ASSISTments Longitudinal Data Mining Competition Special Issue: A Preface

Thanaporn Patikorn
Worcester Polytechnic Institute
tpatikorn@wpi.edu

Ryan S. Baker
University of Pennsylvania
rybaker@upenn.edu

Neil T. Heffernan
Worcester Polytechnic Institute
nth@wpi.edu

This special issue includes papers from some of the leading competitors in the ASSISTments Longitudinal Data Mining Competition 2017, as well as some research from non-competitors, using the same data set. In this competition, participants attempted to predict whether students would choose a career in a STEM field or not, making this prediction using a click-stream dataset from middle school students working on math assignments inside ASSISTments, an online tutoring platform. At the conclusion of the competition on December 3rd, 2017, there were 202 participants, 74 of whom submitted predictions at least once. In this special issue, some of the leading competitors present their results and what they have learned about the link between behavior in online learning and future STEM career development.

Keywords: data challenge, data competition, ASSISTments, longitudinal outcomes, career prediction

1. INTRODUCTION

During the 10th International Conference on Educational Data Mining in Wuhan, China, the ASSISTments Longitudinal Data Mining Competition was announced by the Big Data for Education Spoke of the Big Data Northeast Innovation Hub, a research hub funded by the US National Science Foundation. This competition used a longitudinal dataset collected on students using ASSISTments, a free online tutoring platform, in 2004 - 2006. The ASSISTments team tracked those students longitudinally to see who graduated from high schools, who went on to college, what their majors were, and finally, if they chose a career in STEM (Science, Technology, Engineering, and Math) for their first job, post-college. Several papers have shown that behavior in ASSISTments in middle school can predict high school and college outcomes (Ocumpaugh et al., 2016; San Pedro et al., 2013, 2015). The task given to the participants in this competition was to use deidentified click-stream data to try to predict whether the student pursued a career in STEM or not. This data was provided to participants to analyze before it was used by the research team themselves, an unusual step that enabled participants in the competition to gain first access to a cutting-edge research data set.

In recent years, there has been increasing interest by school districts and state education agencies in predicting student success and dropout (Bowers, 2010; Knowles, 2015). These detectors are used to give early warnings to teachers, guidance counselors, and school leaders when students show signs that they are losing interest or experiencing difficulties. These detectors support teachers making targeted interventions to take necessary actions to help students before it's too late. However, there has thus far been relatively less work to drive K-12 early warning based on students' risk of dropping out of the STEM pipeline. This is particularly concerning, given the current economic context. While there is an increasing demand for STEM workers, substantial numbers of students lose interest in STEM subjects and fields or are insufficiently prepared to participate in these careers (Sass, 2015). Developing automated detection of STEM career participation may help us to identify students who could benefit from an intervention to help to support their interest and readiness for STEM (Reider et al., 2016).

2. ASSISTments Longitudinal Data Mining Competition

The competition ran from June 27th, 2017, to December 3rd, 2017. Registration for the competition and the dataset was entirely free, in line with the goals of promoting 1) STEM education, 2) educational data mining, and 3) open science. The primary condition of accessing the dataset was to not take any action to deanonymize the dataset. Even though the competition has already been concluded, interested researchers are still able to obtain access to the competition dataset[1].

2.1. Dataset

The dataset in this competition was a click-stream dataset collected from the ASSISTments learning platform (Heffernan & Heffernan, 2014) from 2004 - 2006. This dataset contained actions middle-school students took while working on their mathematics assignments. In addition to raw recorded actions, participants were also provided with several distilled measures, for instance, measures of the student's affective state and disengaged behaviors (bored, concentrating, confused, frustrated, off-task, and gaming). These measures were obtained by collecting student affect observations in real classrooms and then using machine learning techniques to train models that replicated those judgments within a click-stream dataset (Pardos et al., 2014). The detectors were validated to ensure that they applied effectively to unseen students from urban, rural, and suburban settings (Ocumpaugh et al., 2014). Data from this same student cohort had been previously used to investigate the relationship between student affect and disengaged behavior within ASSISTments and student outcomes for standardized examinations the same year (Pardos et al., 2014), high school attitudes and course-taking (Ocumpaugh et al., 2016), college enrollment (San Pedro et al., 2013), and college major (San Pedro et al., 2015).

The dataset contains 78 click-stream data predictor variables and the target variable "isSTEM": whether the student's career of choice was in the STEM fields or not, defined using the NSF guidelines for STEM careers and obtained using a LinkedIn Premium Account.

[1] Interested researchers can obtain the dataset at https://goo.gl/forms/seAyF0aHUOxevhfF3 The description of the dataset can be found in the competition website: https://sites.google.com/view/assistmentsdatamining/data-mining-competition-2017

There are 942,816 action-level data rows collected from 1,709 students in total. For the competition, the dataset was split into three sets: the training set, the validation set, and the test set.

2.1.1. Training Set

The training set contained the majority of the students from the full dataset. For each student in this dataset, both the students' action-level ASSISTments usage data and their "isSTEM" variable were available. Participants, as well as any researchers who are interested in STEM education, could make full use of this dataset, using any state-of-the-art data mining technique they chose to find the relationships between the student actions and their career choice (as long as it does not violate the terms of use).

During the data collection, there were many students for whom we collected ASSISTments usage data, but we were unable to retrieve their career information. Specifically, we know the isSTEM for only 591 students out of 1,709 students. We decided to include the ASSISTments usage data of these students in the training set since there are many co-training machine learning approaches that could train a model by using unlabeled data along with labeled data. The training set contains 514 labeled students and 1,118 unlabeled students.

2.1.2. Validation Set

The validation set was mainly used for the public leaderboard on the competition website. This leaderboard let participants know how well they were doing compared to other participants. All click-stream data from students in the validation set were made available to participants. Participants, however, were unable to directly access the "isSTEM" variable for the students in the validation set. When ready, participants could submit their predictions for the validation set's isSTEM students. The system would then evaluate the predictions, inform participants of their scores, and then update the participant's best scores on the leaderboard. The evaluation scheme will be further discussed in the later section.

2.1.3. Test Set

The only purpose of the test set was to determine the winner of the competition. Like the validation set, participants could only access the click-stream data of students in this set and not their isSTEM. The difference between the validation and the test set was that the test set was not used to calculate the leaderboard scores; the results were not visible until after the competition was complete. The reason we chose to separate the test set from the validation set was to make sure that the winners of the competition were not simply participants who overfit using the leaderboard, but who genuinely could predict entirely unseen data.

2.2. Evaluation

For the evaluation of models, participants were required to submit their predictions for students in both the validation set and the test set. Participants, however, were not informed as to which students were in which set. Once a day at noon EST, new submissions were evaluated on the validation set. While participants could submit as many predictions as they wanted, only the participant's latest submission was evaluated, to discourage them from overfitting to the leaderboard. The system then updated each participant's personal submission log with their latest submission's scores as well as the public leaderboard, where each participant's best scores were shown compared to other participants' best scores.

2.2.1. Evaluation Criteria

Both the leaderboard scores and the final scores were calculated by using a linear combination of the Area Under the ROC Curve (AUC) and the root mean squared error (RMSE). Since isSTEM was observed and collected as binary values, AUC was initially chosen as the evaluation criterion. AUC captures the model's ability to differentiate students in the two categories from each other, based on the relative confidence in the predictions. It is most suitable when the variable being predicted is binary and the predictions are numerical. However, after testing, we found that AUC, or any single metric, could be easily overfit to, especially given the small sample size.

Thus, we selected a second evaluation criterion: RMSE. While RMSE is designed for comparing two numbers, it provides an assessment that rewards models that are more certain when they are correct and punishes models that are uncertain with high confidence. It also maps to a context of use where the model provides different recommendations when it is uncertain than when it is highly confident.

For the sake of the competition, we decided to aggregate the two metrics, AUC and RMSE, into one score so that we could determine the winners. Since AUC ranges from 0 (reverse ranking) to 1 (perfect ranking) and RMSE, in this case, ranges from 0 (perfect predictions) to 1 (total opposite predictions), we define Aggregated Score as a linear combination of the two metrics, with one metric inverted:

$$\text{Aggregated Score} = \text{AUC} + (1 - \text{RMSE})$$

2.3. Different Population from Training to Validation and Test Sets

In October 2017, we discovered that the distribution of isSTEM within the training set was not the same as that of validation and test set. Specifically, the ratio of isSTEM = true and isSTEM = false of the validation set and test set were the same, but that ratio of the training set was more than double that of the validation set and test set. We investigated the issue and decided to keep the three sets as they were and announced this information to all participants. The reasons we decided to keep the data sets unchanged were 1) it is not uncommon for models to be applied to a context with different distribution and/or population from the training set. The difference between the sets, while they were not intended, did emulate this possible real application issue. 2) the isSTEM ratio of the validation set and the test set were the same, meaning participants could use the result from the validation set to adjust for the discrepancies between the training and the validation set, which would be reflected in the test set, since the isSTEM distribution of the validation and test sets were the same.

3. Conclusion of the Competition

The competition concluded on December 3rd, 2017. Figure 1 shows the pattern of sign-ups for the competition data set over time. At the conclusion of the competition, 202 participants had signed up for the competition, 74 of whom submitted predictions at least once. However, additional participants continued to sign up even after the competition concluded.

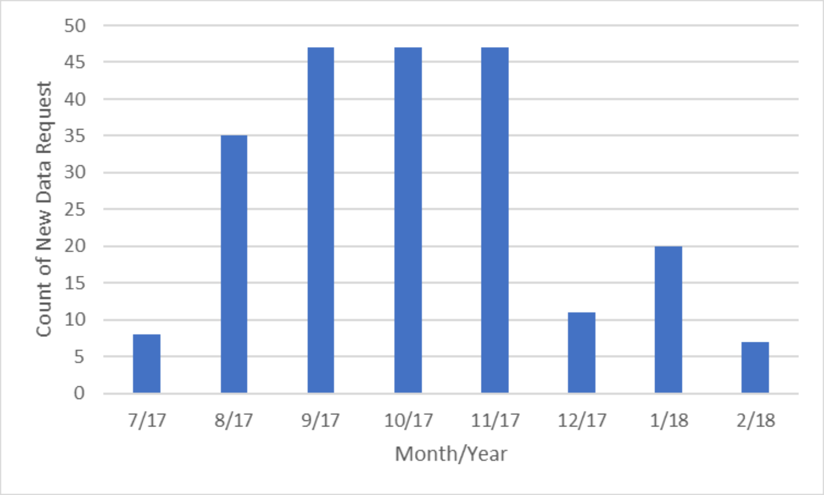

Figure 1: The number of new unique emails that signed up for the competition dataset in each month from July 2017 to February 2018.

3.1. Submissions Over Time

At first glance, the number of submissions, shown in Figure 2, peaked during November 2017, which was the last full month before the competition concluded. However, since the competition concluded on December 3rd, 2017, December 2017 was the month with the most submissions per day of 19.33, more than double the rate in November 2017 (9.19 submissions per day).

Figure 3 shows the number of submissions per participant. Among all participants who submitted predictions at least once, about two-thirds of them submitted more than once. Only about one-sixth submitted more than ten times. Only 8 participants submitted more than 20 times.

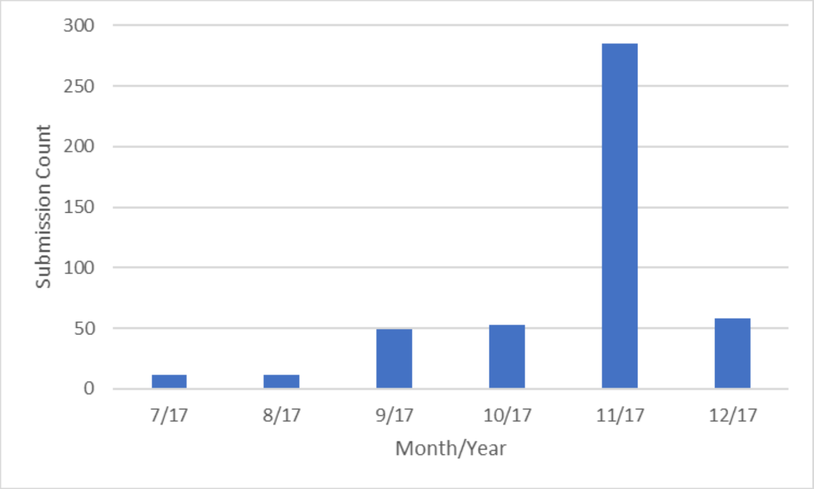

Figure 2: The number of submissions evaluated by the system in each month from July 2017 to December 2017.

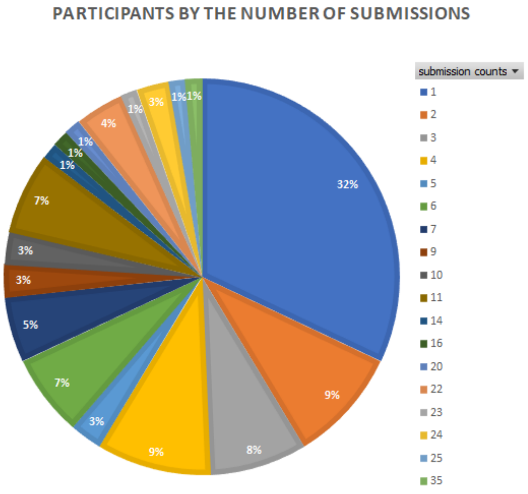

Figure 3: The percentage of participants by the number of submissions they made during the competition.

3.2. Submissions Over Time

Overall, the quality of submitted predictions averaged across all participants appeared to increase slightly over the months, as shown in Figure 4. While the average scores seemed to plateau after October, it is important to note that there were many participants who joined later in the competition. Their scores were averaged together with other participants who had already been working on the competition for several weeks.

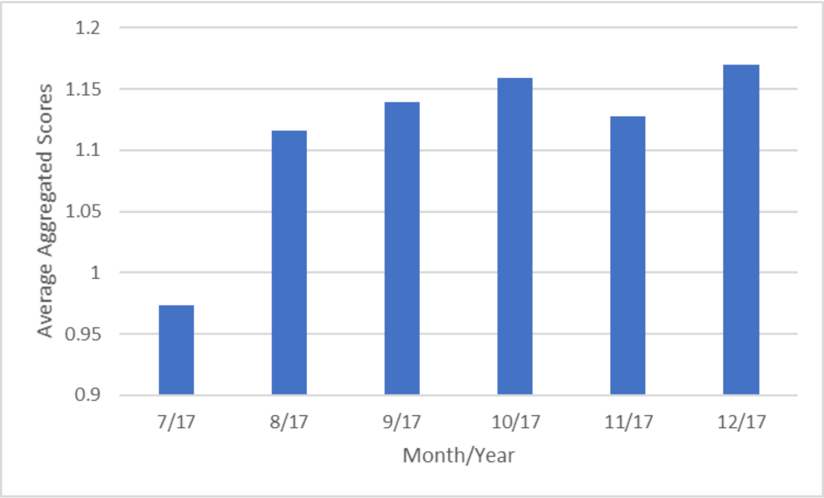

Figure 4: The aggregated scores averaged across all participant predictions submitted and evaluated in each month from July 2017 to December 2017.

Figure 5 shows how competitors' performance shifts as they submit more times; it shows the aggregated score of the 1st, 2nd, 3rd, etc. submissions averaged across all participants. A similar increasing trend to Figure 4 can also be observed in Figure 5. It is important to note that there were only 8 participants who submitted more than 20 times, which could be one of the reasons why the graph fluctuates considerably for x > 20.

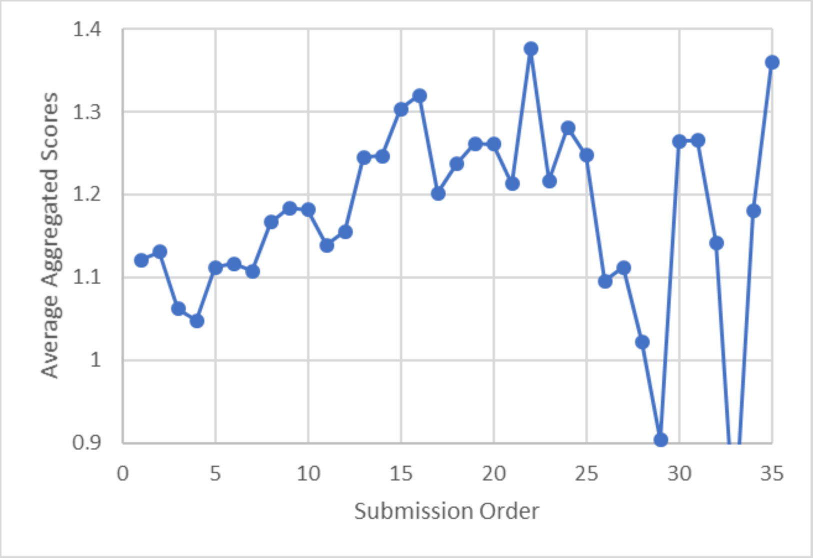

Figure 5: The aggregated scores by the submission order of each participant, averaged across participants from July 2017 to December 2017.

3.3. Winners

The three winners were announced during the Northeast Big Data Spoke Meeting at MIT on February 16th, 2018. The first-place winning team was Chun Kit Yeung, Kai Yang, and Dit-yan Yeung from the Hong Kong University of Science and Technology, who participated in the workshop. A write-up of their work was published prior to this special issue in the International Journal on Artificial Intelligence and Education (Yeung & Yeung, 2018). The second-place winner was Makhlouf Jihed from Japan's Kyushu University, whose paper is found in this issue. The third-place honors went to the University of Michigan Data Science Team, a group that regularly competes in data competitions like this one.

4. Special Issue

This special issue presents articles from three competition participants, as well as one team that did not participate in the competition.

Jihed and Mine (pp. 1-18) investigate what the best grain-size for developing predictive features is – the problem or the skill – and whether it is beneficial to compare students specifically to their classmates. They find that skill-level analysis is more effective than problem-level analysis and that comparing students specifically to their classmates is beneficial. This analysis is useful information – while lay understanding of student decisions to go into STEM careers often suggests that higher-performing schools disproportionately prepare students, Jihed and Mine find that within-school differences matter more than overall comparisons to all students, suggesting that a student's performance *within* their schools can influence the processes that lead to their eventual choice of STEM jobs. It may be an interesting area of future work for this line of research to look at whether student performance within schools exerts this influence by influencing mathematics attitudes and self-efficacy, an analysis that could be conducted using the data from (Ocumpaugh et al., 2016).

Liu and Tan (pp. 19-32) provide a more technically-focused investigation on how to derive and select features in a data set such as this one. They input data sets with different degrees of feature engineering into a variety of feature selection mechanisms, from straightforward forward-backward strategies to penalized logistic regression methods. Ultimately, they find that a straightforward forward-backward strategy on an extensive feature set is the most successful approach to prediction within this data set, among the options they investigate. They discuss how this approach can obtain high predictive performance without sacrificing model interpretability.

Almeda and Baker (pp. 33-47) use this data set to follow up on previous reports on longitudinal follow-up of this cohort (i.e., Pardos et al., 2014; San Pedro et al., 2013, 2015; Ocumpaugh et al., 2016). In doing so, Almeda and Baker adopt a statistical procedure similar to the approach used in those earlier articles rather than using more sophisticated data mining algorithms that could produce deeper understanding of the patterns in student affect or disengagement. They find that even after more than a decade, earlier negative correlations between gaming the system and outcomes (negative) persist, and that an earlier pattern observed where carelessness is positively correlated with outcomes until knowledge is controlled for also persists. As in San Pedro’s (2013, 2015) work on student outcomes in college, but unlike earlier analyses of same-year performance on standardized examinations (Pardos et al., 2014), affective states are not associated with later outcomes.

Chiu (pp. 48-77) breaks down the type of findings in Almeda and Baker further, looking in particular at the role played by gender. Chiu finds that different affective patterns are associated with STEM jobs for female students versus male students; whereas boredom is associated with a lower probability of going into STEM for female students, frustration is associated with a lower probability of going into STEM for male students. Engaged concentration is associated with a higher probability of going into STEM, but only for male students. These findings call into question the decades-long debate about “whether it is better to be frustrated than bored” (Baker et al., 2010). Much of the work in this debate has combined male and female students. Chiu’s findings – along with other work by Arroyo and colleagues (2011) suggesting that different affective strategies are appropriate for female students than male students – call into question this assumption and call for splitting out students by gender in future analyses on student affect.

5. Future Work

There are multiple types of future work that we hope this competition in general, and this special issue in particular, will help to promote. First, this competition and special issue demonstrate that the phenomena that educational data mining researchers are investigating in interactive learning environments have long-term predictive power. There has been considerable research into the longitudinal impacts of more general student orientations towards learning, measured through questionnaires (i.e., Dweck, 2013; Duckworth, 2016). The findings of this competition, and the earlier work it builds on, suggest that fine-grained differences in student skill, engagement, and strategy matter for long-term outcomes, and further study of these constructs has the potential to yield deeper understanding of the phenomena that drive student success and outcomes. In turn, this type of research may enhance theoretical understanding of student career development, an area where theories remain high-level and under-specified.

We also hope to see wider use of the data set that we have shared for this competition. EDM researchers investigate a broad range of within-system behaviors in data sets like this one – not even all of the constructs studied within ASSISTments have made it into this special issue, or other attempts to work with this data set to succeed in the competition. To give just one noteworthy example, wheel-spinning is a behavior that seems important to student success in ASSISTments (Beck & Gong, 2013), but was not explored (to our knowledge) by competitors. This data set creates an opportunity to study which of the many student behaviors we are investigating matter for learner outcomes.

Going forward, we hope that this special issue will encourage the collection – and sharing -- of a greater number of longitudinal data sets linked back to student engagement within interactive learning environments. At the moment, there are literally hundreds of fine-grained data sets publicly available to the educational data mining community, many in repositories like LearnSphere (formerly the PSLC DataShop). However, there are very few cases where these fine-grained data sets are paired with longitudinal outcome data, and we are aware of no cases outside of ASSISTments which have the key attributes of this competition's data set:

- A substantial number of learners
- Fine-grained data on learner behavior in an interactive learning environment
- Longitudinal outcome data from more than a year later
- Publicly available to the broader scientific community

We call on our colleagues worldwide to follow the example of this competition, and both collect longitudinal outcome data and release it to the community so that we can all conduct research on the long-term impacts of the types of constructs being studied in educational data mining. Educational data mining researchers study a broad range of learning systems that differ from ASSISTments in many ways. We need data from several systems available to the community to develop a fuller understanding of how the constructs we can model for a student today link to their outcomes going forward.

Overall, by making a larger amount of longitudinal data available to the EDM community, and conducting a range of analyses on what behaviors/constructs today matter for outcomes tomorrow, we can develop a better theory on the development of learners. In turn, this theory and these findings may help our community and our collaborative partners to design reports for instructors and automated interventions that better support students' long-term growth.

ACKNOWLEDGMENTS

The NSF funded this work under grants to WPI, UPENN, and UMASS (1636782, 1636847, and 1661987). Much of the data collection was supported by NSF grant #1031398.

The ASSISTments platform is made possible via multiple NSF grants (e.g., 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, 1535428, 1440753, 1316736, 1252297, 1109483, & DRL-1031398), the US Department of Education Institute for Education Sciences (e.g., IES R305A170137, R305A170243, R305A180401, R305A120125, R305A180401, & R305C100024) and the Graduate Assistance in Areas of National Need program (e.g., P200A180088 & P200A150306), the Office of Naval Research (N00014-18-1-2768 and other from ONR) and finally Schmidt Futures.

REFERENCES

ALMEDA, M.V.Q., AND BAKER, RS 2020. Predicting student participation in STEM careers: The role of affect and engagement during middle school. *Journal of Educational Data Mining,* 33-47.

ARROYO, I., BURLESON, W., TAI, M., MULDNER, K, AND WOOLF, B.P. 2013. Gender differences in the use and benefit of advanced learning technologies for mathematics. *Journal of Educational Psychology, 105* (4), 957-969.

BAKER, R.S., D'MELLO, S.K., RODRIGO, M.M.T., AND 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. *International Journal of Human-Computer Studies, 68* (4), 223-241.

BECK, JE, AND GONG, Y. 2013. Wheel-spinning: Students who fail to master a skill. *Proceedings of the International Conference on Artificial Intelligence in Education,431-440.*

BOWERS, A. J. 2010. Grades and graduation: A longitudinal risk perspective to identify student dropouts. *The Journal of Educational Research*, 103(3), 191-207.

CHIU, M-S. 2020. Predicting STEM choice by emotional traits and states of online mathematical problem-solving in middle school. *Journal of Educational Data Mining,* 48-77.

DUCKWORTH, A. 2016. *Grit: The Power of Passion and Perseverance.* New York, NY: Scribner.

DWECK, C.S. 2013. *Self-theories: Their Role in Motivation, Personality, and Development.* Hove, UK: Psychology Press.

HEFFERNAN, N.T., AND HEFFERNAN, C.L. 2014 The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence and Education, 24* (4), 470-497.

LIU, R., AND TAN, A. 2020. Towards interpretable automated machine learning for STEM career prediction. *Journal of Educational Data Mining,* 19-32.

KNOWLES, J. E. 2015. Of needles and haystacks: Building an accurate statewide dropout early warning system in Wisconsin. *Journal of Educational Data Mining,* 7(3), 18-67.

JIHED, M., AND MINE, T. 2020 Analysis of click-stream data to predict stem careers from student usage of an intelligent tutoring system. *Journal of Educational Data Mining,* 1-18.

OCUMPAUGH, J., BAKER, R., GOWDA, S., HEFFERNAN, N., AND HEFFERNAN, C. 2014 Population validity for educational data mining models: A case study in affect detection. *British Journal of Educational Technology*, 45 (3), 487-501.

OCUMPAUGH, J., SAN PEDRO, M.O., LAI, H-Y., BAKER, RS, AND BORGEN, F. 2016 Middle school engagement with mathematics software and later interest and self-efficacy for STEM careers. *Journal of Science Education and Technology, 25* (6), 877-887.

PARDOS, Z.A., BAKER, R.S.J.D., SAN PEDRO, MOCZ, GOWDA, SM, AND GOWDA, SM 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), 107–128.

REIDER, D., KNESTIS, K., AND MALYN-SMITH, J. 2016. Workforce education models for K-12 STEM education programs: Reflections on, and implications for, the NSF ITEST program. *Journal of Science Education and Technology*, 25(6), 847-858.

SAN PEDRO, M., BAKER, R., BOWERS, A. AND HEFFERNAN, N. 2013. Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. In S. D'Mello, R. Calvo, & A. Olney (Eds.) *Proceedings of the 6th International Conference on Educational Data Mining*, 177–184.

SAN PEDRO, M.O., BAKER, R., HEFFERNAN, N., AND OCUMPAUGH, J. 2015. Exploring College major choice and middle school student behavior, affect and learning: What happens to students who game the system? *Proceedings of the 5th International Learning Analytics and Knowledge Conference*, 36-40.

SASS, T. R. 2015. Understanding the STEM pipeline. Working Paper 125. National Center for Analysis of Longitudinal Data in Education Research (CALDER).

YEUNG, C.K., AND YEUNG, D.Y. 2018 Incorporating features learned by an enhanced deep knowledge tracing model for STEM/non-STEM job prediction. *International Journal of Artificial Intelligence and Education, 29* (3), 317-341.