Long-Term Prediction from Topic-Level Knowledge and Engagement in Mathematics Learning

Andres Felipe Zambrano*
Graduate School of Education, University of Pennsylvania
afzambrano97@gmail.com

Ryan S. Baker
Graduate School of Education, University of Pennsylvania
ryanshaunbaker@gmail.com

ABSTRACT

During middle school, students' learning experiences begin to influence their future decisions about college enrollment and career selection. Prior research indicates that both knowledge gained and the disengagement and affect experienced during this period are predictors of these future outcomes. However, this past research has investigated affect, disengagement, and knowledge in an overall fashion – looking at the average manifestation of these constructs across all topics studied across a year of mathematics. It may be that some mathematics topics are more associated with these outcomes than others. In this study, we use data from middle school students interacting with a digital mathematics learning platform, to analyze the interplay of these features across different topic areas. Our findings show that mastering *Functions* is the most important predictor of both college enrollment and STEM career selection, while the importance of knowing other topic areas varies across the two outcomes. Furthermore, while subject knowledge tends to be the most relevant predictor for general college enrollment, affective states, especially confusion and engaged concentration, become more important for predicting STEM career selection.

CCS CONCEPTS

• **Human-centered computing**; • **Social and professional topics**; • **Applied computing** → Education; Interactive learning environments;

KEYWORDS

Educational data mining, Affect, Gaming the system, Predictive analytics, Long-term prediction, STEM

ACM Reference Format:
Andres Felipe Zambrano and Ryan S. Baker. 2024. Long-Term Prediction from Topic-Level Knowledge and Engagement in Mathematics Learning. In *The 14th Learning Analytics and Knowledge Conference (LAK '24), March 18–22, 2024, Kyoto, Japan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3636555.3636851

*Corresponding author.

LAK '24, March 18–22, 2024, Kyoto, Japan

ACM ISBN 979-8-4007-1618-8/24/03…$15.00
https://doi.org/10.1145/3636555.3636851

1 INTRODUCTION

The decision to attend college, and the selection of a major, often begins to form in middle school [31]. While factors like economic status, gender, family background, and other demographics influence college decisions, class-level experiences play a pivotal role in determining both whether a student decides to attend college and future persistence in academic pursuits [28]. These beneficial outcomes can largely be attributed to key formative experiences that students undergo. Colleges often emphasize the importance of engaging in a challenging curriculum in high school, which is made possible by middle school experiences [31], in order to cultivate effective study habits crucial for college success. Lent et al. [18] further suggest that student experiences that sharpen essential skills lead to increased self-confidence, refined interests, and clearer goal-setting, key steps towards a career decision. Equipped with this groundwork, students can more confidently navigate their higher education paths and establish career goals. However, a lack of such experiences can diminish self-confidence and motivation for higher education [3].

For the particular case of mathematics, previous studies have shown that middle school students who challenge themselves with courses like Algebra I are better positioned to enroll in advanced courses once they reach high school [31]. Engaging with these advanced courses often heightens students' awareness of opportunities in higher education and increases their propensity to apply to four-year colleges [2]. Among minority and first-generation college students, those enrolled in advanced math courses demonstrate a higher likelihood of attending college [15]. Moreover, proficiency in mathematics has been identified as a reliable predictor of long-term academic outcomes [4].

Courses and classwork influence students beyond just imparting knowledge and skills; they also shape a student's interest in particular topics. Genuine intrinsic interest in math directly influences both the student's effort in math class and overall school attendance, which in turn affects college enrollment opportunities [4]. Students who do not find their courses engaging or associate negative self-perceptions and affective states with those courses are less inclined to pursue further study in that area in college [30]. In contrast, students who develop a passion for specific subjects are more likely to choose related majors in higher education [13]. Therefore, the specific learning experiences and the related affective experiences a student has in different domains can guide their interests and decisions regarding college majors and future careers.

In this context, educational software logs offer a valuable data source for assessing not only a student's knowledge but also their affective states during the learning process. Within STEM education, these logs, when paired with observations of trained human researchers [6], have been instrumental in creating detectors

for a range of affective states, including engaged concentration, boredom, frustration, and confusion [16, 23]. Such data has also been employed to train models to identify disengaged behaviors such as being off-task and gaming the system (trying to succeed in the learning system without genuinely understanding the content [23]). Additionally, log data has been pivotal in developing models that estimate a student's knowledge level [10, 27]. Based on these models and data collected from middle school onwards, previous studies have explored how knowledge, affective states, and disengaged behaviors can influence vocational self-efficacy [23] and longitudinal educational outcomes [1, 9, 25, 26]. Findings indicate that both acquired knowledge and specific affective states and disengaged behaviors can predict students' likelihood of attending college [25], opting for a STEM major [26], and pursuing a STEM career [1, 9], also showing an association with interest and vocational self-efficacy in STEM [22]. However, these works have considered these features in an overall fashion, overlooking the nuanced differences that learning experiences in specific subjects can suppose for these outcomes.

Recognizing that not every major involves the same mathematics, even within STEM learning, there is still a need to understand how knowledge, affective states, and disengaged behaviors in various topic areas might influence and predict these long-term educational outcomes. Therefore, in this study, we investigate the association between students' knowledge of several mathematical topic areas and both university enrollment and career selection (STEM vs non-STEM). Additionally, we explore whether knowledge of these subjects, along with affective states and disengaged behaviors experienced in each area, can predict their likelihood of university enrollment and pursuing a STEM career.

2 LONGITUDINAL OUTCOMES PREDICTION

Previous research has employed ML techniques for predicting longitudinal outcomes such as college and STEM major enrollment using knowledge estimates and detectors, derived from students' interactions with learning platforms, as features. San Pedro et al. [25] developed a Logistic Regression (LR) model that could accurately identify U.S. students who eventually attended college 68.6% of the time using knowledge, affective states, and disengaged behaviors models based on interaction data from the middle school math platform ASSISTments [14], along with the number of actions and correctness as features. Their findings indicated that college enrollment is positively associated with knowledge, engaged concentration, and correctness. On the other hand, boredom and confusion had negative relations with college enrollment (with smaller odds ratios), though when these affective states were integrated into the final prediction model, the direction of these two associations flipped. San Pedro et al. [26] found that the same dataset and features could also differentiate between students who enrolled in STEM majors and students who did not enroll in STEM majors with only slightly worse performance (66%). While the directions of associations persisted when contrasting STEM and non-STEM enrollees, many of them individually lost statistical significance. In fact, their top-performing model only included knowledge and gaming, which were the only variables that remained significant in the comparison.

Extending this scope, other research projects have focused on predicting and analyzing the choice of STEM careers post-college using data from the same learning system. Almeda & Baker [1], using a statistical approach rather than ML models, presented similar results to those related to STEM major enrollment, highlighting knowledge and disengaged behaviors, especially gaming the system, as pivotal predictors, finding the same direction of associations as observed by [26] for STEM major selection. Similarly, Chiu [9], employing Logistic Regression, Ordinary Least Square Regression, and Random Forests, also found that gaming the system is the most important predictor for STEM career selection. Furthermore, Chiu [9] explored potential gender disparities in predictions, discovering that detectors typically perform better for male students than female ones. Chiu [9] also noted that female students who experience less boredom and more off-task behaviors were more inclined towards a STEM career. In contrast, male students were more likely to select a STEM career when they experienced greater concentration and less frustration.

Other papers have focused on conducting further feature engineering in the same data set to maximize the performance indicators of STEM career predictive models. Liu & Tan [19] showed that by appropriately enriching features, considering statistical measures, mathematical transformations, and inter-feature interactions, and adopting a forward-backward strategy for feature selection, the performance of STEM career prediction models can improve by 9.3%. Makhlouf & Mine [20] found that detectors yielded better results when the features were aggregated based on the skills students practiced rather than the problems solved. Their best-performing model (a decision trees classifier) could distinguish between STEM and non-STEM careers 68% of the time when leveraging skill-based and school-aggregated features. This approach used similar features to [25, 26] while adding more nuance regarding help requests and time spent on problems and skills. Similarly, Yeung & Yeung [32] also assessed several features and models to enhance the performance of STEM career predictions. Their results indicated that integrating deep knowledge tracing models for each specific skill as features, rather than treating all knowledge uniformly, improved model metrics. Their top-performing model (logistic regression) achieved an AUC of 0.692.

While the relationships between knowledge, affective states, and disengaged behaviors in STEM education have been extensively explored for several educational outcomes, there seems to be an unaddressed gap in understanding the specific impacts of learning experiences associated with different topic areas. Although previous work has investigated specific skills, this grain-size is too fine to produce understanding of which areas of curriculum are most important to emphasize within interventions targeted at increasing participation in STEM. Based on the improvements in predictions shown when the features are aggregated considering the different skills [20, 32], we hypothesize that the influence of experiences across distinct topics might vary, potentially identifying certain areas with a stronger association with successful educational outcomes. Consequently, this research aims to discern these nuanced differences. Given previous studies emphasizing the primacy of knowledge in predicting college and STEM major enrollment, we begin our exploration by examining the influence of proficiency in each topic area on these two types of enrollment. Subsequently,

we integrate affective states and disengaged behaviors (categorized by topic area) into our analysis to determine if they offer additional insights that could refine our predictions for both general and STEM-specific enrollments.

3 METHODS

This section describes the dataset utilized in this study, enumerates skills and topic areas considered, and discusses the models employed to analyze the potential of each topic area for predicting college enrollment and participation in a STEM career.

3.1 Data

In this study, we employ an interaction dataset from the ASSISTments learning platform [14], gathered between 2004 and 2007 [24, 25]. This dataset contains 78 variables derived from students' interactions while completing mathematics problems using ASSISTments. Data from 1709 students from 4 middle schools in the U.S. (Avg=427.25, SD=209.25) was obtained. The dataset includes measures of students' affective states and disengaged behaviors, specifically boredom, concentration, confusion, frustration, off-task behavior, carelessness, and gaming the system. These measures were calculated by observing students in classrooms [6] and training machine learning models to replicate those judgments from student interactions with the learning system [23]. Validation was conducted to ensure these detectors were applicable across unobserved students from urban, suburban, and rural populations [21]. The dataset also includes student knowledge estimates, calculated using Bayesian Knowledge Tracing (BKT; [9]). Finally, the dataset contains longitudinal outcomes, including MCAS state standardized examination scores [23], college enrollment [25], and students' chosen careers (classified as being in STEM fields or not, according to the NSF guidelines for defining STEM careers; [24]). Although the MCAS test scores and college enrollment data were available for all students, data on career choice was only available for 591 students. The dataset is publicly available at https://sites.google.com/view/assistmentsdatamining/data-mining-competition-2017.

3.2 Grouping Skills

The dataset includes 3162 unique mathematics problems, categorized by the ASSISTments team into 102 distinct skills corresponding to the evaluated topic or knowledge component. This analysis excludes problems unrelated to a specific skill or set of skills (only considering 2210 problems). These skills were then consolidated into 12 topic areas (Cohen's Kappa between authors was calculated for each category to check inter-rater reliability and was above 0.85 for every category). We use these topic areas to investigate the impact of student proficiency (and affect and behavior) in each area on the longitudinal outcome. For instance, addition, subtraction, multiplication, and division were collated under the *Basic Operations* topic area. A detailed description and the inter-rater reliability of each topic area is provided in Table 1.

3.3 Analyses

This study investigates two longitudinal outcomes: college enrollment and STEM career selection. We do not investigate MCAS data, as our findings might simply reflect what topic areas the state chose to emphasize in their test design. We investigate the relationship between student proficiency in each topic area and longitudinal outcomes by first taking each student's final BKT estimate for each skill. Then we average across those final BKT estimates for each topic area, producing a single average of students' estimated knowledge at the end of their use of ASSISTments. We then compare those averages between students achieving the positive longitudinal result and students not achieving that result. Specifically, we compare final knowledge for enrolled and non-rolled students and compare final knowledge for STEM career and non-STEM career students. In making this calculation, for each topic area, we filter out any skill that was not encountered by the student and filter out any student who does not solve at least one problem of any skill in that topic area. For each topic area, we compute effect sizes using Cliff's Delta and conduct a Mann-Whitney U test due to the non-normal distributions of BKT estimations. We apply the Benjamini-Yekutieli correction for controlling the false discovery rate, in line with the original method proposed by [7]. Within this method, only tests with p-values lower than their corresponding alphas are deemed statistically significant.

We also assess the degree to which each topic area's data can predict each longitudinal outcome, utilizing a range of machine learning (ML) techniques. Initially, we develop 12 distinct logistic regression models, each employing a specific single topic area's BKT estimate to predict college and STEM major enrollments. These models are tested via a 4-fold (middle) school-level cross-validation approach and evaluated based on the mean and standard deviation of the Area Under the Receiver Operating Characteristic Curve (AUC ROC; AUC for short). Subsequently, we apply forward feature selection, considering BKT estimates of all topic areas, to identify a set of topic areas that, combined, lead to the best prediction. The ML techniques used in the forward feature selection are Logistic Regression (LR), Random Forests (RF), Support Vector Machines (SVM), and Multi-Layer Perceptron (MLP). We used the factory default settings of all ML models given by the SciKit Learn Library.

Finally, we incorporate affective states and behaviors into the feature forward selection algorithm to examine if supplementary features could improve predictions for college enrollment and STEM careers selection. Carelessness estimates are not considered for this analysis because in the publicly available dataset they were averaged across all the skills of each student and therefore it is not possible to calculate the specific carelessness for each topic area. We evaluate the mean decrease impurity (MDI) feature importance [8] of affective states, behaviors, and knowledge estimates for each topic area, as selected by the top-performing RF predictive model, which is also the ML technique with the best results among the two learning outcomes that we are predicting (See table 6, section 4.3). We selected this feature importance because its calculation is straightforward based on the splits across the decision trees [8]. Moreover, MDI reduces the risk of obscuring the relevance of features that are not uniformly (positively or negatively) associated with the outcome and depend on the interactions with other features, which is already known to be the case when predicting multiple educational outcomes with this set of features [9, 17]. For all the selected features, we compute effect sizes using Cliff's Delta

Table 1: Topic areas description.

Topic area	Skills	Kappa
Numbers	Integers, number sense operations, ordering numbers, number line, interpreting number line, prime numbers, rounding and scientific notation	0.936
Basic Operations	Addition, subtraction, multiplication, division, multiplication of positive and negative numbers, simple calculation, reciprocal, order of operations, multi-column addition, and multi-column subtraction.	0.954
Algebra	Algebra symbolization, algebraic manipulation, making sense of algebraic expressions, comparing expressions, equation concept, equation solving, interpreting linear equations, and algebraic relations.	0.878
Decimals, Fractions, and Percentages	Proportions, fraction concept, comparing fractions, reducing fractions, fraction multiplication, fraction division, adding decimals, subtracting decimals, multiplying decimals, dividing decimals, finding percentages, discount, and fraction, decimals, and percentage equivalence.	0.966
Geometry	Area concept, perimeter, volume, area of the circle, circumference, meaning of pi, triangles, congruence, pythagorean theorem, similar triangles, sum of interior angles, supplementary angles, properties of geometric figures, properties of solids, surface area and volume, rotations and transversals, and triangle inequality.	0.971
Exponents and square root	Exponents and square root.	1.000
Factors and multiples	Divisibility, least common multiple, prime number.	1.000
Inequalities	Inequality solving.	1.000
Functions	Patterns, relations, and functions, pattern finding, evaluating functions, slope, rate, and substitution.	0.865
Graphing	Graph shape, graph interpretation, reading graphs, comparing points, reading points, and finding the slope in a graph.	1.000
Probability	Probability, combinatorics, and Venn diagrams.	0.795
Statistics	Statistics, measurement, mean, mode, median, and stem and leaf plot.	0.928

and conduct a Mann-Whitney U test to reveal the overall direction of associations.

For students who did not attempt any problem in a specific topic area, we applied the mean output from the corresponding detector for that topic, calculated across all students with available data, as the imputation method. While more advanced imputation techniques could have been used to potentially improve model performance, our main objective was to identify the most relevant predictors. Therefore, we chose not to employ these sophisticated methods to avoid the inherent assumptions of relatedness that imputation makes, especially considering that the missing data in our study was likely not random.

4 RESULTS

4.1 Comparison between enrolled and not enrolled students, and STEM/non-STEM students

Table 2 presents a comparison of the mean BKT estimates for each topic area, contrasting students who enrolled in college with those who did not. The mean BKT estimates were statistically significantly higher for enrolled students across all topic areas, except for *Inequalities* (p=0.120). However, the effect sizes vary among the topic areas. The largest effect sizes were seen for *Functions* (cliff's delta=0.306; average BKT 0.389 for enrollees versus 0.271 for non-enrollees), *Decimals, Fractions, and Percentages* (cliff's delta=0.297; average BKT 0.381 for enrollees versus 0.294 for non-enrollees), and *Geometry* (cliff's delta=0.291; average BKT 0.300 for enrollees versus 0.226 for non-enrollees) whereas, apart from *Inequalities*, *Graphing* (cliff's delta=0.169) and *Factors and Multiples* (cliff's delta=0.206) had the smallest effect sizes. Generally, these effect sizes reveal that, while those who enrolled in college tended to have higher knowledge estimates after completing their experience with ASSISTments—statistically significantly so—the differences were nonetheless only moderately large. Comparable outcomes are observed when examining the AUC values. The classifiers that utilized *Functions, Decimal, Fractions, and Percentages,* and *Geometry* as their sole independent variables obtained the highest performance (with AUCs of 0.649, 0.641 and 0.644, respectively). In contrast, models employing *Factors and Multiples, Graphing,* and *Exponents and Square Root* achieve an AUC smaller than 0.6. Finally, for the detector based on the BKT estimation of *Inequalities*, the AUC (0.512) was just above 0.5, indicating that its performance is near to chance.

The results from comparing students pursuing STEM versus non-STEM careers differ from the findings around enrollment. Although STEM students typically have higher mean BKT estimates across all topics, only three differences stand out as statistically significant after the Benjamini-Yekutieli correction (*Functions, Algebra,* and *Geometry*). This diminished statistical disparity between populations could be due simply to the smaller sample size. However, effect sizes were also generally smaller than those seen for college enrollment. Only *Inequalities* topic area (cliff's delta=0.228; average BKT 0.437

Table 2: Comparison of mean BKT estimates for each topic area between enrolled and not enrolled students. For each topic area, the number of students who solved at least one problem in that area is included.

Topic area	Number of Enrolled Students	Number of Not Enrolled Students	Mean BKT Estimation(Enrolled Students)	Mean BKT Estimation(Not Enrolled Students)	Effect Size (Cliff's Delta)	Mann-Whitney U Test (p-value)	Benjamini-Yekutieli correction (alpha)	Logistic Regression Test AUC score	Test AUC Score Standard Deviation
Numbers	**979**	**501**	**0.536**	**0.432**	**0.271**	**<0.001**	**0.004**	**0.622**	**0.049**
Basic Operations	**993**	**508**	**0.517**	**0.422**	**0.248**	**<0.001**	**0.005**	**0.616**	**0.040**
Algebra	**1080**	**600**	**0.334**	**0.250**	**0.238**	**<0.001**	**0.011**	**0.605**	**0.032**
Decimal, Fractions, and Percentages	**1077**	**587**	**0.381**	**0.294**	**0.297**	**0.001**	**0.015**	**0.641**	**0.030**
Geometry	**1077**	**598**	**0.300**	**0.226**	**0.291**	**<0.001**	**0.003**	**0.644**	**0.041**
Exponents and square root	**829**	**416**	**0.688**	**0.580**	**0.241**	**<0.001**	**0.009**	**0.593**	**0.021**
Factors and Multiples	**597**	**284**	**0.647**	**0.541**	**0.206**	**<0.001**	**0.012**	**0.557**	**0.060**
Inequalities	485	233	0.344	0.289	0.072	0.120	0.016	0.512	0.088
Functions	**1062**	**572**	**0.389**	**0.271**	**0.306**	**<0.001**	**0.001**	**0.649**	**0.046**
Graphing	**1027**	**564**	**0.535**	**0.458**	**0.169**	**<0.001**	**0.013**	**0.589**	**0.035**
Probability	**1058**	**574**	**0.270**	**0.178**	**0.241**	**<0.001**	**0.008**	**0.610**	**0.047**
Statistics	**752**	**352**	**0.450**	**0.336**	**0.265**	**<0.001**	**0.007**	**0.603**	**0.039**

Table 3: Comparison of mean BKT estimates for each topic area between STEM and non-STEM students. For each topic area, the number of students who solved at least one problem in that area is included.

Topic area	Number of STEM students	Number of non-STEM students	Mean BKT Estimation(STEM)	Mean BKT Estimation(Non-STEM)	Effect Size (Cliff's Delta)	Mann-Whitney U Test (p-value)	Benjamini-Yekutieli correction (alpha)	Logistic Regression Test AUC score	Test AUC Score Standard Deviation
Numbers	116	414	0.546	0.538	0.006	0.922	0.016	0.457	0.024
Basic Operations	116	424	0.545	0.501	0.112	0.062	0.011	0.510	0.048
Algebra	**124**	**459**	**0.404**	**0.326**	**0.197**	**0.001**	**0.003**	**0.574**	**0.047**
Decimal, Fractions, and Percentages	122	455	0.429	0.379	0.147	0.012	0.007	0.516	0.081
Geometry	**124**	**457**	**0.345**	**0.292**	**0.178**	**0.002**	**0.004**	**0.541**	**0.061**
Exponents and square root	101	358	0.729	0.665	0.123	0.057	0.009	0.521	0.048
Factors and Multiples	71	248	0.684	0.642	0.096	0.215	0.015	0.543	0.066
Inequalities	56	206	0.437	0.337	0.228	0.009	0.006	0.537	0.049
Functions	**122**	**451**	**0.459**	**0.378**	**0.203**	**<0.001**	**0.001**	**0.577**	**0.037**
Graphing	117	443	0.599	0.542	0.125	0.037	0.008	0.527	0.054
Probability	125	445	0.327	0.280	0.099	0.089	0.012	0.514	0.050
Statistics	96	315	0.496	0.448	0.104	0.122	0.013	0.523	0.033

for STEM careers versus 0.337 for non-STEM), and *Functions* (cliff's delta=0.203; average BKT 0.459 for STEM careers versus 0.378 for non-STEM) had effect sizes above 0.2, whereas the effect sizes for three topic areas were below 0.1. The high cliff's delta for *Inequalities* represented a substantial contrast to the lack of significance of this topic area for predicting enrollment. The performance of the classifiers yields similar findings, where ten models perform near to chance (AUC<0.55), and none achieve an AUC value above 0.6. From this perspective, the models employing BKT estimates of *Functions* and *Algebra* show the highest performance (with AUCs of 0.577 and 0.574, respectively).

Table 4: Topic areas selected within the best model for each ML algorithm predicting college enrollment using BKT estimates as features. ML models were tested using 4-fold school-level cross-validation. Topic areas selected in the best-performing model of all the ML algorithms are shown in bold.

Topic Area	Logistic Regression	Random Forests	Support Vector Machine	Multi-Layer Perceptron
Numbers	**X**	**X**	**X**	**X**
Basic Operations	X	X	-	X
Algebra	-	X	X	-
Decimals, Fractions, and Percentages	X	-	-	X
Geometry	X	-	X	X
Exponents and square root	X	-	X	X
Factors and Multiples	-	X	X	-
Inequalities	**X**	**X**	**X**	**X**
Functions	**X**	**X**	**X**	**X**
Graphing	-	X	-	X
Probability	-	X	-	-
Statistics	**X**	**X**	**X**	**X**
Test AUC Score (Standard Deviation)	0.684 (0.050)	0.656 (0.039)	0.655 (0.048)	0.683 (0.046)

4.2 Enrollment and STEM career prediction using forward selection considering BKT estimations for all topic areas

We used four different machine learning techniques to assess the predictive capacity of BKT estimates for college enrollment. Table 4 shows the topic areas utilized in the most effective models derived from each of the four ML techniques. As discussed above, these features were selected using a forward feature selection algorithm with AUC as the performance indicator. The AUC values for these ML models range between 0.655 and 0.684. Logistic Regression and Multi-Layer Perceptron were the techniques with the highest performance with AUCs of 0.684 and 0.683, respectively. This performance is very similar to the AUC of 0.686 reported by [25], who used the same dataset but incorporated data on students' affective states and disengaged behaviors as well as the knowledge estimates.

The topic areas selected by the forward selection are aligned with the hypothesis tests and effect sizes previously observed. *Functions*, the topic area with the largest effect size, was consistently the initial choice in the forward selection of all the ML techniques. Furthermore, *Statistics* and *Numbers*, which both have effect sizes above 0.25, were also selected by the best models across the four ML techniques. Interestingly, even though the knowledge estimate for *Inequalities* didn't display a statistically significant difference between enrollees and non-enrollees, it was deemed beneficial by all ML models when combined with information from other topics. No topic areas were excluded across all techniques.

Table 5 shows the results of the forward feature selection when predicting STEM career selection. *Functions*, which is the topic area with the second highest effect size and the lowest p-value, was chosen in the top models across all the ML techniques. This result highlights *Functions* as a pivotal area for predicting both college enrollment and STEM career. In contrast, other topic areas such as *Statistics, Geometry,* and *Basic Operations*, prominent in models for college enrollment, were absent in the STEM career prediction models. Additionally, all ML models for STEM career selection have lower performances than those for college enrollment (AUCs ranging from 0.595 to 0.618), with Random Forest being the best performing technique, contrasting with the AUC of 0.692 reported by [32] for this task, who included several additional features. This higher difference with the benchmark performance for this task when exclusively considering knowledge estimates for the prediction models compared to college enrollment prediction, suggests that affective states and disengaged behaviors may be more useful for predicting the choice of a STEM career than for predicting college enrollment itself.

4.3 Enrollment and STEM career prediction considering affective states and behaviors detectors as features

To explore the impact of additional variables on predictions for college enrollment and STEM career selection (as in [20, 25, 26, 32]), we incorporated the average levels of each affective state and disengaged behavior from every subject area into the forward feature selection algorithm. Table 6 presents the AUC for the best model across the four ML techniques for predicting both longitudinal outcomes. For college enrollment predictions, our top models (Logistic Regression, AUC=0.693) only very slightly outperformed the previous benchmark set in this dataset (AUC=0.686; [25]). In contrast, when predicting STEM career selection, even our best model (Random Forest, AUC = 0.660) underperformed the best prior result obtained for this task using this dataset (AUC=0.692; [32]), although that paper included a range of other features that are outside the scope of our current research questions.

After including affective states and disengaged behaviors as features, the mean performance improvement for predicting STEM career selection (0.049) was more than twice higher than the observed AUC improvement for predicting college enrollment (0.021).

Table 5: Topic areas selected on the best model for each ML algorithm predicting STEM career selection using BKT estimates as features. ML models were tested using a 4-fold school-level cross-validation. Topic areas selected in the best-performing model of all the ML algorithms are shown in bold.

Topic Area	Logistic Regression	Random Forests	Support Vector Machine	Multi-Layer Perceptron
Numbers	X	X	-	-
Basic Operations	-	-	-	-
Algebra	X	X	-	X
Decimals, Fractions, and Percentages	-	-	X	-
Geometry	-	-	-	-
Exponents and square root	X	X	X	-
Factors and Multiples	-	-	X	X
Inequalities	X	-	X	-
Functions	**X**	**X**	**X**	**X**
Graphing	-	-	-	-
Probability	-	X	X	-
Statistics	-	-	-	-
Test AUC Score (Standard Deviation)	0.597 (0.030)	0.618 (0.026)	0.601 (0.045)	0.595 (0.037)

Table 6: Best-performing model for each ML algorithm predicting college enrollment and STEM career selection using BKT estimations, behavior detections, and affect detections as features. Each cell corresponds to the mean and standard deviation of the AUC for each specific model. ML models were tested using a 4-fold school-level cross-validation. The mean and the standard deviations of the improvement of the models when adding disengaged behavior and affective state detections as features are included.

Topic Area	Logistic Regression	Random Forests	Support Vector Machine	Multi-Layer Perceptron	Mean Improvement
College Enrollment	0.693 (0.050)	0.692 (0.037)	0.691 (0.046)	0.686 (0.041)	0.021 (0.015)
STEM Career	0.649 (0.040)	0.660 (0.043)	0.643 (0.064)	0.654 (0.046)	0.049 (0.009)

This result suggests that data on affective states and disengaged behaviors (broken out by topics) can be more relevant for predicting the choice of a STEM career than for college enrollment.

Table 7 shows the mean decrease in impurity feature importance (FI) of a RF model with factory default settings. As mentioned before, we selected the RF model because it was the best-performing model for STEM career selection prediction, and the feature importance calculation is straightforward based on the splits across the decision trees [8]. Even though it was not the top model for predicting college enrollment, the difference in its AUC with the best-performing model was less than 0.002. The cumulative importance of BKT estimations remains the highest among all the features of the model (cumulative FI of 0.300). Within these BKT estimates, *Functions* (FI of 0.059), *Decimals, Fractions, and Percentages* (FI of 0.055), and *Geometry* (FI of 0.052) are the topic areas with the highest feature importances, mirroring the results observed in the effect sizes between enrollees and non-enrollees. Those are the only features with FI higher than 0.05.

Among affective states, frustration and confusion stand out with the highest cumulative FI (0.159 and 0.147, respectively). Specifically, frustration detectors for *Functions*, *Algebra*, and *Geometry* are the only features surpassing a FI of 0.4, aside from BKT estimates. However, when comparing frustration levels between enrollees and non-enrollees, only the *Algebra* topic displays a statistically significant difference, with a moderate effect size (cliff's delta=0.1; see Table 8), indicating a slight unexpectedly positive association between experiencing frustration in *Algebra* and college enrollment. Meanwhile, confusion detectors for *Algebra*, *Graphing*, and *Basic Operations* are still important for predicting college enrollment (each with FI higher than 0.3). However, among these confusion detectors, only *Inequalities* shows a statistically significant difference, and the effect size is moderately small (cliff's delta of 0.133). Engaged concentration in *Geometry* and *Graphing*, and the boredom detector within *Graphing*, also have a FI surpassing 0.3 in the college enrollment prediction, but do not have statistically significant differences between enrollees and non-enrollees. Although many affect/topic area combinations do not have statistically significant differences between the two groups of students, some of

Table 7: Topic areas selected on the best-performing model for each ML algorithm predicting college enrollment using BKT estimations, behavior detections, and affect detections as features. Each cell corresponds to the Mean and standard deviation of the feature importance for each specific feature. ML models were tested using a 4-fold school-level cross-validation.

Topic Area	BKT	Gaming	Off Task	Engaged Concentration	Confusion	Frustration	Boredom	Total
Numbers	0.045 (0.003)	-	-	-	-	-	-	0.045
Basic Operations	-	0.028 (0.002)	0.031 (0.002)	-	0.031 (0.002)	-	-	0.090
Algebra	-	-	-	-	0.039 (0.003)	0.043 (0.004)	-	0.082
Decimals, Fractions, and Percentages	0.055 (0.003)	0.039 (0.002)	-	-	-	-	-	0.094
Geometry	0.052 (0.004)	0.017 (0.003)	-	0.039 (0.004)	-	0.042 (0.002)	-	0.150
Exponents and square root	0.034 (0.003)	-	-	-	-	0.031 (0.002)	-	0.065
Factors and Multiples	-	-	0.020 (0.001)	-	-	-	-	0.020
Inequalities	0.016 (0.003)	-	0.040 (0.003)	-	0.016 (0.002)	-	-	0.072
Functions	0.059 (0.005)	0.039 (0.002)	-	-		0.043 (0.002)	-	0.141
Graphing	0.039 (0.002)	-	-	0.034 (0.002)	0.033 (0.002)	-	0.031 (0.002)	0.137
Probability	-	0.035 (0.002)	0.031 (0.008)	-	-	-	-	0.066
Statistics	-	-	-	-	0.028 (0.001)	-	-	0.028
Total	0.300	0.158	0.122	0.073	0.147	0.159	0.031	1

them are still valuable for enhancing the prediction performance. This suggests that these affective states may influence educational outcomes in ways that are not straightforward or solely tied to knowledge levels.

The gaming detector is the feature with the third highest FI (cumulative FI of 0.158). As with confusion and frustration, the fact that gaming the system still has high FI, particularly for *Functions* and *Decimals, Fractions, and Percentages* (FI of 0.039 in both cases and cliff's deltas of -0.171 and -0.163 respectively), suggests that gaming impacts performance beyond just through lower knowledge despite the connections between gaming and knowledge (e.g. [1, 25]). Contrasting with the mixed results of confusion, gaming has a negative statistically significant association with college enrollment for all the topic areas included in the model, as well as having the largest (negative) effect sizes among all the features, although still lower than almost all the BKT estimates. Off-task behaviors within *Inequalities, Probability,* and *Factors and Multiples* are also selected by forward feature selection, showing a statistical difference for *Factors and Multiples* between enrollees and non-enrollees with modest positive effect size (cliff's deltas of 0.129). In general, *Geometry*, *Functions*, and *Graphing* are the topic areas with the highest cumulative feature importance among all (cumulative FI of 0.150, 0.141 and 0.137, respectively). The final model incorporated at least one feature from each topic area.

Feature importances shown in Table 9 reveal that confusion detectors (cumulative FI of 0.244) surpass BKT estimates (cumulative FI of 0.229) in importance for predicting STEM career selection. This shift in the prominence of BKT estimates, coupled with the pronounced impact seen when integrating data on affective states and disengaged behaviors, further emphasizes that BKT estimates might play a more critical role in college enrollment predictions than in STEM career choices, while the importance of affective states and disengaged behaviors rise in career selection predictions. Within the model, the topic areas represented by BKT estimates are *Functions* (FI of 0.100), *Algebra* (FI of 0.077), and *Inequalities* (FI of 0.052). As mentioned before, these areas also exhibited the largest effect sizes in the comparison between STEM and non-STEM students. Notably, in the particular case of *Algebra*, the BKT estimate overshadows all affective state or disengaged behavior detectors in importance, contrasting to the trends noted in college enrollment predictions. This underscores that while the relevance of BKT estimates might wane for career choice predictions (among college enrolled students), and the affective states gain more relevance, knowledge remains an important predictor. Depending on the content of each topic area, either the knowledge itself or the

Table 8: Comparison of mean of each feature and topic area selected by the top performing model predicting college enrollment. Statistically significant differences are shown in bold.

Feature	Topic Area	Mean(Enrolled Students)	Mean(Not Enrolled Students)	Effect Size (Cliff's Delta)	Mann-Whitney U Test (p-value)	Benjamini-Yekutieli correction (alpha)
	Functions	**0.234**	**0.278**	**-0.171**	**<0.001**	**0.001**
Gaming the System	**Decimals, Fractions, and Percentages**	**0.251**	**0.287**	**-0.163**	**<0.001**	**0.001**
	Probability	**0.258**	**0.307**	**-0.149**	**<0.001**	**0.002**
	Basic Operations	**0.219**	**0.258**	**-0.136**	**<0.001**	**0.003**
	Geometry	**0.254**	**0.290**	**-0.143**	**<0.001**	**0.003**
	Inequalities	0.287	0.267	0.094	0.013	0.007
Off Task	Basic Operations	0.312	0.309	0.033	0.291	0.010
	Probability	0.310	0.309	0.071	0.017	0.007
	Factors and Multiples	**0.264**	**0.256**	**0.129**	**0.001**	**0.005**
Engaged Concentration	Geometry	0.533	0.529	0.037	0.215	0.009
	Graphing	0.539	0.538	0.001	0.805	0.014
	Algebra	0.125	0.124	0.015	0.600	0.013
	Graphing	0.120	0.120	0.021	0.478	0.012
Confusion	Basic Operations	0.119	0.133	-0.035	0.257	0.010
	Statistics	0.125	0.151	-0.097	0.009	0.006
	Inequalities	**0.132**	**0.108**	**0.133**	**0.004**	**0.005**
	Functions	0.128	0.128	0.018	0.570	0.012
Frustration	**Algebra**	**0.141**	**0.131**	**0.100**	**<0.001**	**0.004**
	Geometry	0.140	0.138	0.043	0.141	0.008
	Exponents and square root	0.116	0.116	-0.012	0.718	0.013
Boredom	Graphing	0.436	0.436	0.029	0.354	0.011

experienced affective states and behaviors can take prevalence for the model.

Within the confusion detectors, the topic areas selected are *Probability* (FI of 0.103), *Basic Operations* (FI of 0.075) and *Exponents and Square Root* (FI of 0.066). For all these topic areas, the BKT estimates are not selected as features. While the omission of the BKT estimates might stem from collinearity between affective states and knowledge, there is not a clear and significant relation between confusion and learning gains [17, 26] to imply the degree of collinearity that would cause the model to exclude the BKT estimates. Thus, effectively navigating this affective state, which already holds relevance in enhancing college enrollment chances, is even more important for selecting a STEM major.

For engaged concentration, often linked to better learning [11, 17, 23], there's a higher potential for greater collinearity with BKT estimates, compared to other affective states, which could influence feature selection. However, results indicate that both engaged concentration and knowledge are pertinent for STEM career selection. For the *Functions* topic area, which is the most relevant for STEM career prediction (cumulative FI of 0.203) and has the highest AUC among the single-feature models (see Table 3), the model equates the importance of both knowledge and the state of engaged concentration (FI of 0.100 and 0.103, respectively). This result suggests that engaged concentration provides insights to the model beyond just promoting learning. In other words, it implies that the importance of affective states stems not only from the learning they may facilitate but also from students' perceptions of their learning journey, especially within specific topic areas, underscoring again the value of scaffolding and supporting positive affect during educational experiences.

The relevance of the frustration (cumulative FI of 0.056) and off-task detectors (cumulative FI of 0.058) diminishes in the STEM career prediction compared to the college enrollment model. On the other hand, the gaming the system detector still retains its high importance, particularly for topic areas of *Probability* and *Statistics* (FI of 0.094 and 0.078, respectively). The boredom detector persists as the least important feature (cumulative FI of 0.055), mirroring observations from the college enrollment model. This may stem from the established association between gaming and boredom [11, 17], leading to potential collinearity between these features. However, Almeda & Baker [1] also found no significant association between boredom and STEM career selection, even when boredom was considered alone.

When comparing the cumulative FIs of topic areas with the effect sizes and statistical differences between STEM and non-STEM careers, *Functions* (cumulative FI of 0.203) remains as the most relevant topic area for distinguishing these two groups. Beyond knowledge estimates, *Probability* and *Exponents and Square Root* also stand out as important topic areas in this differentiation (cumulative FI of 0.197 and 0.179, respectively) over other topic areas, such as *Inequalities* or *Algebra* that showed a statistical difference

Table 9: Topic areas selected on the best-performing model for each ML algorithm predicting STEM major enrollment using BKT estimations, behavior detections, and affect detections as features. Each cell corresponds to the Mean and standard deviation of the logistic regression coefficient for each specific feature. ML models were tested using a 4-fold school-level cross-validation.

Topic Area	BKT	Gaming	Off Task	Engaged Concentration	Confusion	Frustration	Boredom	Total
Numbers	-	-	-	-	-	-	-	-
Basic Operations	-	-	-	-	0.075 (0.007)	-	-	0.075
Algebra	0.077 (0.017)	-	-	-	-	-	-	0.077
Decimals, Fractions, and Percentages	-	-	-	-	-	-	-	-
Geometry	-	-	-	0.083 (0.006)	-	-	-	0.083
Exponents and square root	-	-	0.058 (0.004)	-	0.066 (0.013)	-	0.055 (0.006)	0.179
Factors and Multiples	-	-	-	-	-	0.056 (0.006)	-	0.056
Inequalities	0.052 (0.008)	-	-	-	-	-	-	0.052
Functions	0.100 (0.006)	-	-	0.103 (0.007)	-	-	-	0.203
Graphing	-	-	-	-	-	-	-	-
Probability	-	0.094 (0.005)	-	-	0.103 (0.011)	-	-	0.197
Statistics	-	0.078 (0.008)	-	-	-	-	-	0.078
Total	0.229	0.172	0.058	0.186	0.244	0.056	0.055	1

Table 10: Comparison of mean of each feature and topic area selected by the top performing model predicting STEM career selection.

Feature	Topic Area	Mean(STEM)	Mean(Non-STEM)	Effect Size (Cliff's Delta)	Mann-Whitney U Test (p-value)	Benjamini-Yekutieli correction (alpha)
Gaming the System	Probability	0.234	0.256	-0.058	0.317	0.007
	Statistics	0.218	0.249	-0.124	0.066	0.003
Off Task	Exponents and Square Root	0.275	0.296	-0.019	0.774	0.015
Engaged Concentration	Functions	0.534	0.531	0.046	0.432	0.010
	Geometry	0.518	0.517	-0.001	0.916	0.017
	Probability	0.136	0.138	-0.020	0.723	0.014
Confusion	Basic Operations	0.100	0.125	-0.117	0.054	0.002
	Exponents and Square Root	0.110	0.095	0.046	0.479	0.012
Frustration	Factors and Multiples	0.104	0.100	0.065	0.379	0.009
Boredom	Exponents and Square Root	0.408	0.419	-0.068	0.292	0.005

and larger effect sizes for the STEM career comparison when considering BKT estimates exclusively. This result suggests again the high importance of affective states and disengaged behaviors on each topic area beyond their interplay with gaining knowledge, particularly for predicting career selection. Interestingly, no statistically significant results emerged when performing statistical tests to determine the directions of the aforementioned associations (see Table 10). These findings strongly indicate that while these features (especially confusion) are important for predicting STEM career selection, their impact is neither uniform nor straightforward.

Lastly, three topic areas were filtered out by our top performing model for STEM career prediction: *Decimals, Fractions and Percentages*, *Numbers*, and *Graphing*. We hypothesize that these areas might be pertinent to a range of majors beyond just STEM, rendering them less essential as predictors. However, one could argue that *Basic Operations* area serves a similar broad-based function,

yet the confusion related to this area still holds notable importance for predicting STEM career selection. We previously observed that of these areas, only *Decimals, Fractions and Percentages* showed a significant difference when comparing the BKT estimates of STEM and non-STEM major enrollees, although with a relatively modest effect size. Further studies are required to discern why neither the BKT estimates, the affective states, nor the disengaged behaviors from these three areas were deemed relevant by the predictive model.

5 DISCUSSION AND CONCLUSION

In this study, we investigated how knowledge of several mathematical topic areas, along with affective states and disengaged behaviors within those topic areas, are associated with both college enrollment and career selection (STEM vs non-STEM). Although knowledge is a key predictor for both outcomes, the data reveals that the knowledge gap for STEM careers is smaller than for college enrollment. While all topic areas except *Inequalities* showed a significant difference between college enrollees and non-enrollees, only *Functions*, *Algebra*, and *Geometry* showed a significant difference for the STEM major comparison, being all of them positively associated with the participation in a STEM career. The feature importances of the models also show a higher relevance of knowledge when predicting college enrollment than for career selection. By contrast, affective states, particularly confusion in *Probability* and *Basic Operations*, and engaged concentration in *Functions* and *Geometry*, become more important for STEM career prediction.

These results suggest that the level of knowledge can be the most relevant factor in determining whether or not a student is able to enroll in a college program. However, among those students with the minimum knowledge required to enroll in college, affective states become more relevant when choosing a specific major and future career. In line with our findings, prior research has shown that experiencing confusion negatively affects interest and vocational self-efficacy in STEM [22]. Although in some cases positively resolved confusion can lead to improved learning gains [12], experiencing confusion may also result in reduced motivation to study careers related to the areas where they felt this confusion. This complex relation between confusion and educational and career outcomes was seen in the range of positive, negative, and non-significant associations found within the statistical tests, despite the clear contribution of this variable to model performance (particularly in predicting STEM career selection). By contrast, previous studies have shown that engaged concentration is positively associated with interest and vocational self-efficacy in STEM [22]. In that case, the positive effects of experiencing engaged concentration in class, an affective state associated with the concept of Flow [6], not only helps to improve learning outcomes [17] but also the motivation to study majors related to those subjects. However, the statistical tests for engaged concentration did not reveal significant differences between the groups compared, suggesting that the association between longitudinal outcomes (mainly STEM career selection) and this feature is more nuanced than just a straightforward connection with being more or less concentrated.

Among the two disengaged behaviors considered in this study, gaming the system stands out as the most important predictor for both longitudinal outcomes, showing a negative association with both, and a higher relevance for predicting STEM career choices, (also see [1, 9]). Notably, gaming the system was the only construct where statistical tests revealed significant differences between college enrollees and non-enrollees, perhaps due to the lower sample size for this construct. Specifically, for the prediction of STEM choices, gaming the system in *Probability* and *Statistics* stand out as two of the most important features of the model. This finding again demonstrates that affective states and disengaged behaviors can be more relevant for predicting career of choice than predicting college enrollment. In contrast, off-task behavior is more relevant in predicting college enrollment than STEM career, also showing a positive significant association with college enrollment, although the effect sizes are small. In neither case is it one of the most relevant features. This result is also observed for frustration, which is more relevant to predicting college enrollment than for STEM career selection. However, once again, knowledge estimates, primarily in *Functions*, *Decimal*, *Fractions, and Percentages*, and *Geometry*, are the most important features for predicting college enrollment, rather than off-task behavior or frustration.

One limitation to our analysis comes from the feature importance (FI) calculation. All explainable AI methods have limitations [29] and mean decrease in impurity FI is an appropriate method where relationships are complex and contingent and where variables are not independent. However, mean decrease in impurity FI does not allow us to draw conclusions about positive or negative associations between any of the features and the outcome. For this reason, our findings on the high importance of confusion, engaged concentration, and gaming the system in predicting STEM enrollment do not necessarily indicate that each of these affective states or disengaged behaviors individually are positively or negatively associated with the longitudinal outcome. Indeed, as discussed above, no affective state had significant differences between students who selected a STEM career and those who did not, a major contrast to the significant differences observed in the knowledge estimates of five topic areas when comparing these two populations. However, these FI measures allow us to observe that, when considering knowledge estimates, affective states, and disengaged behaviors in the same model, confusion, engaged concentration, and gaming the system in the previously mentioned areas (confusion in *Probability* and *Basic Operations,* engaged concentration in *Functions* and *Geometry,* and gaming the system in *Probability* and *Statistics*) stand out over other features as the most important predictors, primarily for STEM career choice. For this reason, these features on these topic areas, and their interactions with others, especially with knowledge, should be studied more deeply to recognize their potential interplay in career selection and to propose possible interventions to improve learning outcomes.

For example, the math problems presented to students in areas where knowledge level and negative affect are negatively associated with the desired learning outcomes might be overly difficult, leading to substantial confusion, or too easy, resulting in boredom, hindering productive confusion that often promotes deeper learning. It is also possible that some problems are presented in an overly abstract manner. In such cases, students may benefit from a stronger connection between these subjects and their practical applications in real life, to sustain their motivation. Additionally, students might

be surrounded by a context that prioritizes correct answers over the learning process itself, encouraging behaviors like gaming the system. Overall, the outcomes seen for these students are likely multiply-determined, requiring differentiated interventions.

Finally, it is worth noting that our findings pertain to a single learning system and period of time. Replicating these findings within different platforms, domains and other contexts is essential. However, doing so may be challenging, as the acquisition of a comparable dataset that spans from middle school observations to ultimate college enrollment would span years. Collecting this data set required multiple grants and school-platform data agreements that have become considerably less feasible in the United States since this project was completed [5]. For this reason, the ASSISTments Longitudinal Dataset used in this project, currently the only long-term dataset available with this kind of information, serves as a substantial reference for further analysis of potential factors that promote long-term student achievement. Still, we emphasize again the importance of facilitating the replicability of such studies. By studying and understanding how learning and engagement in different topic areas relate to long-term student achievement, we can identify areas to prioritize in instructional enhancement research. Through the development of more effective strategies to support and engage students, we can inspire students to enroll in college and cultivate a love for STEM fields.

ACKNOWLEDGMENTS

We thank the ASSISTments team for collecting and making available the longitudinal data used in this practice. We also thank the National Science Foundation, award IIS-1917545, for their support. We also thank Xiner Liu for conducting a software review to validate correctness and match to paper. Andres Felipe Zambrano thanks the Ministerio de Ciencia, Tecnología e Innovación and the Fulbright-Colombia commission for supporting his doctoral studies through the Fulbright-MinCiencias 2022 scholarship.

REFERENCES

[1] Ma Victoria Almeda, and Ryan S. Baker. 2020. Predicting Student Participation in STEM Careers: The Role of Affect and Engagement during Middle School. In Journal of Educational Data Mining 12, 2, 33-47.
[2] Robert Atanda. 1999. Do gatekeeper courses expand education options?. National Center for Education Statistics.
[3] Robert Balfanz. 2009. Putting middle grades students on the graduation path. Policy and practice brief.
[4] Robert Balfanz, Liza Herzog, and Douglas J. Mac Iver. 2007. Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions. In Educational Psychologist 42, 4, 223-235.
[5] Ryan S. Baker. 2023. The Current Trade-off Between Privacy and Equity in Educational Technology. In G. Brown III, C. Makridis (Eds.) The Economics of Equity in K-12 Education: Necessary Programming, Policy, and Systemic Changes to Improve the Economic Life Chances of American Students, 123-138. Lanham, MD: Rowman & Littlefield.
[6] Ryan S. Baker, Jaclyn L. Ocumpaugh, and J. M. A. L. Andres. 2020. BROMP quantitative field observations: A review. In Learning Science: Theory, Research, and Practice. McGraw-Hill, New York, NY.
[7] Yoav Benjamini and Daniel Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. Annals of statistics, 1165-1188.
[8] Leo Breiman. 2017. Classification and regression trees. Routledge.
[9] Mei-Shiu Chiu. 2020. Gender Differences in Predicting STEM Choice by Affective States and Behaviors in Online Mathematical Problem Solving: Positive-Affect-to-Success Hypothesis. Journal of Educational Data Mining 12, 2, 48-77.
[10] Albert T. Corbett and John R. Anderson. 1995. Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction 4, 253-278.
[11] 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, 241-250.
[12] Sidney D'Mello, Blair Lehman, Reinhard Pekrun, and Art Graesser. 2014. Confusion can be beneficial for learning. *Learning and Instruction 29*, 153-170.
[13] Jacquelynne S. Eccles and Allan Wigfield. 2002. Motivational beliefs, values, and goals. Annual review of psychology 53, 1, 109-132.
[14] Neil T. Heffernan and Cristina Lindquist Heffernan. 2014. The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. Int'l Journal of Artificial Intelligence in Education 24, 4, 470-497.
[15] Laura Horn and Anne-Marie Nuñez. 2000. Mapping the road to college first-generation students' math track, planning strategies, and context of support. Diane Publishing.
[16] Shiming Kai, Luc Paquette, Ryan S. Baker, Nigel Bosch, Sidney D'Mello, Jaclyn Ocumpaugh, Valerie Shute and Matthew Ventura. 2015. A Comparison of Video-Based and Interaction-Based Affect Detectors in Physics Playground. International Educational Data Mining Society.
[17] Shamya Karumbaiah, Ryan S. Baker, Yan Tao, and Ziyang Liu. 2022. How does Students' Affect in Virtual Learning Relate to Their Outcomes? A Systematic Review Challenging the Positive-Negative Dichotomy. In Proc. of the 12th Int'l Learning Analytics and Knowledge Conference, 24-33.
[18] Robert W. Lent, Steven D. Brown, and Gail Hackett. 1994. Toward a unifying social cognitive theory of career and academic interest, choice, and performance. In Journal of Vocational Behavior 45, 1, 79-122.
[19] Ruitao Liu, and Aixin Tan. 2020. Towards interpretable automated machine learning for STEM career prediction. Journal of Educational Data Mining 12, 2, 19-32.
[20] Jihed Makhlouf, and Tsunenori Mine. 2020. Analysis of click-stream data to predict STEM careers from student usage of an intelligent tutoring system. Journal of Educational Data Mining 12, 2, 1-18.
[21] Jaclyn Ocumpaugh, Ryan Baker, Sujith Gowda, Neil Heffernan, and Cristina Heffernan. 2014 Population validity for educational data mining models: A case study in affect detection. British Journal of Educational Technology 45, 3, 487-501.
[22] Jaclyn Ocumpaugh, Maria O. San Pedro, Huei-yi Lai, Ryan S. Baker, and Fred Borgen. 2016. Middle school engagement with mathematics software and later interest and self-efficacy for STEM careers. Journal of Science Education and Technology 25, 877-887.
[23] Zachary A. Pardos, Ryan S. J. d. Baker, Maria O. C. Z. 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. J. of Learning Analytics 1, 1, 107-128.
[24] Thanaporn Patikorn, Ryan S. Baker, and Neil T. Heffernan. 2020. ASSISTments longitudinal data mining competition special issue: a preface. Journal of Educational Data Mining 12, 2: i-xi.
[25] Maria Ofelia Z. San Pedro, Ryan S. Baker, Alex J. Bowers, and Neil T. Heffernan. 2013. Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. In Educational Data Mining 2013.
[26] Maria O. San Pedro, Jaclyn Ocumpaugh, Ryan S. Baker, and Neil T. Heffernan. 2014. Predicting STEM and Non-STEM College Major Enrollment from Middle School Interaction with Mathematics Educational Software. In EDM 2014, 276-279.
[27] Radek Pelánek. 2017. Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. User Modeling and User-Adapted Interaction, 27, 313-350.
[28] Samantha Schams, Nadya A. Fouad, Stephanie G. Burrows, Kristen Ricondo, Yixing Song. 2022. Effect Of a Class-level Intervention On Career Indecision Variables. The Career Development Quart 2, 70, 162-171.
[29] Vinitra Swamy, Bahar Radmehr, Natasa Krco, Mirko Marras, and Tanaj Käser. 2022. Evaluating the Explainers: Black-Box Explainable Machine Learning for Student Success Prediction in MOOCs. In Proceedings of the 15th International Conference on Educational Data Mining, 98-109.
[30] Cathy Van Tuijl, and Juliette H. W. van der Molen. 2016. Study choice and career development in STEM fields: An overview and integration of the research. International journal of technology and design education 26, 2, 159-183.
[31] George L. Wimberly, and Richard J. Noeth. 2005. College Readiness Begins in Middle School. ACT Policy Report. American College Testing Inc.
[32] Chun-Kit Yeung, and Dit-Yan Yeung. 2019. Incorporating features learned by an enhanced deep knowledge tracing model for stem/non-stem job prediction. International Journal of Artificial Intelligence in Education 29, 3, 317-341.