Correlating Affect and Behavior in Reasoning Mind with State Test Achievement

Victor Kostyuk
Reasoning Mind
200 Bering Drive, Suite 300
Houston, TX 770057
vkostyuk@reasoningmind.org

Ma. Victoria Almeda
Teachers College, Columbia University
525 W. 120th Street
New York, NY 10027
victoriaalmeda@gmail.com

Ryan S. Baker
University of Pennsylvania
3700 Walnut Street
Philadelphia, PA 19104
ryanshaunbaker@gmail.com

ABSTRACT

Previous studies have investigated the relationship between affect, behavior, and learning in blended learning systems. These articles have found that affect and behavior are closely linked with learning outcomes. In this paper, we attempt to replicate prior work on how affective states and behaviors relate to mathematics achievement, investigating these issues within the context of 5th-grade students in South Texas using a mathematics blended learning system, Reasoning Mind. We use automatic detectors of student behavior and affect, and correlate inferred rates of each behavior and affective state with the students' end-of-year standardized assessment score. A positive correlation between engaged concentration and test scores replicates previous studies, as does a negative correlation between boredom and test scores. However, our findings differ from previous findings relating to confusion, frustration, and off-task behavior, suggesting the importance of contextual factors for the relationship between behavior, affect, and learning. Our study represents a step in understanding how broadly findings on the relationships between affect/behavior and learning generalize across different learning platforms.

CCS CONCEPTS

• Applied computing • Applied computing~Education • Applied computing~Computer-assisted instruction

KEYWORDS

high-stakes tests, affection detection, prediction, intelligent tutoring system

LAK '18, March 7–9, 2018, Sydney, NSW, Australia

ACM ISBN 978-1-4503-6400-3/18/03…$15.00
https://doi.org/10.1145/3170358.3170378

ACM Reference format:

V. Kostyuk, M. Almeda, and R. Baker. 2018. Correlating Affect and Behavior in Reasoning Mind with State Test Achievement. In *Proceedings of the International Conference on Learning Analytics and Knowledge, Sydney, Australia, March 2018 (LAK'18)*, 5 pages. DOI: https://doi.org/10.1145/3170358.3170378

1 INTRODUCTION

One of the ongoing questions in research on affect and engagement has been how these aspects of student experience relate to learning outcomes [1, 2, 3, 4], particularly performance on standardized examinations [5, 6, 7]. Understanding how student behaviors correlate with learning outcomes can offer insight into educational practice. In particular, understanding which forms of engagement matter for student outcomes can drive the design of reporting and intervention systems for teachers that help improve academic outcomes [8].

Though there has been research on how engaged and disengaged behaviors relate to outcomes for quite some time (e.g., Lahaderne [9]), considerable recent research in this area has been conducted using fine-grained analyses of student behavior in computer-based learning environments. For example, Feng, Heffernan and Koedinger [6] found that a model which included students' online interactions in the ASSISTments tutoring system (e.g., total number of attempts, time spent solving each question) was significantly better at predicting standardized test scores than a model which only used students' correctness of items.

In another example, Rodrigo and colleagues [2] were able to identify which affective states and behaviors significantly predicted students' midterm scores in an undergraduate programming course. Findings indicate that students who experienced more confusion, boredom, and on-task conversation were more likely to do worse on the midterm exam. A marginally significant positive relationship between on-task behavior and learning was found. Frustration, off-task behavior and gaming the system were not significantly associated with midterm exam scores.

In a third example, Pardos et al. [5] investigated the relationship between affect/behavioral engagement and learning outcomes by correlating students' proportions of affect and engagement in the ASSISTments tutoring system, as assessed by machine-learned detectors with students' scores on the Massachusetts Comprehensive Assessment (MCAS) state standardized test. Engaged concentration was positively associated with standardized test scores, while boredom, confusion, and gaming

the system had negative relationships with test scores. They also found a positive correlation between frustration and learning. In contrast, off-task behavior had the weakest relationship with learning and was not consistently associated with lower standardized test scores. However, it is not yet clear how much these results generalize beyond the specific learning platform and student population in which that study was conducted.

In a fourth example, Ritter, Joshi, Fancsali, and Nixon [7] found that students' pre-test scores and their online behaviors in the Cognitive Tutor learning platform significantly predicted standardized test scores, above and beyond pre-test scores alone.

Lastly, in more recent work, Feng and Roschelle [10] investigated how online homework behaviors in ASSISTments relate to end-of-year standardized test scores. While their results indicate that completing more homework problems in less time is associated with higher scores, the effect size between online behaviors and learning was not as strong as seen in prior work (e.g. [6]).

In this paper, we attempt to determine how general these prior findings are, in particular, those findings relating to affective states and disengaged behaviors in Rodrigo et al. [2] and Pardos et al. [5]. To understand how general those findings are, we do so in the context of a different learning platform, geographic region, ethnic makeup of the student population, and using a different standardized assessment. Specifically, we investigate this question in the context of Reasoning Mind *Foundations*, a mathematics learning system for elementary students. We measure student affect and behavioral disengagement using models developed from a combination of field observations and data mining. We then examine the correlation between students' affect and behavioral disengagement and their scores on the State of Texas Assessments of Academic Readiness (STAAR) standardized examination.

2 REASONING MIND FOUNDATIONS

Reasoning Mind *Foundations* is a mathematics blended learning program that provides an online mathematics curriculum with instructional, practice, and assessment components together with teacher training and support [11]. This blended learning system was developed for elementary students and designed to be implemented in a classroom with a teacher present. Students spend most of the time in the online program in *Guided Study*, a mode that combines interactive instruction with independent practice and review. The system diagnoses gaps in pre-requisite knowledge of struggling students and attempts to remediate those gaps. The teacher receives real-time information on student progress and uses that information to tailor one-on-one or small group interventions with students [12]. The Reasoning Mind learning system is used by approximately 150,000 students a year, mainly in the Southern part of United States.

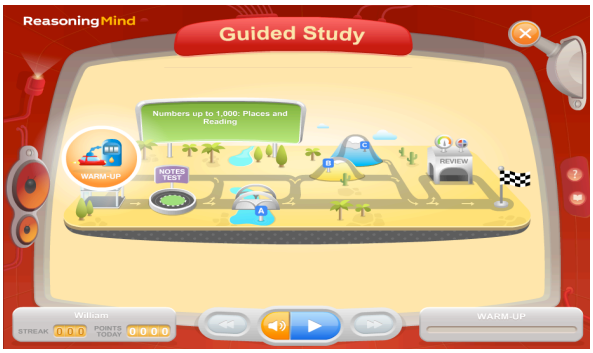

Figure 1: Reasoning Mind Guide Study Objective Map.

3 METHODOLOGY

3.1 Data Collection

For the measure of learning, we used Mathematics State of Texas Assessment of Academic Readiness (Math STAAR) scale scores. These scores were obtained from 15 schools within six school districts in South Texas, including both urban and rural districts. We also obtained system log data from 5th-grade students who used the Reasoning Mind platform during AY 2015-16. Of these, 1444 students used the system for at least 84 hours during the academic year, approximately 2 hours per week, and were included in the study. Because the 5th Grade Reasoning Mind curriculum is designed to be used as a core curriculum (replacing traditional mathematics instruction), low usage of the system typically corresponds to students who have been absent for most of the academic year (e.g., transferred to a different class or school), or to teachers who implement the program as a supplement, with infrequent use targeting particular topics rather the full set of grade-level standards.

Student actions in the system were aggregated into 20-second intervals, or "clips," for application of the affect and behavior detectors, discussed below. The selected students generated a total of 1,948,485 action clips within the instructional mode in the system.

Merging the two data sets resulted in 613 students having both STAAR data and a sufficient quantity of log data. Of the 613 students, 347 are female and 266 male; the majority (547) are economically disadvantaged; few (22) are special education students; the majority (580) are Hispanic.

3.2 Affect and Behavior Detection

Detectors of four affective states were utilized in this study: boredom, frustration, confusion, and engaged concentration. We also utilized detectors of four forms of engaged and disengaged behavior: off-task behavior, solitary on-task behavior, on-task conversation, and proactive remediation [12]. Proactive remediation occurs when a teacher decides to provide specialized

instruction to one or more students about a specific math topic, typically based on reports that those students are struggling with that topic.

Detectors were developed using training data from quantitative classroom observations conducted by human coders using the BROMP protocol [13]. Trained personnel observed students in a pre-set order for up to 20- second intervals ("clips") and recorded, their affect and behavior during the observation using the HART Android app [14]. The observations, conducted in five Texas schools and one school in West Virginia, resulted in a total of 2550 clips. These were synchronized with the log data of the corresponding student actions in the Reasoning Mind system. Feature engineering was conducted, resulting in 277 distinct features. As shown in Table 1, examples of the features generated across detectors were related to the student's rate of slowing down or speeding up his/her work relative to an average rate within a 20-second clip or longer 2-hour window.

Table 1: Examples of features included across detectors

Feature	Feature Description
timeSDlong-windowdiffmax	Maximum difference in standard deviation from the mean of times it took the student to answer the questions in the present 20-second clip and the average standard deviation of the student in answering questions in a 2-hour long window.
timeSDdiff2max	Maximum standard deviation from the mean response time to the past three answers and the standard deviation of mean response times for the 20-second clip.
timeSDdiffmin	Minimum standard deviation from the response time to the past two answers and the standard deviation of mean response times for the 20-second clip.
Time3SDmin	Minimum standard deviation from the mean response time to the past four answers and the standard deviation of the mean response times for the 20-second clip.
3rdPeakPJmin	Minimum standard deviation of the third largest predicted value of the probability of just having learned a skill ("J") on the present item across all skills in the 20-second clip.

For details on the observation procedure in the Reasoning Mind classroom, interrater reliability of observers, and the feature engineering process, see Miller et al. [12].

The resulting labeled training dataset was used to fit the detectors. Student-level 24-fold cross-validation was used to test five algorithms in predicting each affective state and behavior. Before training, upsampling of clips from the smaller group was used to exactly balance the incidence of the target affect/behavior in the training set of the fold. Additionally, feature reduction based on pair-wise correlations among all features was used to select 80 out of the 277 features before training. The following algorithms and R packages/functions were tested: stepwise forward selection linear regression (**glm**), LASSO (**glmnet**), J48 (**RWeka**), random forest (**randomForest**), and gradient boosting machine (**gbm**). AUC ROC [15], the probability that the model can distinguish a member of one class from a member of the other class, was used to evaluate the goodness of fit for each fold, and the algorithm with the highest average AUC across all folds was selected for each affective state and behavior.

Stepwise forward selection linear regression (GLM) had the best AUC for confusion, off-task, and solitary on-task. Random forest performed best for boredom, frustration, and proactive remediation. Gradient boosting machine performed best for engaged concentration and on-task conversation.

As shown in Table 2, all of the affect and behavior detectors performed better than chance, with the detector for proactive remediation achieving the best performance (AUC = 0.79). The overall range of the detectors was AUC from 0.53 to 0.79, with most detectors in the 0.60-0.66 range. These values were somewhat lower than the sensor-free affect detectors used by Pardos et al. [5], which achieved AUC values from 0.63-0.82.

Table 2: Performance of affect and behavior detectors

Detector	Algorithm	AUC
Boredom	RF	0.60
Frustration	RF	0.65
Confusion	GLM	0.53
Engaged Concentration	GBM	0.61
Off-task	GLM	0.64
Solitary on-task	GLM	0.66
On-task conversation	GBM	0.58
Proactive remediation	RF	0.79

3.3 Data Analysis

We correlated students' affect and behavior to their STAAR scores as follows: First, we computed the average predicted probabilities of each affective and behavior across all clips per student, giving us an estimate of each student's affective state and behavior across the entire academic year. Then, we computed Pearson and Spearman correlations between the rates of each affective state/ behavior and the students' STAAR scores.

We calculated the corresponding p-values for the correlations and used the Benjamini-Hochberg [16] method to adjust the alpha values for multiple comparisons.

4 RESULTS AND DISCUSSION

4.1 Correlational Results

We summarize the results of the Spearman and Pearson correlations between each detector and the STAAR test scores in Tables 3 and 4, respectively.

Table 3: Spearman Correlation of Student Affect and Behavior to their STAAR Scores

Detector	ρ	p-value	adjusted alpha
Boredom	-0.48	<0.01	0.016*
Frustration	-0.01	0.85	0.050
Confusion	-0.25	<0.01	0.025*
Engaged Concentration	0.23	<0.01	0.019*
Off-task	-0.54	<0.01	0.022*
Solitary on-task	0.04	0.58	0.047
On-task conversation	-0.06	0.16	0.038
Proactive remediation	-0.18	<0.01	0.031*

Note: *significant after Benjamini-Hochberg correction

Table 4: Pearson Correlation of Student Affect and Behavior to their STAAR Scores

Detector	*r*	p-value	adjusted alpha
Boredom	-0.27	<0.01	0.003*
Frustration	0.03	0.47	0.044
Confusion	-0.18	<0.01	0.009*
Engaged Concentration	0.20	<0.01	0.006*
Off-task	-0.43	<0.01	0.013*
Solitary on-task	0.02	0.38	0.041
On-task conversation	-0.07	0.07	0.034
Proactive remediation	-0.18	<0.01	0.028*

Note: *significant after Benjamini-Hochberg correction

Engaged concentration was positively associated with standardized test scores (ρ = 0.23, *r* = 0.20). This result accords with previous findings, both in Pardos et al. [5] and in Rodrigo et al. [2], where engaged concentration was referred to as "flow."

Boredom showed strong negative correlations with STAAR scores, indicating that more boredom is associated with worse performance (ρ = -0.48, *r* = -0.27), a pattern also seen in Rodrigo et al. [2]. We also found that confusion had a similar relationship with learning, such that more confusion was associated with lower STAAR scores (ρ = -0.25, *r* = -0.18) - another result in common with Rodrigo et al. [2].

However, frustration had no statistically significant correlation with standardized test scores (ρ = -0.01, *r*= 0.03), a pattern in line with Rodrigo et al. [2] but different from Pardos et al. [5].

In terms of behavior related to engagement, we found that on-task conversation was not significantly related to STAAR performance (ρ= -0.06, *r*= -0.07), a result in common with Pardos et al. [5] but different from Rodrigo et al. [2]. We also found that solitary on-task behavior did not have a statistically significant correlation with STAAR performance (ρ= 0.04, *r*=0.02). In contrast, off-task behavior had the strongest correlation across all of the detectors, with a strong negative relationship to standardized test scores (p=-0.54, *r* =-0.43).

Finally, we found that proactive remediation was also significantly negatively correlated with learning, albeit mildly so (ρ =-0.18, *r* = -0.18). This result was unexpected, considering that proactive remediation sessions are typically intended to improve learning by providing students with specialized instruction. A possible explanation for this result, however, is that teachers typically offer proactive remediation to students who are already struggling, potentially explaining those students' lower test scores.

5 CONCLUSIONS

In this article, we have investigated how student affect and behavior relate to achievement outcomes, by correlating several affective states and engagement-related behaviors to their STAAR scores. We find that students who had high rates of engaged concentration were more likely to perform well on the test. By contrast, students who were frequently bored, confused, or off-task were less likely to achieve higher scores on the test. Frustration, on-task conversation, and on-task behavior were not found to have a significant correlation with performance on the standardized test.

Some of these results aligned with Pardos et al.'s [5] findings and Rodrigo et al.'s [2] findings. In particular, the positive relationship between engaged concentration and standardized test scores replicated within the context of Reasoning Mind, as did the negative correlations for boredom and confusion, seen in Rodrigo et al. [2]. Our results indicate that it is potentially beneficial for Reasoning Mind teachers to anticipate or help resolve student confusion - for example, by providing scaffolds that help resolve an impasse during learning. Helping relieve student boredom is also likely to be beneficial for Reasoning Mind students.

However, while Pardos and his colleagues found a positive relationship between frustration and standardized test scores, we did not find a relationship between this affective state and learning, a null effect also seen in Rodrigo et al. [2]. A more surprising finding was the strong negative correlation between off-task behavior and performance on the standardized exam, a pattern not seen in previous work. Rodrigo et al. [2] and Fancsali [17] found a null effect for off-task behavior, and Pardos et al. [5] found a mild positive effect. Previous studies found that Reasoning Mind students have unusually low rates of off-task behavior [18, 19] – as such, it is possible that off-task behavior correlates strongly with performance within Reasoning Mind because a considerable amount of the more benign forms of off-task behavior disappear in Reasoning Mind, leaving only the most problematic off-task behaviors. Better understanding this relationship represents a valuable area for future work. At minimum, however, this work suggests that re-designs focused on reducing off-task behavior are particularly important within the context of Reasoning Mind. Previous research suggests that students using Reasoning Mind engaged in less off-task behavior when e-learning principles, such as personalization and multi-media, were incorporated in a re-design of the system [19]. Exploring other ways to minimize off-task behavior may potentially be useful for teachers in supporting Reasoning Mind students to avoid disengagement and achieve better learning.

Overall, our findings indicate that several published relationships between affect/behavior and learning can replicate in a different math learning platform, with a regionally and demographically different student population, and with a different state standardized assessment. Given the replicable positive correlations between engaged concentration and learning, research on this relationship can potentially inform the design of new interventions. By comparison, the extent to which frustration, boredom, confusion and off-task behavior relate to standardized test scores appears to be contingent on contextual factors. When designing interventions based on these relationships, it is important for teachers and researchers to think about contextual factors that may mediate the effects of frustration, confusion, and off-task behavior on learning.

ACKNOWLEDGMENTS

We would like to thank Matt Labrum and Sergey Yavorsky for help extracting and cleaning the log file data.

REFERENCES

[1] Scotty Craig, Arthur Graesser, Jeremiah Sullins, and Barry Gholson. 2004. Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. *Journal of educational media* 29, 3 (Jul. 2004), 241-250.

[2] Ma Mercedes T. Rodrigo, Ryan S. Baker, Matthew C. Jadud, Anna Christine M. Amarra, Thomas Dy, Maria Beatriz V. Espejo-Lahoz, Sheryl Ann L. Lim, Sheila AMS Pascua, Jessica O. Sugay, and Emily S. Tabanao. 2009. Affective and behavioral predictors of novice programmer achievement. In *Proceedings of the 14th Annual ACM SIGCSE Conference on Innovation and Technology in Computer Science Education*. ACM Press, New York, NY, 156-160.

[3] Maria Ofelia Z. San Pedro, Ryan S. Baker, Sujith M. Gowda, and Neil T. Heffernan. 2013. Towards an understanding of affect and knowledge from student interaction with an intelligent tutoring system. In *Proceedings of the International Conference on Artificial Intelligence in Education*, Springer, Berlin, Heidelberg, 41-50.

[4] Maria Ofelia Z. San Pedro, Erica L. Snow, Ryan S. Baker, Danielle S. McNamara, and Neil T. Heffernan. 2015. Exploring Dynamical Assessments of Affect, Behavior, and Cognition and Math State Test Achievement. In *Proceedings of the 8th International Educational Data Mining Society*, Madrid, Spain, 85-92.

[5] Zachary A. Pardos, Ryan SJD Baker, Maria San Pedro, Sujith M. Gowda, and Supreeth M. Gowda. 2014. Affective states and state tests: investigating how affect and engagement during the school year predict end-of-year learning outcomes. *Journal of Learning Analytics* 1, 1 (2014), 107-128.

[6] Mingyu Feng, Neil Heffernan, and Kenneth Koedinger. 2009. Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction* 19, 3 (Feb. 2009), 243-266.

[7] Steve Ritter, Ambarish Joshi, Stephen Fancsali, and Tristan Nixon. 2013. Predicting standardized test scores from Cognitive Tutor interactions. In *Proceedings of 6th International Conference on Educational Data Mining*, Memphis, Texas, 169-176.

[8] Kimberly E. Arnold, Steven Lonn, and Matthew D. Pistilli. 2014. An exercise in institutional reflection: The learning analytics readiness instrument (LARI). In *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge*, Indianapolis, Indiana, 163-167.

[9] Henriette M. Lahaderne. 1968. Attitudinal and intellectual correlates of attention: A study of four sixth-grade classrooms. *Journal of educational psychology* 59, 5 (Oct. 1968), 320-324.

[10] Mingyu Feng, and Jeremy Roschelle. 2016. Predicting Students' Standardized Test Scores Using Online Homework. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*, ACM Press, New York, NY, 213-216.

[11] George A. Khachatryan., Andrey V. Romashov, Alexander R. Khachatryan, Steven J. Gaudino, Julia M. Khachatryan, Konstantin R. Guarian, and Nataliya V. Yufa. 2014. Reasoning Mind Genie 2: An intelligent tutoring system as a vehicle for international transfer of instructional methods in mathematics. *International Journal of Artificial Intelligence in Education* 24, 3 (Sept. 2014), 333-382.

[12] William L. Miller, Ryan S. Baker, Matthew J. Labrum, Karen Petsche, Yu-Han Liu, and Angela Z. Wagner. 2015. Automated detection of proactive remediation by teachers in Reasoning Mind classrooms. In *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*, ACM Press, New York, NY, 290-294.

[13] Jaclyn Ocumpaugh, Ryan S. Baker, and Ma Mercedes Rodrigo. 2015. Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual. Technical Report. Teachers College, Columbia University, New York, NY. Ateneo Laboratory for the Learning Sciences, Manila, Philippines.

[14] Jaclyn Ocumpaugh, Ryan S. Baker, Ma Mercedes Rodrigo, Aatish Salvi, Martin Van Velsen, Ani Aghababyan, and Taylor Martin. 2015. HART: The human affect recording tool. In *Proceedings of the 33rd Annual International Conference on the Design of Communication*, ACM Press, New York, NY, 24.

[15] James A. Hanley, and Barbara J. McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology 143*, 1 (Apr. 1982), 29-36.

[16] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society Series B (Methodological)*, 289-300.

[17] Stephen Fancsali. 2014. Causal discovery with models: behavior, affect, and learning in cognitive tutor algebra. In *Proceedings of 7th International Conference on Educational data mining*, 28-35.

[18] Jaclyn Ocumpaugh, Ryan S. Baker, Steven Gaudino, Matthew J. Labrum, and Travis Dezendorf. 2013. Field observations of engagement in Reasoning Mind. In *Proceedings of the International Conference on Artificial Intelligence in Education*, Springer, Berlin, Heidelberg, 624-627.

[19] Kevin Malqueeny, Victor Kostyuk, Ryan S. Baker, and Jaclyn Ocumpaugh. 2015. Incorporating effective e-learning principles to improve student engagement in middle-school mathematics. *International Journal of STEM Education* 2, 1(Dec. 2015), 2-14.