Workshop on Feedback from Multimodal Interactions in Learning Management Systems (FFMI)

Virtually all learning management systems and tutoring systems provide feedback to learners based on their time spent within the system, the number, intensity and type of tasks worked on and past performance with these tasks and corresponding skills. Some systems even use this information to steer the learning process by interventions such as recommending specific next tasks to work on, providing hints etc. Often the analysis of learner / system interactions is limited to these high-level interactions, and does not make good use of all the information available in much richer interaction types such speech and video. In the workshop Feedback from Multimodal Interactions in Learning Management Systems (FFMI@EDM'2014) we wanted to bring together researchers and practitioners who are interested in developing data-driven feedback and intervention mechanisms based on rich, multimodal interactions of learners within learning management systems, and among learners providing mutual advice and help. We aim at discussing all stages of the process, starting from preprocessing raw sensor data, automatic recognition of affective states to learning to identify salient features in these interactions that provide useful cues to steer feedback and intervention strategies and leading to adaptive and personalized learning management systems. The contributions presented in this workshop range from work about affect recognition in intelligent tutoring systems to research questions from online learning and collaborative learning.

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

Carles Sierra, IIIA, Spanish Research Council, University of Technology, Sydney
Arvid Kappas, School of Humanities and Social Sciences, Jacobs University Bremen, Germany
Emanuele Ruffaldi, PERCRO, Scuola Superiore Sant'Anna, Pisa, Italy
Sergio Gutierrez-Santos, Birkbeck, University of London, UK
Mark d'Inverno, Goldsmiths, University of London, UK
Manolis Mavrikis, IOE, University of London, UK
Francois Pachet, Sony Computer Science Laboratory Paris, France
Matthew Yee-King, Goldsmiths, University of London, UK
Helen Hastie, Heriot Watt University, Edinburgh, Scotland
Iolanda Leite, Yale University, Connecticut, United States
Luis de-la-Fuente, International University of La Rioja, Spain
Helen Pain, ILCC, Human Communication Research Centre, University of Edinburgh

The FFMI workshop organizers
Lars Schmidt-Thieme
Ruth Janning

Table of Contents FFMI

Interventions during student multimodal learning activities: which, and why?

Beate Grawemeyer
London Knowledge Lab
Birkbeck College
University of London, UK
beate@dcs.bbk.ac.uk

Manolis Mavrikis
London Knowledge Lab
Institute of Education
University of London, UK
m.mavrikis@ioe.ac.uk

Sergio Gutierrez-Santos
London Knowledge Lab
Birkbeck College
University of London, UK
sergut@dcs.bbk.ac.uk

Alice Hansen
London Knowledge Lab
Institute of Education
University of London, UK
a.hansen@ioe.ac.uk

ABSTRACT

Emotions play a significant role in students' learning behaviour. Positive emotions can enhance learning, whilst negative emotions can inhibit it. This paper describes a Wizard-of-Oz (WoZ) study which investigates the potential of Automatic Speech Recognition (ASR) together with an emotion detector able to classify emotions from speech to support young children in their exploration and reflection whilst working with interactive learning environments. We describe a unique ecologically valid WoZ study in a classroom. During the study the wizards provided support using a script, and followed an iterative methodology which limited their capacity to communicate, in order to simulate the real system we are developing. Our results indicate that there is an effect of emotions on the acceptance of feedback. Additionally, certain types of feedback are more effective than others for particular emotions.

Keywords

Affect, emotions, intelligent support

1. INTRODUCTION

Our aim is to build a learning platform for elementary education which integrates speech recognition for children in order to enable natural communication. This paper reports from on a Wizard-of-Oz study which explores the effect of emotions deduced from speech on different feedback types.

The importance of language as both a psychological and cultural tool that mediates learning has long been recognised; from as early as Vygotsky to modern linguists such as Pinker. From a Human Computer Interaction (HCI) perspective, speech recognition technology has the potential to enable more intuitive interaction with a system, particularly for young learners who reportedly talk aloud while engaged in problem solving (e.g. [11]).

Finally, speech provides an additional cue for drawing inferences on students' emotions and attitude towards the learning situation while they are solving tasks. By paying attention to tone and pitch of speech in conjunction with other auditory signs like sighs, gasps etc., we can provide learners with even more individualized help, by detecting emotions and providing support specifically tailored to the emotional state.

As described in [15] emotions interact with and influence the learning process. While positive emotions such as awe, 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 includes a range and combination of positive and negative emotions. For example, a student is motivated and expresses curiosity to explore a particular learning goal, however s/he might have some misconceptions and needs to reconsider her/his knowledge. This can evoke frustration and/or disappointment. However, this negative emotion can turn into curiosity again, if the student gets a new idea on how to solve the learning task.

[9] categorised emotions based on facial expressions. These included, joy, anger, surprise, fear, and disgust/contempt. However, these emotions are not specific to learning. [22] classified achievement emotions that arise in a learning situation. Achievement emotions are emotions that are linked to learning, instruction, and achievement. Emotions are classified into prospective, retrospective and activity emotions. They can be positive or negative. For example, a prospective positive emotion is hope for success, while a negative emotion is anxiety about failure. Retrospective emotions are for example, the positive emotion pride or the negative emotion shame, which the student experienced after receiving feedback of an achievement. Activity emotions arise during learning, such as positive emotions like enjoyment, or negative emotions like anger, frustration, or boredom.

We focus on on a subset of emotions identified by Pekrun and Ekman: enjoyment, surprise, frustration, and boredom. We also add confusion as an emotion, which is placed between enjoyment and frustration.

As described in [29] students can become overwhelmed (very confused or frustrated) during learning, which may increase cognitive load for low-ability or novice students. However, appropriate feedback can help to overcome such problems.

Published in CEUR-WS:
FFMI workshop (Schmidt-Thieme and Janning)
In EDM 2014 Extended Proceedings (Gutierrez-Santos and Santos)

Effective support or feedback needs to answer four main questions: *when*, *what*, *how*, and *why*: (i) *when* to provide the support during learning. (ii) It needs to be decided *what* the support should contain; (iii) *how* it should be presented; and (iv) *why* the feedback needs to be provided.

In this paper we focus on *what* (ii) and *why* (iv) support or feedback should be provided based on the student's emotion. In the area of intelligent tutoring systems or learning environments, the only research we are aware of specifically targeting the question of responding to student affect is [29] and [2]. [29] describes how an embodied pedagogical agent is able to provide different types of interventions, such as praising or mirroring the student's emotional state. [2] looks at the effect of cognitive-affective states on student's learning behaviour. In contrast, in this paper, we investigate the impact of emotions on the effectiveness of different feedback types.

The structure of the paper is as follows: The next section overviews related work on detecting and adapting to emotions in the educational domain. This is followed by a description of the Wizard-of-Oz study, which investigated the effect of emotions on different feedback types. We then discuss the different feedback types. After this, we provide results and discuss the results of the study in respect to adaptive support based on student's emotion. We conclude by outlining directions for future research.

2. BACKGROUND

Different computational approaches have been taken into account in order to detect emotions. These include for example, speech-based approaches (e.g. [6, 27]), using information from facial expressions (e.g. [14]), keystrokes or mouse movements [10], physiological sensors (e.g. [16, 28, 21]), or a combination of these [7].

In the area of education [5] developed a model of emotions (Dynamic Bayesian network) based on students' bodily expressions for an educational game. The system uses six emotional states: joy, distress, pride, shame, admiration and reproach. A pedagogical agent provides support according to the emotional state of the students and the user's personal goal, such as wanting help, having fun, learning maths, or succeeding by oneself.user's personal goal, such as wanting help, having fun, learning maths, or succeeding by oneself.

Another example, is [25] who also used Bayesian Networks to classify students' emotions. Here biophysical signals, such as heart rate, skin conductance, blood pressure, and EEG brainwaves, for the classification of emotions. These include: interest, engagement, confusion, frustration, boredom, hopefulness, satisfaction, and disappointment.

As described earlier, [29] developed an affective pedagogical agent which is able to mirror students' emotional state, or acknowledge a student's emotion if it is negative. They use hardware sensors and facial movements to detect students emotion. The system discriminates between seven emotions: high/low pleasure, frustration, novelty, boredom, anxiety, and confidence. Different machine learning techniques were applied for the classification, including Bayesian Networks and Hidden Markov models.

[17] developed a physics text-based tutoring system called ITSPOKE. It uses spoken dialogue to classify emotions. Acoustic-prosodic and lexical features are used to predict student emotion. They apply boosted decision trees for their classification. Three emotion types are detected: negative, neutral and positive emotions.

Another example is the AutoTutor tutoring system [7], which holds conversations with students in computer literacy and physics courses. The system classifies emotions based on natural language interaction, facial expressions, and gross body movements. The focus is on three emotions, namely frustration, confusion, and boredom. The classification is used to respond to students via a conversation.

Most of the related work in the educational domain focusses on detecting emotions based on different input stimuli, ranging from spoken dialogue to physiological sensors. However, little research has been done on how those detected emotions can be used in a tutoring system to enhance the learning experience. One exception is [29] who describes how an affective pedagogical agent can support students in particular emotional states. Additionally, [2] investigated the impact of student's cognitive-affective states on how they interacted with the learning environment. They found that certain types of emotions, such as boredom, were associated with poor learning and gaming the system. In contrast, we investigate the implications of emotions for different feedback types. We conducted a WoZ study where different kinds of feedback were provided to students in different emotional states. The next section describes the WoZ study in more detail.

2.1 Aims

One of our research aims is to provide adaptive feedback to students during a learning activity which enhances the learning experience by taking into account students' emotion. We were specifically interested in the following questions, which we aimed to address in the WoZ studies:

- Is there an effect of different emotion types upon reaction towards feedback?
- Which interventions were most successful given a particular emotional state?

In order to address these questions we ran an ecologically valid WoZ study which investigated the effect of emotions on different feedback types at different stages of the task.

2.2 Methodology

The studies reported on this paper are part of a methodology referred to as Iterative Communication Capacity Tapering (ICCT). This can be used to inform the design of intelligent support for helping students in interactive educational applications [18]. During the first phase, the facilitator gradually moves from a situation in which the interaction with the student is close, fast, and natural (i.e. face-to-face free interaction) towards a situation in which the interaction is mediated by computer technologies (e.g. voice-over-ip or similar for voice interaction, instant messaging or similar for

textual interaction) and regularised by means of a script. In the second phase, the script is crystallized into a series of intelligent components that produce feedback in the same way that the human facilitator formaly did. The gradual reduction of communication capacity and the iterative nature of the process maximise the probability of the computer-based support being as useful as the facilitator's help. In this paper, we are already starting the second phase, i.e. gradually replacing humans by a computer-based system. Experts ('wizards') are not physically near enough to the students to observe them directly, and therefore must observe them by indirect mediated means: the students' voice was heard by using microphones and headsets and their screen was observed by a mirror screen. The wizards did not have direct access to the students' screens (so e.g. could not point to anything on the screen to make a point), could not see the students' faces (for facial cues), and could not communicate to students by using body language, only by means of the facilities provided by the wizard-of-oz tools that resemble those of the final system.

2.3 Participants and Procedure

After returning informed consent forms signed by their parents 60 Year-5 (9 to 10-year old) students took part in a series of sessions with the learning platform configured for learning fractions through structured tasks from the intelligent tutoring system, together with more open-ended tasks offered by the exploratory learning environment. The sessions were designed to first familiarise all students with the environment, and then to allow them to undertake as many tasks as possible (in a study which has goals outside the scope of this paper). In parallel, we were running the WOZ study by asking two students in each session to work on different computers as described below. In total 12 students took part in the WOZ study but due to data errors we were able to analyse the interaction of only 10 students. At the end of the session the students who participated in the WOZ joined in a focus group discussing their experience with the learning platform. We were particularly interested in students' opinions about the different feedback types provided.

2.4 Classroom setup

The ecological validity of the study was achieved by following the setup depicted in Figure 1, 2 and Figure 3. The classroom where the studies took place is the normal computer lab of the school in which most of the computers are on tables facing the walls in a II-shape, and a few are on a central table. This is the place where the WOZ study took place, while, for ecological validity, the rest of the class was working on the other computers. The students were only told that the computers in the central isle were designed to test the next version of the system and were thus also responding to (rather than just recording as the rest of the computers) their speech. The central isle has two rows of computers, facing opposite directions, and isolated by a small separator for plugs etc. In the central isle the students worked on a console consisting on a keyboard, a mouse, and a screen. Usually, those components are connected to the computer behind the screen; for these studies, they were connected to a laptop on the wizards' side of the table. This allowed the wizard to observe what the students were doing. As the learning platform is a web-based system, and all the students' see is a web browser, the operating system and general look-and-feel of the experience was equivalent to the one that the rest of the students were using. When the wizards wanted to intervene, they used the learning platform's WOZ tools to send messages to the student's machine. These messages were both shown on screen and read aloud by the system to students, who could hear them on their headset.

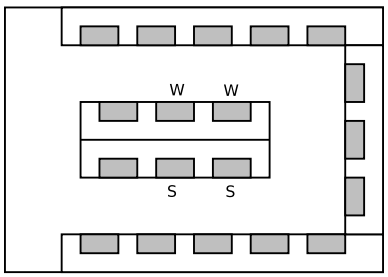


Figure 1: The layout. The Wizard-of-Oz studies took place on the central isle while the rest of the students worked on a version of the system which only sequences tasks and provides minimal support.

Figure 2: The classroom. The children being wizarded in front with wizards at the back.

2.5 The wizard's tools

In line with the ICCT methodology mentioned above, the wizards restricted their 'freedom' in addressing the students by employing a pre-determined agreed script in which the expected interventions had been written. Figure 4 shows a high-level view of this script, the end-points of which require further decisions also agreed in advance in a protocol but not shown here for simplicity. In this study, we limited ourselves to written interventions that could be selected from an online document appropriate for being read aloud by the system. There were no other kinds of interventions (such as sounds, graphical symbols on screen etc.). The intervention had a set of associated conditions that would fire them thus resembling very closely the system under development.

2.6 Feedback types

As outlined in the script (figure 4) 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.

From an HCI perspective speech production and recognition can provide potentially more intuitive interaction. In

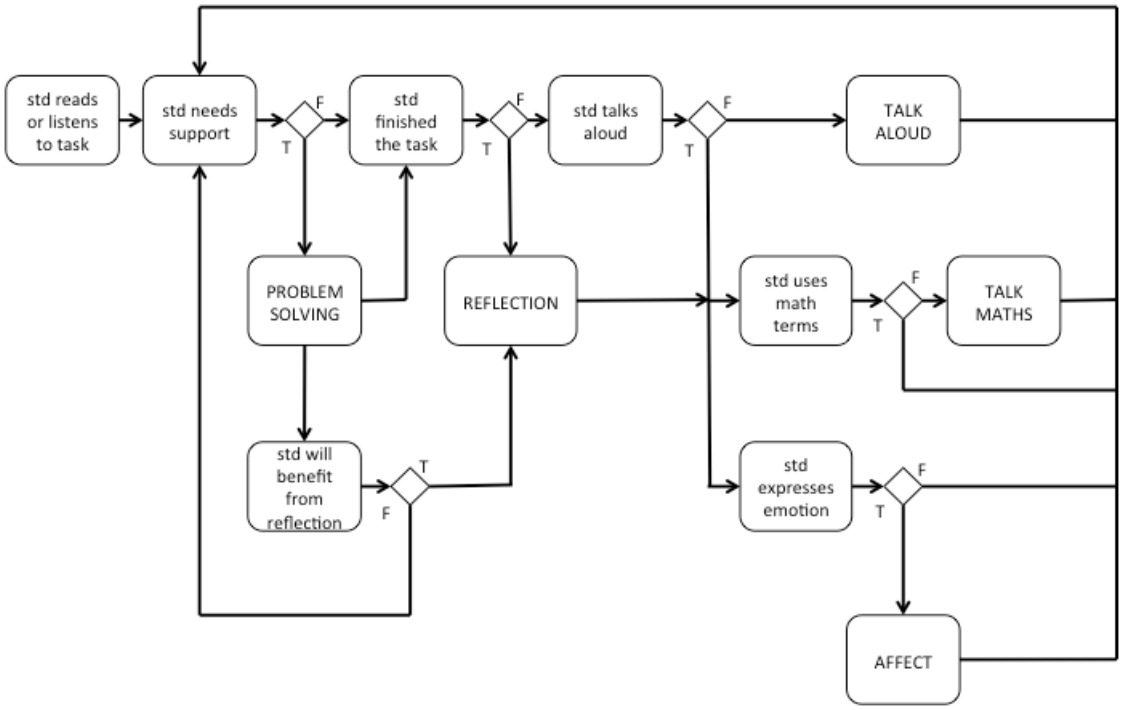


Figure 4: Flowchart representing the wizard's script for support.

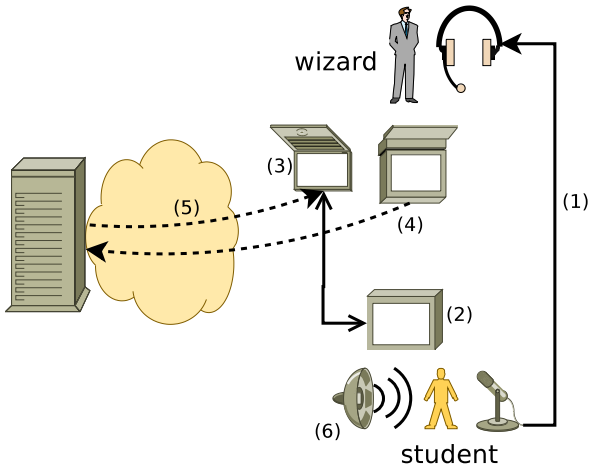


Figure 3: Wizard-of-oz setup. Each student speaks on a headset (mic) which is connected to the wizard's headset (1). The student interacts with a console (i.e. keyboard, mouse, screen) connected to a laptop on the wizard's side (2,3) so that the latter can witness their interaction. The wizard can send messages (4) by using some ad-hoc wizard tools. These messages arrive at the student laptop (5) and are shown on the screen of the student's monitor and read aloud on the student's headset (6).

particular, spoken language input can enable students to communicate verbally with an educational application and thus interact without using human interface devices such as a mouse or keyboard. The following different feedback types were provided:

- **PROBLEM SOLVING - task-dependent feedback**
  This feedback based mainly on the interaction with mouse and keyboard with the learning environment. Here the feedback involved providing support in solving a particular maths problem.
- **TALK MATHS - using particular domain specific maths vocabulary**
  The importance of students' verbal communication in mathematics in particular becomes apparent if we consider that learning mathematics is often like learning a foreign language. Focusing, for example, on learning mathematical vocabulary, [3] encouraged students to talk to a partner about a mathematical text to share confusions and difficulties, make connections, put text into their own words and generate hypotheses. This way, students were able to make their tentative thinking public and continually revise their interpretations.
- **AFFECT - affect boosts**
  As described in [29] affect boosts can help to enhance student's motivation in solving a particular learning task. Higher motivation also implies better performance.
- **TALK ALOUD - 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, 3, 20]).

The possible verbalization effect could be enhanced with ASR since cognitive load theory [26] and cognitive theory of multimedia learning [19] predict that a more natural and efficient form of communication will also have positive learning gains.

The few existing research studies have found mixed results with respect to whether the input modality (speaking vs. typing) has a positive, negative or no effect on learning. In [8], for example, the authors investigated whether student typing or speaking leads to higher computer literacy with the use of AutoTutor. They reported mixed results that highlight individual differences among students and a relationship to personal preferences and motivation.

- **REFLECTION - reflecting on task performance and learning**
  For further consideration is the research about self-explanation; an efficient learning strategy where students are prompted to verbalize their thoughts and explanations about the target domain to make knowledge personally meaningful. Previous research [13] found that the amount of self-explanation that students generated in a computer environment was suppressed by having learners type rather than speaking and the studies. Moreover, some students are natural self-explainers while others can be trained to self-explain [24]. Even when self-explanation is explicitly elicited, it can be beneficial [4] but requires going beyond asking students to talk aloud by using specific reflection prompts [24].

  Self-explanation can be viewed as a tool to address students' own misunderstandings [4] and as a 'window' into students' thinking. While it may be early days for accurate speech recognition to be able to highlight specific errors and misconceptions, undertaking carefully-designed tasks can help identify systematic errors that students make. For example, [12] explores how naming and misnaming involves logic and rules that often aid or hinder students' mathematical learning and relate to misconceptions.

  A lack of mathematical terminology can also be noticed and prompts made to students to use appropriate language as they self-explain.

Table 1 shows examples of the different feedback types. We were interested to explore how emotions impact on the effectiveness of those different feedback types.

3. RESULTS

From the WoZ study we recorded students' screen display and their voices. From this data, we annotated emotions and whether students reacted to feedback.

For the annotation of the emotions and students reactions towards the feedback, we used a similar strategy as described in [23] where dialog between a teacher and a student was annotated according to different feedback types. Also,[2] describe how they coded different cognitive-affective states based on observations of students interacting with a learning environment. Similarly, we annotated student's emotion and if they reacted for each type of feedback provided. Another researcher went through the categories and any discrepancies were discussed and resolved before any analysis took place.

Feedback type	Example
AFFECT	It may be hard, but keep trying. If you find this easy, check your work and change the task.
TALK ALOUD	Remember to talk aloud, what are you thinking? What is the task asking you to do?
TALK MATHS	Can you explain that again using the terms denominator, numerator?
PROBLEM SOLVING	You can't add fractions with different denominators.
REFLECTION	What did you learn from this task? What do you notice about the two fractions?

Table 1: Examples of feedback types

In total 170 messages were sent to 10 students. The raw video data was analysed by a researcher who categorised the emotions and feedback messages. Table 1 shows the different types of messages send to students and the emotions that occurred while the feedback was given. It can be seen that most frequent messages were reminders to talk aloud (66). This was followed by problem-solving feedback (55), and feedback according to students emotions (31). The least frequent messages relates to reflection (13) and using maths terminology (5).

It is not surprising that most of the problem solving feedback was provided when students were confused (35 out of 55). Most of the affect boosts were provided when students enjoyed the activity (15 out of 31), closely followed by students' being confused (11 out of 31). Most of the reflection prompts were given when students enjoyed the activity (10 out of 13). Talk aloud reminders were mainly given when students were confused (30 out of 66). Talk maths prompts were mainly given when students enjoyed the task (3 out of 5) or when they were confused (2 out of 5).

The emotions that were detected by students when feedback was provided and whether students reacted can be seen in figure 5.

Students reacted to all of the feedback when they were bored or surprised (100%).This was followed by reactions to feedback when students were confused (83%) or enjoyed the activity (81%). Students responded the least if they were frustrated (69%).

Looking in more detail at emotions and whether students reacted to the different feedback types, figures 6, 7, and 8 show the percentage of student's reaction towards feedback type for enjoyment, confusion, and frustration.

	emotion					
Feedback type	enjoyment	boredom	confusion	frustration	surprise	total
PROBLEM SOLVING	8	3	35	8	1	55
TALK MATHS	3	0	2	0	0	5
AFFECT	15	2	11	3	0	31
TALK ALOUD	21	1	40	4	0	66
REFLECTION	10	1	1	1	0	13
Total	57	7	89	16	1	170

Table 2: Feedback types, including emotion that occurred while the feedback was provided.

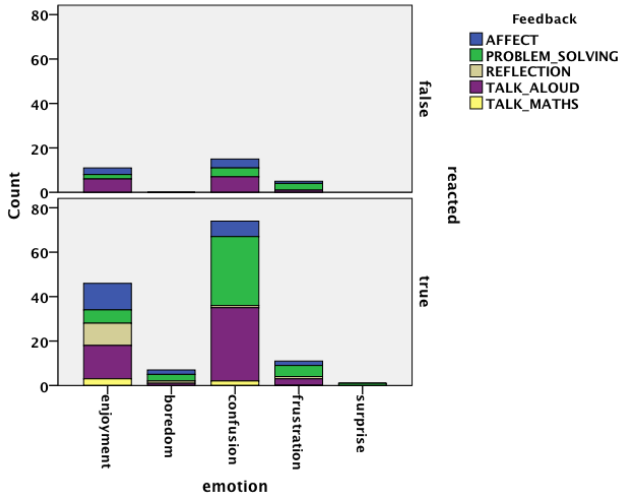


Figure 5: Student's reaction according to feedback types and emotion.

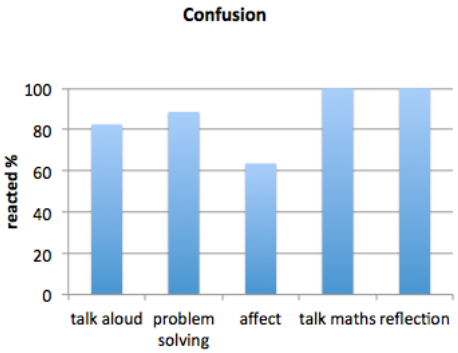


Figure 7: Students' reaction according to feedback types if they were confused.

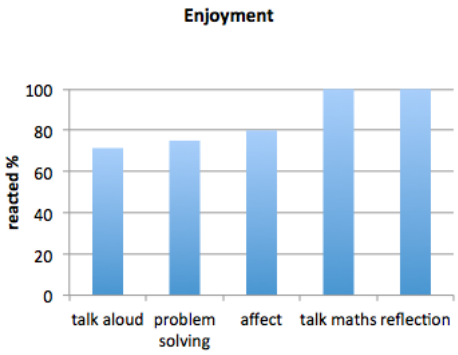


Figure 6: Students' reaction according to feedback types if they enjoyed the activity.

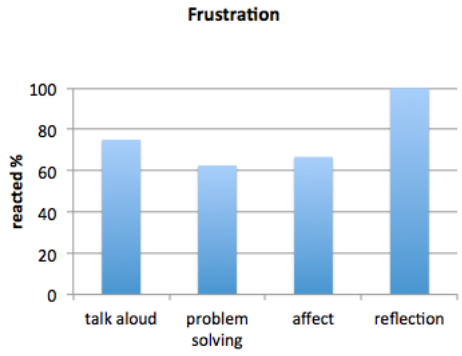


Figure 8: Students' reaction according to feedback types if they were frustrated.

It is interesting to see that while students enjoyed their activity, they responded very well to talk maths (100%) or to reflect on what they have done (100%). The least reaction was given if students were prompted to talk aloud (71%).

If students were confused they responded well again on talk maths (100%) or reflection prompts (100%), followed by problem solving feedback (89%). Surprisingly, least reactions were given when affect boosts were provided (64%).

If students were frustrated most reactions were given for reflection (100%) and prompts to talk aloud (75%). Least responses were given if problem solving feedback was provided (63%).

4. DISCUSSION

The key findings with respect to impact of emotions on the effect of feedback types are listed below in relation to our research aims.

4.1 Is there an effect of different emotion types upon reaction towards feedback?

The results show that for certain types of emotions, such as boredom, any type of feedback is reacted to. This indicates that students may welcome a distraction from their learning and react to feedback if they are bored. As boredom indicates a reduction in learning [2], the feedback provided to students when they are bored should aim to motivate and support the student to continue with the learning task.

Also in most of the cases students reacted to the feedback when they were confused. This implies that students welcome feedback that will help them to get out of their confused state. In designing feedback for learning environments students should be provided with feedback that enables them to overcome their confusion, such as task-dependent problem solving feedback, or feedback to reflect on their learning, which might help to identify and overcome misconceptions.

Additionally, students mainly reacted to feedback when they were enjoying their activity. This is an interesting finding, as in theory this seems to interrupt their learning flow. Here, it seems students' motivation is high and they did not mind being interrupted. Students particularly reacted positively on feedback to reflect.

In contrast, when students were frustrated, they reacted to feedback in only 69% of the cases. This indicates that frustration can reduce motivation and may also increase cognitive load. Here feedback that might help to decrease the frustration, such as reflecting on the difficulty of the learning task might help to motivate the student.

4.2 Which interventions were most successful given a particular emotional state?

The results indicate that for different emotional states, different feedback types are more effective than others.

It is interesting to see that although students enjoyed their activity and reacted to feedback in 81% of the cases, response to talk aloud was only 71%. This was similar when students were frustrated (75%). In contrast when students were confused in 83% of the cases students followed the recommendation to talk aloud. It looks like as if talking aloud might help to identify the problem and might resolve the confusion.

The highest reaction was given to problem solving feedback if students were confused (89%). This is not surprising as students were happy to receive help to perform the task. However, in only 75% of the cases was problem solving feedback reacted to while students enjoyed the activity. This might be because they were interrupted in their learning flow and they needed to switch to a new strategy of answering the learning task based on the problem solving feedback. The number drops even more when students were frustrated (63%). As discussed above, students' motivation might be low when frustrated and also there might be increased cognitive load. Providing problem solving feedback when students are frustrated does not seem to be a very effective strategy.

Providing affect boosts was most effective when students enjoyed their activity (80%). In contrast, students only reacted to affect boosts in 67% of the cases when they were frustrated or 64% when they were confused. From the focus group with the students it emerged that although some students did not react to the emotional boosts when they were confused or frustrated, they liked the encouragement, and that it helped with their motivation to continue to work on the particular learning task.

Providing prompts to talk maths and reflection were very effective across the emotion types. Despite the fact that 5 talk maths prompts and 13 reflection prompt were provided, students seemed to respond to them very well whether confused or frustrated. This implies that reflecting on one's own strategy of solving a task is motivating even if confused or frustrated. We noticed that it may also helped students to identify misconceptions or lead to new ideas on how to solve the learning task.

5. CONCLUSION AND FUTURE WORK

We explored the impact of students' emotional state upon different feedback types. The results indicate that certain types of feedback are more effective then others according to the emotional state of the student. While for some emotional states, such as boredom, a variety of feedback types worked well, for other emotional states, like frustration, only a few types of feedback seem to be effective.

We are now developing and integrating the automatic speech and emotion recognition in our learning platform. Additionally the adaptive support that is able to provide the different feedback types for particular emotional states is under development. At the next stage of our research we are interested to explore how the presentation of the feedback (e.g. high or low intrusive) affects students being interrupted in performing the task and if the presentation has an effect on reaction towards the feedback.

6. ACKNOWLEDGMENTS

This research has been co-funded by the EU in FP7 in the iTalk2Learn project (318051). Thanks to all our iTalk2Learn colleagues for their support and ideas.

7. REFERENCES

[1] M. Askeland. Sound-based strategy training in multiplication. *European Journal of Special Needs Education*, 27(2):201–217, 2012.

[2] R. S. J. d. Baker, S. K. D'Mello, M. T. Rodrigo, and A. C. Graesser. 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, apr 2010.

[3] R. Borasi, M. Siegel, J. Fonzi, and C. Smith. Using transactional reading strategies to support sense-making and discussion in mathematics classrooms: An exploratory study. *Journal for Research in Mathematics Education*, 29:275–305, 1998.

[4] M. Chi. Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser, editor, *Advances in instructional psychology*, pages 161–238. Mahwah, NJ: Lawrence Erbaum Associates, 2000.

[5] C. Conati and H. MacLaren. Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction*, 2009.

[6] R. Cowie, E. Douglas-Cowie, B. Apolloni, J. Taylor, A. Romano, and W. Fellenz. What a neural net needs to know about emotion words. *Computational Intelligence and Applications*, pages 109–114, 1999.

[7] S. D'Mello, S. Craig, B. Gholson, S. Franklin, R. Picard, and A. Graesser. Integrating affect sensors in an intelligent tutoring system. In *Affective Interactions: The Computer in the Affective Loop Workshop at 2005 International Conference on Intelligent User Interfaces*, pages 7–13, 2005.

[8] S. K. D'Mello, N. Dowell, and A. Graesser. Does it really matter whether student's contributions are spoken versus typed in an intelligent tutoring system with natural language? *Journal of Experimental Psychology; Applied 2011*, 17(1):1–17, 2011.

[9] P. Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, 1992.

[10] C. Epp, M. Lippold, and R. Mandryk. Identifying emotional states using keystroke dynamics. In *2011 Annual Conference on Human Factors in Computing Systems*, pages 715–724, 2011.

[11] J. H. Flavell, F. L. Green, E. R. Flavell, and J. B. Grossman. The Development of Children's Knowledge of Inner Speech. *Child Development*, 68(1):39–47, 1997.

[12] H. A. Furani. Misconceiving or misnaming?: Some implications of toddlers' symbolizing for mathematics education. *Philosophy of Mathematics Education Journal*, 17, 2003.

[13] R. G. M. Hausmann and M. T. H. Chi. Can a computer interface support self-explaining? *Cognitive Technology*, 7(1):4–14, 2002.

[14] R. Kaliouby and P. Robinson. Real-time inference of complex mental states from facial expressions and head gestures. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshop*, 2004.

[15] B. Kort, R. Reilly, and R. Picard. An affective model of the interplay between emotions and learning. In *IEEE International Conference on Advanced Learning Technologies*, number 43-46, 2001.

[16] P. Lang, M. Greenwald, M. Bradley, and A. Hamm. Look at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, 30:261–273, 1993.

[17] D. Litman and K. Forbes-Riley. Predicting student emotions in computer-human tutoring dialogues. In *42nd Annual Meeting on Association for Computational Linguistics (ACL '04), Association for Computational Linguistics*, 2004.

[18] M. Mavrikis and S. Gutierrez-Santos. Not all wizards are from Oz: Iterative design of intelligent learning environments by communication capacity tapering. *Computers & Education*, 54(3):641–651, Apr. 2010.

[19] R. E. Mayer and R. Moreno. Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, 38(1):43–52, 2003.

[20] N. Mercer and C. Sams. Tecahing children how to use language to solve maths problems. *Language and Education*, 20(6):507–528, 2007.

[21] F. Nasoz, K. Alvarez, C. Lisetti, and N. Finkelstein. Emotion recognition from physiological signals for presence technologies. *International Journal of Cognition, Technology and Work, Special Issue on Presence*, 6(1):4–14, 2003.

[22] R. Pekrun. The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *J. Edu. Psych. Rev.*, pages 315–341, 2006.

[23] K. Porayska-Pomsta, M. Mavrikis, and H. Pain. Diagnosing and acting on student affect: the tutor's perspective. *User Modeling and User-Adapted Interaction*, 18(1):125–173, Feb. 2008.

[24] M. Roy and M. T. H. Chi. The self-explanation principle in multimedia learning. In R. E. Mayer, editor, *Cambridge handbook of multimedia learning*, pages 271–286. New York: Cambridge University Press, 2005.

[25] L. Shen, M. Wang, and R. Shen. ffective e-learning: Using "emotional" data to improve learning in pervasive learning environment. *Educational Technology & Society*, 12(2):176–189, 2009.

[26] J. Sweller, J. G. van Merrienboer, and G. W. Paas. Cognitive Architecture and Instructional Design. *Educational Psychology Review*, 10:251 – 296+, 1998.

[27] T. Vogt and E. André. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *Multimedia and Expo (ICME05)*, pages 474–477, 2005.

[28] E. Vyzas and R. Picard. Affective pattern classification. In *AAAI Fall Symposium, Emotional and Intelligent: The Tangled Knot of Cognition*, pages 176–182, 1998.

[29] B. Woolf, W. Burleson, I. Arroyo, T. Dragon, D. Cooper, and R. Picard. Affect-aware tutors: recognising and responding to student affect. *Int. J. Learning Technology*, 4(3-4):129–164, 2009.

Multimodal Affect Recognition for Adaptive Intelligent Tutoring Systems

Ruth Janning
Information Systems and Machine Learning Lab (ISMLL)
University of Hildesheim
Marienburger Platz 22, 31141 Hildesheim, Germany
janning@ismll.uni-hildesheim.de

Carlotta Schatten
Information Systems and Machine Learning Lab (ISMLL)
University of Hildesheim
Marienburger Platz 22, 31141 Hildesheim, Germany
schatten@ismll.uni-hildesheim.de

Lars Schmidt-Thieme
Information Systems and Machine Learning Lab (ISMLL)
University of Hildesheim
Marienburger Platz 22, 31141 Hildesheim, Germany
schmidt-thieme@ismll.uni-hildesheim.de

ABSTRACT

The performance prediction and task sequencing in traditional adaptive intelligent tutoring systems needs information gained from expert and domain knowledge. In a former work a new efficient task sequencer based on a performance prediction system was presented, which only needs former performance information but not the expensive expert and domain knowledge. In this paper we aim to support this approach by automatically gained multimodal input like for instance speech input from the students. Our proposed approach extracts features from this multimodal input and applies to that features an automatic affect recognition method. The recognised affects shall finally be used to support the mentioned task sequencer and its performance prediction system. Consequently, in this paper we (1) propose a new approach for supporting task sequencing and performance prediction in adaptive intelligent tutoring systems by affect recognition applied to multimodal input, (2) present an analysis of appropriate features for affect recognition extracted from students speech input and show the suitability of the proposed features for affect recognition for adaptive intelligent tutoring systems, and (3) present a tool for data collection and labelling which helps to construct an appropriate data set for training the desired affect recognition approach.

Keywords

multimodal input, affect recognition, feature analysis, speech, adaptive intelligent tutoring systems

1. INTRODUCTION

Learning management systems like intelligent tutoring systems are an important tool for supporting the education of students for instance in learning fractional arithmetic. The main advantages of intelligent tutoring systems are the possibility for a student to practice any time, as well as the possibility of adaptivity and individualisation for a single student. An adaptive intelligent tutoring system possesses an internal model of the student and a task sequencer which decides which tasks in which order are shown to the student. Originally, the task sequencing in adaptive intelligent tutoring systems is done using information gained from expert and domain knowledge and logged information about the performance of students in former exercises. In [12] a new efficient sequencer based on a performance prediction system was presented, which only uses former performance information from the students to sequence the tasks and does not need the expensive expert and domain knowledge. This approach applies the machine learning method matrix factorization (see e.g. [1]) for performance prediction to former performance information. Subsequently, it uses the output of the performance prediction process to sequence the tasks according to the theory of Vygotsky's Zone of Proximal Development [14]. That is the sequencer chooses the next task in order to neither bore nor frustrate the student or in other words, the next task should not be too easy or too hard for the student.

In this paper we propose to support the task sequencer and performance prediction system of the approach in [12] in a new way by further automatically to get and process multimodal information. One part of this multimodal information, which is investigated in this paper, is the speech input from the students interacting with the intelligent tutoring system while solving tasks. A further part will be the typed input or mouse click input from the students, which will be reported in upcoming works. The approach proposed in this paper extracts features from the mentioned multimodal information and applies to that features an automatic affect recognition method. The output of the affect recognition method indicates, if the last task was too easy, too hard or appropriate for the student. This information matches the theory of Vygotsky's Zone of Proximal Development, hence it is obviously suitable for supporting the performance prediction system and task sequencer of the approach in [12]. However, for the proposed approach we need a large amount

Published in CEUR-WS:
FFMI workshop (Schmidt-Thieme and Janning)
In EDM 2014 Extended Proceedings (Gutierrez-Santos and Santos)

of labelled data. For this reason we developed a tutoring tool which (a) records students speech input as well as typed input and mouse click input and (b) allows the students to label by themselves how difficult they perceived the shown tasks. This tool is presented in the second part of this paper and will be used to conduct further studies to gain the desired labelled data.

The main contributions of this paper are: (1) presentation of a new approach for supporting performance prediction and task sequencing in adaptive intelligent tutoring systems by affect recognition on multimodal input, (2) identification and analysis of appropriate and statistically significant features for the presented approach, and (3) presentation of a new tutoring tool for multimodal data collection and self-labelling to gain automatically labelled data for training appropriate affect recognition methods.

In the following, first we will present some preliminary considerations along with state-of-the-art in section 2. Subsequently, we will describe in section 3 the real data set used for the feature analysis and investigate in section 4 for the data set the correlation between students affects and their performance. In section 5 we will propose and analyse appropriate features for affect recognition and in section 6 we will explain how to support performance prediction and task sequencing in intelligent tutoring systems by affect recognition applied to multimodal input. Before we conclude, we will describe in section 7 the mentioned tool for multimodal data collection and self-labelling.

2. PREPARATION AND RELATED WORK

Before an automatic affect recognition approach can be applied, one has to clarify three things: (1) What kind of features shall be used, (2) what kind of classes shall be used and (3) which instances shall be mapped to features and labelled with the class labels. After deciding which features, classes and instances shall be considered, one can apply affect recognition methods to these input data. In the following subsections we will present possible features, classes, instances and methods for affect recognition supporting performance prediction and task sequencing in adaptive intelligent tutoring systems along with the state-of-the-art.

2.1 Features

The first step before applying automatic affect recognition is to identify useful features for this process. For the purpose to recognise affect in speech one can use two different kinds of features ([13]): acoustic and linguistic features. Further, one can distinct linguistics (like n-grams and bag-of-words) and disfluencies (like pauses). If linguistics features are used, a transcription or speech recognition process has to be applied to the speech input before affect recognition can be conducted. Subsequently, approaches from the field of sentiment classification or opinion mining (see e.g. [10]) can be applied to the output of this process. However, the methods of this field have to be adjusted to be applicable to speech instead of written statements.

Another possibility for speech features is to use disfluencies features like it was done in [17], [7] and [4] for expert identification. The advantage of using such features is that instead of a full transcription or speech recognition approach only for instance a pause identification has to be applied before. 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 affects. For using this kind of features one has to investigate, which particular features are suitable for the special task of affect classification in adaptive intelligent tutoring systems. Because of the mentioned advantage of disfluencies features in this work we focus on features extracted from information about speech pauses as one part of the multimodal input for affect recognition.

As mentioned in the introduction the other part of the multimodal input will be features which are gained from information about typed input or mouse click input from the students. This kind of features is similar to the keystroke dynamics features used in [2]. In [2] emotional states were identified by analysing the rhythm of the typing patterns of persons on a keyboard.

2.2 Classes

The second step before applying automatic affect recognition is to define the classes corresponding to emotions and affective states, which shall be recognised by the used affect recognition approach. According to [6], [5] and [16] it is possible to recognise in intelligent tutoring systems students affects like for instance confusion, frustration, boredom and flow. As mentioned above, we want to use the students behaviour information gained from speech and from typed input or mouse click input for supporting the performance prediction system and task sequencer of the approach in [12], which is based on the theory of Vygotsky's Zone of Proximal Development [14]. That means that the goal is to neither bore the student with too easy tasks nor to frustrate him with too hard tasks, but to keep him in the Zone of Proximal Development. Accordingly, we want to use the output of the automatic affect recognition to get an answer to the question "Was this task too easy, too hard or appropriate for the student?", or with other words we want to find out if the student felt under-challenged, over-challenged or like to be in a flow. However, the mapping between confusion, frustration, boredom and under-challenged, over-challenged is not unambiguous as one can infer e.g. from the studies mentioned in [16]. Hence, we will use instead of the above mentioned affect classes three other classes for supporting performance prediction and task sequencing by automatic affect recognition: under-challenged, over-challenged and flow. One could summarise these classes as *perceived task-difficulty classes*, as we aim to recognise the individual perceived task-difficulty from the view of the student.

2.3 Instances

The third step before applying automatic affect recognition is deciding which instances shall be mapped to features and labelled with the class labels. If the goal of the affect recognition is to provide a student motivation or hints according to his affective state like e.g. in [16], then instances can be utterances. For supporting performance prediction and task sequencing by affect recognition instead one needs at the end of a task the information, if the task overall was too easy, too hard or appropriate for the student. The reason is that this information shall help to choose the next task shown to the student. Hence, an instance for supporting perfor-

mance prediction and task sequencing by affect recognition has to be instead of an utterance the whole speech input of a student for one task.

2.4 Methods

The possible methods for an automatic affect recognition depend on the kind of the features used as input. As mentioned above, for speech we distinct two kinds of features: linguistics features and disfluencies. Linguistics features are gained by a preceding speech recognition process and can be processed by methods coming from the areas sentiment analysis and opinion mining ([10]). Especially methods from the field of opinion mining on microposts seem to be appropriate if linguistics features are considered. State-of-the-art approaches in opinion mining on microposts use methods for instance based on optimisation approaches ([3]) or Naive Bayes ([11]).

The process of gaining disfluencies like pauses is different to the full speech recognition process. For extracting for instance pauses usually an energy threshold on the decibel scale is used as in [4] or an SVM is applied for pause classification on acoustic features as in [9]. Appropriate state-of-the-art methods for automatic emotion and affect recognition on disfluencies features as well as on features from information about typed input or mouse click input are – as proposed e.g. in [13] and [6] – classification methods like artificial neural networks, SVM, decision trees or ensembles of those.

3. REAL DATA SET

After identifying features, classes, instances and methods for affect recognition for supporting performance prediction and task sequencing like above one can collect data for a concrete feature analysis and a training of the chosen affect classification method. We conducted a study in which the speech and actions of ten 10 to 12 years old German students were recorded and students affective states as well as the perceived task-difficulties were reported. The labelling of these data was done on the one hand concurrently by the tutor and on the other hand retrospectively by a second reviewer. Furthermore, a labelling per exercise (consisting of several subtasks) and an overall labelling per student as an aggregation of the labels per exercise was done. During the study a paper sheet with fraction tasks was shown to the students and they were asked to paint (with the software Paint) and explain their observations and answers. 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 speech recordings shall be used to gain the input for affect recognition. The mentioned typed input or mouse click input information we will collect and investigate in further studies with the self-labelling and multimodal data collection tutoring tool described in section 7.1. In this paper we focus on speech features and hence in section 5 we will propose and analyse possible features extracted from speech pauses. But first we will investigate in the following section 4 the correlation between perceived task-difficulty labels and the performance of the students in the real data set.

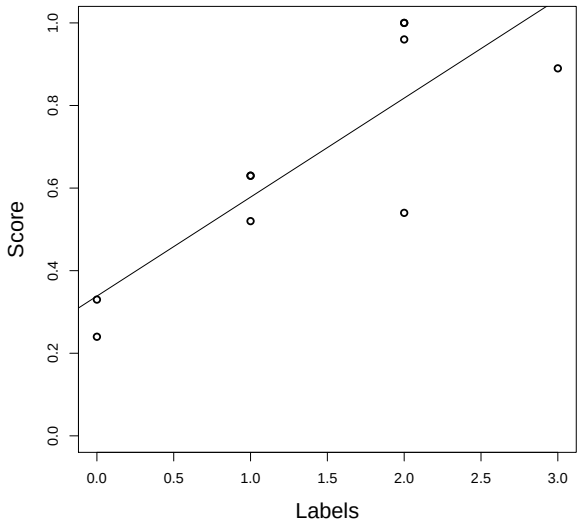


Figure 1: Mapping of the perceived task-difficulty labels to the scores of the students in the real data set.

4. CORRELATION OF PERCEIVED TASK-DIFFICULTY LABELS AND SCORE

Before we present speech features for recognising perceived task-difficulty, we want to show that there is a correlation between the proposed perceived task-difficulty labels and the performance of the students, to underline the suitability of supporting performance prediction and task sequencing by the proposed affect recognition approach. Hence, we mapped the overall perceived task-difficulty labels to the overall score of the students (see figure 1). For this mapping we encoded the different overall perceived task-difficulty class labels as follows:

- 0 = over-challenged
- 1 = over-challenged/flow
- 2 = flow
- 3 = flow/under-challenged
- 4 = under-challenged

The overall score of a student i is computed by

$$\frac{n_{c_i}}{n_{t_i}}\ , \tag{1}$$

where n_{c_i} is the number of correctly solved tasks of student i and n_{t_i} is the number of tasks shown to student i. In figure 1 one can see that there is a clear correlation between perceived task-difficulty labels and score. To substantiate this observation we applied a statistical test by conducting a linear regression and measuring the p-value, indicating the statistical significance, as well as the R^2 and Adjusted R^2 value, indicating how well the regression line can approximate the real data points. This approach delivers a p-value of 0.0027,

Figure 2: Graphic of the decibel scale of an example sound file of a student. The two straight horizontal lines indicate the threshold.

a R^2 value of 0.6966, and an Adjusted R^2 value of 0.6586. The small p-value indicates a strong statistical significance. The significant correlation between perceived task-difficulty labels and scores, which demonstrate the performance, indicates that it makes sense to support performance prediction and task sequencing by perceived task-difficulty classification.

5. SPEECH FEATURE ANALYSIS

The features we propose and analyse in this section are gained from speech pauses. Hence, first one has to identify pauses within the speech input data. The most easy way is to define a threshold on the decibel scale as done e.g. in [4]. For our preliminary study of the data we also used such a threshold, which we adjusted by hand. More explicitly, we extracted the amplitudes of the sound files and computed the decibel values. Subsequently, we investigated which decibel values belong to speech and which ones to pauses (see figure 2). In larger data and in the application phase later on, one has to learn automatically the distinction between speech and pauses by either learn the threshold or train an SVM, which classifies speech and pauses.

5.1 Single Feature Analysis

Before we can introduce the features we want to investigate, we have to define some measurements:

- m: number of students
- p_i: total length of pauses of student i
- s_i: total length of speech of student i
- n_{p_i}: number of pause segments of student i
- n_{s_i}: number of speech segments of student i
- $p_i^{(x)}$: xth pause segment of student i
- $s_i^{(y)}$: yth speech segment of student i
- n_{t_i}: number of tasks shown to student i
- n_{c_i}: number of correctly solved tasks by student i
- Overall score for student i: $\frac{n_{c_i}}{n_{t_i}}$

Table 1: p-value, R^2 and Adjusted R^2 for the feature *Length of maximal pause segment* mapped to score as well as to label.

Mapped to	p-value	R^2	Adjusted R^2
Score	0.1156	0.2802	0.1902
Label	0.0678	0.3577	0.2774

Our data set exists of acoustic recordings from m students, each of which saw n_{t_i} tasks and solved n_{c_i} tasks correctly. The overall score of a student i in this case is the number of correctly solved tasks n_{c_i} divided by the number of seen tasks n_{t_i}. After applying the above mentioned threshold to the data, we get for each student i the total length of pauses p_i and the total length of speech s_i in his acoustic recoding. Furthermore, we can count connected pause and speech segments to get the number of pause segments n_{p_i} and speech segments n_{s_i} of a student i. The xth pause segment is then $p_i^{(x)}$ and the yth speech segment $s_i^{(y)}$. By means of these measurements and their combination we can create a set of features useful for affect recognition supporting performance prediction and task sequencing:

- Ratio between pauses and speech ($\frac{p_i}{s_i}$)
- Frequency of speech pause changes ($\frac{n_{p_i}+n_{s_i}}{\max_j(n_{p_j}+n_{s_j})}$)
- Percentage of pauses of input speech data ($\frac{p_i}{(p_i+s_i)}$)
- Length of maximal pause segment ($\max_x(p_i^{(x)})$)
- Length of average pause segment ($\frac{\sum_x p_i^{(x)}}{n_{p_i}}$)
- Length of maximal speech segment ($\max_y(s_i^{(y)})$)
- Length of average speech segment ($\frac{\sum_y s_i^{(y)}}{n_{s_i}}$)
- Average number of seconds needed per task ($\frac{(p_i+s_i)}{n_{t_i}}$)

The ratio between the total length of pauses and the total length of speech indicates, if one one them is notable larger than the other one, i.e. if the student made much more speech pauses than speaking or vice versa. The frequency of speech and pause segment changes indicates, if there are many short speech and pauses segments or just a few large ones and it is normalised by dividing it by the maximal sum of pause and speech segments over all students. From the percentage of pauses one can see if the total pause length was much larger than the total speech part, i.e. the student did not speak much but was more thinking silently. The length of maximal pause or speech segment indicates if there was e.g. a very long pause segment where the student was thinking silently or a very long speech segment where the student was in a speech flow. The length of average pause or speech segment give us an idea of how much on average the student was in a silent thinking phase or a speech flow. The average number of seconds needed per task indicates how long a student on average needed for solving a task.

To investigate, if these features are suitable to describe perceived task-difficulty as well as performance in our real data

Table 2: p-value, R^2 and Adjusted R^2 for the best combinations of features (with a p-value smaller than 0.05) of a set with 6, 5, 4 or 3 features mapped to the score.

#	Features	p-val.	R^2	Adj. R^2
6	Frequency of changes, seconds per task, max. length of pause, average length of pause, max. length of speech average length of speech	0.0439	0.9516	0.8548
5	Frequency of changes, seconds per task, max. length of pause, average length of pause, average length of speech	0.0105	0.9496	0.8867
4	Frequency of changes, seconds per task, average length of pause, average length of speech	0.0415	0.8207	0.6773
3	Frequency of changes, frequency of changes, average length of speech	0.0431	0.719	0.5786

Table 3: p-value, R^2 and Adjusted R^2 for the best combinations of features (with a p-value smaller than 0.05) of a set with 5, 4, 3 or 2 features mapped to the labels.

#	Features	p-val.	R^2	Adj. R^2
5	Ratio pause speech, frequency of changes, seconds per task, average length of pause, average length of speech	0.0284	0.9158	0.8106
4	Ratio pause speech, frequency of changes, average length of pause, average length of speech	0.0154	0.8818	0.7872
3	Ratio pause speech, frequency of changes, average length of speech	0.0117	0.8207	0.7311
2	Frequency of changes, average length of speech	0.0327	0.6238	0.5163

set, we mapped the values of each feature to the score as well as to the perceived task-difficulty labels. Subsequently, we applied a linear regression to measure the p-value as well as the R^2 and Adjusted R^2 value. However, as expected, single features are not very significant. The feature with the best values for p-value, R^2 and Adjusted R^2 – mapped to score as well as to labels – is the *Length of maximal pause segment*. The statistical values for this feature are shown in table 1. These values are not very satisfactory, as one would desire a p-value smaller than 0.05 and values for R^2 and Adjusted R^2 which are closer to 1. A more reasonable approach is to combine several features instead of considering just one feature. Hence, in the following section we will investigate different combinations of features.

5.2 Feature Combination Analysis

We analysed different combinations of features by applying a multivariate linear regression to them to gain the p-value, R^2 and Adjusted R^2 for these combinations. The investigated combinations are combinations where all features are not strongly correlated, i.e. whenever we had two correlated features we put just one of them into the feature set for that combination. In further steps we removed from the considered feature sets feature by feature. Furthermore, in the multivariate linear regression we mapped the features on the one hand to the score and on the other hand to the labels. The results of the best combinations, i.e. such with a p-value at least smaller than 0.05, are shown in table 2 and 3. For the score there were no combinations with only 2 features with a p-value smaller than 0.05, hence in table 2 we just listed the best combinations with 3 up to 6 features. For the labels instead there were no such combinations, which have a p-value smaller than 0.05, with 6 features, so that in table 3 we only listed the best combinations of 2 up to 5 features. For both (score and labels) there are statistically significant feature combinations. That means that our proposed features are able to describe the score as well as the labels.

6. SUPPORTING PERFORMANCE PREDICTION AND SEQUENCING

As mentioned in the introduction, our goal is to support the performance prediction system and task sequencer of the approach in [12] by affect recognition, or by multimodal input respectively. Hence, in the following we will propose how to realise this support. In figure 3 a block diagram of the approach of supporting performance prediction and task sequencing by means of affect recognition is presented. The approach in [12] is represented in figure 3 by the non-dotted arrows: the performance prediction gets input from former performances and computes by means of the machine learning method matrix factorization predictions for future performances, which are the input for the task sequencer. The task sequencer decides based on the theory of Vygotsky's Zone of Proximal Development from the performance prediction input which task shall be shown next to the student. This process can be supported by the multimodal input as follows:

(1) The additional input for the performance predictor can be the output of the affect recognition, i.e. the perceived task-difficulty labels. In this case the performance predictor can take the perceived task-difficulty of the last task ($T^{(t)}$) to use the following rules for deciding how difficult the next task ($T^{(t+1)}$) should be:
   - If $T^{(t)}$ was too easy (label *under-challenged* or *flow/under-challenged*), then $T^{(t+1)}$ should be harder.
   - If $T^{(t)}$ was appropriate (label *flow*), then $T^{(t+1)}$ should be similar difficult.
   - If $T^{(t)}$ was too hard (label *over-challenged* or *over-challenged/flow*), then $T^{(t+1)}$ should be easier.

(2) The values of the features gained by feature extraction from speech, typed input and mouse click input

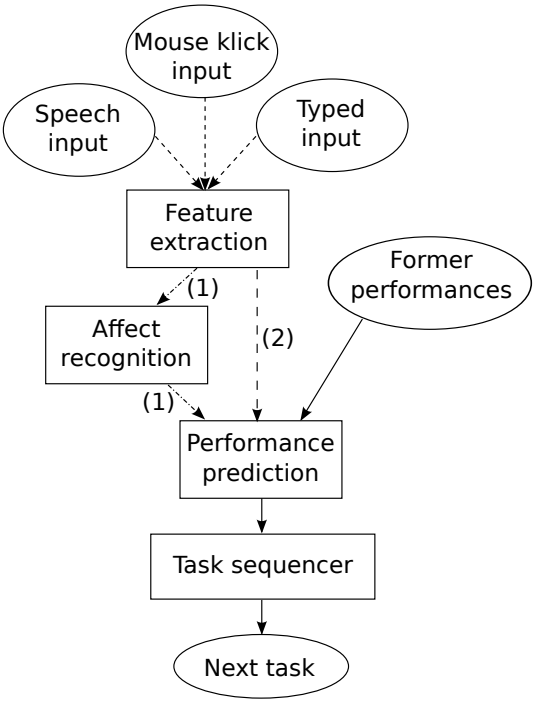


Figure 3: Approach for supporting performance prediction and task sequencing by means of multimodal input and affect recognition.

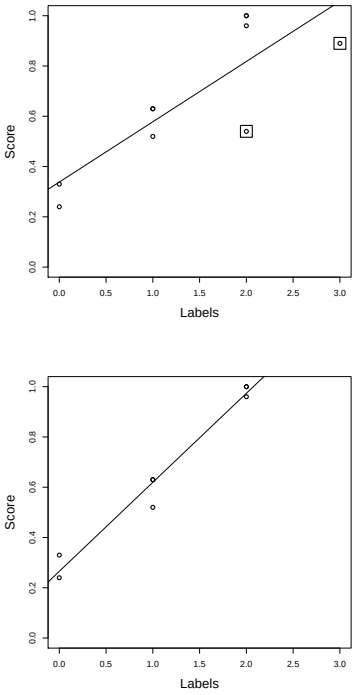


Figure 4: Mapping of the perceived task-difficulty labels to the scores of the students in the real data set (a) with outliers indicated by surrounding rectangles (top) and (b) without outliers (bottom).

can be fed directly into the performance prediction without applying an affect recognition. That means that the features are mapped to scores instead of perceived task-difficulty classes. That this makes sense was shown in section 4 and 5. The performance predictor can then compare e.g. the differences between performances, expressed as *score*, and the scores computed by means of the features ($\widehat{score}$). This difference indicates outliers like if a student felt to be in a flow or under-challenged but his score is worse, i.e. $\widehat{score} > score$. In this case the student may not fully understand the principles of the considered task although he thinks so. Hence, next the system should show the student rather tasks which explain the approach of solving such kind of tasks.

In our studies we observed the behaviour of students described in (2), i.e. the student was labelled as to be in a flow or under-challenged, although he performed worse, as he just thought to understand how the tasks should be solved but he was wrong. In figure 4 this behaviour is indicated by the outliers.

7. LABELLING AND DATA COLLECTION

As mentioned in section 3 the labels of our real data set come from two sources: (a) a concurrent annotation by the tutor and (b) a retrospective annotation by another external reviewer on the basis of the tasks sheet, the sound files and the screen recording. However, in the literature one can find further labelling strategies like self-labelling of the students (see e.g. [5], [6], [8]). The advantage of self-labelling is that one can gain automatically a labelled data set for a subsequent training of an affect recognition method. Furthermore, as we want to recognise the perceived task-difficulty from the view of the student, a label from the student himself seem to be more appropriate than labels from another person only reviewing the behaviour of the student. Hence, for further studies we developed a tool for collecting speech data and typed input and mouse click input data, labelled automatically with the task-difficulty perceived by the student. This tool will be further described in the following section.

7.1 Self-Labelling Fractional Arithmetic Tutor for Multimodal Data Collection

To be able to conduct studies in which the students themselves label the task-difficulty which they perceived, we developed a tutoring tool (*self* - **se**lf-**l**abelling **f**ractional arithmetic tutor for multimodal data collection) written in Java. However, for little children it might be difficult to analyse themselves (see e.g. [8]). Hence, self-labelling is often applied in experiments with at least college students as for instance in [5]. Therefore, we will conduct the experiments with this tool first with older students and more challenging tasks. Later on we will investigate if there is a way to adapt the tool so that a self-labelling is possible also with younger students. Nevertheless, conducting experiments with older students has several advantages besides the possibility of a reasonable self-labelling: older students are able to focus on the tasks longer than young students and the privacy issues are not such strong as for younger students. Both facts lead to more data. Hence, besides investigating the possibility of

adapting *self* for younger students, we have to identify differences and similarities of the data from older and younger students to find out how to exploit older students data to recognise affects from multimodal input from younger students.

In figure 5 one can see the graphical user interface of our self-labelling multimodal data collection tool *self*. To gain more background information, in the beginning *self* asks some information from the students as course of studies, number of terms, age and gender. Subsequently, an instruction with hints how to behave is shown to the students, which they can have a look at also while interacting with the tool (button "Anleitung"). *self* speaks to the students to motivate them to speak with the system and records the speech input of the students. The speech output of *self* is generated by means of *text to speech* realised by the library MARY developed at the DFKI ([18]). While interacting with the system, the student can type in numbers, ask for a hint (button "Hilfe"), skip the task because it is too easy or because it is too hard (left buttons) or submit the solution (button "Endergebnis überprüfen"). Every action of the student, like asking for a hint or submitting the answer, is written – together with a time stamp – into a log file immediately after the action, enabling also the extraction of typed input or mouse click input features. Also a score depending on the number of requested hints h_r and the number of incorrect inputs w is computed according to the approach in [15] and written into the log file. The formula for this score is

$$1 - (\frac{h_r}{h_t} + (w \cdot 0.1)) \,, \tag{2}$$

where h_t is the total number of available hints for the considered task. The meaning behind the formula is that each wrong input $w^{(j)}$ is punished with a factor of 0.1 and every request of a hint $h_r^{(k)}$ is punished with a factor of $\frac{1}{h_t}$, so that if every hint was seen the score will be 0. After the student submitted the correct answer, he is asked to evaluate, if this task was too easy, too hard or appropriate for him (see pop-up window in figure 5). The tasks implemented in *self* for older students cover the following areas:

- Reducing fractions with numbers and variables
- Fraction addition with and without intermediate steps and with numbers and variables
- Fraction subtraction with and without intermediate steps and with numbers and variables
- Fraction multiplication with and without intermediate steps and with numbers and variables
- Fraction division with and without intermediate steps and with numbers and variables
- Distributivity law with and without intermediate steps
- Finite sums of unit fractions
- Rule of Three

After developing *self*, the next step will be to conduct further studies with students to collect an adequate amount of automatically labelled speech input, typed input and mouse click input data for training an affect recognition method and supporting performance prediction and task sequencing. Furthermore, we will investigate if there is a way to adapt *self* so that also younger students can label themselves.

8. CONCLUSIONS

We proposed a new approach for supporting performance prediction and task sequencing in adaptive intelligent tutoring systems by affect recognition on features gained from multimodal input like students speech input. For this approach we proposed and analysed appropriate speech features and showed that there are statistically significant feature combinations which are able to describe students affect, or perceived task-difficulty respectively, as well as the performance of a student. Furthermore, we proved the possibility of supporting performance prediction and task sequencing by perceived task-difficulties by demonstrating that there is a correlation between perceived task-difficulty and performance. Next steps will be to conduct more studies with students by means of the presented self-labelling and multimodal data collection tool to enable a training of an appropriate affect recognition method for supporting performance prediction and task sequencing in adaptive intelligent tutoring systems.

9. ACKNOWLEDGMENTS

The research leading to the results reported here has received funding from the European Union Seventh Framework Programme (FP7/2007 – 2013) under grant agreement No. 318051 – iTalk2Learn project (www.italk2learn.eu). Furthermore, we thank our project partner Ruhr University Bochum for realising the study and data collection as well as the IMAI of the University of Hildesheim for support for the tutoring tool and preparation for future studies.

10. REFERENCES

[1] Cichocki, A., Zdunek, R., Phan, A. H. and Amari, S.I. 2009. Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation, Wiley.

[2] Epp, C., Lippold, M., Mandryk, R.L. 2011. Identifying Emotional States Using Keystroke Dynamics. In Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems (CHI 2011), Vancouver, BC, Canada, pp. 715–724.

[3] Hu, X., Tang, L., Tang, J. and Liu, H. 2013. Exploiting Social Relations for Sentiment Analysis in Microblogging. In Proceedings of the Sixth ACM WSDM Conference (WSDM '13).

[4] Luz, S. 2013. 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, December 2013.

[5] D'Mello, S., Picard, R. and Graesser, A. 2007. Towards An Affect-Sensitive AutoTutor. Intelligent Systems, IEEE, Vol. 22, Issue 4, pp. 53–61.

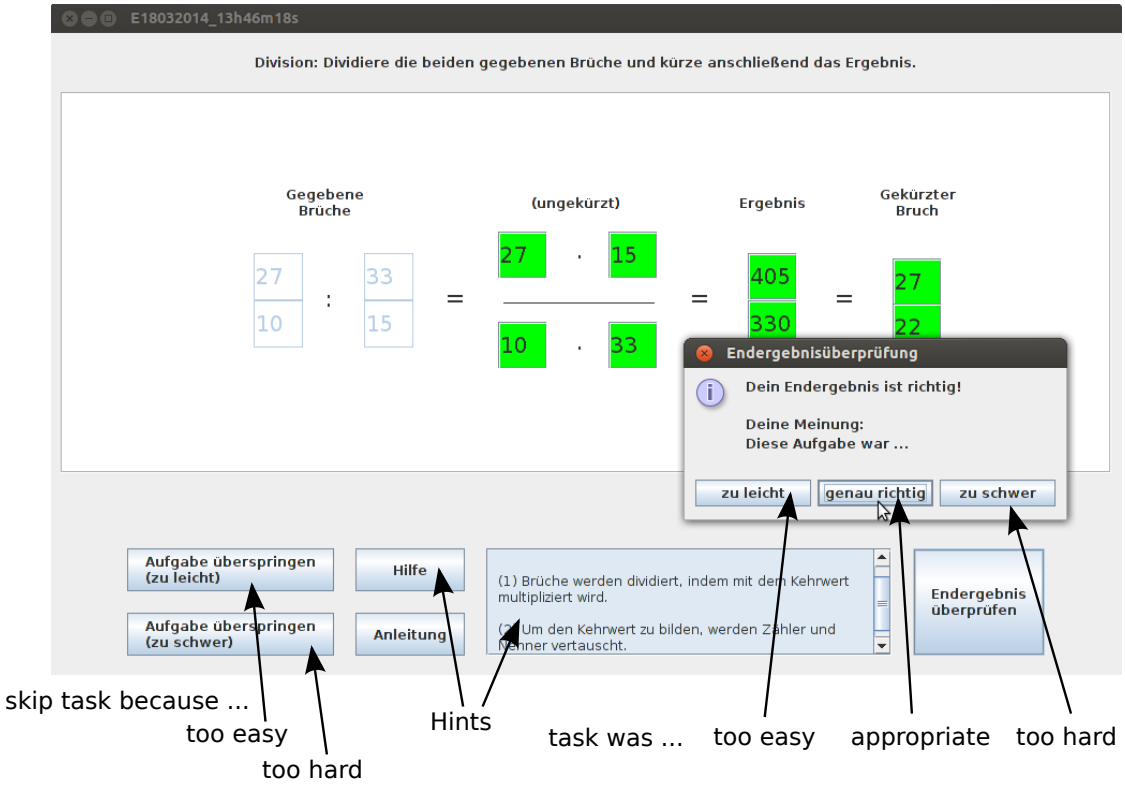


Figure 5: Graphical user interface of the developed fractional arithmetic tutoring tool *self* for self-labelling as well as for speech data and typed input or mouse click input data collection.

[6] D'Mello, S.K., Craig, S.D., Witherspoon, A., McDaniel, B. and Graesser, A. 2008. Automatic detection of learner's affect from conversational cues. User Model User-Adap Inter, DOI 10.1007/s11257-007-9037-6.

[7] Morency, L.P., Oviatt, S., Scherer, S., Weibel, N. and Worsley, M. 2013. 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.

[8] Porayska-Pomsta, K., Mavrikis, M., D'Mello, S., Conati, C. and Baker, R.S.J.d. 2013. Knowledge Elicitation Methods for Affect Modelling in Education. International Journal of Artificial Intelligence in Education, ISSN 1560-4292.

[9] Qi, F., Bao, C., Liu, Y. 2004. A novel two-step SVM classifier for voiced/unvoiced/silence classification of speech. International Symposium on Chinese Spoken Language Processing, pp. 77 – 80.

[10] Sadegh, M., Ibrahim, R., Othman, Z.A. 2012. Opinion Mining and Sentiment Analysis: A Survey. International Journal of Computers & Technology, Vol. 2, No. 3.

[11] Saif, H., He, Y. and Alani, H. 2012. Semantic Sentiment Analysis of Twitter. In Proceedings of the 11th International Semantic Web Conference (ISWC 2012).

[12] Schatten, C. and Schmidt-Thieme, L. 2014. Adaptive Content Sequencing without Domain Information. In Proceedings of the Conference on computer supported education (CSEDU 2014).

[13] Schuller, B., Batliner, A., Steidl, S. and Seppi, D. 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication, Elsevier.

[14] Vygotsky, L.L.S. 1978. Mind in society: The development of higher psychological processes. Harvard university press.

[15] Wang, Y. and Heffernan, N. 2011. Extending Knowledge Tracing to allow Partial Credit: Using Continuous versus Binary Nodes. Artificial Intelligence in Education, Lecture Notes in Computer Science, Vol. 7926, pp. 181–188.

[16] Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D. and Picard, R. 2009. Affect-aware tutors: recognising and responding to student affect. Int. J. of Learning Technology, Vol. 4, No. 3/4, pp. 129–164.

[17] Worsley, M. and Blikstein, P. 2011. 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.

[18] The MARY Text-to-Speech System, http://mary.dfki.de/

Collaborative Assessment

Patricia Gutierrez
IIIA-CSIC
Campus de la UAB
Barcelona, Spain
patricia@iiia.csic.es

Nardine Osman
IIIA-CSIC
Campus de la UAB
Barcelona, Spain
nardine@iiia.csic.es

Carles Sierra
IIIA-CSIC
Campus de la UAB
Barcelona, Spain
sierra@iiia.csic.es

ABSTRACT

In this paper we introduce an automated assessment service for online learning support in the context of communities of learners. The goal is to introduce automatic tools to support the task of assessing massive number of students as needed in Massive Open Online Courses (MOOC). The final assessments are a combination of tutor's assessment and peer assessment. We build a trust graph over the referees and use it to compute weights for the assesments aggregations. The model proposed intends to be a support for intelligent online learning applications that encourage student's interactions within communities of learners and benefits from their feedback to build trust measures and provide automatic marks.

1. INTRODUCTION

Self and peer assessment have clear pedagogical advantages. Students increase their responsibility and autonomy, get a deeper understanding of the subject, become more active in the learning process, reflect on their role in group learning, and improve their judgement skills. Also, it may have the positive side effect of reducing the marking load of tutors. This is specially critical when tutors face the challenge of marking large quantities of students as needed in the increasingly popular Massive Open Online Courses (MOOC).

Online learning communities encourage different types of peer-to-peer interactions along the learning process. These interactions permit students to get more feedback, to be more motivated to improve, and to compare their own work with other students accomplishments. Tutors, on the other hand, benefit from these interactions as they get a clearer perception of the student engagement and learning process.

Previous works have proposed different methods of peer assessment as part of the learning process with the added advantage of helping tutors in the sometimes dauting task of marking large quantities of students [7, 3].

The authors of [7] propose methods to estimate peer reliability and correct peer biases. They present results over real world data from 63,000 peer assessments of two Coursera courses. The models proposed are probabilistic and they are compared to the grade estimation algorithm used on Coursera's platform, which does not take into account individual biases and reliabilities. Differently from them, we place more trust in students who grade like the tutor and do not consider student's biases. When a student is biased its trust measure will be very low and his/her opinion will have a moderate impact over the final marks.

[3] proposes the CrowdGrader framework, which defines a crowdsourcing algorithm for peer evaluation. The accuracy degree (i.e. reputation) of each student is measured as the distance between his/her self assesment and the aggregated opinion of the peers weighted by their accuracy degrees. The algorithm thus implements a reputation system for students, where higher accuracy leads to higher influence on the consensus grades. Differently from this work, we give more weight to those peers that have similar opinions to those of the tutor.

In this paper, and differently from previous works, we want to study the *reliability* of student assessments when compared with tutor assessments. Although part of the learning process is that students participate in the definition of the evalution criteria, tutors want to be certain that the scoring of the students' works is fair and as close as possible to his/her expert opinion.

Our inspiration comes from a use case explored in the EU-funded project PRAISE [1]. PRAISE enables online virtual communities of students with shared interests and goals to come together and share their music practice with each other so the process of learning becomes social. It provides tools for giving and receiving feedback, as feedback is considered an essential part of the learning process. Tutors define *lesson plans* as pedagogical workflows of activities, such as uploading recorded songs, automatic performance analysis, peer feedback, or reflexive pedagogy analysis. The goal of any lesson plan is to improve student skills, for instance, the performance speed competence or the interpretation maturity level. Assessments of students' performances have to evaluate the achievement of these skills. Once a lesson plan is defined, PRAISE's interface tools allow students to navigate through the activities, to upload assignments, to practice, to assess each other, and so on. The tools allow tutors to monitor what students have done and to assess them. In this

Published in CEUR-WS:
FFMI workshop (Schmidt-Thieme and Janning)
In EDM 2014 Extended Proceedings (Gutierrez-Santos and Santos)

work we concentrate on the development of a service that can be included as part of a lesson plan and helps tutors in the overall task of assessing the students participating in the lesson plan. This assessment is based on aggregating students' assessments, taking into consideration the trust that tutors have on the students' individual capabilities in judging each others work.

To achieve our objective we propose in this paper an automated assessment method (Section 2) based on *tutor assessments*, aggregations of *peer assessments* and on *trust measures* derived from peer interactions. We experimentaly evaluate (Section 3) the accuracy of the method over different topologies of student interactions (i.e. different types of student grouping). The results obtained are based on simulated data, leaving the validation with real data for future work. We then conclude with a discussion of the results (Section 4).

2. COLLABORATIVE ASSESSMENT

In this section we introduce the formal model of the method and the algorithms for collaborative assessment.

2.1 Notation and preliminaries

We say an online course has a tutor τ, a set of peer students $\mathcal{S}$, and a set of assignments $\mathcal{A}$ that need to be marked by the tutor and/or students with respect to a given set of criteria $\mathcal{C}$.

The automated assessment state S is then defined as the tuple:

$$S = \langle R, \mathcal{A}, \mathcal{C}, \mathcal{L} \rangle$$

$R = \{\tau\} \cup \mathcal{S}$ defines the set of possible referees (or markers), where a referee could either be the tutor τ or some student $s \in S$. $\mathcal{A}$ is the set of submitted assignments that need to be marked and $\mathcal{C} = \langle c_1, \ldots, c_n \rangle$ is the set of criteria that assignments are marked upon. $\mathcal{L}$ is the set of marks (or assessments) made by referees, such that $\mathcal{L} : R \times A \to [0, \lambda]^n$ (we assume marks to be real numbers between 0 and some maximum value λ). In other words, we define a single assessment as: $\mu_\alpha^\rho = \vec{M}$, where $\alpha \in \mathcal{A}$, $\rho \in R$, and $\vec{M} = \langle m_1, \ldots, m_n \rangle$ describes the marks provided by the referee on the n criteria of $\mathcal{C}$, $m_i \in [0, \lambda]$.

Similarity between marks. We define a similarity function $sim : [0, \lambda]^n \times [0, \lambda]^n \to [0, 1]$ to determine how close two assesments μ_α^ρ and μ_α^η are. We calculate the similarity between assessments $\mu_\alpha^\rho = \{m_1, \ldots, m_n\}$ and $\mu_\alpha^\eta = \{m'_1, \ldots, m'_n\}$ as follows:

$$sim(\mu_\alpha^\rho, \mu_\alpha^\eta) = 1 - \frac{\sum_{i=1}^{n} |m_i - m'_i|}{\sum_{i=1}^{n} \lambda}$$

This measure satisfies the basic properties of a fuzzy similarity [6]. Other similarity measures could be used.

Trust relations between referees. Tutors need to decide up to which point they can believe on the assessments made by peers. We use two different intuitions to make up this belief. First, if the tutor and the student have both assessed some assigments, their similarity gives a hint of how close the judegements of the student and the tutor are. Similarly, we can define the judgement closeness of any two students by looking into the assignments evaluated by both of them. In case there are no assigments evaluated by the tutor and one particular student we could simply not take that student's opinion into account because the tutor would not know how much to trust the judgement of this student, or, as we do in this paper, we approximate that unknown trust by lookig into the chain of trust between the tutor and the student through other students. To model this we define two different types of trust relations:

- *Direct trust*: This is the trust between referees $\rho, \eta \in R$ that have at least one assignement assessed in common. The trust value is the average of similarities on the assessments over the same peers. Let the set $A_{\rho,\eta}$ be the set of all assignments that have been assessed by both referees. That is, $A_{\rho,\eta} = \{\alpha \mid \mu_\alpha^\rho \in \mathcal{L} \text{ and } \mu_\alpha^\eta \in \mathcal{L}\}$. Then,

  $$T_D(\rho, \eta) = \frac{\sum_{\alpha \in A_{\rho,\eta}} sim(\mu_\alpha^\rho, \mu_\alpha^\eta)}{|A_{\rho,\eta}|}$$

  We could also define direct trust as the conjunction of the similarities for all common assignments as:

  $$T_D(\rho, \eta) = \bigwedge_{\alpha \in A_{\rho,\eta}} sim(\mu_\alpha^\rho, \mu_\alpha^\eta)$$

  However, this would not be practical, as a significant difference in just one assessment of those assessed by two referees would make their mutual trust very low.

- *Indirect trust*: This is the trust between referees $\rho, \eta \in R$ without any assignement assessed by both of them. We compute this trust as a transitive measure over chains of referees for which we have pair-wise direct trust values. We define a trust chain as a sequence of referees $q_j = \langle \rho_i, ..., \rho_i, \rho_{i+1}, \ldots, \rho_{m_j} \rangle$ where $\rho_i \in R$, $\rho_1 = \rho$ and $\rho_{m_j} = \eta$ and $T_D(\rho_i, \rho_{i+1})$ is defined for all pairs (ρ_i, ρ_{i+1}) with $i \in [1, m_j - 1]$. We note by $Q(\rho, \eta)$ the set of all trust chains between ρ and η. Thus, indirect trust is defined as a aggregation of the direct trust values over these chains as follows:

  $$T_I(\rho, \eta) = \max_{q_j \in Q(\rho,\eta)} \prod_{i \in [1, m_j - 1]} T_D(\rho_i, \rho_{i+1})$$

  Hence, indirect trust is based in the notion of transitivity.[1]

[1] T_I is based on a fuzzy-based similarity relation sim presented before and fulfilling the $\otimes$-Transitivity property: $sim(u, v) \otimes sim(v, w) \leq sim(u, w)$, $\forall u, v, w \in V$, where $\otimes$ is a t-norm [6].

Ideally, we would like to not overrate the trust of a tutor on a student, that is, we would like that $T_D(a,b) \geq T_I(a,b)$ in all cases. Guaranteeing this in all cases is impossible, but we can decrease the number of overtrusted students by selecting an operator that gives low values to T_I. In particular, we prefer to use the product $\prod$ operator, because this is the t-norm that gives the smallest possible values. Other opertors could be used, for instance the *min* function.

Trust Graph. To provide automated assessments, our proposed method agregates the assessments on a given assignment taking into consideration how much trusted is each marker/referee from the point of view of the tutor (i.e. taking into consideration the trust of the tutor on the referee in marking assignments). The algorithm that computes the student final assessment is based on a graph defined as follows:

$$G = \langle R, E, w \rangle$$

where the set of nodes R is the set of referees in S, $E \subseteq R \times R$ are edges between referees with direct or indirect trust relations, and $w : E \to [0,1]$ provides the trust value. We note by $D \subset E$ the set of edges that link referees with direct trust. That is, $D = \{e \in E | T_D(e) \neq \perp\}$. An similarly, $I \subset E$ for indirect trust, $I = \{e \in E | T_I(e) \neq \perp\} \setminus D$. The w values will be used as weights to combine peer assessments and are defined as:

$$w(e) = \begin{cases} T_D(e) & \text{, if } e \in D \\ T_I(e) & \text{, if } e \in I \end{cases}$$

Figure 1 shows examples of trust graphs with $e \in D$ (in black) and $e \in I$ (in red —light gray) for different sets of assessments $\mathcal{L}$.

2.2 Computing collaborative assessments

Algorithm 1 implements the collaborative assessment method. We keep the notation (ρ, η) to refer to the edge connecting nodes ρ and η in the trust graph and $Q(\rho, \eta)$ to refer the set of trust chains between ρ and η.

The first thing the algorithm does is to build a trust graph from $\mathcal{L}$. Then, the final assessments are computed as follows. If the tutor marks an assignment, then the tutor mark is considered the final mark. Otherwise, a weighted average (μ_α) of the marks of student peers is calculated for this assignment, where the weight of each peer is the trust value between the tutor and that peer. Other forms of aggregation could be considered to calculate μ_α, for instance a peer assessment may be discarded if it is very far from the rest of assessments, or if the referee's trust falls below a certain threshold.

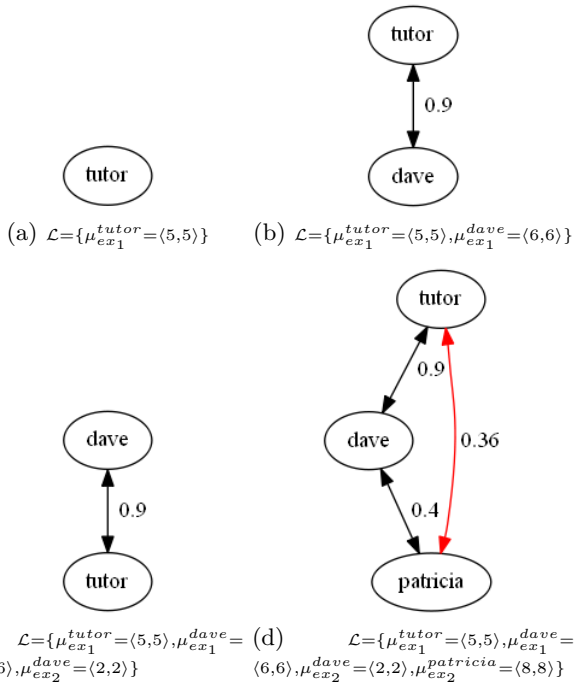
Figure 1 shows four trust graphs built from four assessments histories that corresponds to a chronological sequence of assessments made. The criteria $\mathcal{C}$ in this example are *speed* and *maturity* and the maximum mark value is $\lambda = 10$. For

Algorithm 1: collaborativeAssessments($S = \langle R, \mathcal{A}, \mathcal{C}, \mathcal{L} \rangle$)

▷ Initial trust between referees is zero
$D = I = \emptyset$;
for $\rho, \eta \in \mathcal{R}, \rho \neq \eta$ **do**
  $w(\rho, \eta) = 0$;
end
▷ Update direct trust and edges
for $\rho, \eta \in \mathcal{R}, \rho \neq \eta$ **do**
  $A_{\rho,\eta} = \{\beta \mid \mu^\rho_\beta \in \mathcal{L} \text{ and } \mu^\eta_\beta \in \mathcal{L}\}$;
  if $|A_{\rho,\eta}| > 0$ **then**
    $D = D \cup (\rho, \eta)$;
    $w(\rho, \eta) = T_D(\rho, \eta)$;
  end
end
▷ Update indirect trust and edges between tutor & students
for $\rho \in \mathcal{R}$ **do**
  if $(\tau, \rho) \notin D$ *and* $Q(\tau, \rho) \neq \emptyset$ **then**
    $I = I \cup (\rho, \eta)$;
    $w(\rho, \eta) = T_I(\tau, \eta)$;
  end
end
▷ Calculate automated assessments
$assessments = \{\}$;
for $\alpha \in A$ **do**
  if $\mu^\tau_\alpha \in \mathcal{L}$ **then**
    ▷ Tutor assessments are preserved
    $assessments = assessments \cup (\alpha, \mu^\tau_\alpha)$
  else
    ▷ Generate automated assessments
    $R' = \{\rho | \mu^\rho_\alpha \in \mathcal{L}\}$;
    if $|R'| > 0$ **then**
      $\mu_\alpha = \frac{\sum_{\rho \in R'} \mu^\rho_\alpha * w(\tau, \rho)}{\sum_{\rho \in R'} w(\tau, \rho)}$;
      $assessments = assessments \cup (\alpha, \mu_\alpha)$;
    end
  end
end
return *assessments*;

simplicity we only represent those referees that have made assessments in $\mathcal{L}$. In Figure 1(a) there is one node representing the tutor who has made the only assessment over the assignment ex_1 and there are no links to other nodes as no one else has assessed anything. In (b) student Dave assesses the same exercise as the tutor and thus a link is created between them. The trust value $w(tutor, Dave) = T_D(tutor, Dave)$ is high since their marks were similar. In (c) a new assessment by Dave is added to $\mathcal{L}$ with no consequences in the graph construction. In (d) student Patricia adds an assessment on ex_2 that allows to build a direct trust between Dave and Patricia and an indirect trust between the tutor and Patricia, through Dave. The automated assessments generated in case (d) are: $\langle 5, 5 \rangle$ for exercise 1 (which preserves the tutor's assessment) and $\langle 3.7, 3.7 \rangle$ for exercise 2 (which uses a weighted aggregation of the peers' assessments).

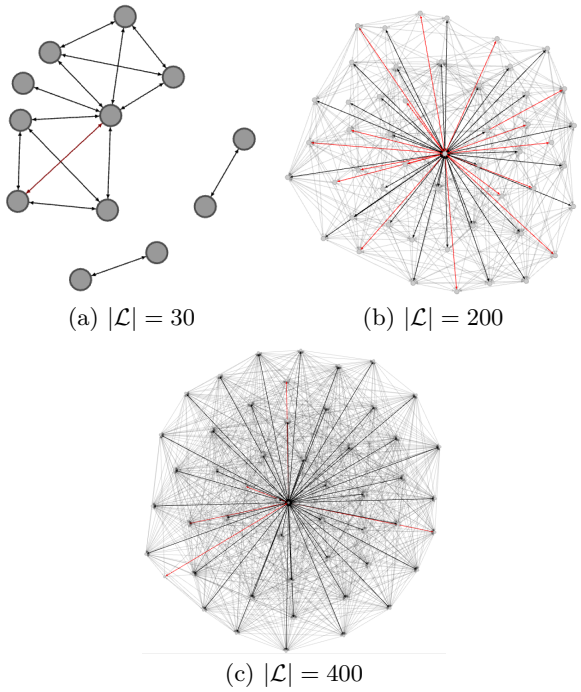
Note that the trust graph built from $\mathcal{L}$ is not necessarily connected. A tutor wants to reach a point in which the graph is totally connected because that means that the collaborative assessment algorithm generates an assessment for every assignment. Figure 2 shows an example of a trust graph of a particular learning community involving 50 peer students and a tutor. When S has a history of 5 tutor assessments and 25 student assessments ($|\mathcal{L}| = 30$) we observe that not all nodes are connected. As the number of assessments in-



(a) $\mathcal{L}=\{\mu_{ex_1}^{tutor}=\langle 5,5\rangle\}$ (b) $\mathcal{L}=\{\mu_{ex_1}^{tutor}=\langle 5,5\rangle,\mu_{ex_1}^{dave}=\langle 6,6\rangle\}$

(c) $\mathcal{L}=\{\mu_{ex_1}^{tutor}=\langle 5,5\rangle,\mu_{ex_1}^{dave}=\langle 6,6\rangle,\mu_{ex_2}^{dave}=\langle 2,2\rangle\}$ (d) $\mathcal{L}=\{\mu_{ex_1}^{tutor}=\langle 5,5\rangle,\mu_{ex_1}^{dave}=\langle 6,6\rangle,\mu_{ex_2}^{dave}=\langle 2,2\rangle,\mu_{ex_2}^{patricia}=\langle 8,8\rangle\}$

Figure 1: Trust graph example 1.



(a) $|\mathcal{L}| = 30$ (b) $|\mathcal{L}| = 200$

(c) $|\mathcal{L}| = 400$

Figure 2: Trust graph example 2

creases, the trust graph becomes denser and eventually it gets completely connected. In (b) and (c) we see a complete graph.

3. EXPERIMENTAL PLATFORM AND EVALUATION

In this Section we describe how we generate simulated social networks, describe our experimental platform, define our benchmarks and discuss experimental results.

3.1 Social Network Generation

Several models for social network generation have been proposed reflecting different characteristics present in real social communities. Topological and structural features of such networks have been explored in order to understand wich generating model resembles best the structure of real communities [5].

A social network can be defined as a graph $\mathcal{N}$ where the set of nodes represent the individuals of the network and the set of edges represent connections or social ties among those individuals. In our case, individuals are the members of the learning community: the tutor and students. Connections represent the social ties and they are usually the result of interactions in the learning community. For instance a social relation will be born between two students if they interact with each other, say by collaboratively working on a project together. In our experimentation, we rely on the social network in order to simulate which student will assess the assignment of which other student. We assume students will assess the assignments of students they know, as opposed to picking random assignments. As such, we clarify that social networks are different from the trust graph of Section 2. While the nodes of both graphs are the same, edges of the social network represent social ties, whereas edges in the trust graph represent how much does one referee trust another in judging others work.

To model social networks where relations represent social ties, we follow three different approaches: the Erdős-Rényi model for random networks [4], the Barabási-Albert model for power law networks[2] and a hierarchical model for cluster networks.

3.1.1 *Random Networks*

The Erdős-Rényi model for random networks consists of a graph containing n nodes connected randomly. Each possible edge between two vertices may be included in the graph with probability p and may not be included with probability $(1-p)$. In addition, in our case there is always an edge between the node representing the tutor and the rest of nodes, as the tutor knows all of its students (and may eventually mark any of those students).

The degree distribution of random graphs follows a Poisson distribution. Figure 3(a) shows an example of a random graph with 51 nodes and $p = 0.5$ and its degree distribution. Note that the point with degree 50 represents the tutor node while the rest of the nodes degree fit a Poisson distribution.

3.1.2 *Power Law Networks*

The Barabási-Albert model for power law networks base their graph generation on the notions of *growth* and *preferential attachment*. The generation scheme is as follows. Nodes are added one at a time. Starting with a small number of initial nodes, at each time step we add a new node with m edges linked to nodes already part of the network. In our experiments, we start with $m+1$ initial nodes. The

edges are not placed uniformly at random but preferentially in proportion to the degree of the network nodes. The probability p that the new node is connected to a node i already in the network depends on the degree k_i of node i, such that: $p = k_i / \sum_{j=1}^{n} k_j$. As above, there is also always an edge between the node representing the tutor and the rest of nodes.

The degree distribution of this network follows a Power Law distribution. Figure 3(b) shows an example of a power law graph with 51 nodes and $m = 16$ and its degree distribution. The point with degree 50 describes the tutor node while the rest of the nodes closely resemble a power law distribution. Recent empirical results on large real-world networks often show, among other features, their degree distribution following a power law [5].

3.1.3 *Cluster Networks*

As our focus is on learning communities, we also experiment with a third type of social network: the cluster network which is based on the notions of *groups* and *hierarchy*. Such networks consists of a graph composed of a number of fully connected clusters (where we believe clusters may represent classrooms or similar pedagogical entities). Additionally, as above, all the nodes are connected with the tutor node. Figure 3(c) shows an example of a cluster graph with 51 nodes, 5 clusters of 10 nodes each and its degree distribution. The point with degree 50 describes the tutor while the rest of the nodes have degree 10, since every student is fully connected with the rest of the classroom.

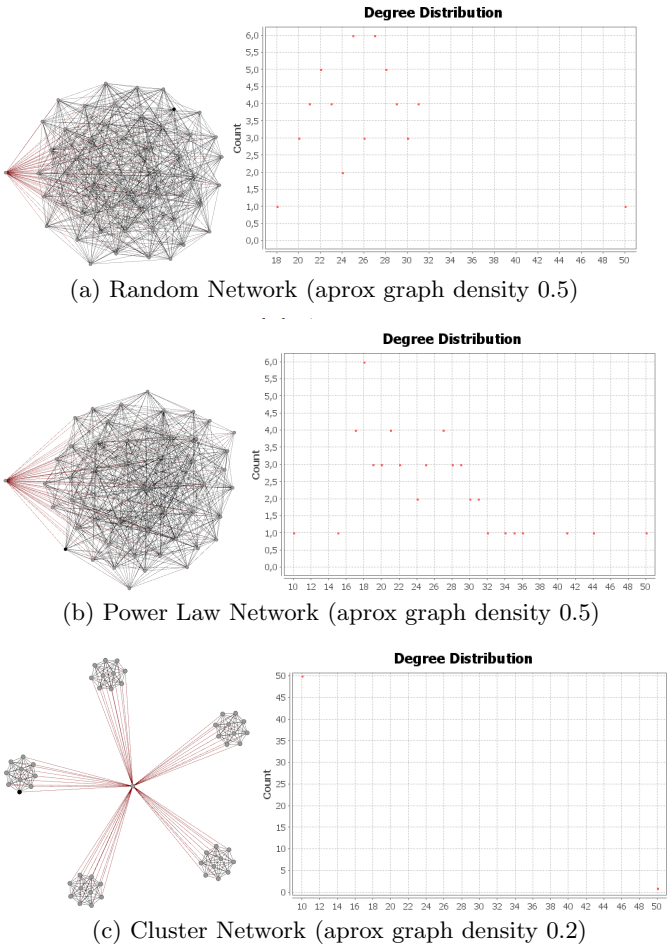
3.2 Experimental Platform

In our experimentation, given an initial automated assessment state $S = \langle R, \mathcal{A}, \mathcal{C}, \mathcal{L} \rangle$ with an empty set of assessments $\mathcal{L} = \{\}$, we want to simulate tutor and peer assessments so that the collaborative assessment method can eventually generate a reliable and definitive set of assessments for all assignments.

To simulate assessments, we say each students is defined by its profile that describes how good its assessments are. The profile is essentially defined by the measure, or distance, $d_\rho \in [0, 1]$ that specifies how close are the student's assessments to that of the tutor.

We then assume the simulator knows how the tutor and each student would assess an assignment. This becomes necessary in our simulation, since we generate student assessments in terms of their distance to that of the tutor's, even if the tutor does not choose to actually assess the assignment in question. This simulator's knowledge of the values of all possible assessments is generated accordingly:

- For every assignment $\alpha \in \mathcal{A}$, we calculate the tutor's assessment, which is randomly generated according to the function $f_\tau : \mathcal{A} \to [0, \lambda]^n$. This assessment essentially describes what mark would the tutor give α, if it decided to assess it.
- For every assignment $\alpha \in \mathcal{A}$, we also calculate the assessment of each student $\rho \in \mathcal{S}$. This is calculated according to the function $f_\rho : A \to [0, \lambda]^n$, such that: $sim(f_\rho(\alpha), f_\tau(\alpha)) \geq d_\rho$ We note that we only need to calculate ρ's assessment of α if the student who submitted the assignment α is a neighbour of ρ in $\mathcal{N}$.



(a) Random Network (aprox graph density 0.5)

(b) Power Law Network (aprox graph density 0.5)

(c) Cluster Network (aprox graph density 0.2)

Figure 3: Social Network generation examples

We note that the above only calculates what the assessments would be, if referees where to assess assignments.

3.3 Benchmark

Given an initial automated assessment state $S = \langle R, \mathcal{A}, \mathcal{C}, \mathcal{L} \rangle$ with an empty set of assessments $\mathcal{L} = \{\}$, a set of student profiles $Pr = \{d_s\}_{\forall s \in \mathcal{S}}$, and a social network $\mathcal{N}$ (whose nodes is the set R), we simulate individual tutor and students' assessments. When does a referee in R assess an assignment in $\mathcal{A}$ is explained shortly. However we note here that the value of each generated assessment is equivalent to that calculated for the simulator's knowledge (see Section 3.2 above).

In our benchmark, we consider the three types of social networks introduced earlier: random social networks (with 51 nodes, $p = 0.5$, and approximate density of 0.5), power law networks (with 51 nodes, $m = 16$, and approximate density of 0.5), and cluster networks (with 51 nodes, 5 clusters of 10 nodes each, and approximate density of 0.2). Examples of these generated networks are shown in Figure 3.

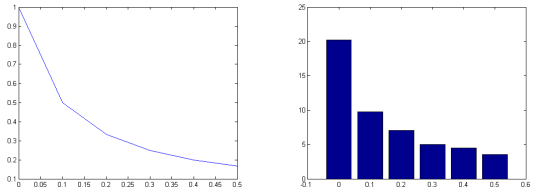
We say one assignment is submitted by each student, resulting in $|\mathcal{S}| = 50$ and $|\mathcal{A}| = 50$. The range that a referee (tutor or student) may mark a given assignment with respect to a given criteria is [0,10]. And the set of criteria is $\mathcal{C} = \langle speed, maturity \rangle$. The criteria essentially measure the *speed* of playing a musical piece, and the *maturity level* of the student's performance.

An assessment profile is generated for each student ρ at the beginning of the execution, resulting in a set of student profiles $Pr = \{d_s\}_{\forall s \in \mathcal{S}}$, where $d \in [0, 0.5]$. We consider here two cases for generating the set of student profiles Pr. A first case where d is picked randomly following a power law distribution (Figure 4(a)) and a second case where d is picked randomly following a uniform distribution (Figure 4(b)).

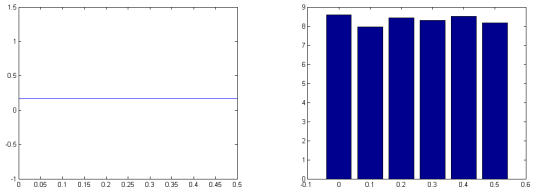
With simulated individual assessments, we then run the collaborative assessment method in order to compute an automated assessment. We also compute the 'error' of the collaborative assessment method, whose range is $[0, 1]$, over the set of assignments $\mathcal{A}$ accordingly:

$$\frac{\sum_{\alpha \in \mathcal{A}} sim(f_\tau(\alpha), \phi(\alpha))}{|\mathcal{A}|}$$

, where $\phi(\alpha)$ describes the automated assessment for a given assignment $\alpha \in \mathcal{A}$



(a) Power law profile generation



(b) Uniform profile generation

Figure 4: Example of the profile distributions (left) and of d counting averaged over 50 instances (right)

With the settings presented above, we run two different experiments. The results presented are an average over 50 executions. The two experiments are presented next.

In experiment 1, students provide their assessments before the tutor. Each student ρ provides assessments for a randomly chosen a_ρ number of peer assigments (of course, where assignments are those of their neighboring peers in $\mathcal{N}$). We run the experiment for 5 different values of $a_\rho = \{3, 4, 5, 6, 7\}$. After the students provide their assessments, the tutor starts assessing assignments incrementally. After every tutor assessment, the error over the set of automated assessment is calculated. Notice that the collaborative assessment method takes the tutor assessment, when it exists, to be the final assessment. As such, the number of automated assessments calculated based on aggregating students' assessments is reduced over time. Finally, when the tutor has assessed all 50 students, the resulting error is 0.

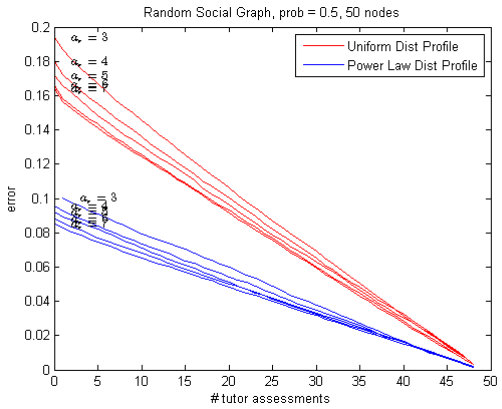
In experiment 2, the tutor provides its assessments before the students. The tutor in this experiment will assess a randomly chosen number of assignments, where this number is based on the percentage a_τ of the total number of assignments. We run the experiment for 4 different values of $a_\tau = \{5, 10, 15, 20\}$. After the tutor provides their assessments, students' assessments are performed. In every iteration, a student ρ randomly selects a neighbor in $\mathcal{N}$ and assesses his assignment (in case it has not been assessed before by ρ, otherwise another connected peer is chosen). We note that in the case of random and power law networks (denser networks), a total number of 1000 student assessments are performed. Whereas in the case of cluster networks (looser network), a total of 400 student assessments are performed. We note that initially, the trust graph is not fully connected, so the service is not able to provide automated assessments for all assignments. When the grap gets fully connected, the service generates automated assessments for all assignments and we start measuring the error after every new iteration.

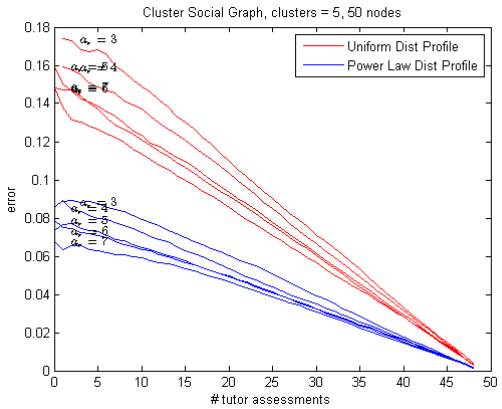
3.4 Evaluation

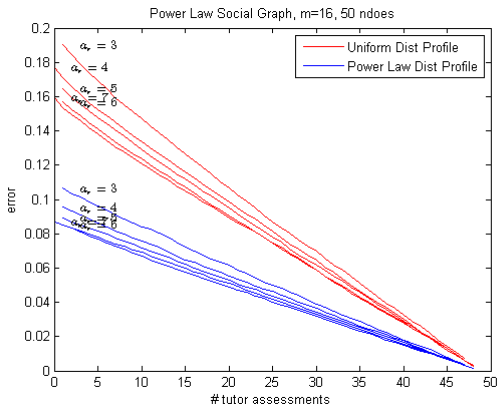
In experiment 1, we observe (Figure 5) that the error decreases when the number of tutor assessments increase, as expected, until it reaches 0 when the tutor has assessed all 50 students. This decrement is quite stable and we do not observe abrupt error variations or important error increments from one iteration to the next. More variations are observed in the initial iterations since the service has only a few assessments to deduce the weights of the trust graph and to calculate the final outcome.

In the case of experiment 2 (Figure 6), the error diminishes slowly as the number of student assessments increase, although it never reaches 0. Since the number of tutor assessments is fixed in this experiment, we have an error threshold (a lower bound) which is linked to the students' assessment profile: the closest to the tutor's the lower this threshold will be. In fact, in both experiments we observe that when using a power law distribution profile (Figure 4(a)) the automated assessment error is lower than when using a uniform distribution profile (Figure 4(b)). This is because when using a power law distribution, more student profiles are generated whose assessments are closer to the tutors'.

In general, the error trend observed in all experiments comparing different social network scenarios (random, cluster or power law) show a similar behavior. Taking a closer look at experiment 2, cluster social graphs have the lowest error and we observe that assessments on all assignments are achieved earlier (this is, the trust graph gets connected earlier). We attribute this to the topology of the fully connected clusters which favors the generations of indirect edges earlier in the graph between the tutor and the nodes of each cluster. Power law social graphs have lower error than random networks in most cases. This can be attributed to the criteria of preferential attachment in their network generation,







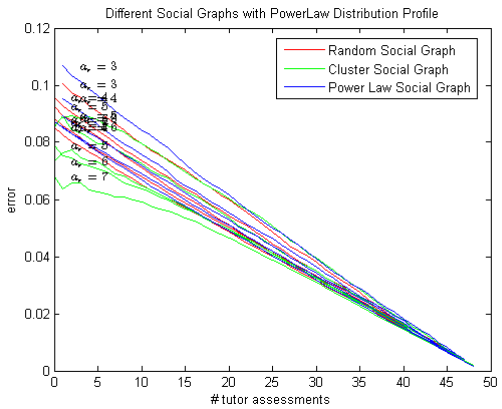
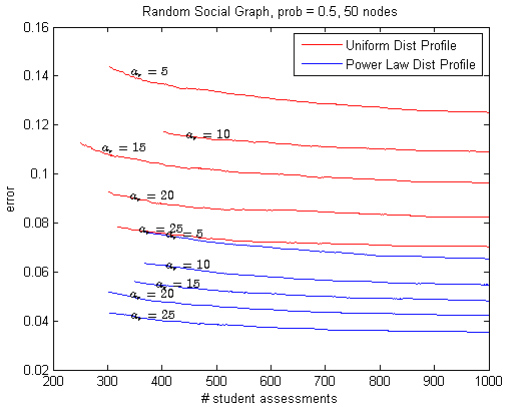
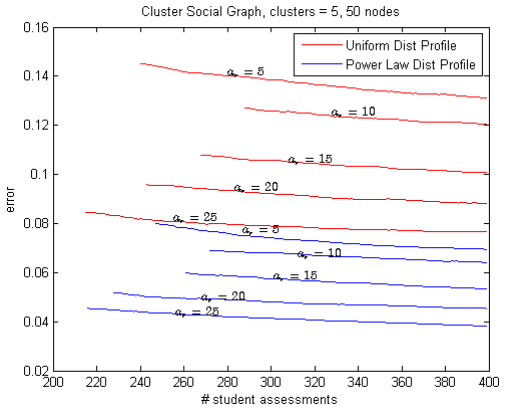
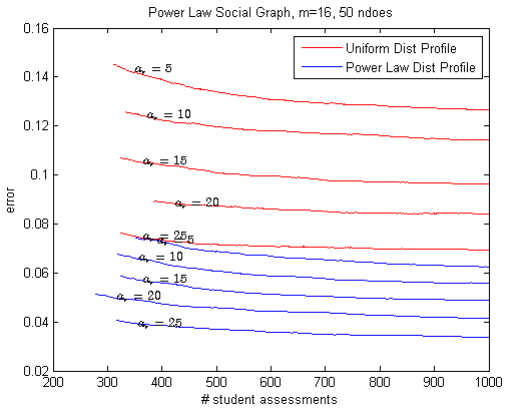


Figure 5: Eperiment 1







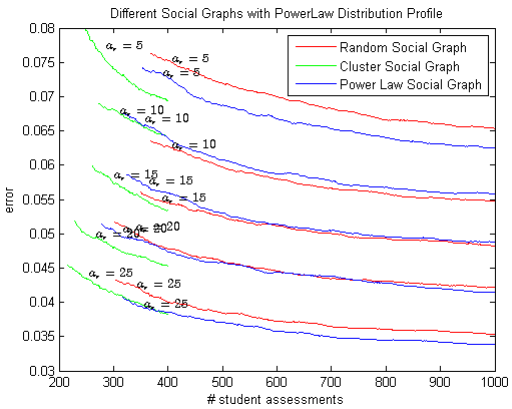


Figure 6: Experiment 2

which favors the creation of some highly connected nodes. Such nodes are likely to be assessed more frequently since more peers are connected to them. Then, the automated assessments of these higly connected peers are performed with more available information which could lead to more accurate outcomes.

4. DISCUSSION

The collaborative assessment model proposed in this paper is thought of as a support in the creation of intelligent online learning applications that encourage student interactions within communities of learners. It goes beyond current tutor-student online learning tools by making students participate in the learning process of the whole group, providing mutual assessment and making the overall learning process much more collaborative.

The use of AI techniques is key for the future of online learning communities. The application presented in this paper is specially useful in the context of MOOC: with a low number of tutor assessments and encouraging students to interact and provide assessments among each other, direct and indirect trust measures can be calculated among peers and automated assessments can be generated.

Several error indicators can be designed and displayed to the tutor managing the course, which we leave for future work. For example the error indicators may inform the tutor which assignments have not received any assessments yet, or which deduced marks are considered unreliable. For example, a deduced mark on a given assignment may be considered unreliable if all the peer assessments that have been provided for that assignment are considered not to be trusted by the tutor as they fall below a preselected acceptable trust threshold. Alternatively, a reliability measure may also be assigned to the computed trust measure T_D. For instance, if there is only one assignment that has been assessed by τ and ρ, then the computed $T_D(\tau, \rho)$ will not be as reliable as having a number of assignments assessed by τ and ρ. As such, some reliability threshold may be used that defines what is the minimum number of assignments that both τ and ρ need to assess for $T_D(\tau, \rho)$ to be considered reliable. Observing such error indicators, the tutor can decide to assess more assignments and as a result the error may improve or the set of deduced assessments may increase. Finally, if the error reaches a level of acceptance, the tutor can decide to endorse and publish the marks generated by the collaborative assessment method.

Another interesting question for future work is presented next. Missing connections might be detected in the trust graph that would improve its connectivity or maximize the number of direct edges. The question that follows then is, what assignments should be suggested to which peers such that the trust graph and the overall assessment outcome would improve?

Additionally, future work may also study different approaches for calculating the indirect trust value between two referees. In this paper, we use the product operator. We suggest to study a number of operators, and run an experiment to test which is most suitable. To do such a test, we may calculate the indirect trust values for edges that do have a direct trust measure, and then see which approach for calculating indirect trust gets closest to the direct trust measures.

Acknowledgements

This work is supported by the Agreement Technologies project (CONSOLIDER CSD2007-0022, INGENIO 2010) and the PRAISE project (EU FP7 grant number 388770).

5. REFERENCES

[1] Praise project: http://www.iiia.csic.es/praise/.

[2] A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 1999.

[3] L. de Alfaro and M. Shavlovsky. Thecnical report 1308.5273, arxiv.org. *Crowdgrader: Crowdsourcing the evaluation of homework assignments*, 2013.

[4] P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae*, 1959.

[5] E. Ferrara and G. Fiumara. Topological features of online social networks. *Communications in Applied and Industrial Mathematics*, 2011.

[6] L. Godo and R. Rodríguez. Logical approaches to fuzzy similarity-based reasoning: an overview. *Preferences and Similarities*, 2008.

[7] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller. Tuned models of peer assessment in moocs. *Proc. of the 6th International Conference on Educational Data Mining (EDM 2013)*, 2013.

Mining for Evidence of Collaborative Learning in Question & Answering Systems

Johan Loeckx
Artificial Intelligence Lab
Vrije Universiteit Brussel
Pleinlaan 2, 1050 Brussel
jloeckx@ai.vub.ac.be

ABSTRACT

Question and Answering systems and *crowd learning* are becoming an increasingly popular way of organising and exchanging expert knowledge in specific domains. Since they are expected to have a significant impact on online education [14], we will investigate to which degree the necessary conditions for collaborative learning emerge in open Q&A platforms like Stack Exchange, in which communities grow organically and learning is not guided by a central authority or curriculum, unlike MOOCs. Starting from a pedagogical perspective, this paper mines for circumstantial evidence to support or contradict the pedagogical criteria for collaborative learning. It is observed that although there are *technically no hindrances towards true collaborative learning*, the nature and dynamics of the communities are not favourable for collaborative learning.

The findings in this paper illustrate how the collaborative nature of feedback can be measured in online platforms, and how users can be identified that need to be encouraged to participate in collaborative activities. In this context, remarks and suggestions are formulated to pave the way for a more collaborative and pedagogically sound platform of knowledge sharing.

1. INTRODUCTION

Computer-assisted instruction (CAI) is one of the hottest topics in education research [9] and often claimed to revolutionise how we teach and learn [6]. Massive Open Online Courses or MOOCs are the newest manifestation of this phenomenon. However, while 2012 was being praised as "the year of the MOOC", more and more critical voices were heard during the last year and MOOCs are under increasing pressure to finally live up to their promise. Spoken in terms of of *Gartner's Hype Cycle* [8], we could say that we're either at the peak of inflated expectations, or already entering the *through of disillusionment* [3, 15, 10].

This however does not mean that online learning isn't advancing in many interesting directions: Kahn's academy emerged more or less organically when Salman Kahn started teaching his cousin mathematics using short videos. When Salman realized a lot more children could benefit from these lessons, he started distributing them on YouTube. Today, Kahn Academy reaches 10 million students per month, according to Wikipedia. Wikipedia itself has become an integral part of traditional education too. Some researchers expect that learning in general will evolve from an individual task centred around the teacher-student dichotomy, to a collaborative social activity, in which online knowledge bases like Wikipedia, forums, social networks and Question & Answering systems are playing an ever more important role [4]. In this paper, we will try to find evidence of the claimed collaborative properties of Q&A systems, more in particular the music forum site of Stack Exchange[1]. Though the analysis is based on text-based feedback, it is expected that the dynamics of feedback in collaborative activities also hold in multi-modal situations.

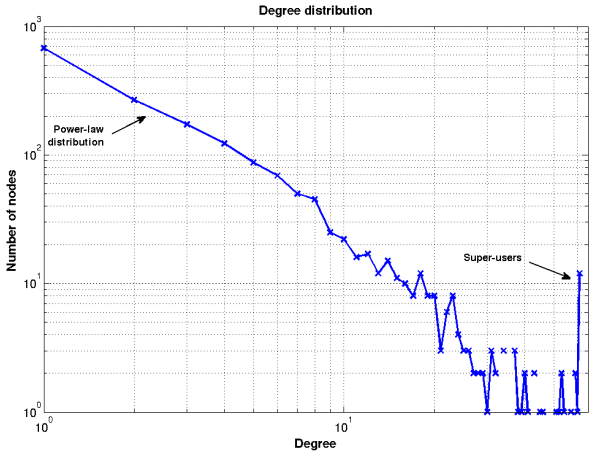


Figure 1: The degree distribution shows that the network of user-interaction is scale-free, which supports the hypothesis that there is no symmetry of knowledge.

This paper is structured as follows. First, the pedagogical background of collaborative learning is set out, based upon the work of Dillenbourg [7] and conditions for and indicators of collaborative learners are introduced. Next,

[1]http://music.stackexchange.com

Published in CEUR-WS:
FFMI workshop (Schmidt-Thieme and Janning)
In EDM 2014 Extended Proceedings (Gutierrez-Santos and Santos)

educational data mining techniques are applied [12] to find evidence of collaborative learning in *crowd learning* systems, more specifically Question and Answering systems like Stack Exchange. Lastly, a critical discussion is performed and suggestions towards more collaborative Q&A systems are proposed, to end with conclusions.

2. COLLABORATIVE LEARNING

2.1 Pedagogical approach

Existing definitions of collaborative learning in the academic fields of psychology, education and computer science, differ significantly and are often vague or subject to interpretation. We thus needed a theory that unified the different theories and was applicable to the online, computerised world as well. Not the least, it had to be easily operationalisable. A review of the literature brought us to the work done by Pierre Dillenbourg [7] that perfectly suited our requirements. Dillenbourg takes a broad view on the subject and argues that collaborative learning is a *situation* in which two or more people *learn* through *interactions*.

This means that *collaborative learning can not be reduced to one single mechanism*: just like people do not learn because they are individual but rather because the activities they perform trigger learning mechanisms, people don't learn collaboratively because they are together. Rather, the interactions between the peers create activities (explanation, mutual regulation,...) that trigger cognitive learning mechanisms (elicitation, internalisation, ...) [7].

For these processes to be effective, some requirements need to be fulfilled. A subset was extracted that could be measured numerically, albeit indirectly, using the information available in our data set (summarized in Table 1). In the next section we will have a closer look at these indicators.

2.2 Indicators

Dillenbourg discriminates three important aspects for collaborative learning to be effective and characterises situations, interactions and processes as *collaborative* if they fulfil the following criteria:

- Peers are more or less at the *same level*, have a *common goal* and *work together*;
- Peers *communicate interactively*, in a synchronous and *negotiable manner*;
- Peers apply mechanisms like *internalisation, appropriation and mutual modelling.*

These high-level criteria have been refined by Dillenbourg into more detailed conditions for collaborative learning, of which a subset has been summarised in Table 1. Each corresponding indicator provides indirect circumstantial evidence for each criterion, as our analysis was limited by the data available in the Stack Exchange. Nevertheless, as we will see, they give useful insight in the formation and dynamics of open online collaborative communities for learning.

The research in this paper can be seen as an extension of previous research in Educational Data Mining, that measured participation and interaction between students [11] and the successful formation of learner's communities [1, 13].

3. QUANTITATIVE ANALYSIS

Stack Exchange can be considered as a distant-learning autodidact platform in which communities are formed organically and learning is not guided by a curriculum or some central authority, but exclusively by the members of the community, in contrast with MOOCs. This paper aims at answering the question whether the necessary conditions for collaborative learning emerge spontaneously in these platforms. As the work is done in the context of the PRAISE project[2], a social media platform for music learning, the Music Stack Exchange data set was chosen.

Stack Exchange provides an open API, from which all data can be exported. The data set consisted of 2400 questions, 1500 active members and 1.7 million page views The platform is basically a forum in which anyone can ask and reply to questions. As a means of quality control, users can give up- and down votes to questions, and answers. People can also comment on questions and answers which is actually some kind of meta-discussion in which feedback on relevance, terminology, etc... is given. In the following paragraphs, the criteria listed in Table 1 will be studied in more detail.

3.1 Symmetry of action

Symmetry of action expresses the extent to which the same range of actions is allowed by the different users. Stack Exchange employs a system of so-called *privileges*, attributed according to your reputation[3]. These privileges are generally connected to *moderation rights*, rather than with the actions of asking and replying to questions – unless you have a negative reputation. The fact that users can exert the same actions, does not imply that this also actually the case. An analysis of the distribution of the ratio of answers over the number of questions, reveals that we can roughly discriminate *three kinds of users*, based upon their activity profile:

- *Silent users (62% of the registered users)* that never answer, e.g. users that don't register or register but do not ask questions nor reply to them;
- *Regular users (37% of registered users)* that give roughly as much as answers as they ask questions, that is, two on average;
- *Super-users (<1% of the registered users)*, these are 'hubs' that give at least 40x more answers than they ask questions.

The largest part (96%) of *regular users*, ask less than five questions, and 76% even asks only one question: there are *no 'parasite' users between the regular users that ask question but do not answer.* From the other side, only 8 'expert' super-users (0.5% of the community) were responsible for answering 25% of the questions. Above findings indicate that **the symmetry in action is highly skewed because of a small group of 'super-users' and a large group of 'silent users'.**

[2]http://www.iiia.csic.es/praise/
[3]http://stackoverflow.com/help/privileges

Aspect	Criterion	Indicator
Situation	Symmetry of action	Ratio of answers and questions per user
	Symmetry of knowledge	Scale-freeness of the user interaction graph
	Symmetry of status	Distribution of reputation within the community
Interactions	Synchronous	Response times of answering to questions
	Division of labour	Distribution of questions and answers in the community

Table 1: Criteria of collaborative learning according to Dillenbourg, with corresponding indicators. The indirect nature of the indicators stems from the fact that only meta data was available from the Stack Exchange data set, and that the criteria in general are very hard to measure quantitatively.

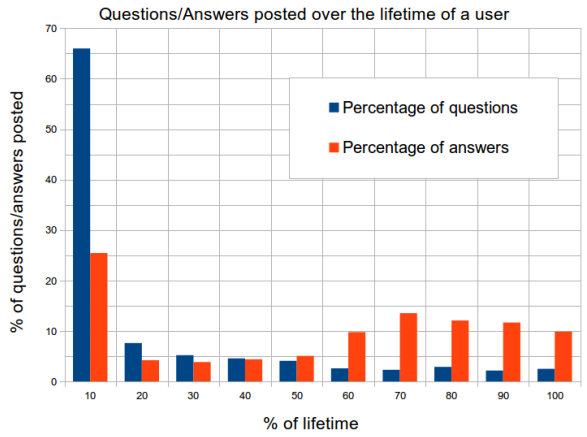


Figure 2: Users tend to ask more questions in the beginning when signing up, and start answering as they have been around some time.

3.2 Symmetry of status

Stack Exchange employs a *reputation system* by which members get rewarded or punished if a peer up- or down votes your answer or question, when your answer gets 'accepted', etc...

We would expect a "healthy" collaborative community to have a strong correlation between reputation and the time a user has been around on the platform: as users spend more time on the platform, their reputation builds up. An inquiry into the Stack Exchange music data set, however, reveals only a correlation of 0.23 between reputation and "time around". We could thus conclude that there is some **odd kind of symmetry, in the sense that no one really builds up reputation**.

3.3 Symmetry of knowledge

Traditionally, these reputation systems are believed to make a good indicator for the knowledge a user possesses. However, there are some problems with this reasoning:

- Knowledge is not a uni-dimensional measure, but is connected to a (sub) domain of expertise;
- Someone's reputation keeps on increasing, even without activity: there is a bias towards old posts and members;
- There is a bias towards "easy answerable questions".

Figuring the knowledge of the members directly is quite an impossible task to perform, especially in a broad and open-ended domain like music. To assess symmetry of knowledge, however, one could argue that *if* everyone in the Stack Exchange music learner's community has more or less the same expertise, *then*, on average, anyone would answer questions asked by anyone.

In other words, there would be no particular hierarchy in answering, rather the network of interaction would be "random" and *not scale-free*. Another way to put this, is to state that *no hubs of people would exist that answer significantly more questions than others*. A network is called *scale-free* if the degree distribution follows a power law[2]:

$$P(k) \sim k^{-\gamma} \tag{1}$$

with $P(k)$ being the fraction of nodes that have a degree k, and γ a constant typically between 2 and 3. Figure 1 reveals a power-law relationship, with exception this special group of "super-users". Above findings therefore suggest that **symmetry of knowledge is not observed.**

3.4 Division of labour

As pointed out before, a small group of super users answer vastly more questions than they ask: a group of 21 users answered half the questions. This is clearly not a balanced situation in which the total labour of answering questions, is equally distributed. Figure 2 shows the relative timing of when users ask and respond to questions over their lifetime.

Users tend to ask questions in the beginning (a visit to the site probably triggered by an urgent need to get a question resolved), but start answering more uniformly after a while. The graph also indicates that engagement is largest in the beginning. This information is relevant when developing platforms with a pedagogical purposes: **users probably need to be "bootstrapped", allowing them to give lesser answers and ask more questions in the beginning, so they get "locked into" the platform**.

Note that a relative plot was preferred, in which the x-axis indicates the % of the lifetime, 0% being the moment of signing up, and 100% the date the data set was obtained. It allowed us to grasp the details of both users that had just signed up, as well as users that have been active for a long time (especially as the rate of signing up is probably not constant but increases with time).

3.5 Synchronous feedback

To keep people engaged in an activity, according to the "theory of flow" [5], immediate feedback is necessary. In the case

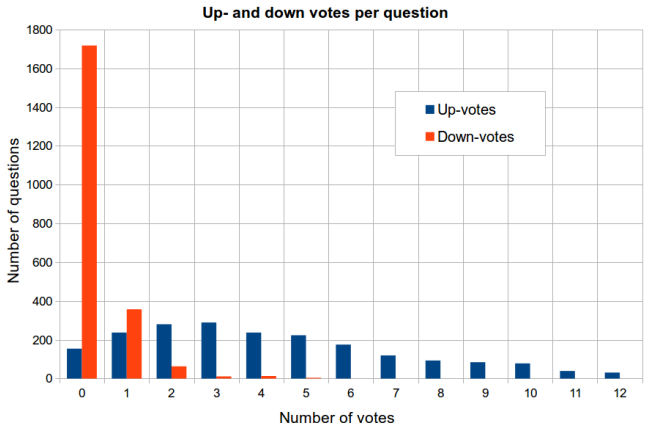


Figure 3: Users tend to give much more up-votes than down-votes to questions. Generally speaking, down-voting is only used to remove off-topic, duplicate questions or questions that are either too specific or broad.

of the music Stack Exchange platform, 68% of the questions received an answer within the day, and 20% even within the hour. This may seem odd, but closer inspection reveals that – once again – this is due to the small-group of "super-users" that are very engaged.

4. CRITICAL DISCUSSION

Based upon the analysis done in the previous section, some critical remarks and suggestions are offered to improve the pedagogical nature and collaborative learning

4.1 Remarks

4.1.1 Limited to no instructional design

The data set on Stack Exchange music's forum, is an amalgam of questions (1) with different levels of granularity, typically with a small scope, (2) on a wide range of topics, for learners (3) with different learning goals and (4) different levels of expertise. The activities are not designed to elicit collaborative learning, and as the data is unstructured, without sufficient scaffolding of the learning content (e.g. through hyper-linking), it is no natural fit for learning but rather provides **ad-hoc answers to appease short-term narrow personal learning goals**.

4.1.2 A heterogeneous community

Above remarks wouldn't be so *problematic for collaborative learning*, if proficient communities existed within the Stack Exchange platform that had more or less the same goals, expertise and engagement. In the current case, there's a risk of frustration and boredom in expert users that don't see their questions answered and who have to answer straightforward questions. For novice members, on the other hand, their learning remains limited because they do not get sufficient guidance and do not really construct knowledge.

Although *the group of super-users* makes sure that questions get answered quickly and perform the largest part of moderation, they are *potentially harmful to collaborative learning* as they distort the natural formation and dynamics of collaborative communities. From the other side, their interventions may bootstrap "young" forums.

4.1.3 Strong preference for "liking"

The dataset revealed a *very strong preference for voting up rather than down*: only two users gave more down votes than up votes and of all the people that have ever cast a down vote (72 users out of the roughly 1500 active users), 80% gave more than five times as much up-votes in return. 80% of the questions had *no down vote*, compared to less than 10% without up-vote. Figure 3 shows the distribution of up- and down-votes. This effect was even more pronounced in the answers: the number of down-votes is typically zero or very small, whereas the up-votes reach a maximum at about 3 up-votes, then slowly attenuates. A further analysis of questions with more down than up-votes, revealed that these questions where either off-topic (40%), too vague, broad or specific (35%), not real questions (10%) or Duplicate questions (8%).

4.2 Suggestions

4.2.1 Sub-communities

Allowing users to organise themselves in smaller active sub-communities with common or similar learning goals, may prove an elegant solution to manage or exploit the variety in expertise of the users. Also, the concept of reputation would make more sense. A similar idea was proposed by Santos [13].

4.2.2 Knowledge construction

Good feedback should provoke critical thinking by asking sensible questions, provide a clue to "what's next" and allow to construct knowledge through scaffolding and coupling back to acquired knowledge. Though the concept of freely asking questions is very accessible, the content stays rather ad-hoc and unstructured. A way to organise and link different questions in order to guide learners would be very useful.

4.2.3 Collaborative interfaces

In the modern ages of web technology, users could benefit from a collaborative interface in which knowledge is constructed together, in a way similar to for example Google Docs where one single entity is shared by all users. So, rather than preserving the strict question/answer or learner/teacher dichotomy, one would go for a situation in which knowledge – not only answers but also questions – is constructed live in an interactive way.

5. CONCLUSIONS

In this paper, the case for collaborative learning in open-ended auto-didact Q&A environments like Stack Exchange is investigated. Based upon the criteria put forward by Dillenbourg, we can state that though there are *technically no hindrances towards collaborative learning, the nature and dynamics of the community that organically form on Stack Exchange, do not support the case for collaborative learning.*

It was observed that the *symmetry of action was distorted* due to a small group of "super-users" that answered the majority of questions and a large group of "silent users" that do not really interact with the platform. Inspection of the

degree distribution of the user interactions reveals that the community network is scale-free, which means that *symmetry of knowledge is very unlikely*. The reputation system seems insufficient as a measure of expertise and a strange kind of symmetry of status is observed, in the sense that *no one really builds up reputation*, except for a small group of users.

Lastly, the limited possibilities to instructional design, elicits *short-term narrow and personal learning goals*. Also, the very heterogeneous nature of the community is not favourable for learning. Suggestions were made to adapt these interesting and popular platforms to learning, like *creating subcommunities with common learning goals*, extend the possibilities for *organising and structuring the content* and employ *collaborative interfaces*.

As future work, these results should be validated by means of other communities on Stack Exchange as well, and on different modes of feedback, rather than only text-based.

6. ACKNOWLEDGEMENTS

This research has been supported by the EU FP7 PRAISE project #318770.

7. REFERENCES

[1] A. R. Anaya and J. G. Boticario. A data mining approach to reveal representative collaboration indicators in open collaboration frameworks. *International Working Group on Educational Data Mining*, 2009.

[2] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[3] C. G. Brinton, C. Mung, S. Jain, H. Lam, Z. Liu, and F. Ming Fai Wong. Learning about social learning in MOOCs: From statistical analysis to generative model. *arxiv.org*, abs/1312.2159, 2013.

[4] M. A. Chatti, M. Jarke, and D. Frosch-Wilke. The future of e-learning: a shift to knowledge networking and social software. *International journal of knowledge and learning*, 3(4):404–420, 2007.

[5] M. Csikszentmihalyi. *The Evolving Self: A Psychology for the Third Millennium*. Harper Collins, New York, 1993.

[6] L. Cuban and L. Cuban. *Oversold and underused: Computers in the classroom*. Harvard University Press, 2009.

[7] P. Dillenbourg et al. Collaborative-learning: Cognitive and computational approaches. Technical report, Elsevier, 1999.

[8] J. Fenn and M. Raskino. *Mastering the Hype Cycle: How to Choose the Right Innovation at the Right Time*. Harvard Business Press, 2008.

[9] J. Hattie. *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge, 2009.

[10] T. Lewin. After setbacks, online courses are rethought. *New York Times*, (December 11), 2013.

[11] T. P. Padilha, L. M. Almeida, and J. B. Alves. Mining techniques for models of collaborative learning. In *Designing Computational Models of Collaborative Learning Interaction, workshop at. ITS*, pages 89–94, 2004.

[12] C. Romero and S. Ventura. Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1):135–146, 2007.

[13] O. C. Santos, A. Rodríguez, E. Gaudioso, and J. G. Boticario. Helping the tutor to manage a collaborative task in a web-based learning environment. In *AIED2003 Supplementary Proceedings*, volume 4, pages 153–162, 2003.

[14] M. Sharples, P. McAndrew, M. Weller, R. Ferguson, E. FitzGerald, T. Hirst, and M. Gaved. Open university: Innovating pedagogy. 2013.

[15] V. Strauss. Are MOOCs already over? *http://www.washingtonpost.com/blogs/answer-sheet/wp/2013/12/12/are-moocs-already-over/*, (December 12), 2013.

Creative Feedback: a manifesto for social learning

Mark d'Inverno
Department of Computing
Goldsmiths, University of London
+44 207 919 7701
dinverno@gold.ac.uk

Arthur Still
Department of Computing
Goldsmiths, University of London
+44 207 919 7701
awstill@btinternet.com

ABSTRACT

Arguably one of the most important activities of a university is to provide environments where students develop the wide variety of social and intellectual skills necessary for giving and receiving feedback. We are not talking here about the kinds of activity typically associated with the term "feedback" - such as that which occurs through individual course evaluation questionnaires or more universal systems such as the National Student Survey, but the profoundly creative and human act of giving and receiving feedback in order validate, challenge and inspire. So as to emphasise we are talking about this kind of feedback, we coin the term "creative feedback" to distinguish it from the pre-conceived rather dreary compliance-inflected notions of feedback and set out in this paper to characterise its qualities. In order to ground and motivate our definition and use of "creative feedback" we take a historical look at the two concepts of creativity/creative and feedback. Our intention is to use this rich history to motivate both the choice two words, and the reason to bring them together. In doing so we wish to emphasise the characteristics of an educational philosophy underpinned by social interaction. By describing those qualities necessary to characterise creative feedback this paper sets out an educational philosophy for how schools, communities and universities could develop their learning environments. What we present here serves not only as a manifesto for designing learning environments generally but as a driver for designing technologies to support online social learning. Technology not only provides us with new opportunities to support such learning but also to investigate and evidence the way in which we learn and the most effective learning environments.

Keywords: Feedback, creative, creativity, learning, technology

1. INTRODUCTION

When the word feedback is mentioned in universities - as happens now with increasing frequency - there are usually one or two winces around the room. The problem it is a word that has become associated with compliance, with checking competency, with measurement and judgement, with having to go through the motions of various government or funding body processes and, perhaps too, with feeling beholden to open up channels of communication so as to hear things that we would rather not have to hear. This is a pity, and especially so at universities, because feedback is central to learning. Not just to learn a discipline, but to learn about the way we are, to learn about the way we think, to learn about the way we interact and about the way in which we produce and value our work. Whether that work is an analytical or interpretive essay, whether it is a poem or a composition, whether it is a new performance or a new artwork, it is only through actively seeking feedback both from others and from ourselves that we learn.

At one level it is clear that without the on-going feedback that we sense and perceive from our environment we could not operate or survive. Without basic perceptual acts such as seeing, hearing and touching we couldn't function for very long. However, feedback is also necessary to experience ourselves as social beings, and especially to understand and investigate the process of social interaction between individuals. Sometimes the communication from one human to another is like an experiment whose result is evidenced by the feedback perceived from the other [22]. For example, shouting "hello?" to check whether anyone is at home, the result might be the perception of a response like "I'm in the kitchen!" or complete silence. This is an example of a simple feedback loop at work providing evidence for a model of the world. At the other extreme feedback loops can be continuous and extremely complex, and often below conscious awareness such as when two jazz musicians are improvising together [54]. In all cases feedback is the way in which we understand the world we are in, and learn about our physical and social place within it.

Suppose you are a learning to play music, for example. If you play a piece of music then the only way you can know how it was heard and experienced by others is to get their feedback on your performance. This feedback will be absolutely critical if you want to understand how you can improve yourself as a performer. Of course in any performance sustained self-feedback is critical too and musicians are skilled enough to give themselves this on-going and continuous feedback as they play. In addition to this, musicians have the option of recording performances and listening to them later in order to provide an entirely new perspective. The distance created in time and space, and moving from performer to listener, provides new opportunities for fresh insights on how to improve ones own performance. In addition, through an understanding of how we come across to others, we can often best advance the quality and precision of the feedback we give ourselves.

If we accept the need for building communities of feedback the issue then becomes how to build the right kinds of learning environments. If students can develop their own skills in giving and receiving feedback at school and university, then they will gain confidence in giving and receiving feedback from friends, colleagues, press and audiences too. Education environments should enable an exploration of how peers and tutors perceive essays, performances, software and artworks and in turn, how we all learn to be open to the feedback from others.

This philosophy is very strong in the Art department at Goldsmiths, where the emphasis is very much focused on developing communities of feedback. This department is especially interesting because of its reputation for producing world-class artists that have become important cultural and creative pioneers in the UK.[1] In our observations, first, second and third year undergraduates come

[1] (Damien Hirst, Malcolm McClaren, Mary Quant, Lucien Freud and Anthony Gormley are all alumni of the Art department. Other alumni include Laurie Provoust who currently holds the Turner prize and Steve McQueen who won a Bafta and Oscar for best film with "12 years a slave". The question to us is whether developing communities of creative feedback is the key to the Art department's success.)

Published in CEUR-WS:
FFMI workshop (Schmidt-Thieme and Janning)
In EDM 2014 Extended Proceedings (Gutierrez-Santos and Santos)

together weekly in order to give feedback on a small selection of undergraduates work. The students clearly worked as a group in balancing praise and criticism, combining the emotional and analytical, and moving from the sociological to the political. In all these open conversations students are learning about how to give and receive feedback to each other and understanding the ever present gap between any intention behind an artwork, and the perception by others. One of the most fascinating aspects observed in these sessions was the ability of students to take a sufficient emotional distance in order to be open to feedback, and to experience it freely without personalising anything. This ability is not only key in terms of learning how others experience their work but becomes an important skill for artists moving into a professional sphere with the free-for-all comment and criticism that social media now encourages.

Arguably then, a learning institution's key objective is to provide the kind of supportive and trusting environments where students can develop their ability to give and receive feedback in a culturally-aware, sensitive, mindful, critical and challenging way. We certainly think so, and would like a label to describe the kind of feedback we have in mind, and for this we choose the term "creative feedback". In this paper we provide a historical account of the notions of creative and creativity in order to justify the use of this term in an educational context. Moreover, by using this term explicitly the hope is we can rescue the concept of feedback from its often rather dreary compliance-inflected interpretation.

In what follows we will call upon our experience as educators spanning mathematics, psychology, psychotherapy, music and computer science, to try to explain what we mean by creative feedback and to justify our use of this term. To do this we need to take a brief historical look at the concepts of "creative" (and the related "creativity") and "feedback" – particularly though not exclusively in an education context - in order to explain exactly what we mean by these terms and why we are bringing them together specifically. The aim of the historical analysis is to give currency to the use of the term and the underlying manifesto for learning. We clearly need to be mindful of using the word "creative" when it is used so loosely, and for so many different educational, marketing and political reasons. We not only have creative writing and creative learning but now we have creative musicianship, creative computing and creative financing, not to mention the growing importance given to "creative industries" and economic arguments about why they are such an important part of our future. The word is in danger of being no more than what is approved of, and we wish to recover an older and fuller meaning for our purposes.

Aims. In this paper we set out to characterise creative feedback as the basis of an educational philosophy that is inspired by the American psychologist, philosopher, and educationalist John Dewey. The idea that follows naturally from this is that we structure schools, learning groups and universities as "communities of discovery". There are a number of motivating factors for the work in this paper described next.

The first is the desire to build educational environments (which include online environments) that give more people access to developing "creative feedback" skills. Creative feedback belongs to what Dewey called "creative intelligence" which is a part of all human thinking and is available to everyone. A strong part of our individual learning journey is gaining an understanding how others see us. The way we think, the way we behave, what we produce. This understanding is such a crucial part of learning that we want to build environments that encourage students to be aware of how others see them. As George Herbert Mead wrote, *"the individual mind can exist only in relation to other minds with shared meanings"* [42: p5]. If this is true, the relation to other people is grounded within a framework of feedback and the individual mind can only exist within such a framework.

Next, we want to emphasise that "creativity" depends on feedback from the world rather than being something that is an intrinsic quality that resides within individuals. It depends on feedback both in the act of creation itself, and also the social feedback that is received once it is made available to others (which may or may not amount to acclamation as great art).

As stated above feedback is not often seen as a creative endeavour but rather as being quite mechanical (tick boxes and scores) and about compliance (such as is often the case when making module feedback forms available to students). The impact of this notion of feedback on tutor/tutee relationships can often be dire. We explicitly introduce "creative feedback" to mitigate against this commonly held view of feedback and, in addition, to move away from another commonly held conception about feedback that it only exists in terms of praise and punishment. Furthermore, we want to emphasise how we are immersed in feedback as biological and social beings and we wish any definition to encompass this.

Most educationalists like us want to promote effective education as available to everyone rather than a middle-class luxury and technology clearly has an important role here. However, technology also provides opportunity to bring communities of learners together and, moreover, serve as a test-bed from which we can start to evidence the benefits of social learning over the individual, rote-learning and exam-based methodology that so dominates current political thinking. It also provides us with exciting new possibilities for understanding the way in which we learn. One of the drivers in our own research, for example, is to develop learning analytics and methodologies that can enable us to correlate creative feedback with learning.

The ability to use technology to understand and support social learning depends on whether we can construct systems that encourage humans to give and receive creative feedback. In order to achieve this we need participatory design methods working with a variety of user groups in order to design software that can support creative feedback across a whole range of disciplines (e.g. poetry, music, design, digital art). We believe a historical and educational underpinning is necessary to drive the principled design of such systems that not only support creative feedback but also allow mixed human and computational societies. One of the practical questions that we are addressing in the design of novel education systems that enable social learning is how to build autonomous artificial systems that can help exemplify creative feedback in a learning community.

2. A HISTORY OF CREATIVITY AND FEEDBACK

The Education Wars. Ever since people started arguing about education, there has been an angry debate that is still not resolved, and is especially marked today in England. On the one hand the Secretary of State for Education crusades for even more frequent and stringent examinations and inspections in the State-based schools, creating what his critics call "*exam factories*" [12], designed to compete with the dauntingly efficient exam factories of the Far East.[2] And on the other hand the popular educationalist

[2] "*Tougher GCSE marks pegged to China scores*". Guardian headline, 3.4.14

Published in CEUR-WS:
FFMI workshop (Schmidt-Thieme and Janning)
In EDM 2014 Extended Proceedings (Gutierrez-Santos and Santos)

Sir Ken Robinson speaks for many when he condemns such an approach for undermining creativity, which is the true goal of democratic education. It may be hard to define creativity, but everyone agrees that it is a good thing, and that it is not fostered by an exclusive focus on training students for success in exams. The emphasis on exam factories may even be self-defeating, since there are studies showing that the success of children in China and Japan depends more on the early nurturance of sociality, than on forced study and rigorous examinations [35] More like what Coffield called "*communities of discovery*" than "*exam factories*", so perhaps Gove is taking us "*ever faster down the wrong road*" [11].

Background to the Conflict. This quarrel occurs at every level of education, from toddlers to adults, and it reflects different views on the nature of children. At one extreme is the active child, full of wonder and curiosity at the world, who needs only skilled guidance from the teacher to flower into a civilized and creative adult. At the other is the resistant child, lazy and easily distracted, whose motivation and attentiveness require firm moulding and sometimes medication in order to learn lessons and become a good citizen. Around 1900 these extremes were given psychological and educational form by two prominent American thinkers [61], and this set the scene for many of the debates on education during the coming century. In the active, curious child camp sat the philosopher, educationalist and psychologist, John Dewey, the great champion of American pragmatism, which is a philosophy based on doing rather than thinking; in the other camp sat Edward Thorndike, famous throughout the 20th century for his puzzle box experiments with cats published in 1898 [56] in which he claimed to show that cats are incapable of reason and learn only through trial and error. During the second half of the 20th century both camps contributed to the new interest in creativity, which has now become a massive and well-funded research industry in Europe especially in relation to technology.

In this paper we aim to show how technology can contribute to the fostering of creativity in education in a way that can satisfy both the jeremiads of Professor Robinson and the ministerial anxieties of Michael Gove. But first we need to be clear about what kind of learner we have in mind, Dewey's or Thorndike's, since this determines what we mean by creative and creativity, and the deployment of these terms has provided a map of the hidden agendas of Psychology and Educational Theory during the 20th century.

E. L. Thorndike: Connectionism, Stimulus-Response And The Importance Of Measurement. In 1911 Thorndike published his puzzle box experiments in *Animal Intelligence*, and developed the theory that learning is initially guided by random trial and error learning, rather than rational intelligence. For Thorndike and later many Behaviourists, the unit of behaviour was the stimulus response (S-R) connection, treated as a kind of reflex. Thorndike's view was that learning takes place by establishing connections in the brain and these connections are stamped in through a system of reward and punishment. Applied to education it was argued that the randomness of the trials in initial learning showed that little is to be gained by relying on the prior capacities of the novice learner.

Connections were treated as "atoms of the mind", and Thorndike speculated that "*the vague gross feelings of the animal sort might turn into the well-defined particular ideas of the human sort, by the aid of a multitude of delicate associations*" [58: p289]. This is Thorndike's Connectionism, and it has been one of the main models guiding studies of learning throughout the 20th century, though it was quickly found that the S-R scheme needed to be extended to S-O-R [68]. In this extended scheme O refers to the state of the organism, which is made up of many variables or factors, including prior knowledge (the multitude of delicate associations), motivation, attentiveness, intelligence and many other variables.

During the second half of the 20th century computers became the new model of the mind, and the language for describing "a multitude of delicate associations" became increasingly sophisticated, eventually leading to a new brand of Connectionism as a model for perception and learning [3]. But even in its most sophisticated form, it is still about the selection of successful acts and the "stamping out" of "profitless" [58: p283] acts by reward and punishment. Nowadays we speak of input and output of information rather than S-R, but whatever the cognitive complexity of what goes on in between, a basic linear structure remains, with the environment operating on the organism, rather than the organism on the environment.

But Thorndike was not only one of the founders of S-R theory, he was also a pioneer of mental testing as a way of classifying individuals for social control, and therefore for assigning numbers to the "O" variables in the S-O-R scheme. Thorndike greatly admired the work of Darwin's cousin Francis Galton (1822-1911) who spent much of his life studying and measuring human variation and its genetic basis after reading *Origin of Species*. As part of this interest Galton became the first to use questionnaires and statistics for the measurement of human differences and Thorndike in turn became a champion of measurement in Psychology and Education. In 1904 he published *An Introduction to the Theory of Mental and Social Measurements* [57] which introduced students to the new statistical methods that were to dominate the scientific practice of Psychology

Deweyan Inquiry. The contrasting philosophy was that of John Dewey, who was one of the first to acknowledge the value of Galton's statistical discoveries [16] but had little faith in the value of measuring the worth of individual human beings [36]. He believed effective education is powered by the child's spontaneous curiosity about the world and is social, taking place in "a community held together by participation in common activities" [20: 55]. This social setting generates inquiry, a process as natural as breathing in all animals. Inquiry is an ongoing process that reveals novelty, which in turn becomes the spur to further inquiry.

In 1896 Dewey had made the revolutionary step of taking the basic S-R reflex studied in the laboratory by physiologists, not as the simple arc of Thorndike, but as a circular structure with neither stimulus nor response being dominant over the other. He argued that the S-R reflex is not an isolable molecule of behaviour, but is inseparable from an ongoing process involving what 50 years later would be called feedback.[3] Dewey was not a laboratory psychologist, and unlike Thorndike's S-R, his scheme did not lend itself to precise control, since it required freedom of action for optimal learning to take place.

The main concern for the teacher therefore is to guide this action toward educational goals, and to avoid stifling freedom through the indiscriminate "stamping out" of what Thorndike referred to as "profitless" acts. For Dewey these "profitless" acts are part of what

[3] Thorndike's S-R connectionism also involved a rudimentary form of feedback. Reward and punishment applied to isolated S-R connections are feedback. But Dewey seemed to have in mind what we now think of as a self-organising system, in which the parts, which we may for convenience label stimulus, response, feedback, etc., cannot usefully be isolated and studied as "laboratory preparations" outside the system. The knowledge gained by an inquiring child involves, not a changing array of S-R connections, but an evolving place within a system that includes its social and physical environment.

Published in CEUR-WS:
FFMI workshop (Schmidt-Thieme and Janning)
In EDM 2014 Extended Proceedings (Gutierrez-Santos and Santos)

he called inquiry and to stamp them out is to suppress inquiry and to stunt human development.

Who Has Won? In Psychology and in Education, Thorndike has won hands down:

One cannot understand the history of education in the United States during the twentieth century unless one realises that Edward L. Thorndike won and John Dewey lost [33: p185].

But as Lagemann goes on to point out, Dewey paradoxically remains a significant figure in education, dominating discussion in schools of education, and pointing to an ideal, even if it is Thorndike who prevails in practice. But occasionally an indirect Deweyan light shines through. A possible example of this was the dramatic reception in the West of Vygotsky's Zone of Proximal Development (ZPD). Dewey had a strong influence on Russian education in the 1920's when Vygotsky was developing his ideas, [39]. Vygotsky had certainly read Dewey's work [63: p53], and there is a close affinity with Dewey's ideal of "a community held together by participation in common activities" [20: p55]. ZPD contrasted the child's developmental level when measured by conventional tests, with the level shown under adult or peer guidance [63: p86] where the ability to follow and imitate comes into play: "*using imitation, children are capable of doing much more in collective activity or under the guidance of adults*" [61: 88]. This presupposes "*a specific social nature and a process by which children grow into the intellectual life of those around them*" [63: p88], which comes close to the collective learning through inquiry described by Dewey. In 1966 Bruner [7] introduced the word "scaffolding" to describe what is going on in ZPD, but this has been often been limited to the capacity to benefit from adult help [67], rather than from the more general sociality of "collective activity", which leads to a form of "social constructivism" [69]. Like an education based on Deweyan inquiry, ZPD in our interpretation goes very deep, and its effects, unlike those of scaffolding (if we take the metaphor literally), cannot be removed once the construction is complete.

In Psychology too, Dewey has been lurking in the background, and his influence became more apparent once the notion of feedback spread after the publication of Norbert Wiener's *Cybernetics* [66]. Later, in 1960, *Plans and the Structure of Behavior* [46] appeared, and brought together feedback of information (rather than reward and punishment) with some of the early influences on Artificial Intelligence. These included Chomsky's generative grammar [9] and Newell, Shaw and Simon on problem solving in computers [47]. The result was the TOTE (test operate, test exit), introduced as a unit of behaviour to replace the S-R model, and the authors were quick to recognise that this was similar to what Dewey had proposed in his 1896 reflex arc paper [46: p30, 43].

More generally, affinity with the Dewey scheme rather than Thorndike's shows itself when the organism, animal or human, is treated as essentially in the world, active and subject to continuous feedback as it acts, rather than a static processor of information. Examples of this Deweyan scheme are Gibson's sensori-motor systems as a model for perception [25]; the move in Robotology from cognitive representions to a focus on sensori-motor activity [6]; Jean Lave's Situated Learning [34]; and more recent work in Psychology and Philosophy on Situated Cognition [48].

Formative Assessment and Feedback. In one respect - through the notion of formative assessment - the Deweyan influence penetrated deep into the heartlands of Thorndikean territory, measurement and educational testing.

The psychologist L.L.Thurstone studied at Chicago with a close colleague of Dewey's, George Henry Mead, and spent most of his career there. Early on in his career he proposed a Deweyan model of ongoing behaviour as an alternative to the S-R scheme [59]. But his main achievements were in test theory and a more careful analysis than was usual of what is typically meant by measurement in Psychology [60]. Lee Cronbach, whose PhD was also from Chicago, continued this critical tradition within psychological measurement. His work with Meehl on Construct Validity [14] showed the limitations of psychological testing, since it measures constructs rather than reality. And he recommended that assessment be part of the learning process, rather than a test given after the learning is over [13]. Later this was labelled "formative" by contrast with the conventional "summative" assessment [50]. Summative assessment was by tests after the course had ended, whereas formative assessment was assessment during the course, designed as part of the learning process. It is closer therefore to a Deweyan rather than a Thorndikian philosophy of education, and the formative assessor joins "a community held together by participation in common activities" [20: p55]. Formative assessment involves what came to be called formative feedback. In formative feedback the student is given ongoing information about performance, and the term has replaced the concepts of reward, punishment and reinforcement. But the old S-R scheme dies hard, and many of the experiments reported on formative feedback seem quite similar to those by Thorndike and others of 80 years ago [51]. They are a long way from the feedback of a sensori-motor system that is the necessary vehicle for Deweyan inquiry. This same pattern - an apparent massive victory by the Thorndike camp, yet a persistent critical or subversive presence from the Deweyans - exists in the field of creativity, where the difference between the two viewpoints is especially marked and important given that the concept of creativity is so dominant in educational discourse.

Creative Intelligence. In literature on Creativity, which spans many disciplines and is now remarkably large and increasing every year, two distinct points of view about its nature have remained unchanged. The first is that it is a puzzling and wonderful property of the human mind that has given rise to all great human achievements.[4] The second is that it is a perfectly ordinary and basic property of all human and perhaps even animal behaviour. The reason for this strange contradiction between the two meanings, which seems to have gone largely unnoticed, may be because the modern word "Creativity" derives from two distinct ways of thinking about novelty and innovation in the world. The first of these, which sees creativity as the basic process of every mind, belongs to the Deweyan view. The second, which came later, sees creativity as a marvellous addition to the mechanical processes of ordinary thinking; this belongs to the Thorndikean view.

Figure 1. Creative and Creativity in Google's nGram

As the diagram above suggests, the popularity of words like "creative" and "creativity" is only quite recent. Originally both words were the prerogative of God, who was unique in being able to make something (the world) out of nothing. This is what

[4] "*Creativity is consensually viewed as one of the most remarkable characteristics of the human mind*." Cardosa (8:147). Creativity "*is the humble human counterpart of God's creation*" Arieti [1: 4].

Published in CEUR-WS:
FFMI workshop (Schmidt-Thieme and Janning)
In EDM 2014 Extended Proceedings (Gutierrez-Santos and Santos)

creation meant, making something out of nothing. With this in mind, "Creative" (though not creativity) was occasionally extended to women giving birth and in the 19th century to refer to the divine and mysterious work of poets and artists[5]. This can be seen clearly in the diagram above.

But after the widespread acceptance of the Theory of Evolution by the end of the 19th century, the world itself could be seen as creative through variation and selection, with no help from God. This is how it is used in the title of Bergson's *Creative Evolution* [4] which was first published in French in 1907, and then translated into English four years later[6]. This was a book that was widely discussed, especially in the pragmatist circles around William James in Harvard and John Dewey in Chicago.

Dewey's *Creative Intelligence* was published later in 1917, and the word "creative" in the title was not being used to pick out one kind of intelligence amongst others, but to emphasise that human intelligence is inherently creative through a natural process of deliberate variation and invention. This could be the herald of a new beginning for education, since according to the traditional philosophies, "*If ever there was creation it all took place at a remote period. Since then the world has only recited lessons."* [21: p23]. Dewey thought that reciting lessons is a way of suppressing the variation that is necessary for creative intelligence to flourish. There was nothing divine about Dewey's view of creative thought, and he made little use of the popular concept of genius, instead seeing art and creativity as present in the most mundane activities: "*The sources of art in human experience will be learned by him who sees how the tense grace of the ball-player infects the onlooking crowd; who notes the delight of the housewife in tending her plants, and the intent interest of her goodman in tending the patch of green in front of the house"* [18: p3].

In this philosophy, education involves social control, but not via rules dictated by authority. Instead Dewey took as a benign paradigm of social control that of children playing games, in which the control is not from on high, but is naturally social from "a community held together by participation in common activities" [20: p55]. This underlies his practical experiments in education in the experimental schools he set up first in Chicago, later at Columbia University.

Creativity. The modern word "Creativity" came into play a little later than "creative," in the mid 1920's [45]. In 1924, around seven years after Dewey's Creative Intelligence was published, the mathematician and philosopher Alfred North Whitehead was invited to Harvard, where he developed the process philosophy for which he is best known. At the centre of this philosophy was his concept of creativity, a term he coined from the Medieval Latin "creare". [63: p208]. This was his word for the evolution of forms or species. Darwin had shown how this could be a property of organic evolution, and Whitehead applied the same basic structure (variation, and a means of fixing change) to the universe as a whole. It was his metaphysical principle through which entities are created out of flow ("*all things flow*" [65: p208]) which is more basic than the things that we experience. New forms (the solar system, new species) emerge and creativity is the power that enables this to happen. Dewey read this as a universal generalisation of his own views of human invention, managed by creative intelligence out of variation, and wrote approvingly about Whitehead and his ideas of creativity in 1937 [19]. On this view, there is nothing special about creativity. It is a basic principle of the world, and human creativity is no more than a reflection of this.

From Creativity to Social Creativity. Dewey's friend and colleague the social psychologist G.H. Mead had contributed one of the chapters in Dewey's *Creative Intelligence* of 1917 writing, "*The individual in his experiences is continuously creating a world which becomes real through his discovery"*. [41: p210] After reading Whitehead, he used the word "creativity" in his lectures during the 1920's, [41: p325], and it appeared in his best known book "*Mind, Self and Society*" [40] which was widely read.

There Mead described how any individual self is constituted by the social and physical environment it inhabits, but at the same time affects the environment in which the it is situated. More generally, the organism is partly determined by its environment, but also *"is determinative of its environment*" a more general version of the circular process described by Dewey [17]. Thus the word "creativity" is will have been familiar to the many readers of Mead and Dewey, and they would have had a common understanding that there was nothing special about it, not linked to genius but essential for the thinking of every human being and animal.[7]

Creativity as Faculty. But when creativity re-emerged in 1950 [26] it had a different meaning, and came from a different tradition of Psychology, that of Psychological measurement, therefore closer to Thorndike than to Dewey. It was not about creativity as the generation of change and novelty in the world, but referred instead to a personality characteristic. Launched by J.P. Guilford in 1950 in a presidential address to the American Psychological Association, he started by expressing astonishment at the lack of work on Creativity. He made no mention of Whitehead, Dewey or Mead, and based his concept of creativity on Factor Analysis, discovered by Charles Spearman [52]. Spearman had actually written a book called *Creative Mind* in 1930 [53], in which the word "creativity" appears, but it is not referred to by Guilford though he is likely to have known it. Spearman was a colleague of Whitehead's at UCL for several years before Whitehead left for Harvard, and may have picked the word up from him.

By partitioning similar correlations in tables from a large number of tests, Spearman had shown how to extract distinct factors of the mind, like intelligence, perseverance, memory and so on, and now creativity, which can be used to form part of the O in the S-O-R scheme. By 1950 Factor Analysis had reached a high level of sophistication, and Guilford had isolated a factor he called Creativity, based on his test of Convergent and Divergent thinking. Convergent thinking is conventional problem solving, converging on the correct solution, divergent is open ended and was thought to allow the free play of imagination, with questions like "in what different ways can you make use of a brick?" Later many other tests of creativity were devised including Torrance's Incomplete Figure Test [62] tests of insight, similar to Duncker's classic candle problem [23] and of "remote associations" Mednick et al [44].

The Creativity Bandwagon. The vastness of the bandwagon launched by Guilford has been extraordinary, and cannot be

[5] "*But this I know; the writer who possesses the creative gift owns something of which he is not always master--something that at times strangely wills and works for itself*." Charlotte Brontë in editorial preface to 1850 edition of Wuthering Heights [5, p 1iii].

[6] Translation of Bergson's *L'Évolution créatrice* from 1907 as *Creative Evolution* in 1911 [4].

[7] Vygotsky had a similar view: "*just as electricity is equally present in a storm with deafening thunder and blinding lightning and in the operation of a pocket flashlight, in the same way, creativity is present, in actuality, not only when great historical works are born but also whenever a person imagines, combines, alters, and creates something new, no matter how small a drop in the bucket this new thing appears compared to the works of geniuses.*" [64: p10-11]

Published in CEUR-WS:
FFMI workshop (Schmidt-Thieme and Janning)
In EDM 2014 Extended Proceedings (Gutierrez-Santos and Santos)

explained only by the happy Utopian vision offered by the definition that runs throughout the literature: "a creative response is novel, good, and relevant." [32: xiii]. From a comfortable seat on board in 1966, Liam Hudson wrote:

'Creativity' . . . applies to all those qualities of which psychologists approve. And like so many other virtues . . . it is as difficult to disapprove of as to say what it means. As a topic for research, 'creativity' is a bandwagon; one which all of us sufficiently hale and healthy have leapt athletically abroad [29: p100-101].

But why, what are the reasons for the astonishing success of the Creativity bandwagon, which continues to gain speed, and has left in its wake a whole set of often quite unrelated "creative industries" (media, advertising, TV, film, design, games). Even banking is given the epithet creative without a trace of irony, as well as the great entrepreneurs, led by Richard Branson. Here are just a few of the possible reasons for this remarkable juggernaut.

A. It is held together by the scientific armour of Factor Analysis, a way of constructing smooth curves from the uncertain data of questionnaires.

B. Protected by this show of rigour, it was able to break away from the aridities of Behaviourism, which had given Psychology its needed scientific respectability but had bored students for years.

C. The giants of Humanistic Psychology got on board, each with a mouth-watering trade mark to draw students to Creativity 101: Carl Rogers' self-actualization in 1954 [49], Csikszentmihalyi's flow in 1975 [15], and Maslow's peak experiences in 1968 [37]. Charles Tart was there with altered states of consciousness in 1969 [55], and Frank Barron, veteran of LSD experiments in 1963 [2]. And even Buddhism, offering an endless stream of books with titles beginning "Zen and Art of . . . ." to say nothing of Kabat-Zinn's introduction mindfulness as an essential component of creativity in 1990 [31]. It all added much needed glamour to Psychology.

D. Artificial Intelligence hitched a lift. As early as 1958 Newell et al [47], had raised the problem of creativity for computers and described a programme on ILLIAC that composed music. Computational creativity has progressed independently (there are remarkably few cross references between the two disciplines) but in parallel with Psychology's version, and has probably added a further bit of hard-nosed scientific respectability to the whole endeavour.

E. Last but not least, there has been massive funding from military and industry. As Guilford wrote in 1959, soon after the launch of Sputnik by the USSR "*The preservation of our way of life and our future security depend upon our most important national resources: our intellectual abilities and, more particularly, our creative abilities. It is time, then, that we learn all we can about those resources*" [27: p469]. The economy and safety of the West is thought to depend on the practical benefits of making things that work, from nuclear weapons to the stylish artefacts of Steve Jobs, and the secret is creativity.

3. CREATIVE FEEDBACK

But in the midst of all this razzmatazz, there was a quiet Deweyan revolution. Some of it took place on the bandwagon itself, where there are researchers who stress that Creativity is an everyday matter, and that we all possess it in our capacity for flow and mindfulness. More recently there are those who have turned away from creativity with a capital C, and looked at how a more modest Deweyan creative intelligence can be encouraged throughout education [10, 24, 30]. Dewey believed that creative intelligence is necessary for democracy to prosper, and it is fostered by what we call creative feedback.

This is the goal of MusicCircle Software project at Goldsmiths; to design an online environment to support communities of creative feedback for learning to play music. It includes the ability to upload performances, share them with others, and then seek and provide creative feedback. It is developed through a process of participatory design, working with students and other users to ensure we build what people want. Through systems such as ours perhaps we can begin to reconcile the conflicting demands of Michael Gove and Ken Robinson through evidencing clearly how learning takes place through creative feedback.

In order to understand how to design learning environments, we now set out to characterise creative feedback in more detail. We do so by describing its qualities along a number of dimensions drawing both upon our historical analysis and our combined backgrounds: teaching, programme development and management in higher education; performance and composition in music; design and implementation in software; and mindfulness and psychotherapy in practice. These qualities of creative feedback are offered in hope of receiving creative feedback to inspire the next steps.

1. CF is social. It comes from one social agent who has perceived the feedback object in some way (whether that is an output or a process of an individual) to another (the originator of the feedback object). Note this definition does not preclude students giving creative feedback to their own work.

2. CF is mindful. This incorporates at least two aspects. a) That the person giving the CF is aware of the cultural and individual context of the receiver (such as an understanding of the individual's artistic or scientific goals/methods/audiences etc.) and b) That individuals are aware of any personal judgments that are being made and can articulate these if required.

3. CF contains a degree of community awareness. a) That CF embodies an awareness of what creative feedback has occurred previously but also that it features as part of a complex and developing system b) That giving and receiving CF should be embraced equally for the community to sustain itself. It would be difficult for communities to thrive if everyone wanted to give more CF than they wanted to receive of course. CF creates a self-sustaining self-organising system where flexibility and robustness need to be balanced. Whilst each learner may have more or less knowledge about what is required to maintain such a system it is clear that it can only exist if individuals in the learning environment actively encourages engagement in CF.

4. CF is clear, the language used being unambiguous and terms used mutually understood.

5. CF is democratic. Being a tutor or student bestows no special right to giving or receiving CF (though of course one might hope that tutors have more experience and skills in giving it).

6. CF is challenging. Underpinning any creative partnership is the notion of the challenge that the each brings to the other. CF that provides the right level of challenge is arguably the most sought after feedback. To do so involves "skill in means", a Buddhist concept meaning that feedback is geared to the level and character of the student, and is always open to the student's needs.

7. CF incorporates generosity of spirit and compassion. It is an act of giving and enabling, itself an essential aspect of skill in means.

8. CF is always open to discussion and further explanation.

9. CF is comparative rather than absolute. No absolute judgment about a feedback object can be made. Comparisons (explicit or implicit) of the feedback object to other existing objects is a mindful tactic in many cases and involves skill in means. (For example, CF to a jazz piano student from a tutor could simply say

how close the student's playing is to another well-known jazz pianist and how they may want to take a listen.)

We believe the key to successful education is about providing the right kinds of environments where skills in creative feedback can develop. The role of technology is both to build new kinds of learning environments but critically to start to evidence how the creative feedback ability is correlated with learning and artistic development more generally. This may have ramifications for the way in which we think about structuring learning in schools, universities and any other kind of learning community.

4. CONCLUDING THOUGHTS

We are designing a new technology at Goldsmiths called Music Circle as part of a European Project (Practice and Performance Analysis Inspiring Social Education) through the technology-enhanced learning Programme. It is designed to allow students to upload and share performances and compositions within learning communities and then by inviting feedback from others. In order to identify the kind of feedback we wish to encourage in our system (which currently operates in a blended learning context at Goldsmiths) we have identified the term "creative feedback" which embodies a range of characteristics including clarity, mindfulness, generosity, challenge and democracy.

At the heart of the motivation for designing this system is the idea that students can learn a huge amount from the creative feedback given by others. Not only that, but that the students can develop their own abilities as musicians through the ability to give creative feedback to others. And there is little doubt that the ability to receive feedback well, to depersonalise it as much as possible and respond to it appropriately, will stand students in good stead for the world of professional musicianship. Moreover, outside the professional music world, employers will be seeking students who have the skills to work in communities that have skills in giving and receiving creative feedback. Indeed one can easily imagine a world where an employer is much more interested in the way in which a student has contributed to and benefitted from being in a community. So our manifesto and agenda for change may result in students leaving universities not with a transcript of module marks but with a detailed account of their sustained engagement with creative feedback in a community of learners.

As part of the design of the system, we are designing "creative feedback agents" that are software systems that can start to provide some aspects of creative feedback on uploaded performances and compositions. With the development of techniques from audio analysis, gesture analysis, and style analysis combined with building models of learners we are looking to build systems that can start to embody some of the CF characteristics we have identified in this paper. What is important to us is that the design of our software is underpinned by a strong educational philosophy that comes from an understanding of the historical precedents and discoveries of many before us. We want to move away from the idea that technologies are designed and built by technologists and we embrace a multi-disciplinary approach where learners, educators, designers, sociologists, philosophers, historians, psychologists and computer scientists come together to build systems but with a clear understanding of the work that has come before. Perhaps more than anything this paper is a call to arms to revive and embed a Deweyian educational philosophy that can now be both supported and evidenced through technology.

5. ACKNOWLEDGEMENTS

Our thanks to Goldsmiths, Harry Brenton, Roger Burrows, Rosie Shepperd, Matthew Yee-King, Francois Pachet, Jon McCormack, Andreu Grimalt-Reynes, Melly, Maisie and Maureen Still, Sarah Khan, Jonathan James, Chris Kiefer, Carles Sierra and Robert Zimmer. This research was supported by the FP7 Technology Enhanced Learning Program Project: Practice and Performance Analysis Inspiring Social Education (PRAISE) which includes Goldsmiths, Sony Computer Science Laboratories in Paris, the Institute of Artificial Intelligence in Barcelona and VUB, Brussels.

6. REFERENCES

[1] Arieti, S. (1976). Creativity: The Magic Synthesis. New York, Basic Books.

[2] Barron, F. (1963). Creativity and Psychological Health. Oxford, Van Nostrand.

[3] Bechtel, W. and A. Abrahamsen (1991). Connectionism and the Mind: an introduction to parallel processing in networks. Oxford, Blackwell.

[4] Bergson, H. (1911). Creative Evolution. London, Macmillan.

[5] Brontë, E. (1995). Wuthering Heights. London, Penguin.

[6] Brooks, R. (1991). "Intelligence without representation." Artificial Intelligence 47: 139-159.

[7] Bruner, J. (1966). Toward a Theory of Instruction. Cambridge, MA, Harvard University Press.

[8] Cardoso, A., et al. (2000). An Architecture for hybrid creative reasoning. Soft Computing in Case Based Reasoning. S. K. Pal, T. S. Dillon and D. S. Yeung, Springer: 147-178.

[9] Chomsky, N. (1957). Syntactic Structures. The Hague, Mouton.

[10] Claxton, G., et al. (2006). "Cultivating creative mentalities: a framework for education." Thinking Skills and Creativity 1(2): 57-61.

[11] Coffield, F. (2007). Running ever faster down the wrong road: An alternative future for education and skills. London, Institute of Education.

[12] Coffield, F. and B. Williamson (2011). From Exam Factories to Communities of Discovery: The democratic route. London, Institute of Education

[13] Cronbach, L. J. (1957). "The two disciplines of scientific psychology." American Psychologist. 12(11): 671-684.

[14] Cronbach, L. J. and P. E. Meehl (1955). "Construct validity in psychological tests." Psychological Bulletin 52: 281-302.

[15] Csikszentmihalyi, M. (1975). Beyond Boredom and Anxiety: Experiencing Flow in Work and Play. San Francisco Jossey-Bass.

[16] Dewey, J. (1889). "Review of Natural Inheritance by Francis Galton." Publications of the American Statistical Association 1(7): 331-334.

[17] Dewey, J. (1896). "The reflex arc concept in psychology." Psychological Review 3: 357-370.

[18] Dewey, J. (1934 (1980)). Art as Experience. New York, Perigree Books.

[19] Dewey, J. (1937). "Whitehead's Philosophy " The Philosophical Review 46(2): 170-177.

[20] Dewey, J. (1938 (1963)). Experience and Education. New York, Collier Books.

[21] Dewey, J. et al (1917). Creative intelligence. New York, Henry Holt.

[22] d'Inverno, M. and M. Luck (2012). "Creativity through Autonomy and Interaction." Cognitive Science, 4(3): 332-346.

[23] Duncker, K. (1945). "On problem solving." Psychological Monographs 58(5, Whole No. 270).

[24] Gauntlett, D. (2011). Making is Connecting: The Social Meaning of Creativity, from DIY and Knitting to YouTube and Web 2.0 London, Polity Press.

[25] Gibson, J. J. (1966). The Senses considered as Perceptual Systems. Boston, Houghton-Mifflin.

[26] Guilford, J. P. (1950). "Creativity." American Psychologist. 5(9): 444-454.

[27] Guilford, J. P. (1959). "Three faces of intellect." American Psychologist. 14(8): 469-479.

[28] Hargreaves, D. J., et al., Eds. (2012). Musical Imaginations. Oxford, Oxford University Press.

[29] Hudson, L. (1966). Contrary Imaginations. London, Methuen.

[30] Johnston, J. S. (2006). Inquiry and Education: John Dewey and the quest for democracy. Albany.

[31] Kabat-Zinn, J. (1990). Full Catastrophe Living. New York, Delacorte.

[32] Kaufman, J. C. and R. J. Sternberg, Eds. (2010). The Cambridge Handbook of Creativity. Cambridge, Cambridge University Press.

[33] Lagemann, E. C. (1989). "The plural worlds of educational research." History of Education Quarterly 29(2): 185-214.

[34] Lave, J. (1988). Cognition in Practice. Cambridge, Cambridge University Press.

[35] Lewis, C. C. (1995). Educating Hearts and Minds. Cambridge, Cambridge University Press.

[36] Manicas, P.T. (2002). "John Dewey and American Psychology." Journal for the Theory of Social Behaviour 33(2): 267-294.

[37] Maslow, A. H. (1968). Toward a Psychology of Being. New York, Wiley.

[38] McCormack, J. and M. d'Inverno (2014). "On the Future of Computers and Creativity", AISB 2014 Symposium on Computational Creativity, London.

[39] Mchitarjan, I. (2000). "John Dewey and the development of education in Russia." Studies in Philosophy and Education 19(1-2): 109-131.

[40] Mead, G. H. (1934). Mind, Self and Society. Chicago, University of Chicago Press.

[41] Mead, G. H. (1936). Movements of Thought in the Nineteenth Century. Chicago, University of Chicago Press.

[42] Mead, G. H. (1964). Selected Writings. Chicago, University of Chicago Press.

[43] Mead, G. H., Ed. (1982). The Individual and the Social Self. Chicago, University of Chicago Press.

[44] Mednick, M. T., et al. (1964). "Incubation of creative performance and specific associative priming." Journal of Abnormal and Social Psychology 69: 84-88.

[45] Meyer, S. (2005). "Introduction: Whitehead Now." Configurations 13(1): 1-33.

[46] Miller, G. A., et al. (1960). Plans and the Structure of Behavior. New York, Holt, Rinehart and Winston.

[47] Newell, A., et al. (1958). The processes of creative thnking. Presented before a symposium at the University of Colorado, May 14, 1958.

[48] Robbins, P. and M. Aydede, Eds. (2009). Situated Cognition. Cambridge, Cambridge University Press.

[49] Rogers, C. R. (1954). "Towards a theory of creativity." ETC: A Review of General Semantics 11: 249-260.

[50] Scriven, M. (1967). The methodology of evaluation. Perspectives of Curriculum Evaluation. R. W. R. M. Tyler, R. M. Gagné and M. Scriven. Chicago, Rand McNally.

[51] Shute, V. J. (2008). "Focus on Formative Feedback." Review of Educational Research 78(1): 153-189.

[52] Spearman, C. (1904). ""General Intelligence", Objectively Determined and Measured." The American Journal of Psychology 15(2): 201-292.

[53] Spearman, C. (1930). The Creative Mind. London, Nisbet & Co.

[54] Sudnow, D. (1978). Ways of the Hand. London, Rooutledge & Kegan Paul.

[55] Tart, C., Ed. (1969). Altered States of Consciousness. New York, Wiley.

[56] Thorndike, E. L. (1898). "Animal Intelligence: an experimental study of the associative processes in animals." Psychological Review Monograph, No 8.

[57] Thorndike, E. L. (1904). An Introduction to the Theory of Mental and Social Measurements. New York, The Science Press.

[58] Thorndike, E. L. (1911). Animal Intelligence. New York, Macmillan.

[59] Thurstone, L. L. (1923). "The Stimulus-Response Fallacy in Psychology." Psychological Review 30: 354-369

[60] Thurstone, L. L. (1927). "A law of comparative judgement." Psychological Review 34(4): 278-286

[61] Tomlinson, S. (1997). "Edward Lee Thorndike and John Dewey on the Science of Education." Oxford Review of Education 23(3): 365-383.

[62] Torrance, E. P. (1962). Guiding Creative Talent. New York, Prentice-Hall.

[63] Vygotsky, L. (1978). Mind in Society. Cambridge, MS, Harvard University Press

[64] Vygotsky, L. (2004). "Imagination and creativity in childhood." Journal of Russian and East European Psychology 42(1): 7-97.

[65] Whitehead, A. N. (1929 (1978)). Process and Reality. New York, The Free Press.

[66] Wiener, N. (1948). Cybernetics. New York, Wiley.

[67] Wood, H. and D. Wood (1999). "Help seeking, learning and contingent tutoring." Computers & Education 33: 153-169.

[68] Woodworth, R. S. (1918). Dynamic Psychology. New York, Columbia University Press.

[69] Young, M. F. D. (2008). Bringing Knowledge Back In: from social constructivism to social realism in the sociology of education. Abingdon, Oxfordshire, Routledge.