Measuring the Effect of ITS Feedback Messages on Students' Emotions

Han Jiang, Zewelanji Serpell, and Jacob Whitehill

[1] Worcester Polytechnic Institute, Worcester, MA 01605, USA
[2] Virginia Commonwealth University, Richmond, VA 23284, USA
[3] Worcester Polytechnic Institute, Worcester, MA 01605, USA

Abstract. When an ITS gives supportive, empathetic, or motivational feedback messages to the learner, does it alter the learner's emotional state, and can the ITS detect the change? We investigated this question on a dataset of $n = 36$ African-American undergraduate students who interacted with iPad-based cognitive skills training software that issued various feedback messages. Using both automatic facial expression recognition and heart rate sensors, we estimated the effect of the different messages on short-term changes to students' emotions. Our results indicate that, except for a few specific messages ("Great Job", and "Good Job"), the evidence for the existence of such effects was meager, and the effect sizes were small. Moreover, for the "Good Job" and "Great Job" actions, the effects can easily be explained by the student having recently scored a point, rather than the feedback itself. This suggests that the emotional impact of such feedback, at least in the particular context of our study, is either very small, or it is undetectable by heart rate or facial expression sensors.

Keywords: intelligent tutoring systems · emotion · feedback · facial expression analysis · heart rate analysis.

1 Introduction

One of the main goals of contemporary research in intelligent tutoring systems (ITS) is to promote student learning by both *sensing* the student's emotions and *responding* with affect-sensitive feedback that is appropriate to the student's cognitive and affective state. For sensing students' emotions, a variety of methods are now available, including physiological measurements [18], facial expression analysis [23], and "sensor-free" approaches [15] based on analyzing the ITS logs. Given an estimate of what the student knows and how they feel, the tutor must then decide how to *respond*. Based on the intuition that good human tutors are often empathetic and supportive, many ITS today provide real-time "empathic feedback" to learners that tries to encourage and motivate them to keep learning. This feedback can range in complexity from short utterances [1, 9, 16, 6] to longer prompts [2, 9, 17, 14] such as growth-mindset [3] messages.

Empathic feedback messages could make learners' interactions with ITS more natural and effective, but they also increase the complexity of designing the ITS

and its control policy, i.e., how it acts at each moment. Moreover, if feedback is given injudiciously, it could become distracting and suppress learning [6]. While affect-aware ITS with empathic feedback have demonstrated some notable success [10, 23, 2], the sum of evidence of their benefit is unclear. Empathic feedback has often been evaluated as part of a treatment condition in which the feedback was not the only variable being manipulated [16, 2]. Moreover, optimistic hypothesis testing that did not account for multiple hypotheses was often used.

In this paper we investigate the instantaneous impact of ITS feedback on each student's emotional state. The context of our study is an iPad-based system for cognitive skills training [11], specifically a task called "Set" (similar to the classic card game) in which the participants must reason about different dimensions (size, color, shape) of the shapes shown on the cards in order to score a point. The participants are African-American undergraduate students at a Historically Black College/University (HBCU). As measures of emotion, we consider facial expression, heart rate, and heart rate variability, all of which can be estimated automatically, in real time, and with a high temporal resolution.

We examine the following **research questions**: Is there an instantaneous change in facial expression and/or heart rate after each ITS feedback message that is consistent across the participants? Does the evidence for such a change persist even after taking possible confounds into account? Is there evidence that at least *some* participants may exhibit a relationship between the sensor readings and the prompts, even if not all of them do? Finally, is there evidence of any non-emotional change in students' behavior as a result of the feedback messages?

2 Related Work

Empathic Virtual Agents: [17] compared an "empathetic" avatar to a "non-empathetic" one. At the start of the experiment, the empathetic avatar would ask the user, "Hopefully, you will get more comfortable as we go along. Before we start, could I please have some of your information?" with the goal of building trust and comforting the participant. In contrast, the non-empathetic one would simply ask, "Have you participated in similar tests before?" They found that the empathetic agent performed no better, in terms of changing students' self-reported mood after the intervention, than the non-empathetic agent. However, they did find in the questionnaire results that participants found the empathetic avatar to be more "enjoyable, caring, trustworthy, and likeable". In another study on virtual agents [19], the researchers compared an "empathic" virtual therapist with a "neutral" one. The empathic therapist was designed to respond to the participant "in a caring manner". For instance, at the start of the session, it would say, "I'm very happy to meet you and hope you'll find our session together worthwhile. Please make yourself comfortable," whereas the neutral therapist would say simply, "Hello, I am Effie a virtual human." The study found that the empathic therapist was beneficial, relative to the neutral therapist, only for a subset of participants; this is reminiscent of the study by [6] who found that the emotionally-adaptive ITS only helped students with less prior knowledge.

Moreover, the benefit of the empathic therapist did not persist after the first meeting between the participant and the agent.

Empathic ITS: In [20], the researchers assessed the impact of ITS empathic feedback on students' emotions by manually coding students' facial expressions (frustration, confusion, flow, etc.). They found that there was a difference, in terms of the transition dynamics of students' affective states (e.g., flow to boredom), between the feedback messages that were rated as "high-quality" versus "low-quality" by the students. [9] compared different types of ITS feedback – epistemic, neutral, and emotional – in terms of their impact on facial emotions. The epistemic feedback was more impactful than the emotional feedback in their study. However, their study did not compare to giving no feedback at all. In [14], feedback of different types – growth mindset, empathy, and success/failure – were compared in terms of students' subsequent self-reported emotions. Their results suggest that the different feedback conditions were associated with different emotions (interest, excitement, frustration, etc.). Widmer [26] employed a Wizard-of-Oz experimental design similar to ours to assess the benefit of prompts in ITS; they measured the impact on learning but not on students' emotions.

Multiple Hypotheses: Most prior studies on ITS feedback messages tested many hypotheses but did not statistically correct for this. It is thus possible that they were overly optimistic when identifying possible impacts.

3 Sensors of Emotion and Stress

In our work we investigate the impact of ITS feedback on emotion as it is expressed by facial muscle movements and changes in heart rate.

Heart Rate: Heart rate (HR) and heart rate variability (HRV) are well known and widely used as a biomarker of stress [24, 4, 18, 8]. To measure HR and HRV, we use a Polar heart monitor chest belt that is connected wirelessly to a laptop to record the inter-beat-interval (IBI) of heartbeats. We measure HR as the inverse of the IBI, and the HRV as the standard deviation of the IBI.

Facial Expression: Behavioral and medical science researchers have used facial expression as a way of assessing various mental states such as engagement [25], driver drowsiness [5], thermal comfort [12], and students' emotional states in ITS [21]. Facial expression sensor toolkits are now also used in several prominent intelligent tutoring systems [13, 23, 10]. In particular, we use the Emotient SDK from iMotions, which can recognize 20 Facial Action Units (1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 18, 20, 23, 24, 25, 26, 28, 43) [7] and 12 emotions (anger, joy, sadness, neutral, contempt, surprise, fear, disgust, confusion, frustration, positive sentiment, negative sentiment). In each frame, the Emotient SDK could provide a numeric value for each facial expression if there is a face detected.

4 Dataset

In our analysis, we examined the HBCU2012 dataset [22] which is an extension of the HBCU dataset from [25]. In HBCU2012, $n = 36$ African-American under-

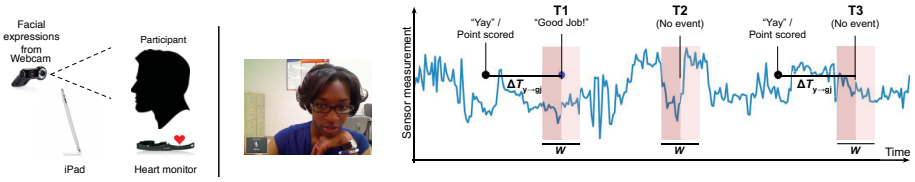

Fig. 1. Left: Experimental setup. **Middle**: View of the student from the camera. **Right**: Methodology: For all message types and sensors (heart rate, facial expressions), we compute the difference in average sensor value $W/2$ sec before vs. after an event (T1), and compare it to the corresponding difference at a random timepoint not near a message (T2). For "Good Job" and "Great Job" messages, to remove a confound due to the "Yay", we compare the difference in average sensor value at T1 to the corresponding difference at another time (T3) that is also $\Delta T_{\mathrm{y}\rightarrow\mathrm{gj}}$ sec after "Yay" but not near a feedback message.

graduate students interacted with iPad-based cognitive skills training software that is designed to strengthen basic cognitive processes such as working memory and logical reasoning. While interacting with the software, their facial expressions and heart rate is being recorded (see Figure 1). Each participant interacted with the ITS for 3-4 periods each, resulting in a total of 108 videos.

Procedure: Each student participated for 3-4 sessions, and each daily session lasted about 40 minutes. Although the system contains several tasks, the main task is called Set, which is similar to the classic card game. In this task, the player scores a point if they correctly group 3 cards together that have a correct configuration of size, shape, and color. When the student scores a point, the software automatically issues a "Yay!" sound. The Set task is highly demanding, particularly at the advanced difficulty levels and given the time pressure. At the start of each daily session, the participant takes a 3min pretest. Then, they undergo 30min of cognitive skills training that is facilitated by the system. In particular, the tutor decides the difficulty level at which the student practices, when to switch tasks to take a break, etc. The tutor also issues hints and prompts of different types (described below). During this practice section (but not during the tests), the student receives various feedback messages (see below). After the practice session, the participant takes a posttest.

Types of feedback: The tutor can issue feedback messages of various types (see Table 1). Some of them are empathetic, some are motivational, and some are goal-oriented. Note that each message type may be expressed with slightly different phrasing, e.g., "Good Job" might be spoken by the tutor as "Good Job" or just "Good"; "Try harder" can be expressed as either "It seems like you are not trying. Please try your hardest." or "Try harder."

Human-assisted ITS: While in many aspects the cognitive skills training software used to collect the HBCU2012 dataset was automated, the decisions of when to issue feedback messages were made by a human tutor (sometimes called the *trainer* in a cognitive skills training regime) who was either in another room (Wizard-of-Oz style) or in the same room (1-on-1 style) as the participant.

For the Wizard-of-Oz setting, the trainer could watch the student's face via a live webcam and also observe the student's practice on a real-time synchronized iPad. Compared with a fully automated ITS, this human-assisted apparatus might actually yield feedback messages that are more appropriately timed and chosen than what an ITS would decide.

Sensor Measurements and Synchronization: Each participant completed the cognitive skills testing and training on an iPad. The inter-beat interval (IBI) of heartbeats was recorded using a Polar heart monitor. Facial expressions were estimated in each frame (30 Hz) of video recorded by a webcam connected to a laptop. The game log was recorded wirelessly from the iPad onto the laptop. Game log, heart rate, and facial expression events were synchronized by finding a common timepoint between the face video and game log.

Prompt	**Total Events**	**Events per Learner: Avg. (s.d.)**
Great Job	1950	18.06 (12.72)
Good Job	2935	27.18 (15.89)
Nice Try	621	5.75 (5.46)
Watch Your Time	522	4.83 (2.75)
Keep Going	655	6.06 (4.99)
Faster	1025	9.49 (8.08)
Unique	55	0.51 (1.08)
Different Dims	220	2.04 (2.80)
Brief Directions	88	0.81 (1.09)
Try Harder	45	0.42 (0.98)
Missing	63	0.58 (1.33)
Extra Card	156	1.44 (4.16)
Take Break	33	0.31 (0.57)

Table 1. Frequency of the various prompts in our system.

5 Methodology

Since all participants received multiple feedback messages, we used a within-subjects design. To assess whether the various messages were associated with any immediate change in students' emotions (see Figure 1), we measured the change in the average value of a specific sensor (heart rate, heart rate variability, or one of the 20 AUs + 12 emotions) around the time (T1) when a specific message was issued. Specifically, we computed the average sensor value within a time window of length $W/2$ just after T1 and subtracted the corresponding average sensor value in the time window of length $W/2$ just before T1; this yields Δv. These values, at different times T1, constitute the treatment group of our study. Then, we computed the difference Δv (after-before) at a *random* timepoint (T2) in the participant's time series that was not within 10 seconds of any other prompt.

These values, at different times T2, constitute the control group. By comparing Δv due to the treatment vs. the control group, we can estimate the effect of the feedback message on the change in the sensor value. While this is not a truly causal inference approach, our methodology does eliminate the confound that could arise, for example, if the average sensor value tended to increase (or decrease) over time, e.g., due to fatigue.

Repeated Measures Design: Since we have multiple feedback messages and multiple days of participation for each student in our study, we use a repeated-measures design based on a linear mixed-effect model, where the student ID is a random effect. We then assess whether the presence (1) or absence (0) of the feedback message is statistically significantly related to the change Δv in a specific sensor value (facial expression or heart rate value). We repeat this for all message types and sensor values.

Hypothesis Correction: Due to many hypotheses (different messages and sensor measurements) that are largely independent of each other and lack of strong prior belief that a relationship exists between any particular sensor and feedback message, we take a conservative approach and perform Bonferonni correction to the p-values: Instead of the traditional $\alpha = 0.05$ threshold, we require $\alpha = 0.05/m$, where m is the total number of hypotheses.

Effect Size: We quantified the effect size in two ways, both of which are a form of Cohen's d statistic: (1) Global effect size: we divided the fixed-effect model coefficient for the treatment by the standard deviation of the sensor value (e.g., happiness value) over the *entire dataset* (all participants, all days, and all times). (2) Local effect size: we divided the fixed-effect model efficient for the treatment by the standard deviation of all Δv in the union of the treatment and control groups. This expresses whether the change due to the feedback message is large compared to changes that occur in other time windows of length W.

6 Analysis

6.1 Facial Expression

Analysis Details: We followed the methodology described above, where we picked 20 time points (T2) per each video such that there are no other event 10s before or after them for the control group. For the time window W, we used 5s and 10s. We allowed for the possibility that the participants' reactions to the ITS feedback messages might be slightly delayed; hence, we conducted analyses with a "right-shift" parameter τ of either 0s or 1s. Finally, for the number of hypotheses m by which we corrected the p-value threshold α, we considered that the 12 *emotions* (happy, sad, angry, etc.) can be considered combinations of individual Facial Action Units (AUs) [7] and are thus not independent of the 20 AUs we already measure. Since there are 13 different ITS feedback messages that we consider, we thus let $m = 13 * 20 = 260$ so that our threshold α for statistical significance by Bonferonni correction is $0.05/260$.

Results: Only 2 of the 13 feedback messages showed any stat. sig. impact, after p-value correction, on *any* of the 32 facial expressions for any of the right-

shift values (0s, 1s) or window sizes (5s, 10s). The two message types were "Great Job" and "Good Job", and the effects were significant across all combinations of W and τ. Table 2 show the facial expression values that have a significant change due to these feedback messages. Note that the effect sizes are generally quite small, especially when assessed at a global level (i.e., relative to the variance of the expression value over the whole dataset). The largest absolute effect size is for AU43 (closing of the eyes) for both "Good Job" and "Great Job", whereby the participants' eyes tend to be more closed before than after the message.

	Great Job			**Good Job**		
Facial Evidence	**p-value**	**Global Effect Size**	**Local Effect Size**	**p-value**	**Global Effect Size**	**Local Effect Size**
Fear	2.59e-12	0.045	0.105	4.6e-10	0.077	0.092
Disgust	9.38e-09	-0.044	-0.197	1.33e-13	-0.104	-0.245
Sadness	6.06e-05	-0.023	-0.081	6.06e-05	-0.023	-0.081
Confusion	8.91e-05	-0.030	-0.150	2.48e-06	-0.062	-0.113
Neutral	-	-	-	5.9e-05	-0.067	-0.160
AU1	-	-	-	5.42e-06	0.035	0.074
AU4	6.75e-08	0.018	0.074	2.71e-07	0.032	0.068
AU5	<2e-16	0.08	0.344	<2e-16	0.120	0.335
AU7	7.11e-08	0.022	0.121	9.87e-12	0.045	0.10
AU15	-	-	-	1.30e-04	0.036	0.08
AU18	-	-	-	3.65e-07	-0.060	-0.106
AU20	7.39e-05	0.025	0.106	4.87e-05	0.034	0.041
AU25	-	-	-	1.60e-04	0.047	0.128
AU26	-	-	-	9.33e-05	0.038	0.148
AU43	<2e-16	-0.086	-0.350	<2e-16	-0.156	-0.736

Table 2. "Great Job/Good Job": Effects on facial expression values which are stat.sig. for $W = 10$s, $\tau = 0$s.

6.2 Heart Rate

Analysis Details: We varied W over 5s and 10s, and the trends were the same. For Bonferonni correction, we let $m = 26$ since we considered two different heart measures (HR, HRV) and there were 13 different message types.

Results: None of the prompts showed a stat. sig. impact on HR or HRV.

7 Effects on Individual Students

Here we consider the hypothesis that the feedback messages may affect *some* students but not others. In particular, we test, for each combination of participant, feedback message, and sensor measurement, whether there is a statistically significant difference *within each student* in the average sensor value $W/2$ seconds

after vs. before the prompt. For each combination of prompt and sensor value, we then calculate the fraction of students for which the difference is statistically significant. Importantly, this analysis allows for a different effect – some positive, some negative – on each student.

Facial Expression: We perform the analysis for $W = 10$s. If, for each student, *any* of the 32 facial expression values were significantly changed due to a feedback message, then we increment our count for that message type. We let m (number of hypotheses) be 20 (the number of unique Facial Action Units we measure) and hence $\alpha = 0.05/m$ =2.5e-03. The results shows that for most messages, less than one quarter of the students showed any effect; only the "Great Job"(18/36) and "Good Job"(19/36) affected at least half of the students

Heart Rate: We varied W over 5s and 10s, and the trends were similar. For Bonferonni correction for each participant, we let $m = 2$ since we considered two different heart measures (HR, and HRV). The trend is similar as for the facial expression measures ("Great Job": 16/36; "Good Job": 19/36).

8 Impact of "Great Job" and "Good Job" Messages

Our analyses have found robust (over multiple sensor measurements, right-shifts, and window sizes) evidence of a relationship between the "Great Job" and "Good Job" messages and facial expression (but not heart rate), despite the conservative Bonferonni correction. However, there was little evidence in support of any other feedback message. Given that these two message types almost always occur shortly after the student has scored a point, we explored whether the change due to the feedback itself or simply because the point scored a point. To examine this, we modify the methodology from Section 5 so that the control group for these messages is taken at times T3 that are $\Delta_{\mathrm{y}\rightarrow\mathrm{gj}}$ after a "Yay"/point scored timepoint but where no such feedback occurs (see Figure 1). Importantly, the decision of whether or not "Good Job"/"Great Job" was given was at the discretion of the human trainer and was essentially random (i.e., quasi-experimental analysis). This allows us to isolate the effect of the feedback itself, rather than of the preceding "Yay" sound. We estimated the value $\Delta_{\mathrm{y}\rightarrow\mathrm{gj}}$ over all the "Great Job" and "Good Job" messages in our dataset (around 1.091s).

Analysis Details: We selected "Great Job" and "Good Job" timepoints T1 such that there is no other message before and after 5 seconds except a "Yay". We also randomly selected a similar number of time points for T3. We varied W as 5s or 10s, and we let τ be 0s or 1s. Since there are now just 2 feedback messages and 20 AUs, we let $m = 40$.

Results: After accounting for the preceding "Yay"/point-scored as described above, we find *no* statistically significant change of any facial expression before vs. after the "Great Job" or "GoodJob", for any W or τ. This indicates that the change in facial expression around these messages is likely due to having scored a point, not the feedback itself.

9 Conclusions

Our analyses of facial expression and heart rate data from 36 African-American students interacting with iPad-based cognitive skills training software suggest that (1) the impact of the short empathic feedback messages on students' emotions was very small. (2) Several of the correlations (for "Good Job" and "Great Job") disappeared after we accounted for the confound that the student's own achievement at having scored a point could explain the impact. (3) When examining the emotional impact on *individual* students, we found that, except for "Great Job" and "Good Job", only a modest fraction of students showed any stat. sig. correlation. Therefore, before trying to optimize an empathic ITS' control policy, it may be worth verifying that the feedback messages have any impact at all. On the other hand, and more optimistically, contemporary emotional recognition systems also offer a pathway forward to measure the impact of the ITS' actions more precisely. Finally, we note that there could be non-emotional effects of the ITS prompts on students' behaviors. For instance, when watching some videos, we noticed that a few participants shifted their eye gaze in response to the "Watch your time" prompt. Future work can explore this issue.

Acknowledgments: This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL 2019805, and also by an NSF Cyberlearning grant 1822768. The opinions expressed are those of the authors and do not represent views of the NSF.

References

1. Andallaza, T.C.S., Jimenez, R.J.M.: Design of an affective agent for aplusix. Undergraduate thesis, Ateneo de Manila University, Quezon City (2012)
2. Arroyo, I., Woolf, B.P., Cooper, D.G., Burleson, W., Muldner, K.: The impact of animated pedagogical agents on girls' and boys' emotions, attitudes, behaviors and learning. In: International Conference on Advanced Learning Technologies (2011)
3. Claro, S., Paunesku, D., Dweck, C.S.: Growth mindset tempers the effects of poverty on academic achievement. Proceedings of the National Academy of Sciences **113**(31), 8664–8668 (2016)
4. De Manzano, Ö., Theorell, T., Harmat, L., Ullén, F.: The psychophysiology of flow during piano playing. Emotion **10**(3), 301 (2010)
5. Dwivedi, K., Biswaranjan, K., Sethi, A.: Drowsy driver detection using representation learning. In: International advanced computing conference (IACC) (2014)
6. D'Mello, S., Lehman, B., Sullins, J., Daigle, R., Combs, R., Vogt, K., Perkins, L., Graesser, A.: A time for emoting: When affect-sensitivity is and isn't effective at promoting deep learning. In: Intelligent tutoring systems (2010)
7. Ekman, R.: What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford UP, USA (1997)
8. Feidakis, M., Daradoumis, T., Caballé, S.: Emotion measurement in intelligent tutoring systems: what, when and how to measure. In: International Conference on Intelligent Networking and Collaborative Systems (2011)
9. Feng, S., Stewart, J., Clewley, D., Graesser, A.C.: Emotional, epistemic, and neutral feedback in autotutor trialogues to improve reading comprehension. In: International Conference on Artificial Intelligence in Education. Springer (2015)

10. Graesser, A., McDaniel, B., Chipman, P., Witherspoon, A., D'Mello, S., Gholson, B.: Detection of emotions during learning with autotutor. In: Proceedings of the 28th annual meetings of the cognitive science society. pp. 285–290. Citeseer (2006)
11. Hill, O.W., Serpell, Z., Faison, M.O.: The efficacy of the learningrx cognitive training program: modality and transfer effects. The Journal of Experimental Education **84**(3), 600–620 (2016)
12. Jiang, H., Iandoli, M., Van Dessel, S., Liu, S., Whitehill, J.: Measuring students' thermal comfort and its impact on learning. Educational Data Mining (2019)
13. Joshi, A., Allessio, D., Magee, J., Whitehill, J., Arroyo, I., Woolf, B., Sclaroff, S., Betke, M.: Affect-driven learning outcomes prediction in intelligent tutoring systems. In: Automatic Face & Gesture Recognition (2019)
14. Karumbaiah, S., Lizarralde, R., Allessio, D., Woolf, B., Arroyo, I., Wixon, N.: Addressing student behavior and affect with empathy and growth mindset. Educational Data Mining (2017)
15. Lan, A.S., Botelho, A., Karumbaiah, S., Baker, R.S., Heffernan, N.: Accurate and interpretable sensor-free affect detectors via monotonic neural networks. In: International Conference on Learning Analytics & Knowledge (2020)
16. Mondragon, A.L., Nkambou, R., Poirier, P.: Evaluating the effectiveness of an affective tutoring agent in specialized education. In: European conference on technology enhanced learning. pp. 446–452. Springer (2016)
17. Nguyen, H., Masthoff, J.: Designing empathic computers: the effect of multimodal empathic feedback using animated agent. In: Proceedings of the 4th international conference on persuasive technology. pp. 1–9 (2009)
18. Pham, P., Wang, J.: Attentivelearner: improving mobile mooc learning via implicit heart rate tracking. In: International conference on artificial intelligence in education. pp. 367–376. Springer (2015)
19. Ranjbartabar, H., Richards, D., Bilgin, A., Kutay, C.: First impressions count! the role of the human's emotional state on rapport established with an empathic versus neutral virtual therapist. IEEE transactions on affective computing (2019)
20. Robison, J., McQuiggan, S., Lester, J.: Evaluating the consequences of affective feedback in intelligent tutoring systems. In: Affective computing and intelligent interaction and workshops (2009)
21. Sarrafzadeh, A., Hosseini, H.G., Fan, C., Overmyer, S.P.: Facial expression analysis for estimating learner's emotional state in intelligent tutoring systems. In: International Conference on Advanced Technologies (2003)
22. Saulter, L., Thomas, K., Lin, Y., Whitehill, J., Serpell, Z.: Detecting affect over four days of cognitive training. Poster presented at the Temporal Dynamics of Learning Center All-Hands Meeting at UCSD (2013)
23. Sawyer, R., Smith, A., Rowe, J., Azevedo, R., Lester, J.: Enhancing student models in game-based learning with facial expression recognition. In: User modeling, adaptation and personalization (2017)
24. Thayer, J.F., Åhs, F., Fredrikson, M., Sollers III, J.J., Wager, T.D.: A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health. Neuroscience & Biobehavioral Reviews **36**(2), 747–756 (2012)
25. Whitehill, J., Serpell, Z., Lin, Y.C., Foster, A., Movellan, J.R.: The faces of engagement: Automatic recognition of student engagementfrom facial expressions. IEEE Transactions on Affective Computing **5**(1), 86–98 (2014)
26. Widmer, C.L.: Examining the Impact of Dialogue Moves in Tutor-Learner Discourse Using a Wizard of Oz Technique. Ph.D. thesis, Miami University (2017)