Predicting Learning from Student Affective Response to Tutor Questions

Alexandria K. Vail[1(✉)], Joseph F. Grafsgaard[2], Kristy Elizabeth Boyer[4], Eric N. Wiebe[3], and James C. Lester[1]

[1] Department of Computer Science, North Carolina State University, Raleigh, NC, USA
{akvail,lester}@ncsu.edu
[2] Department of Psychology, North Carolina State University, Raleigh, NC, USA
jfgrafsg@ncsu.edu
[3] Department of STEM Education, North Carolina State University, Raleigh, NC, USA
wiebe@ncsu.edu
[4] Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL, USA
keboyer@ufl.edu

Abstract. Modeling student learning during tutorial interaction is a central problem in intelligent tutoring systems. While many modeling techniques have been developed to address this problem, most of them focus on cognitive models in conjunction with often-complex domain models. This paper presents an analysis suggesting that observing students' multimodal behaviors may provide deep insight into student learning at critical moments in a tutorial session. In particular, this work examines student facial expression, electrodermal activity, posture, and gesture immediately following inference questions posed by human tutors. The findings show that for human-human task-oriented tutorial dialogue, facial expression and skin conductance response following tutor inference questions are highly predictive of student learning gains. These findings suggest that with multimodal behavior data, intelligent tutoring systems can make more informed adaptive decisions to support students effectively.

1 Introduction

A fundamental goal of the intelligent tutoring systems (ITS) community is modeling student learning during tutoring so that an ITS can effectively adapt its tutorial support [1,2]. Student models often observe the behavior and performance of the student and then use this information to estimate the student's 'hidden' understanding of the material [3,4]. A variety of approaches to student modeling have been investigated and employed successfully, such as cognitive modeling through knowledge tracing [5] and performance factor analysis [6]. Critically, these approaches rely on student task behaviors such as problem-solving traces.

A. Micarelli et al. (Eds.): ITS 2016, LNCS 9684, pp. 154–164, 2016.
DOI: 10.1007/978-3-319-39583-8_15

While problem-solving traces have been shown to indicate student progress or lack thereof, other work has found that multimodal data streams can be highly indicative of students' state during learning. For example, multimodal data such as facial expression, posture, and gestures can predict affective outcomes, such as frustration and engagement [7,8]. Additionally, multimodal data can contribute to inferring incoming student characteristics, including self-efficacy [9], personality [10], and domain expertise [11]. These studies of multimodal behavior during learning pose a critical open question: *what is the relationship between learning gain and students' multimodal behavior during tutoring?*

To investigate this research question, this paper presents an analysis of student multimodal trace data immediately after tutor questions. In the domain of introductory computer science and in the specific context of tutor inference questions, we investigate whether multimodal trace data contributes to accurately predicting student learning gains. The results show that a subset of facial expression events, together with skin conductance response, immediately after tutor questions are highly predictive of students' future performance on a posttest. These results reveal the significant potential of leveraging multimodal trace data for student modeling.

2 Related Work

The work reported in this paper is grounded in research on multimodal data generated during learning, particularly facial expressions and physiological responses. Multiple studies have explored student facial expression during learning activities. For example, D'Mello and Graesser developed a multimodal classifier of expert-tagged student affect using student dialogue, posture, and facial expression features [12]. A multimodal model built upon all three of these categories yielded higher classification accuracy than using a subset of the data streams, achieving a Cohen's $\kappa = 0.33$ for fixed emotion judgments and $\kappa = 0.39$ for spontaneous judgments. A study with Wayang Outpost attempted to predict self-reported affective states using a similar multimodal feature set, with best fit models achieving a correlation coefficient of up to $r = 0.83$ [13].

There is some evidence that physiological response is predictive of student learning. Stein and Levine proposed a theoretical model in which activation of the autonomic nervous system indicates a mismatch between incoming information and existing knowledge, akin to cognitive disequilibrium [14,15]. Further, they suggest that this state is nearly always an indication of learning. Indeed, some preliminary work on physiological responses to learning interactions has indicated support for this theory. Other work has revealed that skin conductance response after negative feedback and student expressions of uncertainty were highly predictive of student learning [16]. Negative feedback and student expressions of uncertainty are both likely to occur in states of cognitive disequilibrium.

3 Study Data

We investigate the relationship between multimodal behavior traces and learning within a tutorial dialogue corpus of computer-mediated human-human tutoring for introductory computer science. The subject matter focus of the tutorial dialogue is Java programming [17,18]. Each session was conducted within an online remote tutoring system, shown in Fig. 1. The interface consists of four panes: the task description, the student's Java source code, the compilation and execution output of the program, and the textual dialogue messages exchanged between tutor and student. The content of the interface was synchronized in real time between the tutor and the student, with the tutor's interactions constrained to sending textual dialogue messages and progressing to the next task.

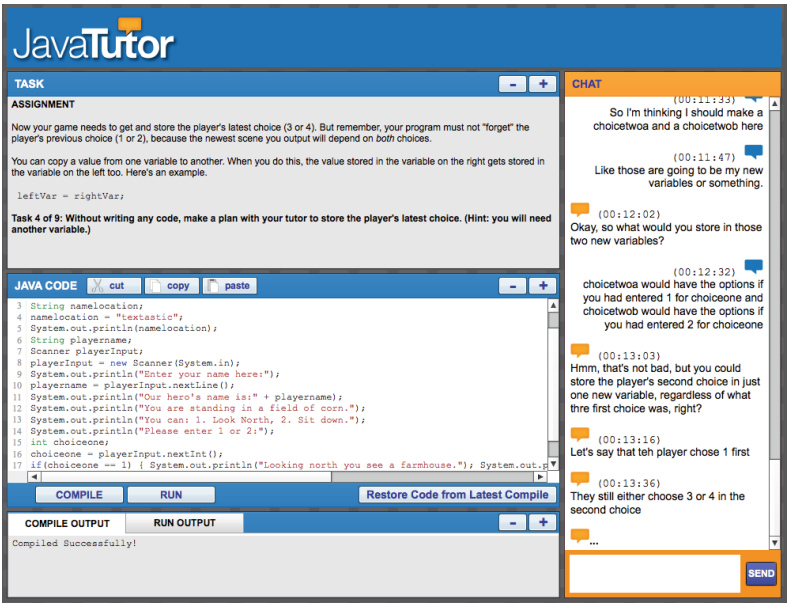

Fig. 1. The web-based tutorial interface for Java programming.

Human tutors ($N = 5$) were primarily graduate students with previous experience in tutoring or teaching introductory programming. Student participants ($N = 67$) were university students in the United States with an average age of 18.5 years ($s = 1.5$ years). Data were collected using multiple multimodal sensors as seen in Fig. 2, including a Kinect depth camera, an integrated webcam, and a skin conductance bracelet (see following subsections for more detail). The data were collected during the fall 2011 and spring 2012 semesters. Each student's participation was distributed over four weeks across six 40-min sessions. (This analysis examines only data from the first lesson.) Before and after each tutorial session, students completed a content-based pretest and identical posttest.

Fig. 2. Multimodal instrumented tutoring session, including a Kinect depth camera to detect posture and gesture, a webcam to detect facial expression changes, and a skin conductance bracelet to detect electrodermal activity.

Normalized learning gain was calculated using the student's pretest and posttest scores, as shown in Eq. 1.

$$norm\_gain = \begin{cases} \dfrac{post - pre}{1 - pre} & post > pre \\ \dfrac{post - pre}{pre} & post \leq pre \end{cases} \tag{1}$$

3.1 Task Event and Dialogue Features

As each student progressed through the session, the tutoring system logged dialogue messages, typing in the code window, and task progress. No strict turn-taking was enforced. Students and tutors could type dialogue messages at any time. All tutor and student dialogue messages were tagged automatically (for details please see [19]) with a dialogue act annotation scheme for task-oriented tutorial dialogue [20].

The present analysis focuses on a key tutor dialogue move: inference questions. Inference questions are questions that require reasoning about content knowledge or formulating a plan. For example, *'How can you fix this error?'*, and *'How do you think this problem can be solved?'* are inference questions. Questions of this nature are known to stimulate cognitive disequilibrium in students [15], which is considered to be a crucial step in knowledge acquisition [21]. The analysis presented here explores the hypothesis that student multimodal traces following tutor inference questions are significantly predictive of student learning gain.

3.2 Facial Expression Features

Facial expression features were automatically identified by a state-of-the-art facial expression recognition and analysis software, FACET (commercial software that was preceded by a research version known as the Computer Emotion Recognition Toolbox, CERT) [22]. FACET provides frame-by-frame tracking of facial action units according to the Facial Action Coding Scheme [23]. These action units include such expressions as AU4 Brow Lowerer, AU15 Lip Corner Depressor, and AU23 Lip Tightener (see Fig. 4 for illustration). Facial features were extracted from webcam videos. The FACET software provides an *Evidence* measure for each facial action unit, indicating the chance that the target expression is present.

3.3 Electrodermal Activity Features

Skin conductance is a type of electrodermal activity [24]. Skin conductance has two components, *tonic*, which changes gradually over time, and *phasic*, which changes in abrupt peaks [25] in response to a stimulus. These peaks represent *skin conductance response (SCR) events*.

A challenge in analyzing SCRs in the context of a series of task and dialogue events is that SCRs occur in close temporal proximity, even overlapping with each other. In order to address this concern, this analysis utilizes Continuous Decomposition Analysis, which decomposes skin conductance data into its tonic and phasic components and detects overlapping SCRs [25]. This analysis was conducted using the Ledalab Matlab software, which additionally supports event-related analysis in the context of SCRs. The threshold for detecting SCRs was set to a minimum change in amplitude of $\delta = 0.02\,\mu$S, based on the results of prior analysis on this corpus of tutorial dialogue [16].

4 Analysis

The primary objective of this analysis is to identify how multimodal signals following tutor questions can predict learning gain. In order to do this, we examine the three seconds (a manually-determined interval) following the delivery of an inference question from a tutor. Student behavior was characterized using the following features, which were all provided to the predictive models reported below. (Only the first two of these were found to be significant predictors, as the results section will describe.)

1. Average Evidence measure for each of the facial expression action units during the interval.
2. Number of skin conductance responses (SCRs) identified during the interval.
3. Percentage of the interval in which a one-hand-to-face or two-hands-to-face gesture was observed.
4. Average student distance from the workstation during the interval.
5. Average difference between the highest and lowest points of the student's body from the workstation during the interval (indicating leaning).

To examine the predictiveness of multimodal traces immediately following tutor inference questions, we averaged the value of each multimodal feature described above across each tutoring session. These features are conditional averages of the form $Avg(Feature|TutorInferenceQ)$. If we built predictive models using only these features, we may identify conditional features that are components in a broader, unconditional association between a multimodal feature and learning gain. To control for this we also included a feature $Avg(Feature)$, which represents the session-wide average value of that multimodal feature (not conditioned on any preceding event). For each student and for each feature type listed above, one value of $Avg(Feature|TutorInferenceQ)$ and one value of $Avg(Feature)$ were generated.

All features were standardized by subtracting the mean and dividing by the standard deviation. This set of features was then used in a stepwise regression modeling procedure that maximizes the leave-one-student-out cross-validated R^2 value (the coefficient of determination), while enforcing a strict p-value cut-off of $p < 0.05$ after Bonferroni correction for the significance of each included feature.

5 Results

The results show that student facial expression features and skin conductance response are significantly predictive of learning gain. The predictive model for normalized learning gain includes six features, four of which are specific to the multimodal traces from the three-second interval following an inference question from the tutor. The other two predictors are session-wide features (Table 1).

Table 1. Predictive model for standardized normalized learning gain after tutor inference questions[a].

Normalized Learning Gain =	R^2	p
+1.4012 * AU23 (Session-wide)	0.0445	< 0.001
+0.1523 * SCRs	0.2457	< 0.001
+0.7548 * AU5	0.2669	< 0.001
−0.3502 * AU15 (Session-wide)	0.0024	0.002
+0.2856 * AU4	0.0789	0.005
−0.4503 * AU23	0.1893	0.004
+0.6440 (Intercept)		1.000
Leave-One-Out Cross-Validated $R^2 = 0.8277$		

[a]This model was built as part of a more expansive exploratory analysis. The p-values reported here have already undergone a Bonferroni correction $p \leq \alpha/n$, where $n = 21$ is the number of statistical tests conducted, in order to reduce the familywise error rate to $\alpha = 0.05$.

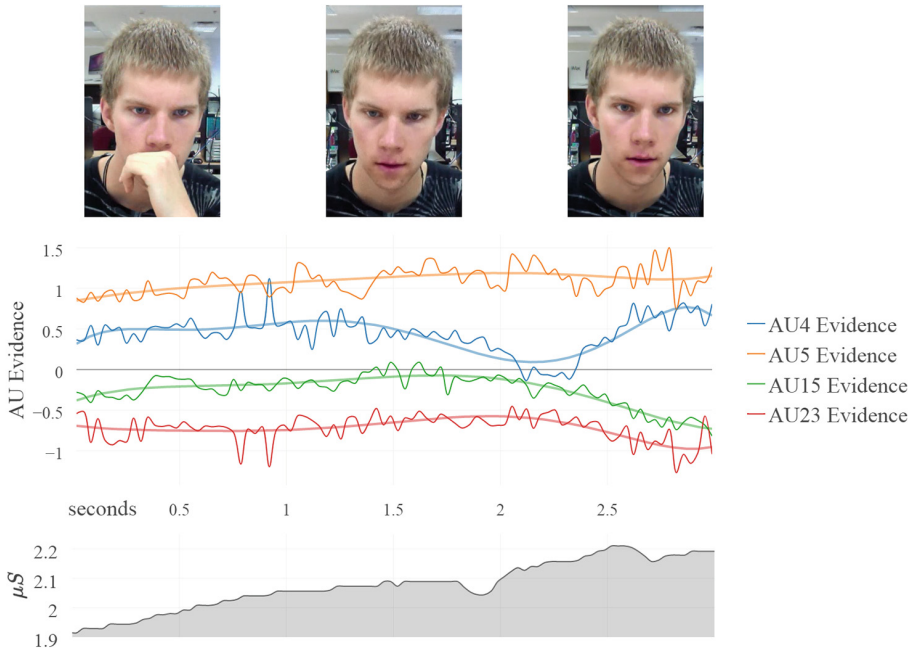

Fig. 3. A segment of the multimodal data collection illustrating a student's response to an inference question from the tutor (*"How can you fix your code?"*). Sample webcam frames are displayed, along with standardized FACET readings of the four significant facial action units and the student's electrodermal activity. Note the overall increase in AU4 and AU5, along with a decrease in AU15, as well as activation of an SCR at approximately two seconds. This student achieved one of the highest learning gains observed in the current study.

(a) AU4 Brow Lowerer (b) AU5 Upper Lid Raiser (c) AU15 Lip Corner Depressor (d) AU23 Lip Tightener

Fig. 4. Sample frames from the student webcam illustrating the four significant facial action unit features appearing in the predictive model, as identified by FACET.

The results show that session-wide features account for a relatively small portion of the variance in learning gain: only two facial action unit features were selected from this set. One of these is frequency of AU15 Lip Corner Depressor (Fig. 4c) which is negatively predictive of student learning gain.

The other session-wide facial action unit feature that is significant in the model is AU23 Lip Tightener (Fig. 4d), the session-wide presence of which is positively correlated with learning. However, this features is also significant as a stimulus-specific feature but in the opposite direction. Higher frequency of AU23 immediately after tutor inference questions is *negatively* predictive of learning gain. The final two facial action unit features that are significantly predictive of learning gain are AU4 Brow Lowerer and AU5 Upper Lid Raiser (Fig. 4a and b, respectively) both positively correlated with learning gain. Finally, the number of skin conductance responses (SCRs) occurring after tutor inference questions is a significant positive indicator of learning gain (Fig. 3).

6 Discussion

Tutor inference questions require students to reason about their knowledge or formulate plans for problem solving. Consequently, student multimodal signals following these pivotal moments offer key insights into the cognitive-affective phenomena that are associated with learning.

Students displaying more frequent AU15 Lip Corner Depressor after tutor inference questions learned less. This action unit has been found in prior task-oriented studies to be a strong predictor of lack of focus [26]. In contrast, prior work has indicated that AU23 Lip Tightener is frequently associated with frustration or focused concentration [26,27]. In the current study, AU23 session-wide was positively associated with learning, but immediately following tutor inference questions it was negatively associated. This finding points to the importance of further study to tease apart frustration from focused concentration, particularly in the context of questions that require reasoning or possibly in the face of cognitive disequilibrium.

AU5 Upper Lid Raiser following tutor inference questions was positively predictive of learning in the current study, and it has previously been found to indicate focused attention in task-oriented domains [26]. Expressing this indicator of engagement directly following tutor questions may suggest that the student is thinking critically about the solution.

AU4 Brow Lowerer following tutor questions was predictive of increased learning in the current study. AU4 has been associated with frustration [7,8], and in general, frustration has been found to be inversely related to learning gains [28,29]. However, these results have been discovered mostly in the context of session-wide features; different analyses have found these features indicative of confusion in shorter time periods [30]. Both frustration and confusion are frequently associated with cognitive disequilibrium [15] which, when resolved, is beneficial to learning [21]. AU4 following tutor inference questions may indicate

cognitive disequilibrium at first, the resolution of which fosters learning. In the tutoring sessions investigated here, AU4 session-wide does not have a significant relationship with learning.

Prior work on this tutorial dialogue corpus has suggested the importance of skin conductance responses following events that indicate cognitive disequilibrium, such as student expressions of uncertainty or encountering negative feedback from the system [16]. We might reasonably infer that inference questions from the tutor may induce cognitive disequilibrium [15], and so skin conductance responses following these questions may indicate heightened response that facilitates learning. Further study is needed to elucidate the causal relationships between tutor questions, student cognitive disequilibrium, skin conductance response, and learning.

7 Conclusion and Future Work

Modeling student learning during tutoring is central to intelligent tutoring systems. The results presented here demonstrate that student multimodal traces can provide insight into cognitive-affective phenomena while yielding accurate predictions of student learning during tutoring sessions. In particular, facial expression and skin conductance responses during tutoring were highly predictive of learning as indicated by improvement from pretest to posttest. These results complement and expand upon prior work investigating these features by decomposing a tutorial session into salient moments and investigating short-term responses versus long-term session features.

Future work should investigate how student multimodal signals at other critical moments in tutoring sessions are related to student learning. For example, introducing new concepts, or when a student reaches an impasse, are likely key moments in tutoring. Another promising direction for future work is to examine affective outcomes such as frustration or engagement, since multimodal signal analysis holds much promise for providing real-time predictions of these phenomena as well. It is hoped that this line of work will lead to powerful, domain-independent predictive measures of learning and other cognitive-affective phenomena that intelligent tutoring systems can use to adaptively support student learning.

Acknowledgements. The authors wish to thank the members of the LearnDialogue and Intellimedia groups at North Carolina State University for their helpful input. This work is supported in part by the Department of Computer Science at North Carolina State University and the National Science Foundation through Grants IIS-1409639, CNS-1453520, and a Graduate Research Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this report are those of the participants, and do not necessarily represent the official views, opinions, or policy of the National Science Foundation.

References

1. Hartley, D., Mitrović, A.: Supporting learning by opening the student model. In: Cerri, S.A., Gouardéres, G., Paraguaçu, F. (eds.) ITS 2002. LNCS, vol. 2363, pp. 453–462. Springer, Heidelberg (2002)
2. Gluga, R.: Long term student learner modeling and curriculum mapping. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010, Part II. LNCS, vol. 6095, pp. 227–229. Springer, Heidelberg (2010)
3. Corbett, A.T., Anderson, J.R.: Student modeling and mastery learning in a computer-based programming tutor. In: Proceedings of the 2nd International Conference on Intelligent Tutoring Systems, pp. 413–420 (1992)
4. Stevens, R., Soller, A., Cooper, M., Sprang, M.: Modeling the development of problem solving skills in chemistry with a web-based tutor. In: Lester, J.C., Vicari, R.M., Paraguaçu, F. (eds.) ITS 2004. LNCS, vol. 3220, pp. 580–591. Springer, Heidelberg (2004)
5. Corbett, A.T., Anderson, J.R.: Knowledge tracing: modeling the acquisition of procedural knowledge. User Model. User-Adap. Inter. **4**, 253–278 (1995)
6. Pavlik Jr., P.I., Cen, H., Koedinger, K.R.: Performance factors analysis - a new alternative to knowledge tracing. In: Proceedings of the 14th International Conference on Artificial Intelligence in Education, pp. 531–538 (2009)
7. Grafsgaard, J.F., Wiggins, J.B., Boyer, K.E., Wiebe, E.N., Lester, J.C.: Automatically recognizing facial expression: predicting engagement and frustration. In: Proceedings of the 6th International Conference on Educational Data Mining, pp. 43–50 (2013)
8. Grafsgaard, J.F., Wiggins, J.B., Boyer, K.E., Wiebe, E.N., Lester, J.C.: Automatically recognizing facial indicators of frustration: a learning-centric analysis. In: Proceedings of the Humane Association Conference on Affective Computing and Intelligent Interaction, pp. 159–165 (2013)
9. Scherer, S., Weibel, N., Morency, L.P., Oviatt, S.: Multimodal prediction of expertise and leadership in learning groups. In: Proceedings of the 1st International Workshop on Multimodal Learning Analytics (2012)
10. Biel, J.I., Teijeiro-Mosquera, L., Gatica-Perez, D.: FaceTube: predicting personality from facial expressions of emotion in online conversational video. In: Proceedings of the 14th International Conference on Multimodal Interaction, pp. 53–56 (2012)
11. Oviatt, S., Cohen, A.: Written and multimodal representations as predictors of expertise and problem-solving success in mathematics. In: Proceedings of the 15th International Conference on Multimodal Interaction, pp. 599–606 (2013)
12. D'Mello, S.K., Graesser, A.C.: Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. User Model. User-Adap. Inter. **20**, 147–187 (2010)
13. Arroyo, I., Cooper, D.G., Burleson, W., Woolf, B.P., Muldner, K., Christopherson, R.: Emotion sensors go to school. In: Proceedings of the 14th International Conference on Artificial Intelligence in Education, pp. 17–24 (2009)
14. Stein, N.L., Levine, L.J.: Making sense out of emotion: the representation and use of goal-structured knowledge. In: Psychological and Biological Approaches to Emotion, pp. 45–73 (1990)
15. Piaget, J.: The Origins of Intelligence. International University Press, New York (1952)

16. Hardy, M., Wiebe, E.N., Grafsgaard, J.F., Boyer, K.E., Lester, J.C.: Physiological responses to events during training: Use of skin conductance to inform future adaptive learning systems. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, pp. 2101–2105 (2013)
17. Ha, E.Y., Grafsgaard, J.F., Mitchell, C.M., Boyer, K.E., Lester, J.C.: Combining verbal and nonverbal features to overcome the 'Information Gap' in task-oriented dialogue. In: Proceedings of the 13th Annual SIGDIAL Meeting on Discourse and Dialogue, pp. 247–256 (2012)
18. Mitchell, C.M., Ha, E.Y., Boyer, K.E., Lester, J.C.: Learner characteristics and dialogue: recognising effective and student-adaptive tutorial strategies. Int. J. Learn. Technol. **8**(4), 382–403 (2013)
19. Vail, A.K., Boyer, K.E.: Adapting to personality over time: examining the effectiveness of dialogue policy progressions in task-oriented interaction. In: Proceedings of the 15th Annual SIGDIAL Meeting on Discourse and Dialogue, pp. 41–50 (2014)
20. Vail, A.K., Boyer, K.E.: Identifying effective moves in tutorial dialogue: on the refinement of speech act annotation schemes. In: Proceedings of the 12th International Conference on Intelligent Tutoring Systems, pp. 199–209 (2014)
21. Graesser, A.C., Olde, B.A.: How does one know whether a person understands a device? The quality of the questions the person asks when the device breaks down. J. Educ. Psychol. **95**(3), 524–536 (2003)
22. Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Javier, M., Bartlett, M.: The computer expression recognition toolbox (CERT). In: Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition and Workshops, pp. 298–305 (2011)
23. Lepper, M.R., Woolverton, M.: The wisdom of practice: Lessons learned from the study of highly effective tutors. Impact of Psychological Factors on Education, Improving Academic Achievement, pp. 135–158 (2002)
24. Boucsein, W.: Electrodermal Activity. Springer Science & Business Media, NewYork (2012)
25. Benedek, M., Kaernbach, C.: A continuous measure of phasic electrodermal activity. J. Neurosci. Methods **190**, 80–91 (2010)
26. Vural, E., Cetin, M., Ercil, A., Littlewort, G., Bartlett, M., Movellan, J.: Drowsy driver detection through facial movement analysis. In: Proceedings of the 12th International Conference on Human-Computer Interaction, pp. 6–18 (2007)
27. Mortillaro, M., Mehu, M., Scherer, K.R.: Subtly different positive emotions can be distinguished by their facial expressions. Soc. Psychol. Pers. Sci. **2**(3), 262–271 (2011)
28. Goldin, I.M., Carlson, R.: Learner differences and hint content. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED 2013. LNCS, vol. 7926, pp. 522–531. Springer, Heidelberg (2013)
29. San Pedro, M.O.Z., Baker, R.S.J., Gowda, S.M., Heffernan, N.T.: Towards an understanding of affect and knowledge from student interaction with an intelligent tutoring system. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED 2013. LNCS, vol. 7926, pp. 41–50. Springer, Heidelberg (2013)
30. D'Mello, S., Lehman, B., Pekrun, R., Graesser, A.: Confusion can be beneficial for learning. Learn. Instr. **29**, 153–170 (2004)