

Self versus Teacher Judgments of Learner Emotions during a Tutoring Session with AutoTutor

Sidney D'Mello¹, Roger Taylor², Kelly Davidson¹, and Art Graesser¹

¹ Institute for Intelligent Systems, University of Memphis, Memphis, TN 38152 USA
{sdmello|kldavdsn|a-graesser}@memphis.edu

² Department of Teaching and Learning, Peabody College, Vanderbilt University
roger.s.taylor@vanderbilt.edu

Abstract. The relationship between emotions and learning was investigated by tracking the affective states that college students experienced while interacting with AutoTutor, an intelligent tutoring system with conversational dialogue. An emotionally responsive tutor would presumably facilitate learning, but this would only occur if learner emotions can be accurately identified. After a learning session with AutoTutor, the affective states of the learner were classified by the learner and two accomplished teachers. The classification of the teachers was not very reliable and did not match the learners self reports. This result suggests that accomplished teachers may be limited in detecting the affective states of learners. This paper discusses the implications of our findings for theories of expert tutoring and for alternate methodologies for establishing convergent validity of affect measurement.

1 Introduction

Researchers in the ITS community have always considered it important to develop a model of the learner. The model parameters can come from different sources, such as static trait measures that are extracted from learner self reports and dynamic measures that are induced from the stream of behaviors and thoughts of the learner during the course of learning. ITSs are expected to adapt their tutoring strategies to the learners' aptitude, personality, prior-knowledge, goals, progress, and a host of other parameters that presumably impact learning. It is also widely acknowledged that the scope of learner modeling need not be restricted to cognitive factors alone, because the affective states (emotions) of learners are inextricably bound to the cognitive states and ultimately linked to learning gains [1-4]. A person's affective response to an ITS can change, depending on their goals, preferences, expectations and knowledge state. For example, academic risk theory contrasts adventuresome learners who want to be challenged with difficult tasks, take risks of failure, and manage negative emotions when they occur, whereas cautious learners want to tackle easier tasks, take fewer risks, and minimize failure and the resulting negative emotions [5].

We know that events that arise during a tutoring session with an ITS cause learners to experience a variety of possible emotions that depend on the learning challenges,

the amount of changes they experience, and whether important goals are blocked. Negative emotions such as confusion and frustration occur when learners confront contradictions, anomalous events, obstacles to goals, salient contrasts, perturbations, surprises, equivalent alternatives, and other stimuli or experiences that fail to match expectations [6-7]. Positive emotions (such as engagement, flow, delight, excitement and eureka) are experienced when tasks are completed, challenges are conquered, insights are unveiled, and major discoveries are made.

There is some evidence that there are significant relationships between affective states and learning gains. Kim [8] conducted a study which demonstrated that the interest and self-efficacy of a learner significantly increased when the learner was accompanied by a pedagogical agent that served as a virtual learning companion that was sensitive to the learner's affect. Linnenbrink and Pintrich [9] reported that the posttest scores of physics understanding decreased as a function of negative affect during learning. Graesser and colleagues have demonstrated that the affective state of confusion, where learners' are in a state of cognitive disequilibrium, with more heightened physiological arousal and with more intense thought, is positively correlated with learning [1], [3]. Of course, it is important to differentiate the state of being productively confused, which leads to learning and positive emotions, from being hopelessly confused, which has no pedagogical value. The affective state of flow, where the learner is so absorbed in the material that time and fatigue disappear [10], is positively correlated with learning, whereas prolonged experiences of boredom seem to negatively impact learning gains [1].

An affect-sensitive tutor would presumably enhance intelligent learning environments [3], [11-13]. Such an ITS would incorporate assessments of the students' cognitive, affective, and motivational states into its pedagogical strategies to keep students engaged, boost self-confidence, heighten interest, and presumably maximize learning. For example, if the learner is frustrated, the tutor would need to generate hints to advance the learner in constructing knowledge, and make supportive empathetic comments to enhance motivation. If the learner is bored, the tutor would need to present more engaging or challenging problems for the learner to work on. We are currently in the process of developing a version of AutoTutor that is sensitive to both the cognitive and affective states of learners [11], [6]. AutoTutor is an intelligent tutoring system that helps learners construct explanations by interacting with them in natural language and helping them use simulation environments [3].

At this point in science, we need to answer several questions about the role of emotions in deep learning before we can build a functional affect-sensitive ITS. One important question needs to be addressed by all theoretical frameworks and pedagogical practices that relate emotions and learning: How are affective states detected and classified?

A first step is to explore a simple measurement question: How reliably can emotions be classified by humans and machines. An emotionally sensitive learning environment, whether it be human or computer, requires some degree of accuracy in classifying the learners' affect states. The emotion classifier need not be perfect, but it must have some degree of accuracy.

We have previously conducted a study that investigated the reliability by which emotions can be classified by the learners themselves versus peers and versus trained

judges [14]. Our results supported a number of conclusions about emotion measurement by humans. First, the interrater reliability between the various pairs of judges (self-peer, self-trained judge 1, self-trained judge2, peer-trained judge 1, peer-trained judge 2, trained judge1– trained judge 2) was quite low, with an average kappa of 0.18. Second, trained judges who are experienced in coding facial actions and tutorial dialogue provided affective judgments that were more reliable ($\kappa = .36$) and that matched the learners' self reports better than the judgments of untrained peers.

The overall low kappa scores between the various judges highlight the difficulty in measuring a complex construct such as emotion. It is illuminating to point out, however, that the kappas for the two trained judges in the Graesser et al [14] study are on par with data reported by other researchers who have assessed the reliability of emotion detection by [15-18]. Statisticians have sometimes claimed that kappa scores ranging from 0.4 – 0.6 are typically considered to be fair, 0.6 – 0.75 are good, and scores greater than 0.75 are excellent [19]. Based on this categorization, the kappa scores obtained in these studies would range from poor to fair. However, such claims of statisticians address the reliability of multiple judges or sensors when the phenomenon is more salient and when the researcher can assert that the decisions are clear-cut and decidable. The present research goal on emotions is very different. Our goal is to use the kappa score as an unbiased metric of the reliability of making affect decisions, knowing full well that such judgments are fuzzy, ill-defined, and possibly indeterminate.

Critics might attribute the low kappa scores achieved in previous studies to various inadequacies of our methodology. Predominant among these concerns is the lack of knowledge about emotions that people have in general, irrespective of whether the affect judges are the participants, their peers, the trained judges, and other researchers conducting field observations on affect. Perhaps people with heightened emotional expertise (i.e., knowledge, intelligence), such as social workers or FBI agents, would provide more accurate models of learners' emotions.

In this paper, we directly investigated the above criticism by measuring the degree to which people with presumably heightened emotion-detection expertise match the judgments of the learner. In particular, we assessed the reliability by which middle and high school teachers judged the emotions of the learner. The notion of teachers having heightened emotion-detection expertise emerges from diverse investigations of accomplished teachers and expert tutors. For example, Goleman [2] stated in his book, *Emotional Intelligence*, that expert teachers are able to recognize a student's emotional state and respond in an appropriate manner that has a positive impact on the learning process. Lepper and Woolverton [13] have claimed that it takes expertise in tutoring before accurate detection of learner emotions can be achieved. This requirement of expertise is apparently quite important because, according to Lepper and Woolverton [13], roughly half of expert tutors' interactions with the student are focused on affective elements. These important claims would be seriously limited if teachers are unable to detect the affective states of the learner. This question motivated the present study.

The present study tracked the affective states that college students experience while interacting with AutoTutor. We investigated the extent to which teachers can accurately identify the affective states of learners who interact with AutoTutor. This

immediate objective feeds into the long-term goal of building a version of AutoTutor that identifies and responds adaptively to the affective states of the learner. AutoTutor will never be able to adapt to the learner's emotions if it cannot detect the learner's emotions. Peer tutors and expert tutors similarly will be unable to adapt to the learner's emotions if they cannot identify such affective states.

2 Methods

The participants were 28 undergraduates at the University of Memphis who participated for extra course credit. After completing a pretest, participants interacted with AutoTutor for 32 minutes on one of three randomly assigned topics in computer literacy: hardware, Internet, or operating systems (see [3] for detailed information about AutoTutor). Two videos were recorded during the participant's interaction with AutoTutor. A video of the participant's face was recorded with a camera and a screen-capturing software program called Camtasia Studio was used to capture the audio and video of the participant's entire tutoring session.

Figure 1 depicts the experimental setup for the study. The participant interacted with the AutoTutor program on the center monitor, while the left and right monitors captured the participants body movements and face respectfully. During the interaction phase, the left and right monitors were turned off.



Fig. 1. Learner interacting with AutoTutor

After the tutorial session, participants completed a posttest on the learning of computer literacy (which is irrelevant data from the standpoint of the present study). Participants subsequently participated in a retrospective emotion judgment procedure.

The videos of the participants' face and screen were synchronized and displayed to the participants (see middle and right monitors in Figure 1). The participants were instructed to make judgments on what affective states were present at 20-second intervals; at each of these points, the video automatically paused (freeze-framed). Participants could also pause the videos at any time in between these 20-second points and make affective judgments at those points.

A list of the affective states and definitions was provided to the participants. The states were boredom, confusion, flow, frustration, delight, neutral and surprise, the emotions that were most frequently experienced during previous research with AutoTutor [1], [20].

In addition to the *self judgments* that were provided by the participants, two middle school teachers judged all of the sessions individually. The teachers were accomplished Master teachers in Memphis middle and high schools who were recognized for their accomplishments in motivating students and promoting student learning. Since affect judgment is a time consuming procedure, both teachers judged either the first half or the second half of each participants session. Specifically, for 14 randomly assigned participants, both teachers made affective judgments on the first half of the participants' AutoTutor session. Both teachers judged the second half of the remaining 14 sessions.

3 Results and Discussion

Interjudge reliability in judging emotions was computed using Cohen's kappa for the three possible pairs of judges (self vs. teacher1, self vs. teacher2, and teacher1 vs. teacher2). The observations included those judgments at the 20-second interval polling ($N = 1459$) and those in-between observations in which at least one judge observed an emotion in between two successive pollings ($N = 329$). Cohen's kappa scores were computed separately for each of the 28 learners.

We performed a repeated measures ANOVA, with the three judge pairs as within subject factors, and the order (first half vs. second half of participants session) as a between subject factor. There were statistically significant differences in kappa scores among the three judges, $F(2, 52) = 6.783$, $MSe = .01$, $p < .01$, partial $\eta^2 = .207$. Bonferroni post-hoc tests indicated that there were no significant differences in the kappa scores between the self and the teachers ($\kappa_{\text{self-teacher1}} = .076$, $\kappa_{\text{self-teacher2}} = .027$). However, kappa score between the self and teacher2 was significantly lower than the kappa between the two teachers ($\kappa_{\text{teacher1-teacher2}} = .123$). Furthermore, the interaction between judge pair and order was not significant $F(2, 52) < 1$, $p = .859$, indicating that kappa scores were the same irrespective of whether the judgments were made on the first or the second half of the learners' AutoTutor session.

These results support the conclusion that teachers are not particularly good at judging the learners emotions. Judgments provided by the two teachers were not very reliable (i.e. the teachers did not agree with each other) and did not match the learners' self reports. Before we accepted this conclusion too cavalierly, we examined whether the different judge types (self vs. teachers) are sensitive to a different set of emotions. We answered this question by examining the proportion of emotions re-

ported by each judge. Table 1 presents means and standard deviations for the proportion scores that were computed individually for each of the 28 learners and 3 judges.

Table 1. Proportion of emotions observed by self and teachers

Emotion	Self		Teacher1		Teacher2		Mean Judges	
	Mean	Stdev	Mean	Stdev	Mean	Stdev	Mean	Stdev
Boredom	0.155	0.137	0.044	0.078	0.074	0.072	0.091	0.057
Confusion	0.186	0.149	0.065	0.051	0.188	0.119	0.146	0.070
Delight	0.031	0.048	0.019	0.027	0.009	0.021	0.020	0.011
Flow	0.192	0.173	0.634	0.107	0.575	0.187	0.467	0.240
Frustration	0.130	0.122	0.145	0.097	0.032	0.044	0.102	0.061
Neutral	0.284	0.248	0.074	0.060	0.108	0.107	0.155	0.113
Surprise	0.022	0.029	0.018	0.027	0.014	0.026	0.018	0.004

We performed a $3 \times 7 \times 2$ factor repeated measures ANOVA on the proportions of emotions observed by the three judges. The two within subject factors were the affect judge with 3 levels (self, teacher1, and teacher2) and the emotion with 7 levels (boredom, confusion, delight, flow, frustration, neutral, surprise). The order (first half vs. second half of participants session) was included as a between subject factor. The proportion scores are constrained to add to 1.0 within order and judge, so it is not meaningful to consider the main effects of order and judge. However, the main effect of emotion and the remaining interactions are not constrained and therefore justifiable.

The main effect for emotion was statistically significant, $F(6, 156) = 118.455$, $MSe = .017$, $p < .001$, $\eta^2 = .820$, as was also the interactions between emotion \times order, $F(6, 156) = 2.627$, $MSe = .017$, $p < .05$, partial $\eta^2 = .092$, judge \times emotion, $F(12, 312) = 32.204$, $MSe = .012$, $p < .001$, partial $\eta^2 = .553$, and the three way interaction of judge \times emotion \times order, $F(12, 312) = 1.997$, $MSe = .012$, $p < .05$, $\eta^2 = .071$. Quite clearly, most of the variance is explained by the main effect of differences in emotions and by the judge \times emotion interaction. Therefore, we performed follow up analyses of simple main effects between three judges within the seven emotions.

Bonferroni post-hoc tests indicated that the proportions of boredom, neutral, and frustration reported by the two teachers were statistically similar and quantitatively lower than the self judgments. Therefore, it appears that teachers have difficulty in detecting states such a boredom and neutral that are accompanied by a generally expressionless face that is devoid of diagnostic facial cues [21]. But what about frustration? This is arguably a state that is expressed through significant bodily arousal and animated facial expressions. We suspect that the difficulty experienced by the teachers in detecting frustration might be explained by the social display rules that people adhere to in expressing affect [22]. Social pressures may result in the learner disguising of negative emotions such as frustration, thus making it difficult for the teachers to detect this emotion.

It appears that the affective state of flow was detected at higher proportions by the two teachers than by the self. However, if self reports of affect are considered to be the ground-truth measure of affect, a majority of the instances of flow that were observed by the teachers would be considered to be false positives. It appears, that in

the absence of sufficient facial and contextual cues, the teachers attribute the learners' emotions to the flow experience. This is clearly attributing too much to the learner.

Confusion, an emotion that is fundamental to deep learning [1], [3] has a facial imprint of a lowered brow and tightened eyelids [21]. This was detected at similar rates by the self and teacher1, but at lower rates than teacher2. This pattern is plausible because judgments provided by teacher1 matched judgments provided by the participants at a somewhat higher rate, although not statistically significant, than teacher2.

Finally, delight and surprise were detected at similar rates by the self and the two teachers. Experiences of delight and surprise are rare, however (2% each when averaged across all judges), and are typically accompanied by highly animated facial activity [21]. Such salient constraints would explain why they were detected at similar rates by all the judges.

The low kappa scores between the self and the two teachers, coupled with the differences in the proportion of emotions experienced by the self and the teachers, suggest that the teachers tend to judge self classified experiences of boredom, confusion, frustration, and neutral as similar to the state of flow. This was verified by conducting a follow-up analyses that focused on isolating the source of errors in the teachers' judgments. Two confusion matrices were computed, each contrasting the self judgments with judgments by teacher1 and teacher 2. Table 2, presents an average of the two matrices.

An analysis on Table 2 revealed two clear sources of discrepancies between the self judgments and the judgments provided by the two teachers. First, the teachers appear to annotate several of the emotions as being in the state of flow or heightened engagement. For example, the teachers classified 41% of self diagnosed experiences of boredom as flow. This miscategorization is heightened for neutral, with 61% self reported neutral instances being classified as flow. The second source of classification errors occurs at instances where the teacher fails to make an emotion judgment, but the self provides a rating (see the None column). This occurs during instances when the learner makes a voluntary affect judgment, in between the 20 second stops, and the teachers fail to detect those points.

Table 2. Confusion matrix contrasting self judgments with average of teachers' judgments

Self Judg- ments	Teachers Judgments							
	Boredom	Confu- sion	De- light	Flo w	Frustra- tion	Neu- tral	Sur- prise	Non e
Boredom	0.13	0.10	0.00	0.41	0.08	0.07	0.01	0.25
Confusion	0.05	0.14	0.01	0.40	0.05	0.06	0.02	0.31
Delight	0.02	0.06	0.03	0.38	0.05	0.01	0.02	0.42
Flow	0.05	0.13	0.01	0.52	0.04	0.09	0.00	0.10
Frustration	0.03	0.08	0.01	0.46	0.05	0.05	0.01	0.28
Neutral	0.05	0.09	0.01	0.61	0.05	0.09	0.01	0.10
Surprise	0.02	0.04	0.05	0.19	0.05	0.05	0.00	0.56
None	0.02	0.03	0.01	0.07	0.04	0.01	0.01	0.83

4 General Discussion

An emotionally sensitive tutor, whether human or artificial, would presumably promote learning gains, engagement, and self-efficacy in the learner. Such a tutor should have different strategies and dialogue moves when the learner is confused or frustrated than when the learner is bored. However, both human and automated tutors can be emotionally adaptive only if the emotions of the learner can be detected. The accuracy of the detection need not be perfect, but it should be approximately on target.

We have previously documented that trained judges who are experienced in coding facial actions and tutorial dialogue provide affective judgments that are more reliable and that match the learner's self reports better than the judgments of untrained peers. [14]. The results of this study support a number of additional conclusions about emotion detection by humans. It appears that accomplished teachers do not seem to be very adept at detecting the learners' emotions. Emotion judgments provided by the two teachers were not very reliable, i.e. the teachers did not agree with each other, and their judgments showed very little correspondence to the learner's self reports. In fact the degree to which the teachers affective judgment matched the self reports of the learner were on par with peer judges and were quantitatively lower than the trained judges. So untrained peers and accomplished teachers do not seem to be very proficient at judging the emotions of the learner.

It is possible that the assessments of learner affect provided by peers and teachers would be more accurate in naturalistic settings such as tutoring sessions or classrooms, where the judgments would occur in real time and the peers and teachers would have established a rapport with the students and have vested interests in their learning. These conditions are difficult to recreate in a laboratory, as it would be difficult to envision a scenario where the learner, a peer, trained judges, and teachers could simultaneously provide online emotion judgments. Nevertheless, our results suggest that, when presented with the identical stimulus (videos of the participants face and screen), judgments by the self and trained judges were more reliable than judgments by the peers and teachers.

It appears that each type of affect judge, be it the self, the untrained peer, the trained judges, or the accomplished teachers, bring a unique set of perspectives, standards, and experience to the affect judgment task. For example, it is reasonable to presume that participants tap into episodic memories of the interaction in addition to the prerecorded facial cues and contextual features when they retrospectively judge their own emotions (self judgments).

Unlike the self, the trained judges are not mindful of the episodic memory traces of the participants. However, they have been extensively trained on detecting subtle facial expressions with the Facial Action Coding System [22], and are more mindful of relevant facial features and transient facial movements. They also have considerable experience interacting with AutoTutor. Our results suggest that training on facial expressions (diagnostic assessment) coupled with knowledge on AutoTutor dialogue (predictive assessment), makes the trained judges robust affect detectors. The trained judges exhibit reliability (they agree with each other) as well as convergent validity (their judgments match self reports). Therefore, from a methodological perspective,

retrospective affect judgments by the participant combined with offline ratings by trained judges, seems to be valuable protocol to establishing construct validity in emotion measurement, at least when compared to untrained observers, peers, and even teachers.

Affect sensitivity is an important requirement for ITSs that aspire to bridge the communicative gap between the highly expressive human and the socially challenged computer. Therefore, integrating sensing devices and automated affect classifiers is an important challenge for next generation ITSs that are attempting to broaden the bandwidth of adaptivity to include the learners' cognitive, affective, and motivational states. Although, a handful of automated affect detection systems operate in an unsupervised fashion, supervised machine learning techniques are at the heart of most of the current affect detection systems. Consequently, providing accurate models of ground-truth for a complex construct such as emotion is an important requirement for such supervised affect classifiers. We hope to have scaffolded the development of automated affect-detection systems by providing a methodology to annotate the emotions of a learner in an ecologically valid setting (randomly selected participants rather than actors and the emotional expressions occurred naturally instead of being induced), and contrasting our methodology of self plus trained judgments with alternatives (peers, teachers, observers [1], and emotive-aloud protocols [20]). We are currently developing such an emotion classifier with an eye for integrating it into an affect-sensitive version of AutoTutor. Whether an automated affect-sensitive AutoTutor has a positive impact on learning awaits future research and technological development.

Acknowledgements

Special thanks to Gloria Williams and Arlisha Darby, the two teachers who provided the affect judgments. We also acknowledge O'meed Entezari, Amy Witherspoon, Bethany McDaniel, and Jeremiah Sullins for their help with the data collection. This research was supported by the National Science Foundation (REC 0106965 and ITR 0325428) and the Institute of Education Sciences (R305B070349). Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF or IES.

References

1. Craig, S.D., Graesser, A. C., Sullins, J., & Gholson, B. (2004). Affect and learning: An exploratory look into the role of affect in learning. *Journal of Educational Media*, 29, 241-250.
2. Goleman, D. (1995). *Emotional intelligence* (New York, Bantam Books).
3. Graesser, A. C. Chipman, P., King, B., McDaniel, B., and D'Mello, S (2007). Emotions and Learning with AutoTutor. *13th International Conference on Artificial Intelligence in Education (AIED 2007)*. R. Luckin et al. (Eds), (pp 569-571). IOS Press.

4. Snow, R., Corno, L., & Jackson, D. (1996). Individual differences in affective and cognitive functions. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 243-310). New York: Macmillan.
5. Meyer, D. K., & Turner, J. C. (2006). Reconceptualizing emotion and motivation to learn in classroom contexts. *Educational Psychology Review*, 18, 377-390.
6. Graesser, A.C., Jackson, G.T., & McDaniel, B. (2007). AutoTutor holds conversations with learners that are responsive to their cognitive and emotional states. *Educational Technology*, 47, 19-22.
7. Mandler, G. (1976) *Mind and emotion* (New York, Wiley).
8. Kim, Y. (2005). Empathetic Virtual Peers Enhanced Learner Interest and Self-Efficacy. Workshop on Motivation and Affect in Educational Software at the 12th *International Conference on Artificial Intelligence in Education*, Amsterdam, The Netherlands.
9. Linnenbrink, E. A., & Pintrich, P. R. (2002). The role of motivational beliefs in conceptual change. In M. Limon & L. Mason (Eds.), *Reconsidering conceptual change: Issues in theory and practice* (pp. 115-135). Dordrecht, The Netherlands: Kluwer Academic Publishers.
10. Csikszentmihalyi, M. (1990). *Flow: The Psychology of Optimal Experience*. New York: Harper-Row.
11. D'Mello, S. K., Picard, R., & Graesser, A. C. (2007). Towards an Affect Sensitive AutoTutor, *IEEE Intelligent Systems*, 22(4), 53-61.
12. Lepper. M.R. and R.W. Chabay (1988). Socializing the intelligent tutor: Bringing empathy to computer tutors. In Heinz Mandl and Alan Lesgold (Eds), *Learning Issues for Intelligent Tutoring Systems* (pp. 242-257). Hillsdale, NJ: Erlbaum.
13. Lepper, M. R., & Woolverton, M. (2002). The wisdom of practice: Lessons learned from the study of highly effective tutors. In J. Aronson (Ed.), *Improving academic achievement: Impact of psychological factors on education* (pp. 135-158). Orlando, FL: Academic Press.
14. Graesser, A.C., McDaniel, B., Chipman, P., Witherspoon, A., D'Mello, S., & Gholson, B. (2006). Detection Of Emotions During Learning With Autotutor. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, Mahwah, NJ: Erlbaum, 285-290.
15. Ang, J., Dhillon, R., Krupski, A., Shriberg, E., & Stolcke, A. (2002). Prosody-Based Automatic Detection Of Annoyance And Frustration In Human-Computer Dialog. *Proceedings of the International Conference on Spoken Language Processing*, Denver, CO, 2037-2039.
16. Grimm, M., Mower, E., Kroschel, K. & Narayan, S. (2006). Combining Categorical and Primitives-Based Emotion Recognition. *14th European Signal Processing Conference (EUSIPCO)*, Florence, Italy.
17. Litman, D. J., & Forbes-Riley, K. (2004). Predicting Student Emotions In Computer-Human Tutoring Dialogues. *Proceedings Of The 42nd Annual Meeting Of The Association For Computational Linguistics*, East Stroudsburg, PA: Association for Computational Linguistics, 352-359.
18. Shafran, I., Riley, M. & Mohri, M. (2003). Voice signatures. *Proceedings IEEE Automatic Speech Recognition and Understanding Workshop*, Piscataway, NJ: IEEE, 31-36.
19. Robson C. (1993). Real word research: A resource for social scientist and practitioner researchers. Oxford: Blackwell.
20. D'Mello, S. K., Craig, S. D., Sullins, J., & Graesser, A. C. (2006). Predicting Affective States Through An Emote-Aloud Procedure From Autotutor's Mixed-Initiative Dialogue. *International Journal Of Artificial Intelligence In Education*, 16, 3-28.
21. McDaniel, B. T., D'Mello, S. K., King, B. G., Chipman, P., Tapp, K., & Graesser, A. C. (2007). Facial Features for Affective State Detection in Learning Environments. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society* (pp. 467-472). Austin, TX: Cognitive Science Society.
22. Ekman, P., & Friesen, W. V. (1978). The facial action coding system: A technique for the measurement of facial movement. Palo Alto: Consulting Psychologists Press.