

Affective states and state tests: Investigating how affect and engagement during the school year predict end of year learning outcomes

Zachary A. Pardos

University of California, Berkeley, USA

Ryan S.J.d. Baker

Columbia University Teachers College, USA

Maria O.C.Z. San Pedro

Columbia University Teachers College, USA

Sujith M. Gowda

Worcester Polytechnic Institute, USA

Supreeth M. Gowda

Worcester Polytechnic Institute, USA

In this paper, we investigate the correspondence between student affect and behavioral engagement in a web-based tutoring platform throughout the school year and learning outcomes at the end of the year, on a high-stakes mathematics exam. The relationships between affect, behavioral engagement, and learning outcomes have been previously studied, but not in a manner that is both longitudinal and finer-grained. Affect and behavior detectors are used to estimate student affective states and behavior based on post-hoc analysis of tutor log-data. For every student action in the tutor the detectors give us an estimated probability that the student is in a state of boredom, engaged concentration, confusion, and frustration, and estimates of the probability that they are exhibiting off-task or gaming behaviors. We ran the detectors on two years of log data from 8th grade student use of the ASSISTments math tutoring system and collected corresponding end-of-year, high-stakes, state math test scores for the 1,393 students in our cohort. By correlating these data sources, we find that boredom during problem solving is negatively correlated with performance, as expected; however, boredom is positively correlated with performance when exhibited during scaffolded tutoring. A similar pattern is unexpectedly seen for confusion. Engaged concentration and, surprisingly, frustration are both associated with positive learning outcomes. In a second analysis we build a unified model that predicts student standardized examination scores from a combination of student affect, disengaged behavior, and performance within the learning system. This model achieves high overall correlation to standardized exam score, showing that these types of features can effectively infer longer-term learning outcomes.

Keywords: Affect, confusion, boredom, high stakes tests, tutoring, automated detectors, prediction, data mining.

Introduction

In recent years, researchers have increasingly investigated the relationship between fine-grained details of student usage of tutoring systems and performance on high-stakes examinations (cf. Feng, Heffernan, & Koedinger, 2009; Pardos, Wang, & Trivedi, 2012). Understanding how different student behaviors correspond to student outcomes can help us to understand the larger implications of student choices that might seem only momentary. This information can be useful both in terms of advancing theory on meta-cognition and engagement (cf. Alevan, McLaren, Roll, & Koedinger, 2004; Baker, Gowda, & Corbett, 2011), and to provide actionable information for teachers about factors potentially influencing their

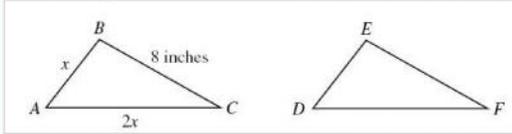
students' learning outcomes (Arnold, 2010). Within this paper, we analyze the relationships between a student's affect, engagement, and their outcomes. Several studies have indicated that affect and behavioral engagement can lead to differences in learning (Craig, Graesser, Sullins, & Gholson, 2004; Pekrun, Goetz, Titz, & Perry, 2002; Rodrigo et al., 2009; Baker, 2007; Cocea, Hershkovitz, & Baker, 2009); however, past research on these relationships has been limited by the use of observational or survey methods, which are either coarse-grained, or can only be applied over brief periods of time (year-long field observations are possible, but prohibitively expensive to conduct for large numbers of students). Longitudinal approaches have been used to predict college attendance (San Pedro, Baker, Gowda, & Heffernan, 2013), suggesting that a similar approach may be feasible to predict long-term learning outcomes. Within this paper, we use automated detectors of affect and behavioral engagement that can be applied to every student action in an entire year's log file data to analyze this question, asking how predictive a student's affect and engagement, throughout the school year, is of his or her end-of-year high-stakes test outcome. In specific, we investigate overall relationships between affect/engagement and learning, and dig deeper to ask: are there some contexts where a particular affect is constructive and others where it is not? We also compare the overall predictiveness of affect and engagement relative to student performance in the learning system. We investigate these questions in the context of two school years of student learning within the ASSISTments tutoring system (Feng et al., 2009), involving over a thousand students.

The Tutor and the Test

ASSISTments is a web-based tutoring platform, primarily for 7th-12th grade mathematics. Within ASSISTments, shown in Figure 1, students complete mathematics problems and are formatively *assessed* – providing detailed information on their knowledge to their teachers – while being *assisted* with scaffolding, help, and feedback. Items in ASSISTments are designed to correspond to the skills and concepts taught in relevant state standardized examinations. Figure 1 shows how after the student answers the original question incorrectly, the system provides scaffolding that breaks the problem down into steps. Hints are provided at each step and the student can ask for a bottom-out hint that eventually tells the answer. Students in the data sets studied within this paper used ASSISTments in classroom computer lab sessions targeted towards preparation for the standardized state test, during school hours. While teachers had the ability to assign students questions of a particular skill, the most popular problem set within the data set that will be analyzed in this paper was one that randomly sampled 8th grade math test prep questions from the system. Because of this, students sometimes received questions with skills they had not encountered in class yet. One data set, which was used to develop models of student affect, represented a few days of software usage. The other data set, used to study the relationship between student affect and learning outcomes, represents an entire year of data of students using the ASSISTments system.

Triangles ABC and DEF are congruent. The perimeter of triangle ABC is 23 inches.

What is the length of side DF in triangle DEF?



Break this problem into steps

Type your answer below (mathematical expression):

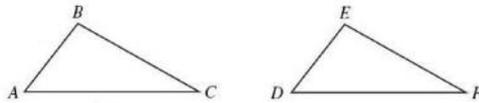
5

Submit Answer

You are almost right, but remember that DF is twice x.

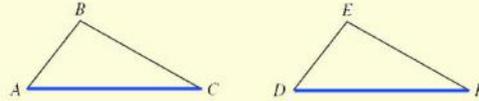
Let's move on and figure out this problem.

Which side of triangle ABC has the same length as side DF of triangle DEF?



Congruent triangles means triangles whose corresponding sides are equal in length.

Look at both triangles and find the pairs of sides that have the same length.



The side that corresponds to DF is AC.
Select AC

Select one:

- AB
- BC
- AC

Submit Answer

Side AB corresponds to side DE of triangle DEF, not DF. Try again, please.

Figure 1: An example of an ASSISTments item where the student answers incorrectly and receives scaffolding help

Near the end of their school year, students took the MCAS (Massachusetts Comprehensive Assessment System) state standardized test. We collected scores for the math portion of the test. Raw scores range from 0 to 54 and are later scaled by the state after all tests are in. The scaling maps to four categories; Failing, Needs Improvement, Proficient, and Advanced. Students must score above Failing to graduate high school and an Advanced score earns them an automatic state college scholarship.

Methodology

In this section we will describe both the methodology for employing the automatic affect detectors to our dataset and the methodology for conducting the correlation analysis.

Affect and Behavior Detection

In order to assess student affect and behavior across contexts, we adopt a two-stage process: first labeling student affect and behavior for a small but reasonably representative sample with field observations (cf. Baker, D'Mello, Rodrigo, & Graesser, 2010), and then using those labels to create automated detectors that can be applied to log files at scale. The detectors are created by synchronizing log files generated by the ASSISTments system with field observations conducted at the same time. To enhance scalability, only log data is used as the basis of the detectors; physical sensors can enhance detector goodness (cf. Conati & Maclaren, 2009; D'Mello, Craig, Witherspoon, McDaniel, & Graesser, 2008), but reduce the applicability of the resultant models to existing log files. The detectors are constructed using log data from student actions within the software occurring at the same time as or before the observations, making our detectors usable for automated interventions, as well as for the type of discovery with models analysis conducted in this paper. Our process for developing sensor-free affect and behavior detectors for ASSISTments replicates a process which has been successful for developing affect detectors for a different intelligent tutor, Cognitive Tutor Algebra (Baker et al., 2012).

Data Collection

Two sets of data from ASSISTments were used in this study.

The first data set was used to develop the automated detectors of affect. This data set was composed of field observations of affect and behavior that were conducted over a few days in an urban middle school in central Massachusetts, sampled from a diverse population of 229 students. Within this school, 40% of students were Hispanic, 14% were African-American, 4% were Asian-American, and 39% were Caucasian. In this school, per capita income was significantly lower than the state average. Information from these observations and the corresponding interaction logs was used to develop and validate the affect detectors discussed below.

The second data set was used to conduct analyses of the relationships between affect and learning. This data set was composed of action log files that were distilled from a diverse population (racially and socio-economically) of 1,393 students that came from middle schools in the same city in central Massachusetts, in 2004-2005 and 2005-2006 (these years were chosen due to the availability of standardized examination data). 629 students used the software in 2004-2005, and 764 students used the software in 2005-2006. This data set involved a whole year of students using the software for two hours, twice a week. As this data set represented whole-year usage of the software, 810,000 student actions (entering an answer or requesting help) were represented in the data. The affect models were applied to this larger dataset.

Affect and Behavior Observations

Student affect and behavioral engagement was coded by a pair of expert field observers as students used ASSISTments in 2010. An observation protocol developed for coding affect during the use of educational software (cf. Baker et al., 2010) was implemented using field observation synchronization software (Baker et al., 2012) developed for Google Android handheld devices. Each observation lasted up to twenty seconds, with elapsed observation time so far displayed by the hand-held observation software. If affect or behavior was labeled before twenty seconds elapsed, the coder moved to the next observation. Each observation was conducted using side glances, to reduce observer effects. To increase tractability of both coding and eventual analysis, if two distinct affective states were seen during a single observation, only the first state observed was coded. Any affect or behavior of a student other than the student currently being observed was not coded. The observers based their judgment of a student's affect or behavior on the student's work context, actions, utterances, facial expressions, body language, and interactions with teachers or fellow students. These are, broadly, the same types of information used in previous methods for coding affect (e.g., Bartel & Saavedra, 2000), and in line with Planalp, DeFrancisco,

and Rutherford's (1996) descriptive research on how humans generally identify affect using multiple cues in concert for maximum accuracy rather than attempting to select individual cues. Affect and behavior coding was conducted on a handheld app previously designed for this purpose (Baker et al., 2012). Student affect or behavior was coded according to the following set of categories: boredom, frustration, engaged concentration, confusion, off-task behavior, gaming, and other (comprising any affective or behavior state not represented by the other categories). These categories were chosen due to past evidence that they are relatively common and are either associated with learning or hypothesized to be associated with learning (cf. Alevan, 2004; Baker, 2007; Baker et al., 2010; Baker et al., 2012; Cocea et al., 2009; Craig et al., 2004; Lee, Rodrigo, Baker, Sugay, & Coronel, 2011; Lehman, D'Mello, & Graesser, 2012; Rodrigo et al., 2009). The affective categories were defined for coding according to the definitions in (Baker et al., 2010), and the behavior categories were defined according to the definitions in (Baker, 2007; Baker et al., 2010).

At the beginning of data collection, an inter-rater reliability session was conducted, where the two coders coded the same student at the same time, across 51 different coding instances across multiple students. With reference to the categories of affect studied in this paper, inter-rater reliability achieved Cohen's Kappa of 0.72, indicating agreement 72% better than chance. For categories of behavior, inter-rater reliability achieved Cohen's Kappa of 0.86, agreement 86% better than chance. This level of agreement is substantially higher than the level of agreement typically seen for video coding of affect (D'Mello et al., 2008; Sayette, Cohn, Wertz, Perrott, & Parrott, 2001). After this session, the observers coded students separately, for a total of 3,075 observation codes.

Within the observations, not counting observations marked as "?" (which represents cases where coding was impossible due to uncertainty, behavior outside the coding scheme, a student leaving the room, impossible positioning, or other factors), boredom was observed 21.7% of the time, frustration was observed 5.4% of the time, engaged concentration 65.0% of the time, and confusion was observed 7.9% of the time. In terms of behavior, off-task behavior was observed 21.9% of the time, and gaming was observed 1.5% of the time. This distribution of affect and behavior corresponds to previous studies, where engaged concentration is the most prevalent affect in a classroom environment (Baker et al., 2010; Baker et al., 2012; Sabourin, Mott, & Lester, 2011).

ASSISTments Interaction Logs

During observations, both the handheld devices and the educational software logging server were synchronized to the same internet time server, using the same field observation data collection software as was used in (Baker et al., 2012). This enabled us to determine which student actions within the software were occurring when the field observations occurred. Interactions with the software during the twenty seconds prior to data entry by the observer were aggregated as a clip, and data features were distilled.

The original log files consisted of data on every student attempt to respond (and whether it was correct), and requests for hint and scaffolding, as well as the context and time taken for each of these actions. In turn, 43 features were distilled from each action (Table 1), including features distilled for detecting other constructs in ASSISTments (cf. Baker, Goldstein, & Heffernan, 2011), and features developed for detecting student behavior and affect in Cognitive Tutors (cf. Baker, 2007; Baker et al., 2012). Many of the distilled features pertained to the student's past actions, such as how many attempts the student had previously made on this problem step, how many previous actions for this skill or problem step involved help requests, how many incorrect actions the student had made on this problem step, and so on. To aggregate individual student actions into twenty-second clips, the sum, minimum, maximum, and average values were calculated across actions for each clip. This relatively simple approach to summarizing features was used due to its success in similar problems in other learning systems (cf. Baker et al., 2012). Thus, for the creation of affect and behavior models, a total of 172 features were used.

Table 1. The 43 features generated for affect detection. The min, max, and avg were also calculated, totaling 173 features

Total problems attempted in the tutor so far.	Problem is original not a scaffolding problem
Bottom-out hint is used	Number of last 8 problems that used the bottom-out hint.
Total number of 2 wrong answers in a row across all the problems	Percent of all past problems that were correct on this KC.
Answer is correct	Wrong answer after hint
Problem ends with automatic scaffolding	Response is chosen from a list of answers (Multiple choice, etc).
Problem ends with scaffolding	Response is filled in (No list of answers available)
First response is a help request	Problem is a scaffolding problem
First response is a help request Scaffolding	Second to last hint is used – indicates a hint that gives considerable detail but is not quite bottom-out
Number of last 5 First responses that included a help request	Long pause after wrong answer
Number of last 5 First responses that were wrong	Long pause after correct answer
Number of last 8 First responses that included a help request	Long pause after help or bug message
Number of last 8 First responses that were wrong	Long pauses after 2 Consecutive wrong answers
First response time taken on scaffolding problems	Time since the current KC was last seen.
Total first response practice opportunities on this skill so far	Time spent on the current step
First response Working during school hours (between 7:00 am and 3:00 pm)	Total first responses attempted in the tutor so far.
Time spent on help was under 10 seconds	Total first responses wrong attempts in the tutor so far.
Time spent on help was under 1 second	Percent of all past problems that were wrong on this KC.
Time spent on help was under 2 seconds	Total first response practice opportunities on this KC so far.
Time spent on help was under 5 seconds	Total first response scaffolding opportunities for this KC so far
Immediate help Request - Help on first response and time spent was under 2 seconds	Total first response time spent on this KC across all problems
Action is a hint response	Total time spent on this KC across all problems divided by percent correct for the same KC
Total number of hints requested so far	

Creation of Affect and Behavior Models

A detector for each affective state or behavior was developed separately, comparing that affective state to all other affective states (e.g., “Bored” was compared to “Not Bored,” “Frustrated” was compared to “Not Frustrated,” “Engaged Concentration” was compared to “Not Engaged Concentration,” and “Confused” was compared to “Not Confused”), or comparing that behavior to all other behaviors (e.g., “Off-task” was compared to “Not Off-task” and “Gaming” was compared to “Not Gaming”). Each detector was evaluated using 5-fold cross-validation, at the student-level (e.g., detectors are trained on four groups of students and tested on a fifth group of students). By cross-validating at this level, we increase confidence that detectors will be accurate for new groups of students. Further, in this student-level cross-validation, students were stratified into fold assignments based on their training labels. This guarantees that each fold has a representative number of observations of the majority and minority class. In addition, for unbalanced classes, re-sampling was used on the training sets to make the class frequency more equal for detector development (but detector goodness was validated on a data set that was not re-sampled, to ensure model validity for data with natural distributions). We attempted to fit sensor-free affect detectors using eight common classification algorithms, including J48 decision trees, step regression, JRip, Naive Bayes, K*, and REP-Trees. These algorithms were chosen as a sample of the space of potential algorithms, which can represent data with different patterns, but each of which is relatively conservative and not highly prone to over-fitting.

Feature selection for machine learning algorithms was conducted using forward selection with stepwise regression. With this technique the feature that most improves model goodness is added to the list of features of the model until no more features can be added which improve model goodness (Table 1). During feature selection, cross-validated kappa on the original (non-re-sampled) data set was used as the goodness metric. Prior to feature selection, all features with cross-validated kappa equal to or below zero in a single-feature model were omitted from further consideration, as a check on over-fitting.

The affect and behavior detectors' performance was evaluated on their ability to predict the presence or absence of each affective state or behavior in a clip. Detectors were evaluated using A' (Hanley & McNeil, 1980), Cohen's Kappa (Cohen, 1960), and F-measure (Van Rijsbergen, 1974) goodness metrics. The A' metric (equivalent to area under the ROC curve) is the probability that the model will be able to discriminate a randomly chosen positive case from a randomly chosen negative case. An A' value of 0.5 for a model indicates chance-level performance, and 1.0 performing perfectly. Cohen's Kappa assesses the degree to which the model is better than chance at identifying the affective state or behavior in a clip. A Kappa of 0 indicates chance-level performance, while a Kappa of 1 indicates perfect performance. A Kappa of 0.45 is equivalent to a detector that is 45% better than chance at identifying affect or behavior. The F-measure of F1-score is a measure of the model's accuracy, computing for the weighted average of the model's precision and recall where the best F1 score is at 1 and the worst score is 0.

All of the affect and behavior detectors performed better than chance (Table 2). Detector goodness was somewhat lower than had been previously seen for Cognitive Tutor Algebra (cf. Baker et al., 2012), but better than had been seen in other published models inferring student affect in an intelligent tutoring system solely from log files (where average Kappa ranged from below zero to 0.19 when fully stringent validation was used) (Baker et al., 2012; Conati & Maclaren, 2009; D'Mello et al., 2008; Sabourin et al., 2011). The best detector of engaged concentration involved the K* algorithm, achieving an A' of 0.678, a Kappa of 0.358, and an F-measure of 0.687. The best boredom detector was found using the JRip algorithm, achieving an A' of 0.632, a Kappa of 0.229, and an F-measure of 0.632. The best frustration detector achieved an A' of 0.682, a Kappa of 0.324, and an F-measure of 0.677, using the Naive Bayes algorithm. The best confusion detector used the J48 algorithm, having an A' of 0.736, a Kappa of 0.274, and an F-measure of 0.667. The best detector of off-task behavior was found using the REP-Tree algorithm, with an A' value of 0.819, a Kappa of 0.506, and an F-measure of 0.693. The best gaming detector involved the K* algorithm, having an A' value of 0.802, a Kappa of 0.370, and an F-measure of 0.687. These levels of detector goodness indicate models that are clearly informative, though there is still considerable room for improvement.

Table 2: Performances of affect and behavior models

Affect	Algorithm	A'	Kappa	F-Measure
Boredom	JRip	0.632	0.229	0.632
Frustration	Naive Bayes	0.682	0.324	0.677
Engaged Concentration	K*	0.678	0.358	0.687
Confusion	J48	0.736	0.274	0.667
Off-Task	REP-Tree	0.819	0.506	0.693
Gaming	K*	0.802	0.370	0.750

Detector features for boredom include the total number of actions, the total time spent on the last action before the clip and the first action after the clip, and the student's history of help requests and correct answers. For example, students were deemed bored when they spent over 83 seconds inactive immediately before or after the observation (lengthy pauses are also an excellent predictor of off-task behavior (cf. Baker, 2007), a behavior thought to be associated with boredom). Students were also deemed bored when they worked on the same problem during the entire observation but did not provide any correct answers either during the observation or immediately afterwards (a serious and actively working student will generally obtain some correct answers in ASSISTments, as increasingly easy scaffolding is given when students make errors).

The detector's features for frustration involve the percent occurrence of incorrect answers on a skill in the past, the largest hint count in that clip, the average correct actions in that clip, the largest number of scaffolding for a problem in that clip, the total number of past help request for that clip, the total number of actions that were second to the last hint for that clip, the largest number of consecutive errors in that clip, and least sum of right actions in that clip. The resulting model showed students that had low average of correct actions were frustrated.

Features used in the engaged concentration detector included the number of correct answers during the clip, the proportion of actions where the student took over 80 seconds to respond, whether the student followed scaffolding with a hint request, whether the student received scaffolding on the first attempt in a problem, and how many of the student's previous five actions involved the same problem. The model was created using the K* algorithm, which is an instance-based classifier. Instance-based classifiers predict group membership based on similarities to specific cases in the training set, rather than general rules, enabling them to identify constructs which can manifest in several distinct ways. For example, one group of students in engaged concentration repeatedly answered correctly in less than 80 seconds. Another group of students in engaged concentration answered incorrectly on their first attempt at a problem but then spent considerable time making their first response to the scaffolding question they received.

For confusion, detector features included the total number of consecutive incorrect actions for that clip, number of hints used for that clip, the number of correct actions in the clip, total number of past incorrect actions for a skill in that clip, correct actions that took time to answer, actions for a skill that the student got incorrect previously and that took time to answer. The resulting model was fairly complex, but one relationship leveraged in the model is that students who commit consecutive errors in a row for a skill are deemed confused. Another relationship is when students committed a number of incorrect actions in the past for a skill and took a long time to answer the current one, they are seen as confused.

The off-task detector included the total number of attempts made for a skill in that clip, the time taken by a student to answer, if a student has a correct action for that clip, average number of scaffold in that clip, and total number of incorrect actions in the past in the clip. The resulting model also was complex, but one relationship shows that if there were few attempts for a problem, and it took them a long time to answer, the student exhibits off-task behavior.

The features for the gaming detector included the use of a bottom-out hint in the clip, the number of hint usage for that clip, the average hint counts for a skill in that clip, the total number of actions for that clip that were answered incorrectly, and the occurrence of scaffold in that clip. The resulting model for gaming, like engaged concentration, used the K* algorithm. Hence, similarities that resulted to the group of gaming students included those that usually used bottom-out hints, scaffolding and hints.

Application of Models to Broader Data Set

Once the detectors of student affect and behavioral engagement were developed, they were applied to a broader data set consisting of two school years of student usage of the ASSISTments system by Worcester middle schools, 2004-2005 and 2005-2006. As discussed above, these schools represented a diverse sample of students in terms of both ethnicity and socio-economic status. This data set included 1,393 students and around 810,000 student actions within the learning software. The same features as discussed above were distilled for these data sets. Using these detectors, we were able to predict student affect and behavior for each student action within the ASSISTments system.

Correlation Analysis

In order to correlate students' affect estimates with their raw state test scores we first had to summarize their affect during the year, calculating one number per affective state per student. For each affective state we calculated the mean of the predicted probabilities for that state during performance on each skill in the system. This list of means for each skill was then averaged to produce summarized overall proportion of affect for the student. This averaging gives equal weighting of affect for each skill. This procedure was used because the MCAS test, which we are correlating to, consists of a random selection of skills. The weighting prevents a more frequently studied skill from having an influence on the students summarized affect that is disproportionate to its representation on the test.

Table 3: Example student affect dataset to be summarized

Student	Skill	Probability of Bored	Is Original?
Tricia	Subtraction	0.20	Yes
Tricia	Subtraction	0.50	No
Tricia	Subtraction	0.50	No
Tricia	Addition	0.90	Yes
Tricia	Addition	0.70	Yes

Table 3 shows example affect data for calculating the summary of the bored affective state for one student. To calculate the degree of boredom during the year for the student in Table 2, the following calculation would be used:

$$Tricia's P(Bored) = \frac{\left[\frac{(0.20 + 0.50 + 0.50)}{3} + \frac{(0.90 + 0.70)}{2} \right]}{2} = 0.60$$

We also calculate the summarized affect for each student for original and scaffold questions separately. In ASSISTments, scaffold questions are given when a student asks for help or answers an original question (main question) incorrectly. The scaffolding often consists of several sub questions and students know that they will be required to go through the scaffolding if a question is answered incorrectly; therefore we wanted to allow for the possibility of observing different affect during original questions than scaffolds.

Correlation Results

After summarizing the estimates of each student's affect, we used Pearson correlation to observe the correspondence between their affect and their end-of-year state test score. The results below show the correlation of affect to test score for the two years of data. We report separately on the affect experienced by students while answering original questions and the affect while answering scaffold questions, as the patterns of affect were substantially different in these two cases. Across tests, the high sample size resulted in most correlations being statistically significant (using the standard t-test for correlation coefficients, two-tailed).

Table 4: Correlation of student affect to their raw state test score. Statistically significant results ($p < 0.05$) are given in boldface; results where $p < 0.01$ are also italicized.

Correlation	ORIGINAL		SCAFFOLD	
	'04-'05	'05-'06	'04-'05	'05-'06
AFFECT				
Boredom	<i>-0.11930</i>	<i>-0.27977</i>	<i>0.32082</i>	<i>0.26884</i>
Engaged Concentration	<i>0.44923</i>	<i>0.25794</i>	<i>0.20988</i>	<i>0.09238</i>
Confusion	<i>-0.16538</i>	<i>-0.08912</i>	<i>0.37370</i>	<i>0.23457</i>
Frustration	<i>0.30524</i>	<i>0.20376</i>	<i>0.26182</i>	<i>0.22418</i>
Off-Task	<i>0.14820</i>	-0.00662	<i>0.16985</i>	<i>-0.10793</i>
Gaming	<i>-0.43083</i>	<i>-0.30125</i>	<i>-0.32933</i>	<i>-0.24688</i>

The strongest positive correlation, as shown in Table 4, was for engaged concentration on original questions. For 2004-2005, $r = 0.45$, $t(624) = 12.56$, two-tailed $p < 0.01$. For 2005-2006, $r = 0.26$, $t(760) = 7.36$, two-tailed $p < 0.01$. This finding is unsurprising, and maps to previous results showing a positive relationship between this affective state and learning (cf. Craig et al., 2004; Rodrigo et al., 2009). Even on scaffolding items, this relationship remained positive. For 2004-2005, $r = 0.21$, $t(624) = 5.36$, two-tailed $p < 0.01$. For 2005-2006, $r = 0.09$, $t(760) = 2.56$, two-tailed $p = 0.01$.

Boredom on original questions was negatively associated with learning outcomes, again matching previous research (cf. Craig et al., 2004; Pekrun et al., 2002; Rodrigo et al., 2009). For 2004-2005, $r = -0.12$, $t(624) = -3.00$, two-tailed $p < 0.01$. For 2005-2006, $r = -0.28$, $t(760) = -8.03$, two-tailed $p < 0.01$. However, boredom on scaffolding questions was associated with better learning. For 2004-2005, $r = 0.32$, $t(624) = 8.46$, two-tailed $p < 0.01$. For 2005-2006, $r = 0.27$, $t(760) = 7.69$, two-tailed $p < 0.01$. In interpreting this finding, it is worth considering why a student would become bored on a scaffolding question. One possibility is that the student knew the skill in the original question, but was careless (cf. San Pedro, Baker, & Rodrigo, 2011), which would explain these positive correlations. Another possibility is that high scoring students may know most of the skills involved with an original problem but not enough to answer correctly. When they are forced into the scaffolding, which breaks the main problem into individual skill sub questions, they become bored because they are being made to work on simpler questions that they already know the answers to.

Confusion had a similar pattern to boredom, with weak negative associations for original questions. For 2004-2005, $r = -0.17$, $t(624) = -4.19$, two-tailed $p < 0.01$. For 2005-2006, $r = -0.09$, $t(760) = -2.47$, two-tailed $p = 0.01$. By contrast, positive associations were found for scaffolding questions. For 2004-2005, $r = 0.37$, $t(624) = 10.06$, two-tailed $p < 0.01$. For 2005-2006, $r = 0.23$, $t(760) = 6.65$, two-tailed $p < 0.01$. Recent work has suggested that confusion impacts learning differently, depending on whether it is resolved (Lee et al., 2011), and that in some situations, confusion can be beneficial for learning (Lehman et al., 2012). The finding here accords with those papers, suggesting that confusion can be positive if it occurs on items designed to resolve that confusion.

Frustration had a positive correlation to learning, both for original items and scaffolding items. For original items, for 2004-2005, $r = 0.31$, $t(624) = 8.01$, two-tailed $p < 0.01$. For 2005-2006, $r = 0.20$, $t(760) = 5.74$, two-tailed $p < 0.01$. For scaffolding items, for 2004-2005, $r = 0.26$, $t(624) = 6.78$, two-tailed $p < 0.01$. For 2005-2006, $r = 0.22$, $t(760) = 6.34$, two-tailed $p < 0.01$. This finding is unexpected. Past research has suggested little relationship between frustration and learning (Craig et al., 2004; Rodrigo et al., 2009), contrary to hypotheses of a negative correlation. One possibility is that frustration in ASSISTments shows up in teacher reports in terms of negative performance, and that these students

receive greater support from their teachers. Clearly, it will be valuable to follow up and study this unexpected result further.

Gaming the system had a negative correlation with learning outcomes. For original items, for 2004-2005, $r = -0.43$, $t(624) = -11.92$, two-tailed $p < 0.01$. For 2005-2006, $r = -0.30$, $t(760) = -8.71$, two-tailed $p < 0.01$. For scaffolding items, for 2004-2005, $r = -0.33$, $t(624) = -11.92$, two-tailed $p < 0.01$. For 2005-2006, $r = -0.25$, $t(760) = -8.71$, two-tailed $p < 0.01$. These findings match previous evidence that gaming is associated with poorer learning (Aleven et al., 2004; Cocea et al., 2009).

The relationship between off-task behavior and learning was unstable between years, and weak in all cases. It varied between positive and negative, between years. For original items, for 2004-2005, $r = 0.15$, $t(624) = 3.74$, two-tailed $p < 0.01$. For 2005-2006, $r = -0.01$, $t(760) = -0.18$, two-tailed $p = 0.86$. For scaffolding items, for 2004-2005, $r = -0.17$, $t(624) = 4.31$, two-tailed $p < 0.01$. For 2005-2006, $r = -0.11$, $t(760) = -2.99$, two-tailed $p < 0.01$. It is not clear why the relationships between off-task behavior and learning were inconsistent between years.

Affect by Test Proficiency Category

Within this section, we ask: based on the results above (as well as prior research), are successful students mostly in a state of engaged concentration? Are unsuccessful students mostly gaming the system? To answer these questions we plot the affective state estimates by test proficiency category to reveal the dominant affective states with respect to test outcomes.

Figure 2 plots the state test proficiency category against the average estimate of affect on original questions for all students in that proficiency category. This is an average of the same probability estimates calculated in section 2.2.1. Note that these are the summarized affect estimates and therefore do not necessarily add to one. Non-summarized estimates may also not add to one because separate classifiers were used for each affect detector. While a multi-nominal classifier would guarantee a summing to one of predictions for each clip, it would not guarantee a more accurate prediction overall, particularly for underrepresented classes. In this analysis, we applied a second step of offset correction to the affect predictions that was applied in the original test of classifiers (San Pedro et al., 2013). This correction provides a more accurate scaling of the affect summaries but does not change the correlations from the first report of these results (e.g., Pardos, Baker, San Pedro, Gowda, & Gowda, 2013).

Affect on Original Questions

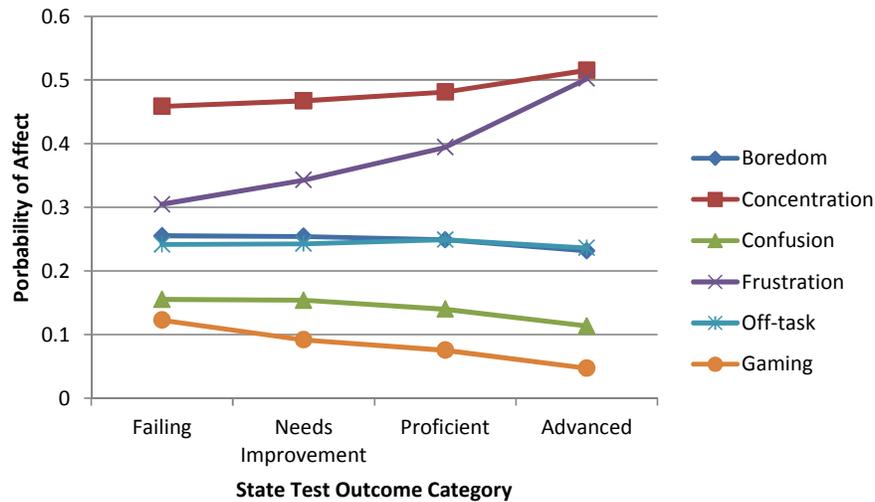


Figure 2: Probability of affect on original questions by test score category (average of both years' data)

We can observe from Figure 2, that the top affective state on original questions among failing students was concentration followed by frustration and boredom. The margin between concentration and frustration narrows as proficiency class increases until there is nearly equal parts of the two among students scoring in the Advanced category. For students scoring in Advanced, a category which earns the students a college scholarship, frustration is unexpectedly tied for the most probable affective state. The position of frustration, on the other hand, is somewhat surprising, raising the question of whether students react with frustration instead of boredom in response to material they find too easy.

Affect on Scaffold Questions

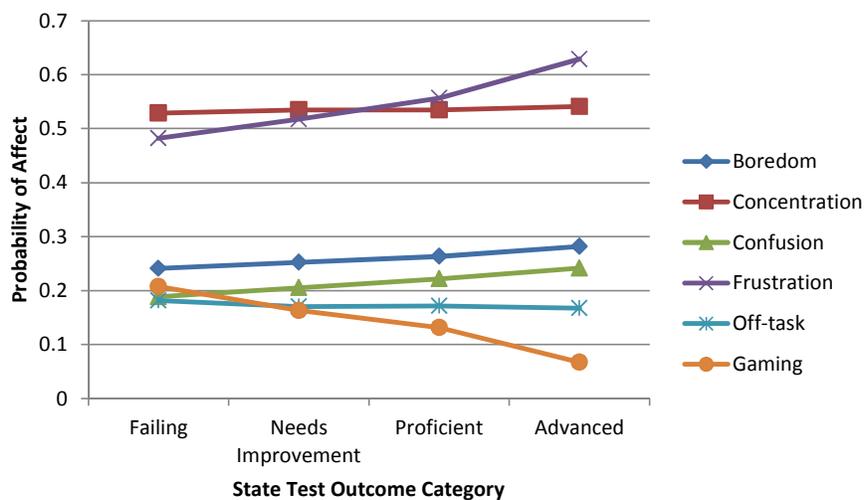


Figure 3: Probability of affect on scaffolds by test score category (average of both years' data)

The breakdown of affective state estimation on scaffold questions, shown in Figure 3, shows similarities to Figure 2 with frustration, engaged concentration and boredom being the most probable affective states. One difference is that frustration becomes the most prominent affect, instead of concentration, in the proficient and advanced categories, and engaged concentration and boredom showing little to no difference in probability between one another. On original questions, the interesting interaction was engaged concentration and frustration increasing in probability over boredom with higher scoring students. On scaffolds, the interesting interaction is among gaming, off-task behavior, and confusion. Among failing students, gaming is strongest, followed by off-task behavior, and then confusion. As the proficiency level increases, off-task and confusion become more probable as gaming becomes less common than they are. There are equal parts of these three states at the proficient level much like there were equal parts frustration, engaged concentration, and boredom at the proficient level for original questions. The takeaway for teachers here may be that gaming is generally undesirable, but confusion is not entirely problematic – successful students experience confusion on scaffolding items (perhaps because they are engaging with the material rather than disengaging by gaming the system).

Curiously, once again, highly successful students become frustrated more often on scaffolding items than less successful students. It may be that in these cases, students become annoyed and then frustrated at receiving scaffolding after making a mistake; or it may be that they are frustrated with themselves when they do not succeed. Higher levels of frustration may reflect a higher level of student emotional investment or pride in mastering the knowledge required to answer the problem. Since the problem sets used by students in these years of the tutor gave a random sampling of 8th grade skills, it is conceivable that this random ordering was a significant source of reasonable frustration for high and low proficiency students alike.

There is an observable difference in the magnitudes of affect estimates on originals and scaffolds. Table 5 quantifies this difference by calculating the estimate on scaffolds subtracted by the estimate on originals for each proficiency category. The average of these values across categories is shown in Table 5 along with the standard deviation among the four categories. If the shape of the trend line curve stays the same but is off-set from Figure 1 to Figure 2 uniformly across categories, this will result in an average difference but zero standard deviation. A high standard deviation indicates that the change in affect between scaffolds and originals is not of uniform magnitude across categories.

Table 5: Scaffold estimate subtracted by Original affect estimate and standard deviation across proficiency categories

Affect	Std.	Avg.
Frustration	0.0142	0.1543
Confusion	0.0404	0.0566
Concentration	0.0165	0.0365
Boredom	0.0301	0.0333
Gaming	0.0262	-0.0286
Off-task	0.0067	-0.0778

Table 6: Difference between Scaffold and Original affect estimates with the highest standard deviation across the proficiency categories

Affect	Failing	Needs Imp.	Proficient	Advanced	Std.
Confusion	0.0205	0.0323	0.0626	0.1111	0.0404
Boredom	0.0037	0.0183	0.0376	0.0735	0.0301
Gaming	-0.0008	-0.0191	-0.0313	-0.0631	0.0262

Table 5 shows that students are more likely to be frustrated in scaffolding than when answering original questions. Frustration increases by 0.1543 on average, the most of the affective states. This increase is fairly uniform across proficiency categories with a standard deviation of only 0.0142. The estimates of Confusion, Concentration and boredom increase in the Scaffolds but to a far lesser degree than Frustration. Gaming and Off-task behavior estimates decrease in Scaffolding. The change in these estimates was uniform across proficiency categories, indicated by the low standard deviation. The states with the highest standard deviation (shown in Table 6), although still low, were Confusion, Boredom, and Gaming. The increase in Confusion on Scaffolds was greater as the proficiency level increased, with Failing students showing a 0.0205 increase and Advanced students showing a 0.1111 increase. A similar, lower magnitude, trend was observed for Boredom. A decrease in Gaming was observed with increasing magnitude as proficiency level increased. Boredom and Confusion change from being negatively correlated with proficiency on Originals to being positively correlated with proficiency in Scaffolds, as shown in Table 4. With this kind of change we would expect to see a variance in the change in estimates across proficiencies for these states.

Prediction

In previous sections we have trained affect and behavioral engagement detectors and correlated these constructs with end-of-year outcomes. In this section we investigate how well student outcomes can be predicted by affect and behavior as compared to student performance. Prior work has shown that student usage choices while receiving tutoring in ASSISTments can predict as much of the variance in students' end-of-year state test scores as student performance can on items designed to assess test-related knowledge (Feng et al., 2009), a result replicated in (Ritter, Joshi, Fancsali, & Nixon, 2013). It may be worth also trying to understand the role that affect and behavior play in predicting student learning outcomes, in the form of end-of-year standardized examinations.

Methodology

In this section, we predict student performance on the math standardized state exam, the Massachusetts Comprehensive Assessment System, from three potential sets of features: an affect/behavior feature set, a performance set, and a combined set. Each of these feature sets was compared to a baseline model, the average test score in the training set. A detailed description of each set follows:

Table 7: Description of the four feature sets used in our prediction

Feature set	# features	Feature description
Affect/ Behavior	12	Six summative affect/behavioral engagement measures for the student on original questions and six on scaffolds
Performance	2	Percent correct of the student on

		original questions and scaffolds
Both	14	Combination of affect/behavioral and performance features
Baseline	N/A	Average test score in training set

By comparing the affect/behavior and performance feature sets to predict test scores, we can determine which has more predictive power. Using the combined feature set can tell us if the two sets are capturing the same variance or novel variance between them. The baseline measure gives us a simple prediction heuristic to compare to, the average test score for all students. These feature sets are described in Table 7. The four sets are generated for all students in both of years of data.

We use a five-fold cross-validation for each year of data separately and then also a train/test hold out where the training set is the previous year's data. In both validation experiments, we use standard linear regression to learn coefficients for each feature in the feature set that maximizes fit to the target variable of test score. Mean absolute error is used as the error metric and statistical significance between predictions is tested on the absolute errors with a two-tailed paired t-test.

Prediction Results

In this section we present the results of predicting the end-of-year scores based on features generated from tutor data collected from students during the school year. The baseline calculates the average test score in the training set and uses that prediction for every student in the test set. We use the different feature sets to compare the predictive power of each. Different datasets are used to observe whether predictive performance of the sets is consistent across years. Finally, a validation is conducted using data from one year as the test set and data from the previous year as the training set in order to test longitudinal model consistency.

Table 8: Prediction results for the four feature sets on the two datasets and a one year train / next year test holdout.

Mean Absolute Error / Pearson Correlation		Validation dataset/holdout		
		'04-'05 5-Fold CV	'05-'06 5-Fold CV	'04-'05 train '05-'06 test
Feature set	Aff/Eng	6.48 / 0.736	7.41 / 0.650	8.56 / 0.587
	Performance	6.24 / 0.753	7.56 / 0.693	7.87 / 0.692
	Both	6.08 / 0.765	6.20 / 0.762	7.67 / 0.694
	Baseline	10.15 / NA	10.29 / NA	10.67 / NA

Table 8 shows predictive performance results in terms of Mean Absolute Error (MAE).

Overall, predictive performance of regression on the combination of the 12 *affect* features and the 2 *performance* features was better than either feature set alone. Specifically, the *both* model was best for all comparisons, although the difference was only marginally significant between *performance* and *both* in the '04-'05 dataset ($p = 0.084$). The difference was statistically significant in the '04-'05 ($p < 0.001$) and the different year hold out set ($p = 0.035$).

Overall, the *baseline* model was worst for all comparisons, significantly at the $p < 0.001$ level in all cases.

The difference between *affect/behavior* and *performance* was not stable. It was only significant in the case where the '04-'05 model was used on the '05-'06 data, where *performance* performed better than *affect/behavior*. In the other comparisons, this difference was not significant, $p = 0.129$ in the '04-'05 dataset and $p = 0.515$ in the '05-'06 dataset.

Overall, then, it can be argued that *affect/behavior* and *performance* are each good predictors of the state test. Furthermore, a combined feature set generally performs better than either of the feature sets alone. This suggests that while affect and performance provide similar predictive ability, they capture significantly different variance.

The third dataset, using the '04-'05 dataset as training and '05-'06 dataset as testing, served as a validation that more closely fit how the detectors and prediction might be used in a real-world scenario, where scores of other students within a year can not be used to train prediction within the same year but instead are used to train a model applied to the next year. With this validation, *affect/behavior* features performed 9% less accurately than *performance* features but the combination of features resulted in a statistically significant improvement.

The overall model, combining both *affect/behavior* and *performance* features, trained on the combined '04-'05 and '05-'06 data set, is shown in Table 9.

Table 8: Features of the stepwise regression model in the order they were added to the model. An “(o)” denotes “on originals” and an “s” denotes “on scaffolds.”

#	Feature description	Coefficient
1*	Gaming (o)	-8.27
2	Percent correct (o)	52.09
3	Confusion (o)	12.81
4	Frustration (s)	10.94
5	Concentration (o)	-65.74
6	Concentration (s)	48.36
7	Bored (o)	-48.61
8	Bored (s)	61.11
9	Offtask (s)	-35.13
10	Offtask (o)	13.28

Conclusions

In this paper, we evaluate the relationship between affect and behavioral engagement in a tutoring system over the course of a year, to performance on an end-of-year high-stakes test. Differentiating affect/behavioral engagement on original problems versus scaffolding help problems elicited interesting results, in terms of boredom and confusion. Students who were bored or confused while answering the main problems, tended to do poorly on the test; however, boredom and confusion on scaffolding were associated with positive performance on the test. Gaming the system was, as expected, associated with poorer learning, while off-task behavior was not consistently associated with poorer learning. One unexpected finding was a positive relationship between frustration and learning, which should be investigated further. These findings are clearly not yet conclusive, representing just a single online

* Gaming (o) was the first feature added to the model; however, it was removed from the model in the last step of the regression

learning environment; but the methodological step that they represent – enabling analysis of affect that is both longitudinal and fine-grained, in the service of understanding the relationships between affect and learning – is a potentially valuable step. The data set produced through the application of these detectors is amenable to considerable further analysis of the ways that the context of affect influences learning. This will be a productive and valuable area for future work.

Overall, we find that a model integrating across multiple measures of affect and behavioral engagement can effectively predict student in the high-stakes exam. Such a model performs even better if measures of performance are also considered. As such, we can infer not just which affective states matter, but make an integrated prediction of how successful a student will be on a standardized examination.

Overall, these findings may be useful in the design of reporting on student behavior and affect for teachers using digital learning and assessment platforms. When reporting on student boredom and confusion, it will be important to report context as well. For example, it may be useful to recommend interventions to teachers if a student is bored or confused on original questions, but not if these affective states occur during scaffolding. We see this work as leading in the direction of better support for teachers on intervening based on students' affect. Real time integration of affect detection into a teacher's tutor dashboard along with an expanded understanding of the conditions that can make an affective state constructive or not, could greatly assist a teacher in signaling when to intervene in a crowded classroom.

Acknowledgments

We would like to thank Neil Heffernan for sharing the ASSISTments data with us, and access to ASSISTments classes, Adam Nakama, Adam Goldstein, and Sue Donas, for their participation and support in the original data collection, Lisa Rossi for copy editing assistance, and support from the Bill and Melinda Gates Foundation, award #OPP1048577 and the NSF, award #DRL-1031398.

References

- Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2004). Toward tutoring help seeking: Applying cognitive modeling to meta-cognitive skills. In J. C. Lester, R. M. Vicario, and F. Paraguaçu (Eds.), *Proceedings of Seventh International Conference on Intelligent Tutoring Systems, ITS 2004*, 227-239.
- Arnold, K. E. (2010). Signals: Applying Academic Analytics. *Educause Quarterly*, 33, 1.
- Baker, R. S. J. d. (2007). Modeling and Understanding Students' Off-Task Behavior in Intelligent Tutoring Systems. In *Proceedings of ACM CHI 2007: Computer-Human Interaction*, 1059-1068.
- Baker, R. S. J. d., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to Be Frustrated than Bored: The Incidence, Persistence, and Impact of Learners' Cognitive-Affective States during Interactions with Three Different Computer-Based Learning Environments. *Int'l. J. Human-Computer Studies*. 68, 4, 223-241.
- Baker, R. S. J. d., Goldstein, A. B., & Heffernan, N. T. (2011). Detecting Learning Moment-by-Moment. *International Journal of Artificial Intelligence in Education*, 21(1-2), 5-25.
- Baker, R. S. J. d., Gowda, S., & Corbett, A. T. (2011). Towards predicting future transfer of learning. In *Proceedings of 15th International Conference on Artificial Intelligence in Education*, 23-30.
- Baker, R. S. J. d., Gowda, S. M., Wixon, M., Kalka, J., Wagner, A. Z., Salvi, A., Aleven, V., Kusbit, G., Ocumpaugh, J., & Rossi, L. (2012). Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra. In *Proceedings of the 5th International Conference on Educational Data Mining*, 126-133.
- Bartel, C. A., & Saavedra, R. (2000). The Collective Construction of Work Group Moods. *Administrative Science Quarterly*, 45(2), 197-231.
- Cocca, M., Hershkovitz, A., & Baker, R. S. J. d. (2009). The Impact of Off-task and Gaming Behaviors on Learning: Immediate or Aggregate? In *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, 507-514.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37-46.
- Conati, C., & Maclaren, H. (2009). Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction*, 19(3), 267-303.
- 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.
- D'Mello, S. K., Craig, S. D., Witherspoon, A. W., McDaniel, B. T., & Graesser, A. C. (2008). Automatic Detection of Learner's Affect from Conversational Cues. *User Modeling and User-Adapted Interaction*, 18(1-2), 45-80.
- Feng, M., Heffernan, N. T., & Koedinger, K. R. (2009). Addressing the assessment challenge in an Intelligent Tutoring System that tutors as it assesses. *Journal of User Modeling and User-Adapted Interaction*, 19, 243-266.
- Hanley, J., & McNeil, B. (1980). The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology* 143, 29-36.
- Lee, D. M., Rodrigo, M. M., Baker, R. S. J. d., Sugay, J., & Coronel, A. (2011). Exploring the Relationship Between Novice Programmer Confusion and Achievement. In *Proceedings of the 4th bi-annual International Conference on Affective Computing and Intelligent Interaction*, 175-184.

- Lehman, B., D'Mello, S. K., & Graesser, A. C. (2012). Confusion and Complex Learning during Interactions with Computer Learning Environments. *The Internet and Higher Education*, 15(3), 184-194.
- Pardos, Z. A., Baker, R. S. J. d., San Pedro, M. O. C. Z., Gowda, S. M., & Gowda, S. M. (2013). Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. In *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge*, 117-124.
- Pardos, Z. A., Wang, Q. Y., & Trivedi, S. (2012). The real world significance of performance prediction. In *Proceedings of the 5th International Conference on Educational Data Mining*, 192-195.
- Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of quantitative and qualitative research. *Educational Psychologist*, 37, 91-106.
- Planalp, S., DeFrancisco, V. L., & Rutherford, D. (1996). Varieties of Cues to Emotion in Naturally Occurring Settings. *Cognition and Emotion*, 10(2), 137-153.
- Ritter, S., Joshi, A., Fancsali, S. E., & Nixon, T. (2013). Predicting Standardized Test Scores from Cognitive Tutor Interactions. In *Proceedings of the 6th International Conference on Educational Data Mining*, 169-176.
- Rodrigo, M. M. T., Baker, R. S., Jadud, M. C., Amarra, A. C. M., Dy, T., Espejo-Lahoz, M. B. V., Lim, S. A. L., Pascua, S. A. M. S., Sugay, J. O., & Tabanao, E. S. (2009). Affective and Behavioral Predictors of Novice Programmer Achievement. In *Proceedings of the 14th ACM-SIGCSE Annual Conference on Innovation and Technology in Computer Science Education*, 156-160.
- Sabourin, J., Mott, B., & Lester, J. (2011). Modeling Learner Affect with Theoretically Grounded Dynamic Bayesian Networks. In *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction*, 286-295.
- San Pedro, M. O. Z., Baker, R. S. J. d., Gowda, S. M., & Heffernan, N. T. (2013). Towards an Understanding of Affect and Knowledge from Student Interaction with an Intelligent Tutoring System. In *Proceedings of the 16th International Conference on Artificial Intelligence and Education*, 41-50.
- San Pedro, M. O. C., Baker, R., & Rodrigo, M. M. (2011). Detecting Carelessness through Contextual Estimation of Slip Probabilities among Students Using an Intelligent Tutor for Mathematics. In *Proceedings of 15th International Conference on Artificial Intelligence in Education*, 304-311.
- Sayette, M. A., Cohn, J. F., Wertz, J. M., Perrott, M. A., & Parrott, D. J. (2001). A psychometric evaluation of the facial action coding system for assessing spontaneous expression. *J. Nonverbal Behavior*, 25(3), 167-185.
- Van Rijsbergen, C. J. (1974). Foundation of evaluation. *Journal of Documentation*, 30, 365-373.