallardo-Hernández, G., Ortiz-Hernández, L., Compeán-Dardón, S., Verde-Flota, E., Delgado-Sánchez, G., Tamez-González, S. Intersection between gender and socioeconomic status in medical sciences career choice. Gaceta Médica Mex. 2006 Nov-Dec; 142(6):467-76 (2006)
24. Gorodnichenko, Y. and Roland, G. Culture, Institutions and the Wealth of Nations, CEPR Discussion Paper No 8013 (2010). http://eml.berkeley.edu/~ygorodni/gorrol_culture.pdf
25. Losey, K. M. Mexican American students and classroom interaction: An overview and critique. Review of Educational Research, 65, 283-318 (1995)
26. Ojeda, L. and Flores, L. The Influence of Gender, Generation Level, Parents' Education Level, and Perceived Barriers on the Educational Aspirations of Mexican American High School Students. The Career Development Quarterly 57, 1, 84–95. (2008)
27. Flores, Y.N., Romero A., and Arbona, C. Effects of Cultural Orientation on the Perception of Conflict Between Relationship and Educational Goals for Mexican American College Students. Hispanic Journal of Behavioral Sciences 22, 1: 46-63 (2000)
28. POV (2011). Programa de Orientación Vocacional para el bachillerato general, tecnológico y profesional técnico. www.dgb.sep.gob.mx/04-m2/.../Programa_Orientacion_Vocacional.pdf

The potential of Ambient Intelligence to deliver Interactive Context-Aware Affective Educational support through Recommendations

Olga C. Santos[1], Mar Saneiro[1], M. Cristina Rodriguez-Sanchez[2], Jesus G. Boticario[1], Raul Uria-Rivas[1], Sergio Salmeron-Majadas[1]

[1] aDeNu Research Group. Artificial Intelligence Dept. Computer Science School, UNED.
Calle Juan del Rosal, 16. Madrid 28040. Spain
http://adenu.ia.uned.es
{ocsantos,marsaneiro,jgb,raul.uria,sergio.salmeron}@dia.uned.es
[2] Electronics Department, Universidad Rey Juan Carlos.
Calle Tulipán s/n. Móstoles 28933 (Madrid), Spain.
cristina.rodriguez.sanchez@urjc.es

Abstract. There is a challenge and opportunity to research if the ambient intelligent support that can be deployed with a recommender system extended with an open hardware infrastructure that can sense and react within the learners' context is of value to supports learners' affectively. In this paper, we summarize the status of our research on eliciting an interactive recommendation for a stressful scenario (i.e., oral examination of a foreign language) that can be delivered through the Ambient Intelligence Context-aware Affective Recommender Platform (AICARP), which is the infrastructure we have designed and implemented with Arduino, an open-source electronic prototyping platform.

1 Eliciting Interactive Recommendations with TORMES

We have reported elsewhere [1] our progress on analyzing the potential of Ambient Intelligence to deliver more interactive educationally oriented recommendations that can deal with the affective state of the learner. In particular, following the TORMES methodology [2], we elicited an educational **scenario** focused on helping the learner when preparing for the oral examination in a second language learning course, which is widely considered as a stressful situation.

The **recommendation** identified in this scenario consists in suggesting the learner to breathe slowly (at a rate of 4 breaths/minute) and is aimed to calm her down when she is nervous. The *applicability conditions* that trigger the recommendation take into account physiological (i.e., heart rate, pulse, skin temperature, skin conductance) and behavioral (facial/body movements and speech speed) information that show evidence of restlessness. The recommendation *output* has been coded in a multisensory way by simultaneously modulating light, sound and vibration behavior at aforementioned breath rate, so the learner can perceive the recommended action through alternative sensory channels (i.e., sight, hearing and touch) without interrupting her activity.

2 Delivering Interactive Recommendations with AICARP

To deliver the aforementioned recommendation elicited with TORMES, the Ambient Intelligence Context-aware Affective Recommender Platform (AICARP) is being implemented with open source software and open hardware following a modular design controlled by an Arduino board (see [1] for details). In the current version, AICARP receives information from physiological **sensors** regarding changes in the learner affective state through corresponding physiological signals. The sensors are integrated into the e-Health platform [3] and a wireless electrocardiogram system [4]. Taking into account this information, AICARP is able to provide the elicited interactive recommendation to the learner by modulating the output of alternative sensorial **actuators** with the recommended breath rhythm. In particular, the following actuators have already been integrated into AICARP: i) white and red flashlights, ii) an array of blue LEDs, iii) a buzzer that vibrates and sounds, and iv) a speaker reproducing a pure tone at 440 Hz (i.e., "La" musical note).

To get some insight on the users' perception on the recommendation delivery, we have deployed the educational scenario outlined in Section 1 in order to deliver the corresponding recommendation elicited with TORMES. So far, in this context we have carried out **2 pilot studies**, one with 6 university students with various interaction needs -including a blind participant-, and another with 4 participants within the 2014 Madrid Science Week. Since we wanted to test the potential of this approach in detecting not only the physiological information but also the behavioral information, we used the Wizard of Oz method [5]. In this way, the recommendation was triggered by the wizard (in our case, a psycho-educational expert) considering participants' information on both physiological evidences detected with AICARP, as well as body/facial movements and speech speed that the wizard observed while the participants carried out the two tasks defined in the pilots (i.e., talking aloud in English about two specific given topics selected from those usually considered in oral exams).

3 Evaluation Outcomes and Open Issues identified

We evaluated AICARP in the 2 pilot studies with the analysis of the participants' responses to the System Usability Scale [6] and to a post-study consisting in a semi structured interview led by the psycho-educational expert. This **evaluation** showed that the implemented infrastructure can actually sense the physiological state of the learner (which seems to be related to some affective state) and deliver ambient intelligent interactive feedback aimed to transform a negative affective (i.e., nervousness) state into a positive one (i.e., relaxation) (see [1] for details on the evaluation results). To the latter, actuators considered aim to provide a natural interaction support not interfering with the participant's task, and consisted of visual, audio and/or tactile feedback.

As discussed in [1], the analysis of the evaluation outcomes has identified several **open issues** to be addressed in future research, as follows:

1. **How to deliver interactive recommendations:** this issue deals with selecting the preferred sensory channels from those available, the format to display the recommendation, the support to understand the purpose of the recommendation and the intrusion level.
2. **When recommendations are to be provided:** in terms of physiological and behavioral changes, while interfering as less as possible with the task. Here, and following TORMES methodology, data mining techniques can be explored to automatically identify the criteria that characterize the appropriate moment to deliver the recommendation [7].
3. **Learners' features of potential relevance in order to design other recommendations:** such as domain dependent attributes (i.e., the English level) and personality traits.
4. **Social aspects involved when collaboration takes place:** in the current scenario, collaboration can occur when learners are asked to perform the oral examination in pairs by dialoging a given situation. The training can be done using a videoconferencing system. In this context, other issues should be considered, such as the intensity of collaboration, the type of collaborative task, the individual acceptance of the technology used to support the collaboration, as well as specific personality traits.

Acknowledgements

This work was supported by the Spanish Ministry of Economy and Competitiveness (MINECO) under Grant TIN2011-29221-C03-01 (MAMIPEC project).

References

1. Santos, O.C., Saneiro, M., Rodriguez-Sanchez, M.C. and Boticario, J.G. (2015) Towards Interactive Context-Aware Affective Educational Recommendations in Computer Assisted Language Learning. New Review of Hypermedia and Multimedia, in press.
2. Santos, O.C., Boticario, J.G. (2015) Practical guidelines for designing and evaluating educationally oriented recommendations. In Computers and Education, vol. 81, 354–374.
3. Cooking Hacks. E-Health Platform. Available from: http://www.cooking-hacks.com.
4. Torrado-Carvajal, A., Rodriguez-Sanchez, M.C., Rodriguez-Moreno, A., Borromeo, S., Garro-Gomez, C., Hernandez-Tamames, J. A., and Luaces, M. (2012) Changing communications within hospital and home health care. In 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 6074-6077.
5. Dahlbäck, N., Jönsson, A., and Ahrenberg, L. (1993) Wizard of Oz studies: why and how. In Proceedings of Intelligent User Interfaces, 193–200.
6. Brooke, J. (1996) SUS: a 'quick and dirty' usability scale. In Jordan, P.W., Thomas, B., Weerdmeester, B.A. and McClelland, A.L. Usability Evaluation in Industry. London: Taylor and Francis.
7. Salmeron-Majadas, S., Arevalillo-Herráez, M., Santos, O.C., Saneiro, M., Cabestrero, R., Quirós, P., Arnau, D. and Boticario, J.G. (2015) Filtering of Spontaneous and Low Intensity Emotions in Educational Contexts. 17th Int. Conf. on Artificial Intelligence in Education (AIED 2015). Lecture Notes in Artificial Intelligence, vol. 9112, 429-438.

The impact of feedback on students' affective states

Beate Grawemeyer[1], Manolis Mavrikis[2], Wayne Holmes[2], Alice Hansen[2], Katharina Loibl[3], and Sergio Gutiérrez-Santos[1]

[1] London Knowledge Lab, Dep of Computer Science and Information Systems, Birkbeck, London, UK
beate@dcs.bbk.ac.uk, sergut@dcs.bbk.ac.uk
[2] London Knowledge Lab, UCL Institute of Education, University College London, London, UK
m.mavrikis@ioe.ac.uk, w.holmes@ioe.ac.uk, a.hansen@ioe.ac.uk
[3] Institute of Educational Research, Ruhr-Universität Bochum, Germany
katharina.loibl@rub.de

Abstract. Affective states play a significant role in students' learning behaviour. Positive affective states can enhance learning, while negative affective states can inhibit it. This paper describes a Wizard-of-Oz study that investigates the impact of different types of feedback on students' affective states. Our results indicate the importance of providing feedback matched carefully to the affective state of the students in order to help them transition into more positive states. For example when students were confused affect boosts and specific instructive feedback seem to be effective in helping students to be in flow again. We discuss this and other ways to adapt the feedback, together with implications for the development of our system and the field in general.

1 Introduction

This paper reports the results of a set of two Wizard-of-Oz studies which explore the effect of different feedback types on students' affective states.

It is well understood by now that affect interacts with and influences the learning process [9, 6, 2]. While positive affective states such as surprise, satisfaction or curiosity contribute towards constructive learning, negative ones including frustration or disillusionment at realising misconceptions can lead to challenges in learning. The learning process is indeed full of transitions between positive and negative affective states and regulating those is important. For example, a student may seem interested in exploring a particular learning goal, however s/he might have some misconceptions and need to reconsider her/his knowledge. This can evoke frustration and/or disappointment. However, this negative affective state may turn into deep engagement with the task again. D'Mello et al., for example, elaborate on how confusion is likely to promote learning under appropriate conditions [6].

It is important therefore, to deepen our understanding of the role of affective states for learning, and to be able to move students out of states that inhibit learning. Pekrun [13] discusses achievement emotions or affective states, which arise in a learning situation. Achievement emotions are states that are linked to learning, instruction, and achievement. We focus on a subset of affective states identified by Pekrun: flow/enjoyment, surprise, frustration, and boredom. We also add confusion, which has been identified elsewhere as an important affective state during learning [15] for tutor support and for learning in general [6].

As described in Woolf et al. [20] students can become overwhelmed (very confused or frustrated) during learning, which may increase cognitive load [19] for low-ability or novice students. However, appropriate feedback might help to overcome such problems. Carenini et al. [3] describe how effective support or feedback needs to answer three main questions: (i) when the support should be provided during learning; (ii) what the support should contain; and (iii) how it should be presented.

In this paper we focus on the question of *what* the support should contain with respect to affect i.e. the types of feedback that are able to induce a positive affective state.

In related work students' affective states have been used to tailor motivational feedback and learning material in order to enhance the learning experience. For example, Santos et al. [17] show that affect as well as motivation and self-efficacy impact the effectiveness of motivational feedback and recommendations. Additionally, Woolf et al. [20] developed an affective pedagogical agent which is able to mirror a student's affective state, or acknowledge a student's affective state if it is negative. Another example is Conati & MacLaren [5], who developed a pedagogical agent to provide support according to the affective state of the students and the user's personal goal. Also, Shen et al. [18] recommend learning material to the student based on their affective state. D'Mello et al. [7] developed a system that is able to respond to students via a conversation that takes into account the affective state of the student.

In contrast, in this paper, we investigate the impact of different types of feedback on students' affective state and how and whether they can help students regulate their affect and thus improve learning. In what follows we present two sets of Wizard-of-Oz studies where feedback was provided to students interacting with an exploratory learning environment designed to learn fractions. From these studies, the affective states of the students were carefully annotated in order to address our research questions.

2 The Wizard-of-Oz studies

2.1 Aims

One of our research aims is to develop intelligent support that enhances the learning experience by taking into account the student's affective state. We were specifically interested in identifying how different feedback types modify affective states.

In order to address this question we conducted two sets of ecologically valid Wizard-of-Oz studies (e.g. [11, 8]) which investigated the effect of affective states on different feedback types at different stages of the task.

2.2 Participants and Procedure

In total, 26 Year-5 (9 to 10-year old) students took part in the Wizard-of-Oz studies. Each session lasted on average 20 minutes. Each student participated in one Wizard-of-Oz session.

The sessions were run in an ordinary classroom with multiple computers, where additional children were working with the learning platform (not wizarded) in order to support ecological validity. This was important particularly as in early settings we identified that children would not speak that much to the platform if they felt that they were monitored [10]. Figure 1 shows the setup of the studies. Wizards followed a script with pre-canned messages to send mes-

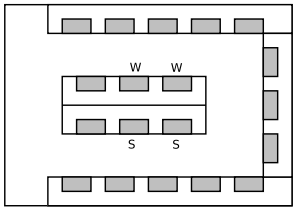

Fig. 1. The layout. The Wizard-of-Oz studies took place on the central isle while the rest of the students worked on a version of the system which only sequences tasks and provides minimal support (W=wizard, S=student).

sages to the students through the learning platform and deliberately limited their communication capacity in order to simulate the actual system. To achieve that wizards were only able to see students' screen. An assistant was able to hear students' reactions to reflections or talk-aloud prompts (as prompted by the 'system') and provide recommendations to the wizard with respect to the detected affective state. Any feedback provided was both shown on screen and read aloud by the system to students.

2.3 Feedback types

Different types of feedback were presented to students at different stages of their learning task. The feedback provided was based on interaction via keyboard and mouse, as well as speech.

We explore different types of feedback that are known from the literature to support students in their learning and fit our context. The following different feedback types were provided:

– **AFFECT BOOSTS - affect boosts**. As described in [20] affect boosts can help to enhance student's motivation in solving a particular learning

task. These included prompts that acknowledged for example that a task is difficult or that the student may be confused but they should keep trying.
- **INSTRUCTIVE FEEDBACK - instructive task-dependent feedback**. This feedback provided detailed instructions, what subtask or action to perform in order to solve the task.
- **OTHER PROBLEM SOLVING FEEDBACK - task-dependent feedback**. This support was centred on helping students to solve a particular problem that they are facing during their interaction by providing either questions to challenge their thinking or specific hints designed to help them identify the next step themselves.
- **TALK ALOUD PROMPTS - talking aloud**. With respect to learning in particular, the hypothesis that automatic speech recognition (ASR) can facilitate learning is based mostly on educational research that has shown benefits of verbalization for learning (e.g., [1]).
- **REFLECTIVE PROMPTS - reflecting on task performance and learning**. Self-explanation can be viewed as a tool to address students' own misunderstandings [4] and as a 'window' into students' thinking.
- **TALK MATHEMATICS PROMPTS - using particular domain specific mathematics vocabulary**. The aim of this prompt was to encourage students to use mathematical vocabulary in order continually revise their interpretations. In early studies [10] we found that students' reflections were often procedural and pragmatic (e.g. talking about the user interface) rather than mathematical.
- **TASK SEQUENCE PROMPTS - moving to the next task**. This feedback is centred on providing support regarding what action to perform next in order to change the task, such as clicking the 'Next' button.

Table 1 shows examples of the different feedback types.

3 Annotation of affective states and feedback reactions

From the Wizard-of-Oz studies we recorded the students' screen display and their voices. From this data, we annotated affective states (e.g. screen interaction and what the students said) before and after feedback was provided.

As described earlier, for the affective state detection we discriminated between five different affective types: enjoyment, surprise, confusion, frustration, and boredom. For the annotation of those affective states we used a similar strategy to that described in [15], where a dialogue between a teacher and a student was annotated retrospectively by categorising utterances in terms of different feedback types. Also, [2] describe how they coded different affective states based on observations of students interacting with a learning environment. Similarly, we annotated student's affective states for each type of feedback provided. In addition to the student's voice we also used the video of the screen capture to support the annotation process. Students' affective states were annotated as follows:

- **FLOW:** Engagement with the learning task. Statements like 'I am enjoying this task' or 'This is fun'. Sustained interaction with the system.

Feedback type	Example
AFFECT BOOSTS	You're working really hard! Keep going!
INSTRUCTIVE FEEDBACK	Use the comparison box to compare your fractions.
OTHER PROBLEM SOLVING FEEDBACK	If you add fractions, they need to have the same denominators first.
REFLECTIVE PROMPTS	What do you notice about the two fractions?
TALK ALOUD PROMPTS	Remember to talk aloud, what are you thinking?
TALK MATHEMATICS PROMPTS	Can you explain that again using the terms denominator, numerator?
TASK SEQUENCE PROMPTS	Well done. When you are ready click 'next' for the next task.

Table 1. Examples of feedback types

- **SURPRISE:** Gasping. Statements like 'Huh?' or 'Oh, no!'.
- **CONFUSION:** Failing to perform a particular task. Statements such as 'I'm confused!' or 'Why didn't it work?'. Uncertain interaction with the system.
- **FRUSTRATION:** Tendency to give up, repeatedly clicking or deleting of objects in the system or repeatedly failing to perform a particular task, sighing, statements such as, 'What's going on?!'.
- **BOREDOM:** Inactivity or statements such as 'Can we do something else?' or 'This is boring'.

4 Results

In total 396 messages were sent to 26 students. The video data in combination with the sound files were analysed independently by three researchers (one was independent of the project) who categorised the affective states of students before and after the feedback messages were provided.

The data is combined from two sets of Wizard-of-Oz studies. We use kappa statistics to measure the degree of the agreements of the annotations for reliability. Kappa was .46, $p<.001$. This is generally expected from retrospective annotation of naturalistic affect experiences [14]. We consolidated the annotations based on discussion between the annotators and the rest of the authors of the paper in order to agree upon the annotations that did not match originally. In the second set we had resources to introduce the Baker-Rodrigo Observation Method Protocol (BROMP) and the HART mobile app that facilitates the coding of students affective states in the classroom [12]. Kappa based on the retrospective annotation was still .56, $p<.001$. We first consolidated the data with the same approach as before and then compared against the field annotations. Kappa between the consolidated annotation and the HART data was .71,

$p<.05$ (note that is may appear low but we did not expect the retrospective annotation to get surprise and frustration accurately). We used the HART data to improve the annotation by mapping feedback actions against the observation for 20 seconds prior to the delivery of the feedback to 20 seconds after the student had closed the corresponding feedback window. We marked the changes for an independent annotator to revisit the first set of annotations.

The student's affective states, that occurred before and after the different types of feedback was provided, can be seen in figure 2. Each block shows an affective state *before* feedback was provided. The colour within the bars indicates the type of affective states that occurred *after* the feedback was provided. The number within the bars indicate the number of times the affective state occurred.

In order to investigate whether there was an effect of the feedback on the learning experience, we looked at whether a student's affective state was enhanced, stayed the same or worsened. An affective state was enhanced for example, when it was changed from confusion to flow, or (given the findings about confusion [6]) from frustration to confusion, frustration to flow, boredom to flow etc. An affective state was worsened if it moved for example, from flow to frustration or confusion, or from confusion to frustration.

As the data is categorical [16], we apply chi-square tests to investigate statistical significant differences between the groups. We present them below and discuss in more detail in the next section.

Flow When students were in flow, there was no significant difference between the feedback types on whether the affective state stayed in the same flow state ($X^2(6, N=169) = 4.31$, $p>.05$) or worsened ($X^2(6, N=169) = 4.89$, $p>.05$). As flow is the most positive affective state, the affective state in this sub-sample cannot be enhanced.

Confusion When students were confused, there was a significant effect of the feedback type on whether students' affective state was enhanced into a flow state ($X^2(6, N=181) = 13.65$, $p<.05$). The most effective feedback types were affect boosts with 68% of the cases, followed by guidance feedback with 67%, and task sequence prompts with 63%. Reflective prompts resulted in a flow state in 48% of the cases, talk aloud prompts 38%, and problem solving support with 34%. Talk maths prompts were the least effective with only 25% of the cases.

There was also a significant effect of the feedback type and whether the affective state stayed the same ($X^2(6, N=181) = 14.34$, $p<.05$). Talk maths prompts were highest associated with a continuing confused state with 75% of the cases. This was followed by problem solving support with 66%, talk aloud prompts with 59%, reflective prompts with 52%, task sequence prompts with 37%, affect boosts with 32%, and the least feedback type that was associated with a continuing confused state were guidance feedback with 29% of the cases.

There was no significant association between the feedback type and whether the affective state worsened ($X^2(6, N=181) = 4.65$, $p>.05$).

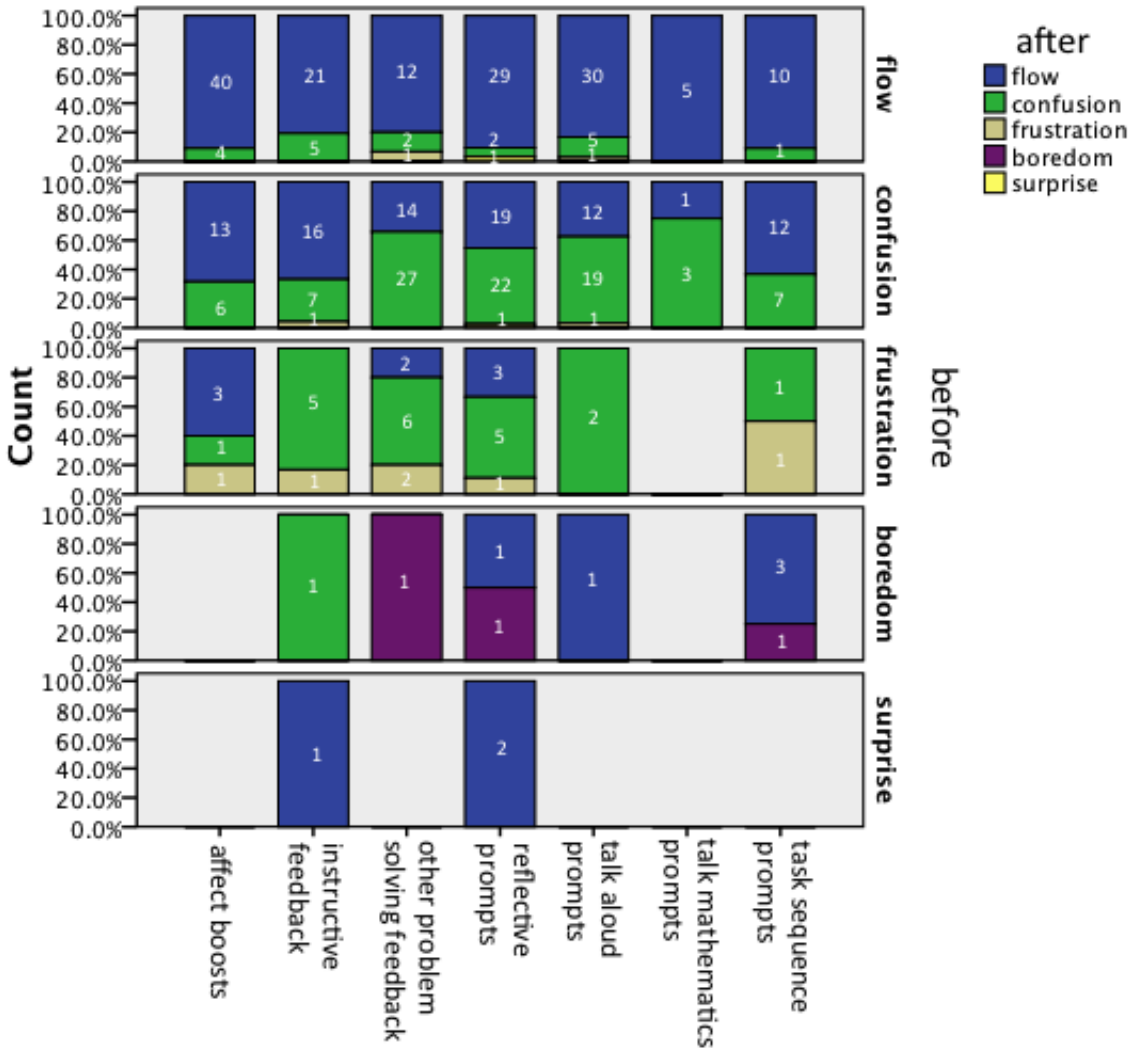

Fig. 2. Students' affective states before and after feedback was provided. Each block shows an affective state *before* feedback was provided. The colour within the bars indicates the type of affective states that occurred *after* the feedback was provided. The number within the bars indicate the number of times the affective state occurred.

Frustration, boredom and surprise There was not sufficient data available when students were frustrated (36 cases), nor when they were bored (9 cases), or surprised (3 cases) to run a statistical test across the different affective states and feedback types.

However, the data indicates that some of the provided feedback types were better able to change the affective state of the student when they were frustrated, bored or surprised, as can be seen in figure 2. For example, 60% of the affect boosts were able to change frustration into flow, followed by reflective prompts 33% and problem solving support 20%.

5 Discussion

The results presented in the previous section show that feedback can enhance students' affective states, and that the impact of the various feedback types mostly depends on the students' affective state before the feedback was provided.

When students were in flow there was no significant difference between the feedback types on whether or not the affective state stayed the same or worsened. This suggests that, when students are in flow, challenging feedback can be provided without negative implications.

However, when students were confused there was a difference between the feedback types on whether the affective state was enhanced, stayed the same or worsened. The feedback types that most effectively moved the student out of a confusion state were affect boosts, instructive, and task sequence prompts. When they were struggling to overcome problems, affect boosts appeared to encourage some students to redouble their efforts without the need for task specific support. We can hypothesise that this enabled students to self-regulate their affect and move forward. As expected, instructive feedback appears to have given the students the next steps that they needed, whereas other problem solving was less successful. Other problem solving feedback seems to have led students to be more confused because of the increased cognitive load caused by them having to understand the hint or the question provided.

While talk aloud prompts and talk maths, encouraged them to vocalize what they are trying to achieve, they appear not to have helped the students address their confusions. Instead, when they were confused, students appeared to have welcomed a new task (the opportunity to abandon the cause of their confusion). While as a strategy this can be pedagogically debatable, there is scope to provide tasks aimed to help them at the same concepts in a different, simpler way or to allow them to practice first some skills in a practice-based rather than exploratory task.

Although there was insufficient data to analyse the impact of the different feedback types on students' affective state when they were frustrated, some tentative observations can be made. For example, it was evident that the affect of students who were frustrated was enhanced whatever the feedback they were provided with. However, it is notable that the frustrated students who were provided affect boosts were most likely to move to a flow. We have other anecdotal evidence in the same scenario with different students that suggest that explicitly

addressing affect and helping students to think of their emotions during learning can help them move to confused or to flow state without need for immediate problem solving support.

It is worth noting that compared to other research we may have been unable to detect more negative states, especially boredom, because of the nature of the environment that the students were using – an exploratory learning environment that encouraged them to speak. The combination of unstructured learning and speech might prevent students from becoming bored.

6 Conclusion and future work

The affective state of students can be modified with feedback. There is a difference in the impact of different feedback types according to the affective state the student is in before the feedback was provided. Although there seems not to be too much of a difference when students are in flow, when students were confused different feedback types seem to matter more. While, for example, affect boosts and instructive feedback were able to change confusion into flow, prompting students to use mathematical vocabulary or providing other problem solving support, were associated with the same confused state or even lead to frustration.

In the light of findings like D'Mello et al. [6] for example of the importance of confusion under appropriate conditions in learning, our findings have important implications for learning and teaching in general, and AIED in particular. Problem solving support specifically in exploratory learning environments is difficult to achieve successfully, particularly when students are in a situation that was not previously encountered during a system's design. However, detecting affect may be relatively easier in certain contexts particularly in speech-enabled software like in our case and therefore affective support matters as much, if not more than, problem solving support. In addition, the exact type of support provided when students are frustrated is important. To understand this better we need to investigate more the different types of problem solving support and their combination with affective feedback that can act both as a way to self-regulate affect and take student into a more positive state like confusion or flow.

In our current study we are implying that learning performance is enhanced when students are in a positive affective state. In the future we are planning to evaluate if learning performance will be enhanced when students are moved out of a negative into a positive affective state. Our next step is to train an intelligent system that is able to tailor the type of feedback according to the affective state of the student in order to enhance the learning experience.

References

1. Askeland, M.: Sound-based strategy training in multiplication. European Journal of Special Needs Education 27(2), 201–217 (2012)
2. Baker, R.S.J.d., DMello, S.K., Rodrigo, M.T., Graesser, A.C.: Better to be frustrated than bored: The incidence, persistence, and impact of learners cognitive-affective states during interactions with three different computer-based learning environments. Int. J. Hum.-Comput. Stud. 68(4), 223–241 (2010)

3. Carenini, G., Conati, C., Hoque, E., Steichen, B., Toker, D., Enns, J.: Highlighting interventions and user differences: Informing adaptive information visualization support. In: Proceedings of CHI 14. pp. 1835–1844 (2014)
4. Chi, M.: Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In: Glaser, R. (ed.) Advances in instructional psychology, pp. 161–238. Mahwah, NJ: Lawrence Erbaum Associates (2000)
5. Conati, C., MacLaren, H.: Empirically building and evaluating a probabilistic model of user affect. User Modeling and User-Adapted Interaction (2009)
6. DMello, S.K., Lehman, B., Pekrun, R., Graesser, A.C.: Confusion can be beneficial for learning. Learning & Instruction 29(1), 153–170 (2014)
7. DMello, S., Craig, S., Gholson, B., Franklin, S., Picard, R., Graesser, A.: Integrating affect sensors in an intelligent tutoring system. In: Affective Interactions: The Computer in the Affective Loop Workshop at IUI 2005. pp. 7–13 (2005)
8. Eynon, R., Davies, C., Holmes, W.: Supporting older adults in using technology for lifelong learning: the methodological and conceptual value of wizard of oz simulations. In: Proceedings of NLC 2012. pp. 66–73 (2012)
9. Kort, B., Reilly, R., Picard, R.: An affective model of the interplay between emotions and learning. In: Proceedings of ICALT 2001. No. 43–46 (2001)
10. Mavrikis, M., Grawemeyer, B., Hansen, A., Gutiérrez-Santos, S.: Exploring the potential of speech recognition to support problem solving and reflection. In: ECTEL 2014. pp. 263–276 (2014)
11. Mavrikis, M., Gutiérrez-Santos, S.: Not all wizards are from Oz: Iterative design of intelligent learning environments by communication capacity tapering. Computers & Education 54(3), 641–651 (2010)
12. Ocumpaugh, J., Baker, R.S.J.d., Rodrigo, M.M.T.: Baker-Rodrigo Observation Method Protocol (BROMP) 1.0. Training Manual version 1.0. Tech. rep., New York, NY: EdLab. Manila, Philippines: Ateneo Laboratory for the Learning Sciences. (2012)
13. Pekrun, R.: The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. J. Edu. Psych. Rev. pp. 315–341 (2006)
14. Porayska-Pomsta, K., Mavrikis, M., DMello, S., Conati, C., de Baker, R.S.J.: Knowledge elicitation methods for affect modelling in education. I. J. Artificial Intelligence in Education 22(3), 107–140 (2013)
15. Porayska-Pomsta, K., Mavrikis, M., Pain, H.: Diagnosing and acting on student affect: the tutors perspective. UMUAI 18(1), 125-173 (2008)
16. Rosenthal, R., Rosnow, R.: Essentials of Behavioral Research: Methods and data analysis. McGraw Hill, 3rd edn. (2008)
17. Santos, O., Saneiro, M., Salmeron-Majadas, S., J.G., B.: A methodological approach to elicit affective educational recommendataions. In: Proceedings of ICALT 2014 (2014)
18. Shen, L., Wang, M., Shen, R.: Affective e-learning: Using emotional data to improve learning in pervasive learning environment. Educational Technology & Society 12(2), 176–189 (2009)
19. Sweller, J., van Merrienboer, J.G., Paas, G.W.: Cognitive Architecture and Instructional Design. Educational Psychology Review 10, 251–296+ (1998)
20. Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., Picard, R.: Affect-aware tutors: recognising and responding to student affect. Int. J. Learning Technology 4(3-4), 129–164 (2009)

Recognising Perceived Task Difficulty from Speech and Pause Histograms

Ruth Janning, Carlotta Schatten, and Lars Schmidt-Thieme

Information Systems and Machine Learning Lab (ISMLL), University of Hildesheim
{janning,schatten,schmidt-thieme}@ismll.uni-hildesheim.de

Abstract. Currently, a lot of research in the field of intelligent tutoring systems is concerned with recognising student's emotions and affects. The recognition is done by extracting features from information sources like speech, typing and mouse clicking behaviour or physiological sensors. In former work we proposed some low-level speech features for perceived task difficulty recognition in intelligent tutoring systems. However, by extracting these features some information hidden in the speech input is loosed. Hence, in this paper we propose and investigate speech and pause histograms as features, which preserve some of the loosed information. The approach of using speech and pause histograms for perceived task difficulty recognition is evaluated by experiments on data collected in a study with German students solving mathematical tasks.

Keywords: Intelligent tutoring systems, perceived task difficulty recognition, low-level speech features, speech and pause histograms

1 Introduction

Automatic cognition, affect and emotion recognition is a relatively young and very important research field in the area of adaptive intelligent tutoring systems. Some research has been done to identify useful information sources and appropriate features able to describe student's cognition, emotions and affects. Those information sources can be speech input, written input, typing and mouse clicking behaviour or input from physiological sensors. In former work ([5], [6], [7]) we proposed low-level speech features for perceived task difficulty recognition in intelligent tutoring systems. These features are extracted from the amplitudes of speech input of students interacting with the system and contain for instance the maximal and average length of speech phases and pauses. However, by extracting those features some more fine granulated information contained within the sequence of speech and pause segments is loosed and the question arises if there is a way to create features which preserve the loosed information. Histograms contain much more information than only the maximal, minimal and average value. Hence, in this work we propose and investigate speech and pause histograms as features for perceived task difficulty recognition, i.e. for recognising if a student feels *over-challenged* or *appropriately challenged* by a task. Speech and pause histograms share the advantages of low-level speech features

(they do not inherit the error from speech recognition and there is no need that students use words related to emotions or affects, see also sec. 2) and avoid to lose information hidden in the sequences of speech and pause segments.

2 Related Work

For the purpose to recognise emotion or affect in speech one can distinct linguistics features, like n-grams and bag-of-words, and low-level features like prosodic features, disfluencies, e.g. speech pauses ([5], [6]), (see e.g. [17]) or articulation features ([7]). If linguistics features are not extracted from written but from spoken input, a transcription or speech recognition process has to be applied to the speech input before emotion or affect recognition can be conducted. Linguistic features for affect and emotion recognition from conversational cues were presented and investigated e.g. in [10] and [11]. Low-level features are used in the literature for instance for expert identification, as in [18], [13] and [8], for emotion and affect recognition as in [12] and [5], [6], [7] or for humour recognition as in [15]. The advantage of using low-level features like disfluencies is that instead of a full transcription or speech recognition approach only for instance a pause identification has to be applied before computing the features. That means that one does not inherit the error of the full speech recognition approach. Furthermore, these features are independent from the need that students use words related to emotions or affects. Another kind of features which is independent from the need that students use words related to emotions or affects are features gained from information about the actions of the students interacting with the system (see e.g. [9]) like features extracted from a log-file (see e.g. [2], [16], [14]). In [9] such kind of features is used to predict whether a student can answer correctly questions in an intelligent learning environment without requesting help and whether a student's interaction is beneficial in terms of learning. Also the keystroke dynamics features used in [4] belong to this kind of features. In [4] emotional states were identified by analysing the rhythm of the typing patterns of persons on a keyboard. A further possibility of gaining features is using the information from physiological sensors as for instance in [1]. However, bringing sensors into classrooms is time consuming and expensive and one has to cope with students' acceptance of the sensors.

3 Speech and Pause Histograms

As mentioned above, in this paper we investigate the ability of speech and pause histograms for perceived task difficulty recognition. How these speech and pause histograms are created from students' speech input is described in sec. 3.2 and the data which we used for our experiments is described in the next section.

3.1 Data

We conducted a study in which the speech and actions of ten 10 to 12 years old German students were recorded and their perceived task-difficulties were

Fig. 1. Graphic of the decibel scale of an example sound file of a student. The two straight horizontal lines indicate the threshold.

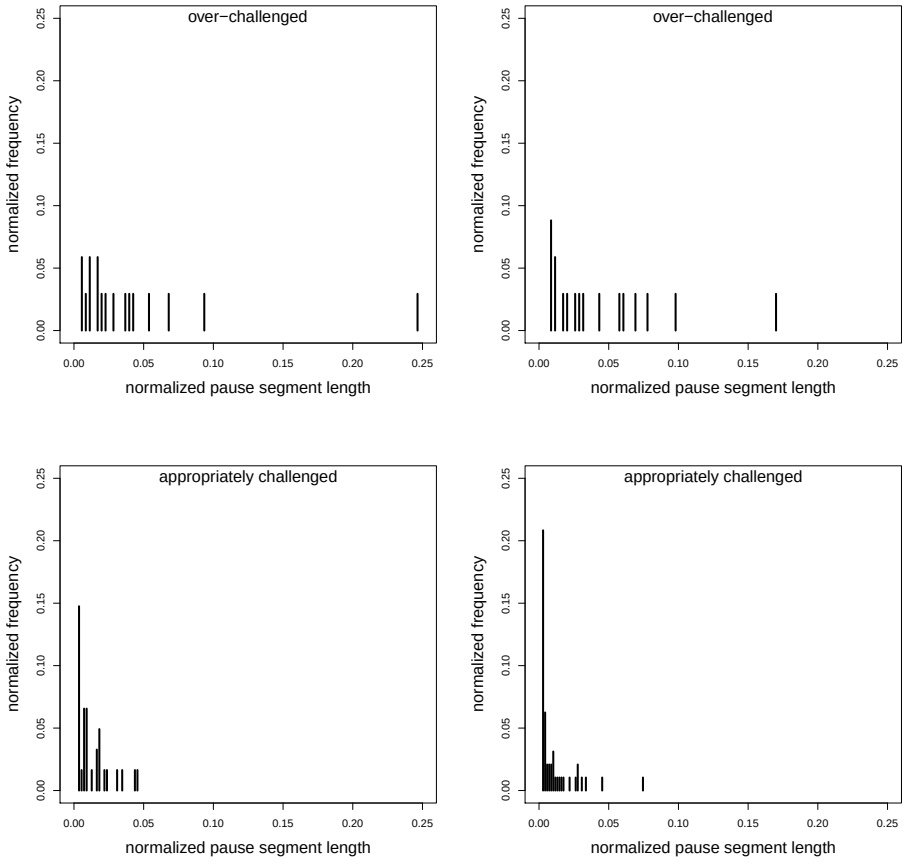

Fig. 2. Normalised pause histograms for a task of four different students, where two are labelled as *over-challenged* and the other two as *appropriately challenged*.

reported per task. The labelling of these data was done on the one hand concurrently by a human tutor and on the other hand retrospectively by a second reviewer (with a Cohen's kappa for inter-rater reliability of 0.747, $p < 0.001$). Divergences in the both labellings were clarified later on by discussions between the reviewers. During the study a paper sheet with fraction tasks was shown to the students and they were asked to paint – by means of a software for painting with a computer – their solution and they were prompt to explain aloud their observations and answers. The fraction tasks were subdivided into similar subtasks and covered exercises like assigning fractions to coloured parts of a circle or rectangle, reducing, adding or subtracting fractions and fraction equivalence. Originally, there were 10 tasks with 1 up to 10 subtasks but not each task was seen by each student. We made a screen recording to record the painting of the students and an acoustic recording to record the speech of the students. The screen recordings were used for the retrospective annotation. The acoustic speech recordings, consisting of 10 wav files with a length from 15 up to 20 minutes, were used to gain the speech and pause histograms. The data collection resulted in 36 examples (tasks) labelled with *over-challenged* (12 examples) or *appropriately challenged* (24 examples), respectively 48 examples (24 of class *appropriately challenged*, 24 of class *over-challenged*) after applying oversampling to the smaller set of examples of class *over-challenged* to eliminate the unbalance in the data.

3.2 Histograms for Classification

In the above mentioned study we observed that the children often exhibited longer pauses of silence while thinking about the problem when they were *over-challenged* or produced fewer and shorter pauses while communicating when they were *appropriately challenged*. Hence, in this paper we investigate information about pauses and speech segments within the speech input of students in connection with the perceived task difficulty. The first step to gain this information is to segment the acoustic speech recordings for identifying segments containing speech and segments corresponding to pauses. The most easy way to do this is to define a threshold on the decibel scale as done e.g. in [8]. For our study of the data we also used a threshold, which was estimated manually. The manual threshold estimation was done by extracting the amplitudes of the sound files, computing the decibel values and generating a graphic of it like the one in fig. 1. Subsequently, it was investigated which decibel values belong to speech and which ones to pauses to create from this information an appropriate threshold. By means of this threshold the pause and speech segments can be extracted. From the pause segments the pause histogram is generated by counting how often each possible pause length occur. This pause histogram is then normalised, to make the pause histograms of different speech inputs (of different students, different tasks and different lengths) comparable. The normalisation is done by dividing each occurring pause length by the length of the whole speech input as well as dividing the frequency of each occurring pause length by the number of all speech and pause segments, so that the resulting values stem

from the interval between 0 and 1. The same is done with the speech segments for generating the speech histogram. Examples of normalised pause histograms and speech histograms are shown in fig. 2 and fig. 3. The examples stem from the speech input for a task of four different students, where two were labelled as *over-challenged* and the other two as *appropriately challenged*. One can see some

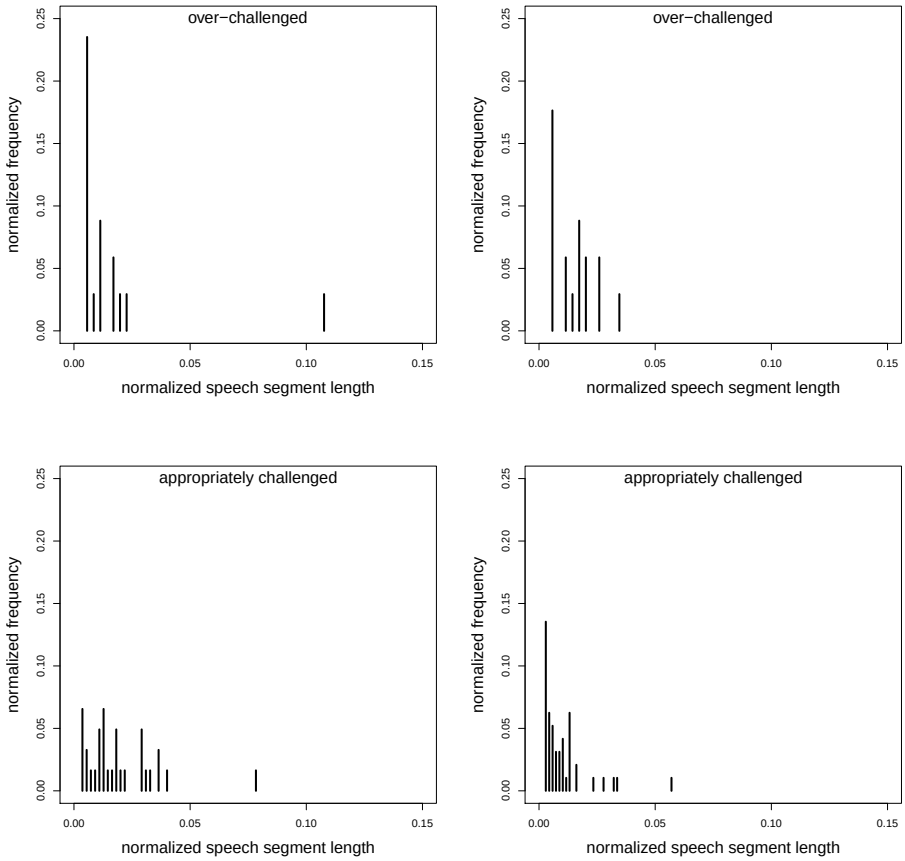

Fig. 3. Normalised speech histograms for a task of four different students, where two are labelled as *over-challenged* and the other two as *appropriately challenged*.

differences between the histograms of the *over-challenged* students and the *appropriately challenged* students as well as some similarities of the examples with the same label. The pause histograms of the *appropriately challenged* students show that there are a lot of very small pauses within their speech, but no very large pauses. The pause histograms of the *over-challenged* students in contrast

report long pauses and less smaller pauses than for the *appropriately challenged* students. In the speech histograms one can see that the *over-challenged* students used a lot of very small speech segments of the same length whereas for *appropriately challenged* students there is a large variance in the speech segment length. In the following section we investigate how these histograms can be used for classifying the speech input of a student for a task as either *over-challenged* or *appropriately challenged*.

4 Experiments

To investigate if the above described speech and pause histograms are applicable for distinguishing *over-challenged* and *appropriately challenged* students we conducted experiments with the perprocessing and settings described in the following section. The experimental results are reported in sec. 4.2.

4.1 Preprocessing and Experimental Settings

To be computationally comparable the normalised histograms still need to be preprocessed, or more explicitly generalised, as the set of possibly occurring segment lengths is infinite (it is a real value between 0 and 1). Hence, we divide the x-axis (the different normalised lengths of pause or speech segments) into a number of equal sized intervals, the *buckets*. Each occurring normalised segment length is then put into the bucket to whose interval it belongs. The number of buckets, or the bucket size respectively, is a hyper parameter and in the experiments we investigated different values for that parameter, i.e. we conducted experiments with 2 up to $1,000,000$ buckets (bucket size 0.5 up to 1.0E-6) where the numbers of buckets are multiples of the numbers by which 100 is divisible without remainder. A comparison of two different histograms can now be done by comparing the content of each bucket in both histograms, that means that for each bucket the normalised frequencies of segments belonging to that bucket are compared. In our experiments we compute the difference between two histograms by computing the differences between the frequencies in all buckets by means of the root mean square error (RMSE):

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{b}(b_i(H_x) - b_i(H_y))^2}{b}}\,, \tag{1}$$

where H_x and H_y are the two histograms to compare, $b_i(H_x)$ and $b_i(H_y)$ are the normalised frequency values belonging to bucket b_i of H_x and H_y and b is the number of buckets. For deciding to which class (*over-challenged* or *appropriately challenged*) a histogram belongs we applied the K-Nearest-Neighbour (KNN) approach. KNN (see e.g. [3]) classifies an example by a majority vote of its neighbours, that is the example is assigned to the class most common among its K nearest neighbours. These K nearest neighbours are the K closest training examples in the feature space. The *closeness* in our case is measured by means

of the RMSE. That is a histogram is assigned to that class to which the majority of the K closest (in terms of RMSE) histograms belongs. K is a further hyper parameter and also for that parameter we tried out different values, i.e. we conducted experiments with a number of 1 up to 35 neighbours where that value is an odd number less than the number of unique examples. For the evaluation we used a Leave-one-out cross-validation in the experiments. The results of our experiments with pause and speech histograms are discussed in the next section.

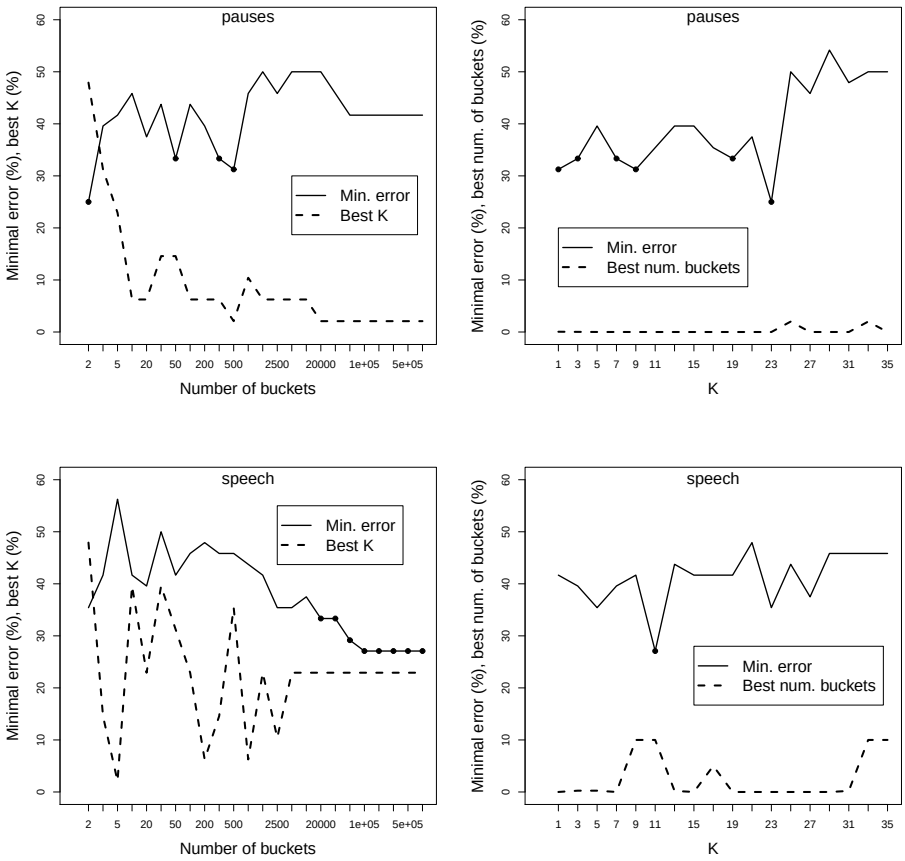

Fig. 4. Different numbers of buckets and different numbers K of neighbours mapped to the minimal classification error (%) and the belonging best value for K (% of the number of examples) and for the number of buckets (% of the max. number of buckets) for pause and speech histograms.

4.2 Experiments with Speech and Pause Histograms

As mentioned above, we conducted experiments with different numbers of buckets and different values for the K nearest neighbours. In fig. 4 we report the minimal classification error and the belonging best value of K for each bucket number as well as the the minimal classification error and the belonging best number of buckets for each value of K for the pause and the speech histograms. The *classification error* is the number of incorrectly classified histograms divided by the number of all histograms. The black dots in fig. 4 indicate the best results which are also reported in tab. 1 and 2. As one can see in fig. 4 for the

Table 1. Number of buckets, bucket size, K, classification error and F-measures of class *over-challenged* & *appropriately challenged* of the experiments with pause histograms with best result (classification error $< 34\%$, black dots in fig. 4).

Number of buckets	2	2	2	50	250	500
Bucket size	0.5	0.5	0.5	0.02	0.004	0.002
K	9	19	23	7	3	1
Error (%)	31.25	33.33	25.00	33.33	33.33	31.25
F-measure	0.57, 0.82	0.55, 0.80	0.67, 0.83	0.59, 0.63	0.59, 0.57	0.60, 0.71

Table 2. Number of buckets, bucket size, K, classification error and F-measures of class *over-challenged* & *appropriately challenged* of the experiments with speech histograms with best result (classification error $< 34\%$, black dots in fig. 4).

Number of buckets	20000	25000	50000	100000	200000	250000	500000	1000000
Bucket size	5.0E-5	4.0E-5	2.0E-5	1.0E-5	5.0E-6	4.0E-6	2.0E-6	1.0E-6
K	11	11	11	11	11	11	11	11
Error (%)	33.33	33.33	29.17	27.08	27.08	27.08	27.08	27.08
F-measure	0.57, 0.73	0.57, 0.73	0.62, 0.78	0.64, 0.77	0.64, 0.77	0.64, 0.77	0.64, 0.77	0.64, 0.77

pause histograms a smaller number of buckets delivers the best results whereas for the speech histograms the number of buckets has to be large, i.e. a more fine granulated division of the x-axis is needed for good results. The reason might be that the pause histograms of *over-challenged* and *appropriately challenged* students are easier distinguishable as in the pause histogram of an *over-challenged* student there are typically long pause segments which usually do not occur in the speech of *appropriately challenged* students (see also fig. 2). As fig. 3 shows, speech histograms of *over-challenged* and *appropriately challenged* students are not so easy to distinct. Tab. 1 and 2 show the results of the best choices for hyper parameter K and number of buckets and reports the classification error as well as the F-measures of both classes (*over-challenged* and *appropriately challenged*).

The F-measure is a value between 0 and 1 and the closer it is to 1 the better. It is the harmonic mean between the ratio of examples of a class c which are correctly recognised as members of that class (*recall*) and the ratio of examples classified as belonging to class c which actually belong to class c (*precision*). In our experiments the F-measures of class *appropriately challenged* are better than those of class *over-challenged*. The reason could be that originally there were more examples of class *appropriately challenged* and we just oversampled class *over-challenged* to receive a balanced example set. Nevertheless, the best classification errors of 25% and 27.08% and F-measures $0.67, 0.83$ and $0.64, 0.77$ in tab. 1 and 2 indicate that speech and pause histograms are applicable for perceived task difficulty recognition.

5 Conclusions and Future Work

We proposed and investigated speech and pause histograms, build from the sequences of speech and pause segments within the speech input of students, as features for perceived task difficulty recognition. To evaluate the approach of using the histograms for distinguishing *over-challenged* and *appropriately challenged* students we applied a K-Nearest-Neighbour classification delivering a classification error of 25% for pause histograms and 27.08% for speech histograms. Next steps will be to try out other classification approaches, for instance from time series classification. Furthermore, the information from the speech histograms and pause histograms could be combined to reach a better classification performance, e.g. by ensemble methods.

Acknowledgements. This work is co-funded by the EU project iTalk2Learn (www.italk2learn.eu) under grant agreement no. 318051.

References

1. Arroyo, I., Woolf, B.P., Burelson, W., Muldner, K. , Rai, D. and Tai, M.: A Multimedia Adaptive Tutoring System for Mathematics that Addresses Cognition, Metacognition and Affect. In International Journal of Artificial Intelligence in Education, Springer, Vol. 24, pp. 387–426 (2014)
2. Baker, R.S.J.D., Gowda, S., Wixon, M., Kalka, J., Wagner, A., Salvi, A., Aleven, V., Kusbit, G., Ocumpaugh, J. and Rossi,L.: Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra. In Proceedings of the 5th International Conference on Educational Data Mining (EDM 2012), pp. 126–133 (2012)
3. Cover, T. and Hart, P.: Nearest neighbor pattern classification. EEE Transactions on Information Theory, Vol. 13(1), pp. 21–27, doi:10.1109/TIT.1967.1053964 (1967)
4. Epp, C., Lippold, M. and Mandryk, R.L.: Identifying Emotional States Using Keystroke Dynamics. In Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems (CHI 2011), pp. 715–724 (2011)
5. Janning, R., Schatten, C., Schmidt-Thieme, L.: Multimodal Affect Recognition for Adaptive Intelligent Tutoring Systems. In Extended Proceedings of the 7th International Conference on Educational Data Mining (EDM 2014), pp. 171–178 (2014)

6. Janning, R., Schatten, C., Schmidt-Thieme, L.: Feature Analysis for Affect Recognition Supporting Task Sequencing in Adaptive Intelligent Tutoring Systems. In Proceedings of the European Conference on Technology Enhanced Learning (EC-TEL 2014), pp. 179–192 (2014)
7. Janning, R., Schatten, C., Schmidt-Thieme, L. and Backfried, G.: An SVM Plait for Improving Affect Recognition in Intelligent Tutoring Systems. In Proceedings of the IEEE International Conference on Tools with Artificial Intelligence (ICTAI) (2014)
8. Luz, S.: Automatic Identification of Experts and Performance Prediction in the Multimodal Math Data Corpus through Analysis of Speech Interaction. Second International Workshop on Multimodal Learning Analytics, Sydney Australia (2013)
9. Mavrikis, M.: Data-driven modelling of students interactions in an ILE. In Proceedings of the International Conference on Educational Data Mining (EDM 2008), pp. 87–96 (2008)
10. DMello, S.K., Craig, S.D., Witherspoon, A., McDaniel, B. and Graesser, A.: Automatic detection of learners affect from conversational cues. User Model User-Adap Inter, DOI 10.1007/s11257-007-9037-6 (2008)
11. D'Mello, S.K. and Graesser, A.: Language and Discourse Are Powerful Signals of Student Emotions during Tutoring. IEEE Transactions on Learning Technologies, Vol. 5(4), pp. 304–317, IEEE Computer Society (2012)
12. Moore, J.D., Tian, L. and Lai, C.: Word-Level Emotion Recognition Using High-Level Features. Computational Linguistics and Intelligent Text Processing (CICLing 2014), pp. 17–31 (2014)
13. Morency, L.P., Oviatt, S., Scherer, S., Weibel, N. and Worsley, M.: ICMI 2013 grand challenge workshop on multimodal learning analytics. In Proceedings of the 15th ACM on International conference on multimodal interaction (ICMI 2013), pp. 373–378 (2013)
14. Pardos, Z.A., Baker, R.S.J.D, San Pedro, M., Gowda, S.M. and Gowda, S.M.: Affective States and State Tests: Investigating How Affect and Engagement during the School Year Predict End-of-Year Learning Outcomes. Journal of Learning Analytics, Vol. 1(1), Inaugural issue, pp. 107–128 (2014)
15. Purandare, A. and Litman, D.: Humor: Prosody Analysis and Automatic Recognition for F * R * I * E * N * D * S *. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pp. 208–215 (2006)
16. San Pedro, M.O.C., Baker, R.S.J.D., Bowers, A. and Heffernan, N.: Predicting College Enrollment from Student Interaction with an Intelligent Tutoring System in Middle School. In Proceedings of the 6th International Conference on Educational Data Mining (EDM 2013), pp. 177–184 (2013)
17. Schuller, B., Batliner, A., Steidl, S. and Seppi, D.: Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication, Elsevier (2011)
18. Worsley, M. and Blikstein, P.: What's an Expert? Using Learning Analytics to Identify Emergent Markers of Expertise through Automated Speech, Sentiment and Sketch Analysis. In Proceedings of the 4th International Conference on Educational Data Mining (EDM '11), pp. 235–240 (2011)

Analyzing Student Action Sequences and Affect While Playing Physics Playground

Juan Miguel L. Andres[1], Ma. Mercedes T. Rodrigo[1],
Ryan S. Baker[2], Luc Paquette[2], Valerie J. Shute[3], Matthew Ventura[3]

[1] Ateneo de Manila University, Quezon City, Philippines
[2] Teachers College, Columbia University, New York, NY, USA
[3] Florida State University, Tallahassee, FL, USA
{mandres, mrodrigo}@ateneo.edu,
baker2@exchange.tc.columbia.edu, luc.paquette@gmail.com,
{vshute, mventura}@fsu.edu

Abstract. Physics Playground is an educational game that supports physics learning. It accepts multiple solutions to most problems and does not impose a stepwise progression through the content. Assessing student performance in an open-ended environment such as this is therefore challenging. This study investigates the relationships between student action sequences and affect among students using Physics Playground. The study identified most frequently traversed student action sequences and investigated whether these sequences were indicative of either boredom or confusion. The study found that boredom relates to poor performance outcomes, and confusion relates to sub-optimal performance, as evidenced by the significant correlations between the respective affective states, and the student action sequences.

Keywords: Affect modeling, action sequences, boredom, confusion, Physics Playground

1 Introduction

Physics Playground (PP) is an educational game that immerses learners in a choice-rich environment for developing intuitive knowledge about simple machines. As the environment does not impose a stepwise sequence on the learner, and because some problems can have multiple solutions, learners have the freedom to explore, attempt to solve, or abort problems as they wish. The challenge these types of environments impose on educators is that of assessment. Within such an open-ended system, how do educators and researchers assess learning as well as the quality of the learning process?

This study focuses its attention on two main phenomena: student learning and student affect. Student learning within PP refers to how well a player can understand the concepts surrounding four simple machines through their efficient execution in attempting to solve levels, as evidenced by the badges they earn.

Student affect refers to experiences of feelings or emotions. In this study, the affective states of interest are confusion and boredom, as prior studies have shown them to relate significantly with learning [4, 10]. Confusion is uncertainty about what to do next [5]. Confusion is scientifically interesting because it has a positive and negative dimension, wherein it either spurs learners to exert effort deliberately and purposefully to resolve cognitive conflict, or leads learners to become frustrated or bored, and may lead to disengagement from the learning task altogether [7].

Boredom, on the other hand, is an unpleasant, transient affective state in which the individual feels a pervasive lack of interest in and difficulty concentrating on the current activity [8]. Boredom has been a topic of interest because of the negative effects usually associated with it, such as poor long-term learning outcomes when students are not provided any scaffolding [10] and its being characteristic of less successful students [11].

A study conducted by Biswas, Kinnebrew, and Segedy [2] investigated frequently traversed sequences of student actions using bottom-up, data-driven sequence mining, the results of which contributed to the development of performance- and behavior-based learner models. The analyses in this paper seek to perform similar sequence-mining methods in order to find student sequences that inform either of the affective states of interest.

This study conducted data-driven sequence-mining analyses to answer the following research questions:

1. What were the frequently traversed student action sequences among students playing Physics Playground?
2. Are these action sequences indicative of either boredom or confusion?

The analyses in this study are limited to the data collected during gameplay of Physics Playground from six data gathering sessions conducted at a public school in Quezon City in 2013. Data is limited to the interaction logs generated by the game as well as human observation of affect as logged by two coders trained in the Baker-Rodrigo-Ocumpaugh Monitoring Protocol [9].

2 Methodology

2.1 Participant Profile

Data were gathered from 60 eighth grade public school students in Quezon City, Philippines. Students ranged in age from 13 to 16. Of the participants, 31% were male and 69% were female. As of 2011, the school had 1,976 students, predominantly Filipino, and 66 teachers. Participants had an average grade on assignments of B (on a scale from A to F).

2.2 Physics Playground

Physics Playground (PP) is an open-ended learning environment for physics that was designed to help secondary school students understand qualitative physics. Qualitative physics is a nonverbal, conceptual understanding of how the physical world operates [12].

PP has 74 levels that require the player to guide a green ball to a red balloon. An example level is shown in Fig. 1. The player achieves this goal by drawing agents (ramps, pendulums, springboards, or levers) or by nudging the ball to the left or right by clicking on it. The moment the objects are drawn, they behave according to the law of gravity and Newton's 3 laws of motion [12].

Fig. 1. Example PP level.

Performance Metrics. Gold and silver badges are awarded to students who manage to solve a level. A gold badge is given to a student who is able to solve the level by drawing a number of objects equal to the particular level's par value (i.e., the minimum number of objects needed to be drawn to solve the level). A student who solves a level using more objects will earn a silver badge. A student earns no badge if he was not able to solve the level. Many levels in PP have multiple solutions, meaning a player can solve the level using different agents.

2.3 Interaction Logs

During gameplay, PP automatically generates interaction log files. Each level a student plays creates a corresponding log file, which tracks every event that occurs as the student interacts with the game. Per level attempt, PP tracks begin and end times, the agents used, and the badges awarded upon level completion. PP also logs the *Freeform Objects* that player draw, or objects that cannot be classified as any of the four agents. The physics agents within PP are as follows:

- Ramp, any line drawn that helps to guide a ball in motion,
- Lever, an agent that rotates around a fixed point, usually called a fulcrum,
- Pendulum, an agent that directs an impulse tangent to its direction of motion,
- Springboard, an agent that stores elastic potential energy provided by a falling weight.

2.4 The Observation Protocol

The Baker-Rodrigo-Ocumpaugh Monitoring Protocol (BROMP) is a protocol for quantitative field observations of student affect and engagement-related behavior, described in detail in [9]. The affective states observed within Physics Playground in this study were engaged concentration, confusion, frustration, boredom, happiness, delight, and curiosity. The affective categories were drawn from [6].

BROMP guides observers in coding affect through different utterances, body language, and interaction with the software specific to each affective state. A total of seven affective states were coded, however, this study focuses on three: concentration, confusion, and boredom. These were identified as follows:

1. Concentration — immersion and focus on the task at hand, leaning toward the computer and attempting to solve the level, a subset of the flow experience described in [5].
2. Confusion — scratching his head, repeatedly attempting to solve the same level, statements such as "I don't understand?" and "Why didn't it work?"
3. Boredom — slouching, sitting back and looking around the classroom for prolonged periods of time, statements such as "Can we do something else?" and "This is boring!"

Following BROMP, two trained observers observed ten students per session, coding students in a round-robin manner, in 20-second intervals throughout the entire observation period of 2 hours. During each 20-second window, both BROMP observers code the current student's affect independently. If the student exhibited two or more distinct states during a 20-second observation window, the observers only coded the first state. The inter-coder reliability for affect for the two observers in the study was acceptably high with a Cohen's Kappa [3] of 0.67. The typical threshold for certifying a coder in the use of BROMP is 0.6, a standard previously used in certifying 71 coders in the use of BROMP (e.g., [9]).

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

2.6 Data Collection Process

Before playing PP, students answered a 16-item multiple-choice pretest for 20 minutes. Students then played the game for 2 hours, during which time two trained observers used BROMP to code student affect and behavior on the HART application. A total of 4,320 observations were collected (i.e., 36 observations per participant per each of the two observers). After completing gameplay, participants answered a 16-item multiple-choice posttest for 20 minutes. The pretest and posttest were designed to assess knowledge of physics concepts, and have been used in previous studies involving PP [12].

To investigate how students interacted with PP, the study made use of the interaction logs recorded during gameplay to analyze student performance. Of the 60 participants, data from 11 students were lost because of faulty data capture and

corrupted log files. Only 49 students had complete observations and logs. As a result, the analysis in this paper is limited to these students, and the 3,528 remaining affect observations. Engaged concentration was observed 72% of the time, confusion was observed 8% of the time, and boredom and frustration were observed 7% of the time. Happiness, delight, and curiosity comprise the remaining 6% of the observation time.

3 Analyses and Results

3.1 Agent Sequences

All PP-generated logs were parsed and filtered to produce a list containing only the events relevant to the study. Sequences were then separated into one of two categories: 1) silver sequences, or the sequences that ultimately led to a silver badge, which comprised 44% of all level attempts, and 2) unsolved sequences, or the sequences that led to the student quitting the level without finding a solution, which comprised 39% of all level attempts. Sequences that ended in gold badges were dropped from the analysis because they only comprised 17% of all level attempts.

Every time a student earns a badge after solving a level, the badge is awarded for one of the four agents (e.g., a player is awarded a silver ramp badge for solving the level using a ramp, and another player is awarded a gold pendulum badge for solving another level using a pendulum). We tracked the agents the badges were awarded for per level, and used this list of badges to relabel the sequences based on correctness. If the level awarded a badge for an agent, that agent was labeled as `correct` for that level; if not, the agent was labeled as `wrong` for the level. For example, on a level that awarded badges for springboards and levers, a sequence of `Lever > Ramp > Springboard > Level End (silver-springboard)` would be relabeled as `correct > wrong > correct > Level End (silver)`.

The relabeling was done because most of the sequences were level-dependent, that is, a majority of some sequences appeared on only one or two levels. By relabeling based on correctness, we were able to ensure level-independence among sequences. Sequences were tabulated and their frequencies calculated (i.e., how many times each of the 49 students traversed each of the sequences). We calculated for distribution of sequence frequencies, and the sequences we found to occur rarely (i.e., less than 30% of the population traversed them) were dropped from the analysis. We found that the gold sequences occurred rarely, which was another reason they were dropped from the analysis. The resulting silver and unsolved sequences can be found in Tables 1 and 2, respectively, along with the frequency means and standard deviations.

Table 1 lists the top 7 silver sequences within PP, which were traversed by more than 30% of the study's population. The Sequences column shows what the respective sequences look like, and the Frequency column shows the average number of times the 49 students traversed them and the standard deviations.

Highlighted sequences showed significant correlations with either boredom or confusion, as discussed further in Section 3.2. Table 2 is presented in the same manner.

Table 1. Top 7 silver sequences, their traversal frequency means, and standard deviations.

	Sequences	Frequency	
		Mean	SD
1	`correct>Level End (silver)`	3.53	2.34
2	`Level End (silver)`	2.61	2.33
3	`wrong>Level End (silver)`	1.90	1.37
4	`correct>correct>Level End (silver)`	1.61	1.15
5	`wrong>correct>Level End (silver)`	0.90	1.01
6	`correct>correct>correct>Level End (silver)`	0.80	1.00
7	`wrong>correct>correct>Level End (silver)`	0.69	0.77

The silver sequences in Table 1 show signs of experimentation, with students playing around with the correct and incorrect agents to solve the levels, as seen in sequences 5 and 7. Sequences 1, 4, and 6 show students using the correct agents, but are unable to earn gold badges. This suggests that students, while knowing which agents to use, do not have a full grasp of the physics concepts surrounding the agents' execution. Sequence 3 shows students using wrong objects to solve the levels. While this may suggest that students are still struggling to understand how the agents work and which agent would best solve a level given the ball and the balloon's positions, this may have also been caused by the PP logger labeling the objects they drew as freeform objects, and not one of the correct agents.

Sequence 1 shows the students drawing only the correct agent, but are still unable to earn a gold badge. The sequence-mining algorithm only pulled events related to drawing any of the four main agents, which are enumerated in Section 2.3. Drawing a lever or a springboard, for example, would require drawing more than one component. A lever requires the fulcrum, the board, and the object dropped on the board to project the ball upwards. In order for the agent to work, it has to be executed correctly (i.e., the board must be long enough, with the fulcrum in the right position, and the object dropped on the board must be heavy enough to propel the ball into the air). Sequence 1 may have been caused by students drawing the correct agent, but improperly executing it. For example, the student may not have drawn the right-sized weight to drop on the lever, and thus had to draw another. While drawing another weight to drop on the lever counts towards the level's object count, it was not logged as a separate event by the sequence mining analysis because the player did not draw another agent, only a component of it. Sequence 2, on the other hand, is suspect because despite the student drawing no objects to solve a level, he ends up with only a silver badge. This was most likely caused by the improper logging of the game. The top 7 most frequently traversed silver sequences account for 58% of the total number of silver sequences.

Table 2. Top 6 unsolved sequences, their traversal frequency means, and standard deviations.

	Sequences	Frequency	
		Mean	SD
1	`Level End (none)`	10.69	8.17
2	`wrong>Level End (none)`	1.55	1.65
3	`correct>Level End (none)`	1.29	1.50
4	`wrong>wrong>Level End (none)`	0.45	0.65
5	`correct>correct>Level End (none)`	0.41	0.73
6	`wrong>correct>Level End (none)`	0.39	0.57

Table 2, which shows the top 6 unsolved sequences, shows signs of students giving up. Sequence 1 shows students giving up without even drawing a single object, which could have been caused by one of two things: 1) the student saw the level and decided to quit without attempting to solve it, or 2) again, the logger did not log the objects correctly. This sequence is similar to one of the silver sequences in that no objects were drawn. What makes them different, however, is what the sequences ultimately led to. The silver sequences ended in a silver badge, and the unsolved sequences ended in the student earning no badge. The majority of the sequences listed in Table 2 show students experimenting mainly with wrong objects, whether agents or freeform objects. This implies that the students are lacking in the understanding of how to solve the levels. Sequences 3 and 5 are interesting because it is unclear whether or not the students understood the concepts of the agents. That is, students were drawing the correct agents, but could not get the ball to reach the balloon. Despite drawing one or two correct agents, the students decided to give up and quit. The top 6 unsolved sequences account for 81% of the total number of unsolved sequences.

3.2 Relationship with Affect

We computed frequencies for each of the 13 sequences that the 49 students traversed. Correlations were then run between each of the 13 arrays and the incidences of confusion and boredom. Because the number of tests introduces the possibility of false discoveries, Storey's adjustment [13] was used as a post-hoc control, which provides a q-value, representing the probability that the finding was a false discovery. Tables 3 and 4 show the results. Highlights and asterisks (*) were used on significant findings ($q \leq 0.05$).

Table 3 lists the top 7 most frequently traversed silver sequences, from left to right. The sequences these header numbers represent can be found in Table 1. The table shows the correlation between each of the top 7 silver sequences using a metric that represents the percentage of all attempts that match each of the sequences, the percentage of time the students were observed to be confused (r, con), and the percentage of time the students were observed to be bored (r, bor).

Table 4 is presented in the same manner, with sequence information in Table 2 for the top 6 unsolved sequences.

Table 3. Correlations between top 7 silver sequences, confusion, and boredom.

	Top 7 silver sequences						
	1	2	3	4	5	6	7
r, con	-0.33	0.23	0.41*	0.03	0.17	0.54*	0.28
r, bor	-0.20	-0.17	-0.19	-0.05	0.14	-0.19	-0.20

Table 3 shows two significant positive correlations between confusion and the silver sequences. The two sequences showed signs of lesser understanding of the agents. Sequence 3 shows students using only a wrong object to solve a level, which may have been caused either by incorrect object labeling (e.g., PP logged a ramp as a Freeform Object), or the student found a different way of solving the level. Like in most learning environments, players are able to game the system – or systematically misuse the game's features to solve a level [1] – within PP through stacking. Stacking is done when players draw freeform objects to either prop the ball forward or upward, which may have been the case in sequence 3. Sequence 6 shows students drawing only correct agents. These sequences having significant correlations with confusion may imply lesser understanding among confused students as the they are not only dealing with proper agent execution, but also with deciding which agent would best solve the level. Despite the challenges faced by these students, however, they still managed to find a solution to the level. Our findings suggest that the inability to grasp the physics concepts surrounding the agents is a sign of confusion.

Table 4. Correlations between top 6 unsolved sequences, confusion, and boredom.

	Top 6 unsolved learning sequences					
	1	2	3	4	5	6
r, con	-0.17	0.00	-0.12	-0.01	-0.06	0.04
r, bor	-0.12	0.13	0.12	-0.03	0.48*	0.06

Table 4 shows that one of the most frequently traversed unsolved sequences has a significant positive correlation with boredom. This sequence shows students using only correct agents, but ultimately deciding to give up. This may have been caused by the inability to execute the agents correctly, which may imply that, unlike confused students, bored students were not likely to exert additional effort to try to solve the level or understand proper agent execution. As mentioned previously, boredom has been found to have significant relationships with negative performance outcomes. In this case, sequences all ultimately led to disengagement: students quitting the level before finding a solution, showing signs of giving up and lack of understanding of any of the four agents.

4 Conclusions and Future Work

This study sought to identify the most frequently traversed student action sequences among eighth grade students while interacting with an education game for physics called Physics Playground. Further, the study sought to investigate how these sequences may be indicative of affective states, particularly boredom and confusion, which have been found to significantly affect student learning.

Data-driven sequence mining techniques were conducted to identify most frequently traversed actions sequences in two categories: the sequences that would eventually lead the student to a silver badge, and the paths that would eventually lead the student to not earning a badge.

In the silver sequences, students played around with freeform objects and some of the four agents in attempting to solve the level. The study found confusion to correlate significantly with two of the silver sequences, which supports previous findings regarding the relationship between confusion and in-game achievement, which suggest that because students are unable to grasp the concepts surrounding the agents and their executions, students resort to finding other solutions.

In the unsolved sequences, students would give up and quit without finding a solution, despite already using the correct agents to solve the level. The study found boredom to correlate significantly with one of the unsolved sequences. This finding supports the literature that has shown that boredom relates to poor learning outcomes. This work provides further evidence that boredom and disengagement from learning go hand-in-hand.

This study provides specific sequences of student actions that are indicative of the boredom and confusion, which has implications on the design and further development of Physics Playground. This study also contributes to the literature by providing empirical support that boredom and confusion are affective states that influence performance outcomes within open-ended learning environments, and are thus affective states that learning environments must focus on detecting and providing remediation to. We found that both bored and confused students will tend to continuously use correct agents in attempting to solve levels, but execute them incorrectly. The difference between the two, however, is that confused students tend to end up solving the level, while bored students give up.

The analyses run in this paper were part of a bigger investigation, and as such, there are several interesting ways forward in light of our findings. The paper aims for its findings to contribute to the creation of a tool that can automatically detect affect given a sequence of student interactions, and provide necessary remediation in order to curb student experiences of boredom.

Relationship analyses run between student action sequences and incidences of affect in this paper were done through correlations. However, findings were not able to determine whether boredom or confusion occurred more frequently during specific action sequences. We want to find out whether boredom or confusion occurred before, during, or after the students' execution of the action sequences, and in doing so, see whether or not the affective states were causes or effects of the action sequence executions. We are currently investigating this relationship in a separate study.

Acknowledgements. We would like to thank the Ateneo Center for Educational Development, Carmela C. Oracion, Christopher Ryan Adalem, and the officials at Krus Na Ligas High School, Jessica O. Sugay, Dr. Matthew Small, and the Gates Foundation Grant #OP106038 for collaborating with us.

References

1. Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task behavior in the cognitive tutor classroom: when students game the system. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 383-390). ACM.
2. Biswas, G., Kinnebrew, J. S., & Segedy, J. R. (2011). Using a Cognitive/Metacognitive Task Model to analyze Students Learning Behaviors.
3. Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 20 (1960), 37-46.
4. Craig, S., Graesser, A., Sullins, J., & Gholson, B. (2004). Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. Journal of Educational Media, 29(3), 241-250.
5. Csikszentmihalyi, M. (1990). Flow: The psychology of optimal experience. New Y ork: Harper Perennial.
6. D'Mello, S. K., Craig, S. D., Witherspoon, A., McDaniel, B., & Graesser, A. (2005). Integrating affect sensors in an intelligent tutoring system. In Proceedings of the Workshop on Affective Interactions: The computer in the affective loop workshop, International conference on intelligent user interfaces (pp. 7- 13). New York: Association for Computing Machinery.
7. D'Mello, S., Graesser, A. (2012). Dynamics of affective states during complex learning. Learning and Instruction, 22(2): 145-157.
8. Fisherl, C. D. (1993). Boredom at work: A neglected concept. Human Relations, 46(3), 395-417.
9. Ocumpaugh, J., Baker, R.S., Rodrigo, M.M.T. (2015) Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual. Technical Report. New York, NY: Teachers College, Columbia University. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.
10. Pardos, Z. A., Baker, R. S., San Pedro, M., Gowda, S. M., & Gowda, S. M. (2014). Affective States and State Tests: Investigating How Affect and Engagement during the School Year Predict End-of-Year Learning Outcomes. Journal of Learning Analytics, 1(1), 107-128.
11. San Pedro, M. O. Z., d Baker, R. S., Gowda, S. M., & Heffernan, N. T. (2013, January). Towards an understanding of affect and knowledge from student interaction with an Intelligent Tutoring System. In Artificial Intelligence in Education (pp. 41-50). Springer Berlin Heidelberg.
12. Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and Learning of Qualitative Physics in Newton's Playground. The Journal of Educational Research, 106(6), 423-430.
13. Storey, J.D. (2002). A direct approach to false discovery rates. Journal of the Royal Statistical Society, Series B, 64: 479-498.

La Mort du Chercheur: How well do students' subjective understandings of affective representations used in self-report align with one another's, and researchers'?

Wixon[1], Danielle Allessio[2], Jaclyn Ocumpaugh[3], Beverly Woolf[2], Winslow Burleson[4] and Ivon Arroyo[1]

[1]Worcester Polytechnic Institute, Worcester Massachusetts
{mwixon, iarroyo}@wpi.edu
[2]University of Massachusetts, Amherst Massachusetts
{allessio@educ, bev@cs}.umass.edu
[3]Teachers College, Columbia University, New York, New York
jocumpaugh@wpi.edu
[4]New York University, New York, New York
Wb50@nyu.edu

Abstract. We address empirical methods to assess the reliability and design of affective self-reports. Previous research has shown that students may have subjectively different understandings of the affective state they are reporting [18], particularly among younger students[10]. For example, what one student describes as "extremely frustrating" another might see as only "mildly frustrating." Further, what students describe as "frustration" may differ between individuals in terms of valence, and activation. In an effort to address these issues, we use an established visual representation of educationally relevant emotional differences [3, 8, 25]. Students were asked to rate various affective terms and facial expressions on a coordinate axis in terms of valence and activation. In so doing, we hope to begin to measure the variability of affective representations as a measurement tool. Quantifying the extent to which representations of affect may vary provides a measure of measurement error to improve reliability.

Keywords: Affective States; Intelligent Tutoring Systems; Reasons for Affect

1 Introduction

The evaluation of students' affective states remains an incredibly difficult challenge. While recognized as a key indicator of student engagement [14, 17, 26], there remains no clear gold-standard for identifying an affective state, leading to researchers such as Graesser & D'Mello [13] to call for greater attention to the theoretical stances that certain research methods entail. A full theoretical review is beyond the scope of this paper; instead, the current work presents a pilot study designed to empirically evaluate the reliability of two different types of affective self-reports in an educational

context. Reliability is measured both in terms of inter-rater reliability (the degree of agreement between students), and "inter-method" reliability (i.e. given words or facial expressions as representations of affective states, which representation produces more consistent results).

A considerable body of research has been devoted to affect computing, and in particular to affect detection in educational software [9]. Progress has been made with methods that include self-report [8, 10], physiological sensors [1, 24], video-based retrospective reports [5, 15], text-based [11, 19], and field observation [16, 23] data. However, much of this research evaluates success based on the ability of a model to predict when a training label is present or absent, without giving deeper consideration to questions about the appropriateness of the training label itself.

Even within limited to the body of research that relies on self-report research, there are serious concerns about how methodological decisions might impact student responses. In addition to issues about the frequency and timing of surveys, one primary area of concern is that students may have subjectively different understandings of the state they are reporting [19], an effect that is likely to be even greater among younger students [10]. For example, Graesser and D'Mello [13] have suggested that a students' tolerance of cognitive disequilibrium (e.g., confusion or frustration) is probably conditioned by their knowledge and prior success with the topic they are interacting with. Further, what students describe as "frustration" in itself may differ between individuals in terms of dimensional component measures of affect: valence, activation, and dominance. The former two dimensions are typically used to differentiate affective states [4], and the latter used in some cases [7].

In this study, we explore these interpretative issues using three different types of representations that have been employed in previous self-report studies: words, facial expressions, and dimensional measures. In particular, we are interested in verifying that students' understanding of the meaning of these representations aligns with interpretations of these labels that are present in the literature (as constructed by experts). To this end, we use dimensional measures (valence & activation) to compare how students respond to both linguistic representations and pictorial representations, further testing hypotheses that the latter might be more appropriate for surveying students [19, 21, 22]. Our goal is to determine the extent to which this student population shows variance in the interpretation of these two different types of representations, since substantial variation in student perception should be accounted for in subsequent research. Last, while we might achieve researcher agreement in terms of methods and terminology for self-reported affects, that will do little good if there is a large degree of variance in terms of our subject pool's agreement on the meaning of these constructs.

1.1 Methods

Students surveyed included eighty one 7th graders from two Californian middle schools in a major city (among the 30 most populous cities in California), where a majority of census respondents identified as Hispanic or Latino and median household

income was within one standard deviation of California's overall median household income. They were surveyed at the end of the academic year.

Fig. 1. Blank Valence & Activation Sheet given to Students

Students were asked to place both textual and facial representations of affect on an XY axis of Activation=X Valence=Y. Textual representations of affect were selected based on the affective states that have been used in the past [2, 12], that corresponded to quite different levels of activation x valence according to us researchers, so that words would theoretically cover all quadrants. These terms and their researcher-hypothesized valence x arousal placements included: Angry (low valence x high activation), Anxious (low valence x high activation), Bored (low valence x low activation), Confident (high valence x low activation), Confused (med-low valence x med-high activation), Enjoying (high valence x medium activation), Excited (high valence x high activation), Frustrated (low valence x high activation), Interested (high valence x medium activation) and Relieved (high valence x med-low activation). In general, it was clear to the researchers which word corresponded to which face, with a few exceptions, such as the level of activation that should be associated to enjoying and interest. An established set of emoticons were chosen from previous affective research [8] that corresponded to extreme emoticon states of activation x valence x dominance. While the emoticons possessed these three attributes, our participants were asked only to orient them based on activation and valence.

Each student was presented with a sheet of paper depicting a coordinate axis with activation from "sleepy" to "hyper" on the x-axis and "bad" to "great" on the y-axis. These terms were used to express what valence and activation mean experientially, using language that children are familiar with and could relate to. Activation is then expressed more as a physical experience of arousal, while Valence is expressed not as much as a physical experience but as a judgment of the positive or negative nature of

the experience. Later, during coding, these axes were mapped discretized into a seven point scale of -3 to 0 to +3 at either extreme of each axis, defining a grid of 7 x 7.

Students were also given stickers for the 10 separate affective terms: Angry, Anxious, Bored, Confident, Confused, Enjoying, Excited, Frustrated, Interested, & Relieved, see Figure 2; as well as 8 stickers to depict each extreme emoticon expression from the ends of each of the 3 axis coordinate systems including: pleasure, activation, and dominance [8]. Students placed each of these stickers on their coordinate axes according to where they felt each term or emoticon should be placed with respect to valence and activation.

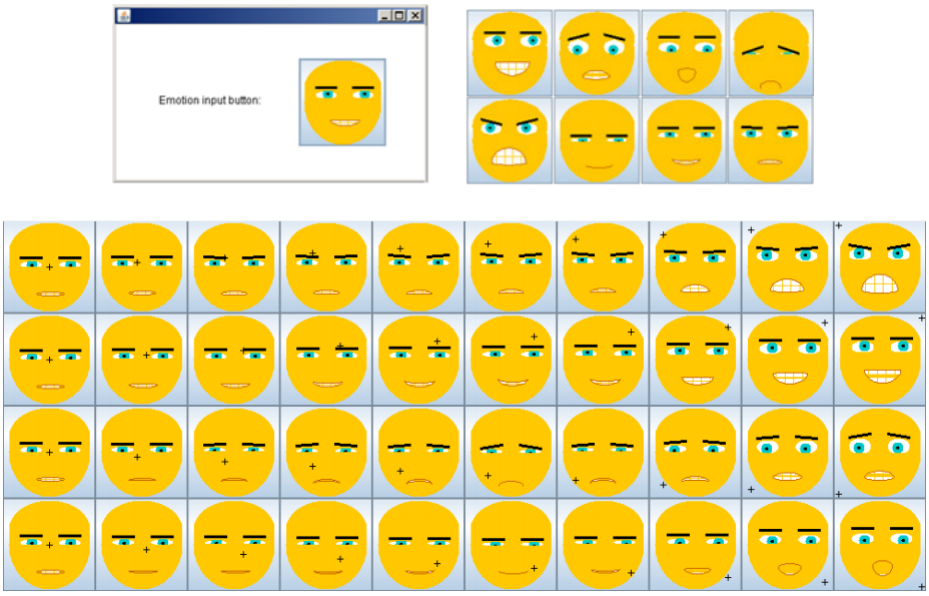

Fig. 2. Directly from Broekens, & Brinkman, 2013 [8]. Top left displays the affect button interface. Students use the cursor to change the expression in the inter-face. Depending on their actions, one of 40 affective expressions may be displayed; these expressions, shown across the bottom of this figure, are designed to vary based on pleasure (valence), activation, and dominance (PAD for brevity). From left to right first row: elated (PAD=1,1,1), afraid (-1,1,-1), surprised (1,1,-1), sad (-1,-1,-1). From left to right second row: angry (-1,1,1), relaxed (1,-1,-1), content(1,-1,1), frustrated (-1,-1,1). Top right displays PAD extremes, which serve as the basis for this research.

2 Results

Mean positioning results are displayed visually in figure 3, corresponding to the position that each word or emoticon sticker was placed averaged across all respondents. Missing data occurred in which students may not have placed every sticker. On average any given term or emoticon was missing 16.6 reports, with a maximum of 23 students of 81 missing reports for boredom, frustration, and relief. The average stu-

dent was only missing 3.7 out of 18 terms and emoticons from their sheet, and there were 5 students who turned in completely blank sheets.

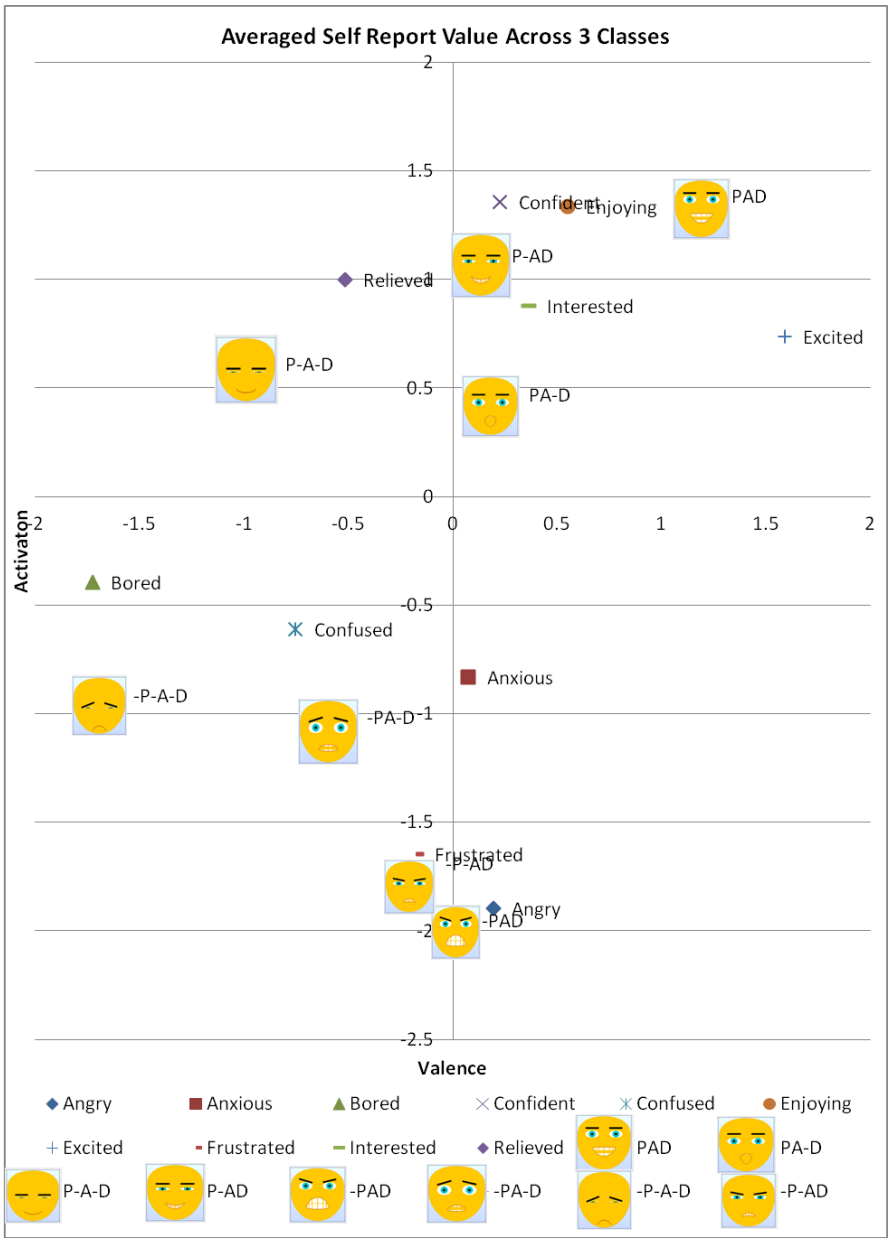

Fig. 3. Averaged Placement of Text and Emoticon Stickers

Interestingly, the placement of -PAD and -P-AD (negative sign indicating most extreme negative activation, pleasure, dominance, lack of a negative indicating most extreme positive, see figure 2 caption) match up with their respective terms "Angry" and "Frustrated" very closely. However, while both seem to be at the extreme end of negative valence, on average both seem to be viewed as fairly neutral in terms of activation by students. Although all emoticons and terms fall under the expected half of the coordinate axes in terms of valence (i.e. those we would expect to be pleasurable are categorized as above the origin, those displeasurable below it), activation does not follow this trend. For example anxiety is rated as neutral activation. One possible explanation, consistent with the results, is that students may be grouping activation and dominance together as a single measure. Emoticons with both negative activation and dominance were rated most negatively in terms of activation, those with either negative activation or dominance tended to fall in the middle, and the rating with all positive PAD was the emoticon with the highest rated activation.

Text or Emoticon	Activation Mean (StdDev)	Valence Mean (StdDev)
Angry	0.19 (1.09)	-1.9 (0.99)
Anxious	0.07 (1.78)	-0.87 (1.19)
Bored	-1.72 (1.28)	-0.4 (1.02)
Confident	0.23 (1.22)	1.35 (0.99)
Confused	-0.75 (1.36)	-0.61 (1.12)
Enjoying	0.55 (1.18)	1.34 (1.14)
Excited	1.59 (1.04)	0.74 (1.26)
Frustrated	-0.17 (1.33)	-1.65 (1.05)
Interested	0.36 (1.34)	0.88 (0.98)
Relieved	-0.52 (1.43)	1 (1.12)
Face_PAD	1.25 (1.3)	1.38 (1.13)
Face_PA-D	0.28 (1.86)	0.47 (0.93)
Face_P-A-D	-0.89 (1.57)	0.61 (0.91)
Face_P-AD	0.2 (1.26)	1.11 (1.08)
Face_-PAD	0.05 (0.95)	-1.95 (0.93)
Face_-PA-D	-0.5 (1.39)	-1.01 (1.01)
Face_-P-A-D	-1.61 (1.41)	-0.91 (1.11)
Face_-P-AD	-0.12 (1.15)	-1.69 (0.89)
Average	-0.08 (1.33)	-0.12 (1.05)

Table 1. Means and Standard Deviations of Students' placement of stickers.

One key goal of this work was to determine the degree of variance between students in terms of where they placed each term or emoticon. Given any affective term, there was little difference between the standard deviation for terms (mean S.D for terms = 1.20) and faces (mean S.D. for faces = 1.18). However, there was a larger

difference between the standard deviation in activation (mean S.D for activation of terms or faces = 1.33) and valence (mean S.D for valence of terms or faces = 1.05), suggesting that students may have a greater degree of agreement in regarding rating the valence of affective representations than the activation it produces in them, which is consistent with the finding that affective representations fall on the division between positive and negative valence as we would categorize them, but not necessarily in terms of activation.

3 Discussion

The results presented in this article highlight a few different conclusions: a) students did not necessarily match emoticons or affective terms to the quadrants where researchers would have placed them, mostly in relation to activation; b) there is a large variation across these middle-school students in terms of where they placed a specific emotion within the axes of valence x arousal.

Characterizing researcher common expectations for arousal or activation is difficult, as many researchers only tentatively suggest how emotional states may be characterized in terms of activation. Pekrun found data to support boredom being somewhat deactivating, [18]. Russell [25] explores the components of affect and offers a few hypotheses which are summarized in figure 1 of Baker et al 2010 [3] wherein boredom is characterized as deactivating, while frustration, surprise, and delight are characterized as activating. Broekens' [8] emoticons follow the scheme outlined in the figure 2 caption: elation, fear, surprise, and anger are seen as activating, while sadness, relaxation, contentment, and frustration are seen as deactivating.

Students seem to agree that delight or elation is highly activating along with excitement, and boredom is deactivating along with sadness and relaxation. However, we found that students viewed an emoticon of fear as deactivating, and other affective states placed relatively close to neutral in terms of activation.

There are a few points of methodological concern. Firstly, the order that the students' place their stickers may be important: beyond a simple priming effect of considering one term/emoticon before another, by placing one item first students are changing the affordance of the coordinate axis itself by adding a milestone in the form of a term or emoticon. In future research, we could consider including fewer stimuli for placement or giving students a clean chart for each stimuli.

A second point of concern is one of validity. The terms, emoticons, and even the coordinate axis itself are abstract descriptors of affective states, which in this experiment are divorced from the actual experiences students may be having.

By placing our study outside the experimental environment we are likely reducing the validity of this work in exchange for simplicity of study design (i.e. not requiring students to respond with faces and words on the axis at various points in their experience).

The work of Bieg et al. [6] tells a much larger story than recommending against self-reports out of context. Out of context self-reports were found to bias in a consistent direction as compared to in context self-reports. However we maintain this

method is “less valid” rather than “invalid”. Further, if we take into consideration the savings in class time an out of context self-report may actually be a better study design choice in some cases. It is our position that establishing more quantitative comparisons of reliability will yield better relative comparisons of validity and allow for improved study design.

This argument can be extended to affective research in general in the distinction between emotional experience and appraisal. We conceptualize the experience itself as the construct, and the cognitive appraisal process as a means of communicating that experience. The appraisal may be performed to send communication (e.g. having an experience and generating a representation of that experience for others), or receive communication (e.g. identify a representation as signifying an emotional state).

From this standpoint we suggest that the fewer steps of appraisal exist, the greater the face validity of an appraisal is in terms of reflecting an emotional experience. This is consistent with the findings of [6] wherein aggregate appraisal may differ from immediate contextual appraisal and we tend to view immediate appraisal as having greater face validity. This hypothesis also lends credence to the belief that external appraisal of an unconsciously generated representation (which may still be unconsciously meant to communicate an experience), in the form of facial expressions may be more valid than self-report measures wherein experiences are appraised by both subject and researcher. However, while passing through multiple appraisals may risk loss of information, the quality and richness of the appraisal may also play a role.

While validity remains very difficult to establish with regard to affect by testing “inter-method” or “representational” reliability perhaps we can building convergent and discriminant validity: multiple representations indicating the same construct across multiple participants. We maintain that reliability and validity are continuous rather than discrete traits of models. Therefore, we wish to reach consensus on methods of determining reliability and validity and then begin applying them to methods of inferring the experience of emotion. This work is a means of determining reliability between appraisals of representations of emotion rather than reliability of appraisals of emotions themselves. This is to say that matching particular facial expression to their personal lexicon of categorical affective terms, a high degree of agreement may validate the relationship between depictions of affect textually and facially, but not between either of those representations and the experience of an emotion.

A potential way towards greater validity and reliability could be to cognitively induce an emotional experience by asking students to respond to how they would feel given a particular situation (e.g. “Report on how you’d feel if you failed a math test.”). Of course there may be a distinction between induced affect and “organic” affect, further there will be a broad degree of subjectivity based on how individual students might feel about any given situation. Therefore the variance in responses could be attributed at least to two types of factors: those pertaining to both how students’ believe they would feel in a given context, and those pertaining to students’ ability to report that subjective experience through self-report measures. While there isn’t a clear way to disambiguate between which type of factor is responsible for the variance here, such an approach might be able to establish a conservative maximum of error in self-report measurements, because two students might have very different

feelings about failing a math exam. In essence, we have measured variance in reliability here, not validity.

Finally, while reliability of self-report measures should inform their design, there may be cases of diminishing returns where a slight improvement in reliability has heavy costs for implementation workload, response time, or other practical concerns. We need not pick the measure with the highest available reliability; however it would be good to have some empirical handle on the relative reliabilities of different types of self-report measures. Perhaps the greatest thing to come out of this work would be future collaborations which might better address these concerns.

4 References

1. AlZoubi, O., D'Mello, S. K., & Calvo, R. A. (2012). Detecting naturalistic expressions of nonbasic affect using physiological signals. Affective Computing, IEEE Transactions on, 3(3), 298-310.
2. Arroyo, I., Woolf, B.P., Royer, J.M. and Tai, M. (2009b) 'Affective gendered learning companion', Proceedings of the International Conference on Artificial Intelligence and Education, IOS Press, pp.41–48.
3. Baker, R.S.J.d., D'Mello, S.K., Rodrigo, M.M.T., Graesser, A.C. (2010) Better to Be Frustrated than Bored: The Incidence, Persistence, and Impact of Learners' Cognitive-Affective States during Interactions with Three Different Computer-Based Learning Environments. *International Journal of Human-Computer Studies*, 68 (4), 223-241.
4. Barrett, L. F. (2004). Feelings or Words? Understanding the Content in Self-Report Ratings of Experienced Emotion. Journal of Personality and Social Psychology, 87(2), 266–281.
5. Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., Shute, V., Ventura, M., & Zhao, W. (2015). Automatic Detection of Learning-Centered Affective States in the Wild. In Proceedings of the 2015 International Conference on Intelligent User Interfaces (IUI 2015). ACM, New York, NY, USA.
6. Bieg, M., Goetz, T., & Lipnevich, A.A. (2014). What Students Think They Feel Differs from What They Really Feel – Academic Self-Concept Moderates the Discrepancy between Students' Trait and State Emotional Self-Reports. PLoS ONE 9(3): e92563.
7. Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the Self-Assessment Manikin and the Semantic Differential. Journal of Behav Ther Exp Psychiatry, 25, 49-59.
8. Broekens, J., & Brinkman, W.-P. (2013). AffectButton: a method for reliable and valid affective support. International Journal of Human-Computer Studies, 71(6), 641-667.
9. Calvo, R. A., D'Mello, S., Gratch, J., & Kappas, A. (Eds.) (2015). The Oxford Handbook of Affective Computing. Oxford University Press: New York, NY.
10. Conati, C., & Maclaren, H. (2009). Empirically building and evaluating a probabilistic model of user affect. User Modeling and User-Adapted Interaction, 19(3), 267-303.
11. D'Mello, S. , Craig, S. D., Sullins, J., & Graesser, A. C. (2006). Predicting Affective States expressed through an Emote-Aloud Procedure from AutoTutor's Mixed-Initiative Dialogue. International Journal of Artificial Intelligence in Education, 16(1), 3-28.
12. D'Mello, S., & Graesser, A. C. (2012). Language and Discourse Are Powerful Signals of Student Emotions during Tutoring. IEEE Transactions on Learning Technologies, 5(4): 304–317.

13. Graesser, A., & D'Mello, S. (2011). Theoretical perspectives on affect and deep learning. In New perspectives on affect and learning technologies (pp. 11-21). Springer New York.
14. Linnenbrink-Garcia, L., & Pekrun, R. (2011). Students' emotions and academic engagement. Introduction to the special issue. Contemporary Educational Psychology, 36, 1–3.
15. McDaniel, B. T., D'Mello, S., King, B. G., Chipman, P., Tapp, K., & Graesser, A. C. (2007). Facial features for affective state detection in learning environments. In Proceedings of the 29th Annual Cognitive Science Society (pp. 467-472).
16. Ocumpaugh, J., Baker, R.S., Rodrigo, M.M.T. (2015) Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual.. Technical Report. New York, NY: Teachers College, Columbia University. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.
17. Pardos, Z. A., Baker, R. S., San Pedro, M. O., Gowda, S. M., & Gowda, S. M. (2013). Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. Proc. 3rd Int.Conf. Learning Analytics & Knowledge, 117-124.
18. Pekrun, R., Goetz, T., Daniels, L. M., Stupnisky, R. H., & Perry, R. P. (2010). Boredom in achievement settings: Exploring control–value antecedents and performance outcomes of a neglected emotion. *Journal of Educational Psychology, 102*(3), 531-549.
19. Porayska-Pomsta, K., Mavrikis, M., D'Mello, S., Conati, C., Baker, R.S.J.d. (2013) Knowledge Elicitation Methods for Affect Modeling in Education. International Journal of Artificial Intelligence in Education, 22 (3), 107-140.
20. Porayska-Pomsta, K., Mavrikis, M., & Pain, H. (2008). Diagnosing and acting on student affect: the tutor's perspective. User Modeling and User-Adapted Interaction, 18(1-2), 125-173.
21. Read, J., McFarlane, S., and Cassey, C. (2002). Endurability, engagement and expectations: Measuring children's fun. In Proceedings of International Conference for Interaction Design and Children.
22. Read J. C. and MacFarlane, S.(2006). Using the fun toolkit and other survey methods to gather opinions in child computer interaction. In Proceedings of the 2006 conference on Interaction design and children (IDC '06). ACM, New York, NY, USA, 81-88.
23. Rodrigo, M. M. T., Baker, R. S. J. d., Lagud, M. C. V., Lim, S. A. L., Macapanpan, A. F., Pascua, S. A. M. S., et al. (2007). Affect and Usage Choices in Simulation ProblemSolving Environments. In R. Luckin, K. R. Koedinger & J. Greer (Eds.), Proceeding of the 2007 conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts that Work (Vol. Frontiers in Artificial Intelligence and Applications 158). Amsterdam: IOS Press.
24. Rowe, J. P., Mott, B. W., & Lester, J. C. (2014) It's All About the Process: Building Sensor-Driven Emotion Detectors with GIFT. In Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym2) (p. 135).
25. Russell J.A, Barrett L,F. (1999) Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. J. Pers. Soc. Psychol. 76(5):805–19.
26. San Pedro, M.O.Z., Baker, R.S.J.d., Bowers, A.J., Heffernan, N.T. (2013) Predicting College Enrollment from Student Interaction with an Intelligent Tutoring System in Middle School. Proceedings of the 6th International Conference on Educational Data Mining, 177-184.