Gender differences in Predicting STEM Choice by Affective States and Behaviors in Online Mathematical Problem Solving: Positive-Affect-to-Success Hypothesis

Mei-Shiu Chiu
National Chengchi University
chium@nccu.edu.tw

This study aims to identify effective affective states and behaviors of middle-school students' online mathematics learning in predicting their choices to study science, technology, engineering, and mathematics (STEM) in higher education based on a *positive-affect-to-success hypothesis*. The dataset (591 students and 316,974 actions) was obtained from the ASSISTments project. In the ASSISTments intelligent tutoring system, students completed mathematical problem-solving tasks, and the data was processed to infer their action-level affective states and behaviors, which were averaged to form student-level measures. The students' future STEM choice was predicted by the student- and action-level affective states and behaviors using logistic regression (LR), ordinary least squares regressions with standardized scores (ORz), and random forest with permutation importance and SHAP values (RFPS). The results revealed that student- and action-level gaming behavior consistently predict STEM choice. In addition to gaming, female students are more likely to study STEM if they are less bored and more off-task, and male students if more concentrated and less frustrated. ORz generates theoretically plausible results and identifies sufficiently distinguishable affective states and behaviors. Suggestions for educational practice and research are provided for adaptive teaching.

Keywords: affect, gender differences, intelligent tutoring systems, mathematical problem solving, STEM choice

1. INTRODUCTION

Science is about not only understanding but prediction. Using students' past learning phenomena (e.g., past online learning) to predict long-term outcomes (e.g., future educational choices) is one of the greatest benefits of conducting relevant scientific research (e.g., cohort study). For traditional educational researchers, however, using middle-school students' affective states and behaviors during online mathematical problem solving to predict students' choice to study science, technology, engineering, and mathematics (STEM) in higher education is a scientific challenge. The challenges begin even before developing literature-based hypotheses and delving into data analysis. They start when identifying the subtle online process (affective states and behaviors) and linking it to the long-term effect (STEM choice) and continue throughout the research process. Yet these challenges can all be resolved.

Firstly, the issue of identifying subtle online behavior has been resolved by recent advancements in computer science for learning purposes. On the one hand, educational researchers have developed methods to examine their data that mainly focus on student-level

affect as a relatively long-term tendency in approaching mathematical learning. On the other hand, advancements in online learning have led to the development of new relevant concepts and data-science methods to analyze the new forms of data (De Witte, Haelermans, & Rogge, 2015; Kai, Almeda, Baker, Heffernan, & Heffernan, 2018). The main new form of data is action-level data; that is, students' direct or inferred behaviors recorded by information and communication technologies (e.g., intelligent tutoring systems and learning management systems) provide authentic data regarding learning processes (Tempelaar, Rienties, & Giesbers, 2015). Related research advancements in data analysis methodologies have also emerged, including educational data mining and learning analytics (Baker & Inventado, 2014). For example, students' online learning actions can be accessed, detected, and computed to form psychological constructs such as learning affect by affective computing (Baker, D'Mello, Rodrigo, & Graesser, 2010).

Secondly, justifying students' online learning processes leading to STEM choice may need support from a theoretical basis, conceptual reasoning, and educational practices. For the theoretical basis, the *positive-affect-to-success (PAS) hypothesis* assumes that "positive affect engenders success," as suggested by Lyubomirsky, King, and Diener (2005, p. 803) and vice versa. The positive affects include long-term positive affective traits (i.e., happiness) and short-term frequently experienced positive affective states (e.g., joy, interest, and pride), although happy people occasionally have negative affective states (e.g., anger, anxiety, and sadness) when receiving negative feedback about their performance. The occasional negative affective states in response to negative feedback, however, play a functional role for happy people to focus on solving current problems aiming to return to their generally long-term trait of positive affects (or happiness) and, in turn, for success, an experience like playing a challenging but solvable game (Gee, 2005a, 2005b). For conceptual reasoning, students' STEM choice can serve as a criterion or learning outcome for educational researchers to identify effective affective and behavioral factors in mathematical problem solving. It is because affective states and behaviors are interwoven with cognitive processes in mathematical learning (McLeod, 1994), and mathematics are the basis for studying in STEM (Chiu, 2007), which may link to future STEM choice (Meece, Wigfield, & Eccles, 1990; Chiu, 2017). For educational practice, linking students' online-learning action-level data to student-level data (e.g., future related educational choices) may serve as a basis for understanding students' longitudinal learning processes (Banerjee, 2016) and identify effective factors for educational intervention. When affect becomes the focus, gender differences are of concern because there is a stronger relationship between affect and both achievement and participation in advanced STEM studies for female students than for male students (Glynn, Taasoobshirazi, & Brickman, 2007; Zeldin & Pajares, 2000).

The ASSISTments project provides necessary measures for the present investigation: (1) To predict students' higher-education STEM choices using students' affective states and behaviors in online mathematical problem solving in middle school, (2) to explore whether there are gender differences in the prediction patterns, and (3) to use typical data analysis methods from both the fields of education and data science. The following literature review will first provide the theoretical basis for the affective states and behavioral measures in the ASSISTments dataset. The second section focuses on empirical studies of factors predicting STEM choice. In the final section, the literature review focuses on gender differences. The rationales for selecting suitable data analysis methods are presented in Section 2.3 on data analysis.

1.1. AFFECTIVE STATES AND BEHAVIORS IN ASSISTMENTS

The affective states and behavior measures in the ASSISTments dataset were initially developed on the basis of the *Baker Rodrigo Ocumpaugh Monitoring Protocol* (BROMP; Ocumpaugh, Baker, & Rodrigo, 2015; Shute et al., 2015). The BROMP uses a dual coding scheme, by which observers record one of the students' affective states and one of the students' affective behaviors at one particular point of time (i.e., at the action level) if any. The BROMP records were synchronized to the log-file data of how students solved problems online. As a later development, the present data on students' affective states and behaviors were inferred by automated detecting, modeling, or computing on the basis of student actions in using the ASSISTments intelligent tutoring platform, where students solved mathematical problems and obtained hints or scaffolding questions if requested (San Pedro, Baker, Gowda, & Heffernan, 2013b; Pardos, Baker, San Pedro, Gowda, & Gowda, 2014). The ASSISTments dataset contains four constructs on affective states (i.e., boredom, concentration, confusion, and frustration) and two constructs on affective behaviors (i.e., being off-task and gaming the system). The meanings of the six constructs and their relationships with related learning outcomes are depicted as follows.

Boredom. Being bored is an aversive state that draws students' attention and engagement away from participating in productive activities, is attributed to external settings, and relates to affective states such as anxiety, sadness, emptiness, and perception of slow time passing (Eastwood, Frischen, Fenske, & Smilek, 2012). Boredom relates to low student mathematical skill or knowledge (San Pedro et al., 2013b).

Concentration. (Engaged) concentration refers to paying full attention to learning tasks (e.g., furrowing one's brow while working) regardless of being on-task, off-task, or multitasking (Ocumpaugh et al., 2015). Concentration is a state of flow in psychology (Nakamura & Csikszentmihalyi, 2002) and was named flow in earlier affective computing research (D'Mello, Picard, & Graesser, 2007). Concentration in online mathematical learning during middle school positively relates to STEM vocational self-efficacy in high school, which is a predictor for future STEM career choice (Ocumpaugh, San Pedro, Lai, Baker, & Borgen, 2016). Concentration also relates to high mathematical knowledge (San Pedro et al., 2013b; Pardos et al., 2014).

Confusion. Confusion occurs when students have difficulty in understanding learning tasks of noticeable concerns, which may be observed as facial expression, verbal requests for explanations, or body language for help (Ocumpaugh et al., 2015). Confusion during online mathematics learning in middle school relates negatively to STEM vocational interest in high school (Ocumpaugh et al., 2016). Prolonged confusion in computer programming relates to low course grades (Lee, Rodrigo, Baker, Sugay, & Coronel, 2011).

Frustration. Frustration is manifested by students' expression of annoyance, sorrow, and distress, which, however, may be cognitively interpreted and expressed differently or reversely in different situations (Ocumpaugh et al., 2015). For example, challenging but solvable tasks may lead to pleasurable frustration and engage students (Gee, 2005a). Real-time frustration experience through interacting with the computer on easy tasks elicited smiles for 90% of post-graduate students in a US study (Hoque, McDuff, & Picard, 2012). Frustration relates to extremely high or low mathematical skill or knowledge (San Pedro et al., 2013b). A surprising finding is that frustration positively relates to higher online problem-solving test scores (Pardos et al., 2014).

Off-task. Being off-task refers to engaging in behaviors other than the assigned learning tasks on the intelligent tutoring system (Baker & Rossi, 2013). However, off-task behaviors may not mean boredom and may re-engage students perhaps because there are many forms of off-task behaviors, some apparently disruptive (e.g., threatening other students and sleeping) and

some not (e.g., staring into space, interacting with peers, playing with objects like pencils, and seeking teachers' attention by putting heads on desks; Ocumpaugh et al., 2015, pp. 36 and 39). Given the diverse forms of off-task behaviors, it is hard to infer whether or not off-task behavior as a whole predicts STEM choice, as evidenced by a research finding that there are unstable relationships between off-task behavior and test scores during online problem solving (Pardos et al., 2014).

Gaming. Gaming the system refers to students playing around with the system but not engaging with learning tasks (Ocumpaugh et al., 2015). Example gaming behaviors include sustained guessing (Baker et al. 2010), requesting hints, or responding too quickly depending on the degree of task difficulty, with successful problem-solvers gaming on easy tasks and unsuccessful problem-solvers gaming on difficult tasks (Baker & Rossi, 2013). Given that gaming the system has a confounding factor (i.e., task difficulty), it is hard to hypothesize gaming's direction in predicting STEM choice even if gaming itself is negative in meaning, but one study has found a negative relationship between gaming behavior and online problem-solving scores (Pardos et al., 2014).

In summary, the above literature review on ASSISTments and affective computing suggests the six constructs have the following characteristics: (A) Concentration is positive in semantic meaning, and the other five constructs are negative. (B) The constructs are assessed by criteria of traditional educational learning outcomes such as student knowledge (task difficulty or correctness), engagement, interest, and self-efficacy. (C) According to the PAS hypothesis, the positive affective state (i.e., concentration) will positively predict STEM choice and the negative affective states (i.e., boredom, confusion, and frustration) negatively predict STEM choice (Lyubomirsky et al., 2005). The two behaviors are relatively uncertain because of diverse meanings and confounding factors. (D) As stated in the PAS hypothesis (in Section 1), happy people occasionally have a negative affect when facing negative feedback about their performances; the occasional negative affect actually positively relates to long-term happiness and then success. This is evidenced by the phenomenon that a brief period of confusion and frustration positively relate to learning gains, but lengthy-period confusion and frustration negatively relate to learning gains (Liu, Pataranutaporn, Ocumpaugh, & Baker, 2013).

1.2. Online affective states and behaviors predicting STEM choice

There appear to be few empirical studies predicting students' STEM choice by students' action-level affective states and behaviors during online mathematical problem solving. The most relevant study is the research conducted by San Pedro, Ocumpaugh, Baker, and Heffernan (2014). They used independent t-test and logistic regression to identify effective factors distinguishing STEM and non-STEM college majors using student-level data on online mathematical learning from ASSISTments. The only significant and stable independent variable among the six constructs (Section 1.1) over the two algorithms (*t*-test and logistic regression) was gaming in a negative direction. (The other effective factor is student knowledge.) Another related study, using ASSISTments data and similar algorithms, indicated that students' college enrollment could be "positively" predicted by boredom and confusion, controlling for mathematics knowledge, number of first actions, and carelessness (San Pedro, Baker, Bowers, & Heffernan, 2013a), which is hard to interpret given the negative essence of boredom and confusion (cf. Section 1.1). A note to make is that the two studies actually used student-level data by averaging action-level data on each construct for each student.

Predicting students' STEM achievements and choices has long been a research interest for educational researchers, who, however, only focus on student-level factors. Qualitative research

in education has identified factors relating to students' STEM choice by interviewing STEM students and their teachers or parents. The most important factors are student affect, such as interest, curiosity, identity, and values. The next are school, family, and informal learning experiences (Cerinsek, Hribar, Glodez, & Dolinsek, 2013; Maltese & Tai, 2010). From a psycho-socio-cultural perspective, students' STEM choice relates to affective factors within different cultural contexts, such as interest or optimism with learning materials or tasks, confidence or self-efficacy with grades, resilience or control with learning strategies, value with authorities in the society, and hope or goal with educational designs (Chiu, 2017). A quantitative study using structural equation modeling finds similar results: Interest plays a major, mediating role and is influenced by peers, family, educators, and prior knowledge; interest, in turn, influences self-efficacy and career outcome expectancy and then knowledge and career orientation for STEM (Nugent, Barker, Welch, Grandgenett, Wu, & Nelson, 2015).

1.3. GENDER DIFFERENCES IN FACTORS PREDICTING STEM CHOICE

There appear to be no studies to date focusing on gender differences in the patterns of factors predicting STEM choice. Most related studies focus on gender differences in STEM achievement, which has long been viewed as the major reason for the persistent underrepresentation of females in STEM. Recent cross-cultural or meta-analysis studies, however, indicate that social-cultural factors address gender differences in STEM choices and achievements (Else-Quest, Hyde, & Linn, 2010). A salient example is that from pre-K to high school, gender differences in STEM achievement are small and subject to gender equality in a certain culture or society, with gender-equal societies having fewer gender differences in STEM achievements or mainly mathematics achievement (Guiso, Monte, Sapienza, & Zingales, 2008). The diminishing gender differences in STEM or mathematics achievements lend support to the *gender similarities hypothesis*, which contends that gender similarities tell more stories than gender differences (Hyde, 2005).

For the present study, it is interesting to extend the debate to whether there are more gender differences or similarities in problem-solving affective states and behaviors, which can serve as key precedents for achievements (Zhu, 2007). In terms of affective states, traditional educational research normally uses the term "affects" and defines "affects" as beliefs, attitudes, and emotions toward a particular school subject (e.g., mathematics), social context (e.g., learning environment), or learning task (e.g., geometric proof); detailed affective measures include self-concept (e.g., "I am able to solve a problem"), interest (e.g., "I enjoy solving problems"), and anxiety (e.g., "I feel anxious about making mistakes when solving problems"; Clifford, 1988; McLeod, 1992). Educational research indicates that boys generally have more positive attitudes, affects, or emotions toward STEM than girls do (Barkatsas, Kasimatis, & Gialamas, 2009) with only some exceptions, especially for primary school students (Yüksel-Şahin, 2008). Males' more positive affects (e.g., higher self-efficacy and lower anxiety) in turn may lead to higher mathematics achievements (Pajares & Miller, 1994) or directly lead to STEM choice controlling for achievements (Carli, Alawa, Lee, Zhao, & Kim, 2016; Organization for Economic Cooperation and Development [OECD], 2014). These studies appear to suggest that affects may play different roles in predicting STEM choices for different genders. Whether the affect and related behavior measures included in the ASSISTments dataset can serve the function suggested by educational literature is worth investigating.

In terms of affect-related or affective behaviors, the negative relationships between off-task behavior and mathematics achievement were stronger for boys than for girls, especially for low-level mathematical tasks (e.g., computation; Peterson & Fennema, 1985). Competitive

mathematics activities positively engage male low-achievers but negatively engage female low-achievers; in contrast, cooperative mathematics activities positively engage female low-achievers but negatively engage male high-achievers (Koehler, 1990). These results suggest that engaging in socially off-task activities may be irrelevant to or even supportive of learning for girls.

1.4. THE PRESENT STUDY

The above literature review suggests that, in predicting STEM choice, student- and action-level affective states and behaviors may perform differently for students as a whole and for both female and male students. Linking action-level data from online learning (i.e., affective states and behaviors during problem solving) with future student-level data (i.e., STEM choice) invites different data analysis methods from the fields of both education and data science. This methodology triangulation (i.e., multiple algorithm uses for the same phenomenon) can increase the understanding, accuracy, validity, and credibility of research results (Hussein, 2015).

This study used data from the ASSISTments project, which provided necessary measures for the present investigation (cf. Sections 2.1 and 2.2). As suggested by related literature (Section 1.1), desirable affective states (e.g., concentration) positively predict STEM choice and undesirable affective states (e.g., boredom) negatively predict STEM choice. The prediction directions are relatively uncertain for the two behaviors (i.e., off-task and gaming). As such, it was difficult to propose a hypothesis for the two affective behaviors.

Given the above condition, this study poses two research questions (RQs), with RQ1 having one embedded hypothesis, as suggested by the PAS. The affective states are boredom, concentration, confusion, and frustration, and the behaviors are being off-task and gaming the system, terms used in ASSISTments (Section 1.1).

RQ1: What student- and action-level affective states and behaviors in online mathematical problem solving predict STEM choice? [Hypothesis: Positive affective states (e.g., concentration) predict STEM choice positively, and negative affective states (e.g., boredom, confusion, and frustration) predict it negatively.]
RQ2: Are there gender differences in the prediction pattern?

2. METHOD

2.1. DATA SOURCE AND PARTICIPANTS

The data was obtained from the ASSISTments project. ASSISTments is a free online tutoring platform that provides students with mathematical problems designed by their teachers. Students solve the problems in school or as homework, and when students do not correctly solve problems, students can request hints and scaffolding questions to support their learning (Botelho, Baker, & Heffernan, 2017). While ASSISTments assists student learning, it can also assess student performance and record student actions (Heffernan & Heffernan, 2014). The dataset used in this study came from middle school students working on ASSISTments during 2004–2005 (58% of the total student actions) and 2005–2006 (42% of the total student actions).

The project also collected offline data on the students' gender and whether the students studied STEM in higher education (i.e., variable name: isSTEM). This study only used observations where isSTEM had no missing data. This data selection procedure resulted in a final dataset of 591 students and their 316,974 action records in solving the mathematical

problems. Among the 591 students, there were 247 females and 237 males, with the others as missing data. The action numbers were 132,684 for females and 131,087 for males.

2.2. MEASURES

The outcome measure was at the student level, indicating whether the students study in the STEM fields or not. Its column name was "isSTEM" in the ASSISTments dataset with a dummy-coding scale (0 = no, 1 = yes). The mean of isSTEM was 0.212, indicating that 21.2% of the participants studied STEM in higher education.

As partially indicated in Section 1.1, the affective state and behavior features in the present ASSISTments dataset were automatically detected, and the original data that trained the detectors was BROMP-based (San Pedro et al., 2013b; Pardos et al., 2014). The predictors included four affective states (i.e., boredom, concentration, confusion, and frustration) and two behaviors (i.e., off-task and gaming the system), in total six concepts, each at both the student level and the action level, which resulted in 12 (= 6 * 2) predictor measures. The action-level affective states and behaviors indicated each student's affective states and behaviors while solving a particular problem, labeled "the student affect prediction of the current response" (column names: 'RES_BORED,' 'RES_CONCENTRATING,' 'RES_CONFUSED,' 'RES_FRUSTRATED,' 'RES_OFFTASK,' 'RES_GAMING') in the ASSISTments dataset. The student-level affects and behaviors (column names: 'AveResBored,' 'AveResEngcon,' 'AveResConf,' 'AveResFrust,' 'AveResOfftask,' 'AveResGaming') were the averages of the students' action-level affects and behaviors; that is, for example, taking the average of 'RES_BORED' values for a student would be the student's 'AveResBored.' The 12 predictors used a continuous scale ranging from 0 to 1.

2.3. DATA ANALYSIS

2.3.1. Overview

The research questions (RQs) were mainly a binary classification task with the goal to predict students' STEM choice in higher education (0 = no; 1 = yes, i.e., a binomial, binary, or dichotomous dependent variable or outcome). The independent variables or predictors were the students' affective states and behaviors during solving mathematical problems in middle school, which were continuous variables with diverse patterns of data distribution. Regression or tree classification methods could be used to answer the RQs.

The data was analyzed using Python and its related packages, including pandas, numpy, seaborn, statsmodels.api, scipy.stats, sklearn, eli5, and shap. The code and results of data analysis, preparation, and exploration or binning (including descriptive statistics, data distribution plots of the measures, and related data analyses and results without being presented in this paper) were made available for public use on five Kaggle kernels (e.g., https://www.kaggle.com/meishiuchiu1/assistmentsaffecttrait-student-level-data).

2.3.2. The three algorithms

The RQs were answered by three data analysis methods: logistic regression (LR), linear or ordinary least squares regression (OR), and random forest with feature selection. All of the three algorithms were typical methods for predicting or classifying dependent variables by using multiple predictors or features. The differences between the three algorithms can be summarized as follows: LR is a typical method for predicting binominal dependent variables, which suited

the present student-level data structure; OR is a typical method for predicting continuous dependent variables, which did not suit the present student-level data structure but might suit the present action-level structure; random forest is a typical non-parametric regression tree algorithm and can identify degrees of feature importance in predicting dependent variables. The detailed rationales and procedures for using the three algorithms are addressed for each algorithm as follows.

Firstly, LR was used because LR is the most common statistical method for identifying effective predictors (using maximum likelihood estimation) to distinguish a dichotomous outcome (Allison, 2012) and tends to perform suitably for individual-level outcomes. To interpret the regression coefficients for each predictor, given the large sample sizes of this study, it was easy to obtain coefficients that were significant but actually had little importance. Effect sizes, therefore, were proper criteria to assess the importance of the predictors. In LR, odds ratios were used as the effect sizes for the predictors.

Secondly, linear or ordinary least squares regression (OR) was used, with all the measures being transformed into standardized z-scores (ORz). The rationale for using OR in this study was that the outcome was at the student level, but the predictors were initially recorded at the action level and then aggregated to the student level. By transforming all the outcome and predictor measures into z-scores, the data could be dealt with as continuous measures. The best choice might have been to analyze the dataset using multilevel modeling, but this is a complex modeling method and time-consuming in terms of data processing, especially given the large dataset at the action level and the large number of groups at the student level. ORz, therefore, served as a compromise for dealing with the present dataset.

A major concern for using OR to analyze dichotomous dependent variables is that this algorithm might violate two of the five assumptions of OR (i.e., homoscedasticity and normality; Allison, 2012). The concern, however, can be released if the datasets have large sample sizes (or even small sample sizes and skewed distribution). OR can generate robust results similar to the results obtained by LR in empirical studies, especially for testing causal hypotheses or classifying cases, though LR provides more accurate predictions than OR for student-level data (Pohlmann & Leitner, 2003).

One major merit of using OR is that interpreting OR results is more intuitive and meaningful than interpreting LR results, which can facilitate communicating research results to the general public (Hellevik, 2009). For example, the significant regression coefficient of an independent variable (predictor) in LR should be interpreted by log-odds (e.g., a logit coefficient of 0.300 refers to log-odds increase by 0.300 for every 1-unit increase in the predictor). However, it is easier to understand a regression coefficient in OR (e.g., 0.300), which can be interpreted as the probability of the outcome increases by 0.300 units for every 1-unit increase in the predictor. Using ORz (as used in the study) can further facilitate the interpretation because after transforming all the measures into z-scores, the ORz regression coefficients are standardized (i.e., betas), in which the "unit" becomes the "standard deviation" of the measures. These can be interpreted as correlation (*r*) between the predictor and the outcome variable controlling for all other predictors in the ORz model. As such, the betas could serve as the effect sizes and use the effect size metrics for correlations: *r*s = 0.100 are small effect sizes, *r*s = 0.300 are medium effect sizes, and *r*s = .500 are large effect sizes (Cohen, 1992). Further, for example, the beta of 0.300 for a predictor (e.g., concentration) could be interpreted as 9% (= 0.300*0.300) of the total variance of students' STEM choice having come from concentration controlling for all other predictors in the ORz model.

Thirdly, random forest is a typical, efficient, and accurate non-parametric regression tree algorithm to perform classification tasks especially for data with missing data or many

predictors (Strobl, Malley, & Tutz, 2009), which suited the aim of this study using six affective state and behavior predictors to predict whether students studied STEM or not (a classification task). Another reason is that random forest can present predictor importance, which can facilitate the comparison with regression coefficients (typical statistics indicating the relative importance of predictors) obtained by LR and ORz. Thus, permutation importance and SHAP values were used to identify the weights and direction for the predictors (Becker, 2019). Permutation importance (PI) was a performance metric on accuracy (how model performance decreases in prediction by randomly shuffling the cases of a predictor) for each predictor, with higher positive PI indicating higher accuracy and negative PI as a sign of small sample sizes. The variance for accuracy was calculated by the results of multiple shuffling. The summary plots of SHAP values (i.e., the impact of a predictor for a case on the model output, SHAP value = 0 as no impact, < 0 as negative impact, and > 0 as positive impact) could tell the direction of each predictor in the model. For example, if most cases with high concentration had high SHAP value, then concentration had a positive impact on isSTEM. The judgment on the directions of predictors, however, relied on visualization of the SHAP-value summary plot (e.g., Figure 1), which might be an unreliable task, especially with large sample sizes. The whole process was called the RFPS (Random Forest, Permutation-importance, and SHAP) procedure in this study, which might play similar roles to LR and ORz in identifying effective predictors or features and reduce concerns about "black box" in random forest algorithms.

This study focused on identifying the importance of predictors or features. Although the global performance of the three algorithms was not the focus, basic evaluation metrics were partially considered. OR's R-squared indicates the total variance of the outcome explained by all the predictors in the regression model. A smaller than .050 *F*-statistic *p*-value for OR indicates that at least one of the regression coefficients is not zero (Allison, 2012). LR uses pseudo R-squared as the evaluation statistics, which is explained as a pseudo R-squared with a smaller than .050 LLR (likelihood ratio chi-squared statistic) *p*-value indicating that at least one of the regression coefficients is not zero. LR's pseudo R-squared, however, is not as robust as OR's R-squared. Because both LR and random forest perform classification tasks, classification metrics were considered. Classification accuracy was an intuitive measure but might have been misleading for this study because the outcome variable ("isSTEM") did not have roughly equal numbers in the two classes (cf. Section 2.1). AUC (or area under ROC [receiver operating characteristic] curve) was a relatively robust performance metric for skewed class distribution (Fawcett, 2006), normally ranging from 0.500 (random classification) to 1.000 (completely correct classification). For this study, higher AUC indicated a better algorithm for distinguishing between students choosing STEM and not.

2.3.3. Multicollinearity in regression analysis

Regression analysis should pay attention to the problem of multicollinearity, in which the results obtained by the individual-feature model (placing only one predictor into a regression analysis) will be different from the all-feature model (placing all the predictors into a regression analysis). Therefore, the problem of multicollinearity could be resolved by comparing the directions (signs) of the regression coefficients between the all- and individual-feature models. A note to make was that a simple relationship (or correlation) between a feature and its outcome in the individual-feature model might involve many confounding factors and does not allow for identifying the relative importance among the features, which could be achieved by the all-feature model. The criterion for the sign change was that, for a particular construct, the sign of its significant regression coefficient (e.g., significantly positive) in its individual-feature model

changed to the opposite sign (e.g., significantly negative) in its all-feature model. In the case of non-significant coefficients, it is unclear whether a sign change actually represents a multicollinearity issue since the coefficients were not significant in either case.

The problem of multicollinearity, however, would not be serious, and the all-feature model would be more suitable than the individual-feature model in this study. The two claims were justified as follows.

Firstly, two measures could check for the problem of multicollinearity. Correlations between the predictors larger than 0.900 and the variance inflation factor (VIF) values of the predictors larger than 10 would suggest the existence of multicollinearity (Hair, Black, Babin, Anderson, & Tatham, 2006). The correlations between the predictors ranged from -0.724 to 0.886 on the student-level data (https://www.kaggle.com/meishiuchiu1/assistmentsaffectstate-action-level-data) and ranged from -.455 to 0.509 on the action-level data (https://www.kaggle.com/meishiuchiu1/assistmentsaffecttrait-student-level-data) for the three samples (all, female, and male students) in this study. All the correlation coefficients were smaller than 0.900. The VIF values ranged from 1.101 to 9.982 on the student-level data and ranged from 1.008 to 1.786 on the action-level data (Tables 1-3). All of the VIF values were smaller than 10. VIF can be explained using an example from this study. In the student-level data in Table 1, the VIF of boredom is 8.724, which means that the standard errors of boredom in the all-feature model would have been increased by 4.362 (= square root of 8.724) times.

Secondly, the predictors were collected based on the BROMP (cf. Section 1.1), which used a dual coding scheme (coding affective states and behavior simultaneously and separately) and assumed that affective states and behaviors should be partially orthogonal or uncorrelated (Ocumpaugh et al., 2015). The design of the BROMP justified the use of an all-feature regression model. Using an all-feature regression model could not only reflect that there was a co-occurrence of affective states and behaviors but also advance our knowledge of the relative importance of the affective states and behaviors in predicting STEM choice. For example, boys and girls might have had different patterns of the relative importance among the affective states and behaviors, which could facilitate the interpretation of the multicollinearity as an actual phenomenon, based on which to form a new theory. An example is that the internal/external frame of reference model (Chiu, 2012; Marsh & Hau, 2004) and the dimensional comparison theory (Jansen, Schroeders, Lüdtke, & Marsh, 2015) were supported by the results of multicollinearity, where there were sign changes from the models with individual predictors to the models with multiple highly correlated predictors (e.g., using mathematics and science achievements to predict mathematics and science self-concept). As such, this study would answer the research questions primarily using the results from the all-feature models. The individual-feature model was only used for discussing multicollinearity.

3. RESULTS

3.1. AFFECTIVE STATES AND BEHAVIORS PREDICT STEM CHOICE FOR ALL STUDENTS (RQ1)

3.1.1. Multicollinearity checking

On the student-level data, both LR and ORz did not reveal salient sign changes from the individual to all-feature models (Table 1). In LR, one construct's regression coefficients remained the same significant signs, and five constructs changed from being significant in the individual-feature models to non-significant in the all-feature model. As "non-significant

coefficients," they were therefore exempt from the problem of multicollinearity. ORz's all- and individual-feature models had the same regression coefficient signs: Five constructs' signs remained non-significant, and one construct's sign remained the same sign (and significant) across the all- and individual-feature models.

Table 1: Analysis results for all students.

algorithm \ predictors		boredom	concentration	confusion	frustration	off-task	gaming
		Student-level data					
	VIF	8.724	2.346	1.339	1.123	4.527	2.190
LR (individual -feature model)	log odds	-5.155	-2.020	-11.566	-9.946	-5.601	-8.138
	std err	0.399	0.155	0.948	0.809	0.466	0.902
	p > |z|	0.000	0.000	0.000	0.000	0.000	0.000
	odds ratio	0.006	0.133	0.000	0.000	0.004	0.000
LR (all-feature model)	log odds	-12.182	2.875	0.205	-2.119	2.300	-3.618
	std err	6.284	1.693	3.059	2.141	2.592	1.212
	p > |z|	0.053	0.089	0.947	0.322	0.375	0.003
	odds ratio	0.000	17.730	1.227	0.120	9.970	0.027
ORz (individual -feature model)	beta	0.000	0.045	-0.005	-0.056	0.011	-0.083
	std err	0.041	0.041	0.041	0.041	0.041	0.041
	p > |t|	0.996	0.271	0.910	0.176	0.792	0.043
	beta-squared	0.000	0.002	0.000	0.003	0.000	0.007
ORz (all-feature model)	beta	-0.165	0.039	-0.002	-0.045	0.082	-0.169
	std err	0.121	0.063	0.047	0.043	0.087	0.061
	p > |t|	0.172	0.531	0.967	0.302	0.348	0.005
	beta-squared	0.027	0.002	0.000	0.002	0.007	0.029
RFPS	accuracy	0.011	0.019	0.000	0.015	-0.015	0.000
	variation	0.007	0.013	0.024	0.016	0.013	0.017
	important	yes	yes	no	yes	no	no
	impact direction	negative?	positive?	?	negative?	?	?
		Action-level data					
	VIF	1.776	1.276	1.068	1.008	1.363	1.141
LR (individual -feature model)	log odds	-5.005	-1.969	-1.883	-1.705	-4.372	-2.210
	std err	0.019	0.007	0.017	0.014	0.026	0.015
	p > |z|	0.000	0.000	0.000	0.000	0.000	0.000
	odds ratio	0.007	0.140	0.152	0.182	0.013	0.110
LR (all-feature model)	log odds	-1.789	-1.234	-0.013	-0.113	0.047	-0.633
	std err	0.037	0.012	0.018	0.015	0.023	0.015
	p > |z|	0.000	0.000	0.480	0.000	0.046	0.000
	odds ratio	0.167	0.291	0.987	0.893	1.048	0.531
ORz (individual -feature model)	beta	0.009	0.000	0.006	-0.001	0.011	-0.048
	std err	0.002	0.002	0.002	0.002	0.002	0.002
	p > |t|	0.000	0.844	0.001	0.746	0.000	0.000
	beta-squared	0.000	0.000	0.000	0.000	0.000	0.002
ORz (all-feature model)	beta	-0.009	0.002	0.002	-0.002	0.004	-0.051
	std err	0.002	0.002	0.002	0.002	0.002	0.002
	p > |t|	0.000	0.400	0.212	0.237	0.077	0.000
	beta-squared	0.000	0.000	0.000	0.000	0.000	0.003
RFPS	accuracy	0.010	0.073	0.019	0.030	0.039	0.061
	variation	0.001	0.002	0.001	0.001	0.001	0.001
	important	yes	yes	yes	yes	yes	yes
	impact direction	?	?	?	?	?	?

Note. Green cells indicate the significant (important) results consistent with past literature and pink cells indicate the inconsistent results (base on the Hypothesis). The cells without colors indicate non-significant or uncertain results. The value "0.000" refers to "< 0.0005". LR = linear

regression; ORz = ordinary least squares regressions with standardized scores; RFPS = random forest with permutation importance and sharp values; important = result judged by permutation importance; impact direction = result visually judged by Figure 1; "?" = uncertain direction in prediction.

On the action-level data, both LR and ORz each had only one construct with a change in the signs of regression coefficients. In LR, off-task behavior changed from significantly negative in the individual-feature model to positive in the all-feature model (-4.372 to 0.047; Table 1). In ORz, boredom changed from positive (0.009) to negative (-0.009). The likely reason was that boredom had a relatively higher correlation with being off-task (0.503) than with all other predictors (-0.455 ~ -0.015), although none of the correlations were high enough (i.e., 0.900) to create a serious problem of multicollinearity. A note to make was that, for the student-level data, there were no salient sign changes even though the correlation between boredom and off-task behavior was much higher (0.868; https://www.kaggle.com/meishiuchiu1/assistmentsaffecttrait-student-level-data). This may have been because, compared with the action-level data, the student-level data had a smaller sample size, which resulted in more non-significant regression coefficients and thus exempted them from being identified as sign changes (e.g., off-task behavior with a sign change from significantly negative (-5.601) to non-significant (2.300); Table 1).

In summary, the problem of multicollinearity in terms of sign changes from the individual- to all-feature models occurred only at the constructs of boredom and off-task behavior and only on the action-level data. The results were inconsistent with the low correlation coefficients and VIF values on the action-level data, which suggested a low possibility of multicollinearity (cf. Section 2.3.3). As has been stated, the BROMP coded affective states and behavior simultaneously and separately and assumed that affective states and behaviors were partially orthogonal or uncorrelated (Ocumpaugh et al., 2015). The simultaneous co-existence of the affective states and behaviors was obvious even though there was some possibility of multicollinearity in the all-feature models. In the all-feature models, the regression coefficients should be explained as the effect of the focused predictor on the outcome controlling for all other predictors in the model, which might reveal that there were co-occurrence and interactions between human affective states and behaviors.

3.1.2. Student-level data

Controlling for all other predictors in the models, the only significant student-level predictor of isSTEM was gaming, which predicted it in a negative direction, as indicated by the results obtained by LR (log odds = -3.618) and ORz (beta = -0.169) (Table 1). The effect sizes of gaming were the odds ratio of 0.027 for LR and the beta-squared of 0.029 for ORz.

As indicated in Section 2.3.2, RFPS used accuracy measures with variation to assess the predictors' degree of importance. The impact direction of a particular predictor was visually judged by the summary plots of SHAP values (Figure 1). For example, the measure "frustration" on the student-level data for the all-student sample had the most red dots (cases/students) (red indicating high in frustration) on the left-hand side (with negative SHAP values, indicating negative impacts of frustration on isSTEM) and the most blue dots (blue indicating low in frustration) on the right-hand side (with positive SHAP values). This result indicated that high frustration (red dots) had negative impacts on STEM choice (left-hand side) for students and vice versa, meaning that frustration was negatively related to STEM choice. RFPS obtained different results from those obtained by LR and ORz; that is, the outcome variable, isSTEM,

was negatively predicted by three important features: boredom (accuracy = 0.011; variation = 0.007) and frustration (0.015; 0.016) and positively predicted by concentration (0.019; 0.013; Table 1; Figure 1), which matched the predictions of the hypothesis and literature (to be discussed in Section 4.1).

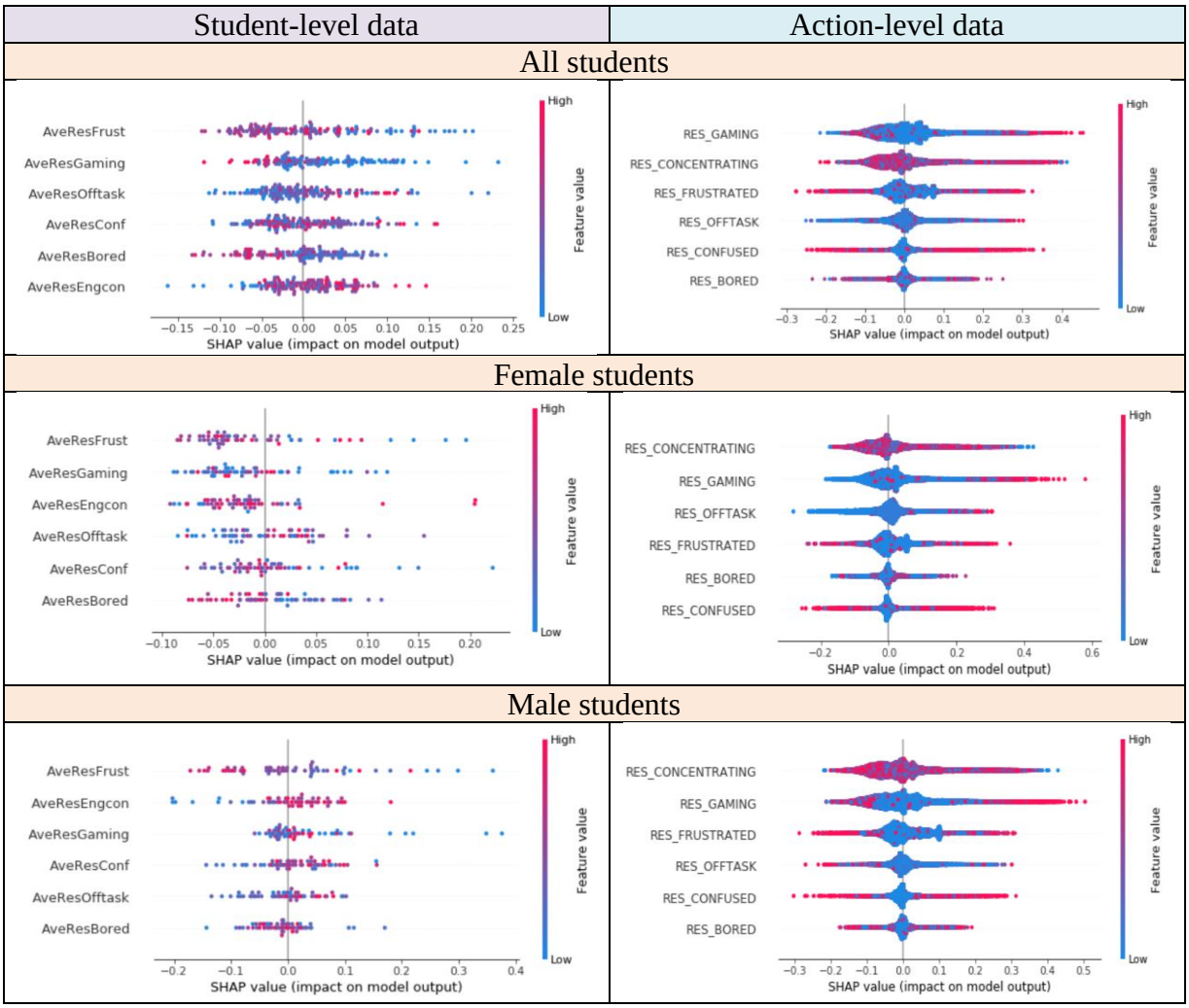

Figure 1: The summary plots of SHAP values using RFPS on student- and action-level data for different student samples. Section 2.2 presents the information about the measure names in the ASSISTments dataset (e.g., "AveResFrust" and "RES-BORED").

3.1.3. Action-level data

There were more significant predictors for the action-level data than for the student-level data, perhaps due to the large sample size for the action-level data. LR generated four significant predictors in the negative direction, boredom (log odds = -1.789; odds ratio = 0.167), concentration (-1.234; 0.291), frustration (-0.113; 0.893), and gaming (-0.633; 0.531), and one in the positive direction: off-task (0.047; 1.048; Table 1). The result that concentration

negatively predicted isSTEM was theoretically non-plausible (cf. the hypothesis). The non-plausible result might come from the unsuitable use of LR for the action-level data.

ORz generated theoretically plausible results but with only two significant predictors: Both boredom (beta = -0.009; beta-squared < 0.0005) and gaming (-0.051; 0.003) negatively predicted isSTEM (Figure 2). The results might have been plausible because using z-scores reduced the problem of multicollinearity, the original data collection design, and low correlations between the predictors (cf. Section 2.3.3).

RFPS results indicated that all six predictors were important in predicting isSTEM: boredom (accuracy = 0.010; variation = 0.001), concentration (0.073; 0.02), confusion (0.019; 0.001), frustration (0.030; 0.001), off-task (0.039; 0.001) and gaming (0.061; 0.001; Table 1). However, the impact direction of the predictions was difficult to visually recognize (Figure 1). For example, the gaming behavior of the all-student sample revealed that a few red dots (cases being high in gaming behavior) were on the right-hand (positive SHAP values, indicating positive impacts of gaming on isSTEM), many blue dots (cases being low in gaming behavior) in the middle (near zero SHAP values), and some blue and red dots on the left-hand side (negative SHAP values). Even though the accuracy measure indicated that gaming was an important variable in predicting isSTEM, it was hard to judge whether gaming was positively or negatively related to isSTEM. The reasons might have been the large sample size and large variations in the SHAP values on the action-level data.

3.2. GENDER DIFFERENCES (RQ2)

3.2.1. Multicollinearity checking

The female-student data had the same patterns of sign changes in the regression coefficient as the all-student data; that is, the problem of multicollinearity occurred only on the action-level data at the constructs of boredom and off-task behavior. In LR, off-task behavior changed from significantly negative in the individual-feature model to positive in the all-feature model (-5.293 to 0.119; Table 2). In ORz, boredom changed from positive (0.008) to negative (-0.016). The changes were larger for the female-student data than for the all-student data.

For male students, no sign changes occurred for either the student- or action-level data (Table 3). Combining all the results for the samples of all, female, and male students suggested that the major sign changes occurred for the female-student data. The results also suggested that different patterns of how online-learning affective states and behaviors predicted STEM choice between female and male students would be found if analyzing female and male data separately, as done in this study.

3.2.2. Student-level data

For female students, isSTEM was negatively predicted by gaming using both LR (log odds = -4.320; odds ratio = 0.013) and ORz (beta = -0.198; beta-squared = 0.039) all-feature models (Table 2). The results for females replicated the results obtained by LR and ORz for all the students (Table 1). RFPS obtained only one important predictor: off-task behavior (accuracy = 0.016; variation = 0.050) positively predicted isSTEM. The seemingly positive prediction direction was as indicated in Figure 1: The ‘AveResOfftask’ for the female student-level data had mostly red dots (cases being high in off-task behavior) on the right-hand side (showing positive impacts on isSTEM) and mostly blue dots (cases being low in off-task behavior) on the left-hand side (showing negative impacts on isSTEM).

For male students, LR and ORz failed to identify any significant predictors (Table 3). However, RFPS obtained four important predictors: isSTEM was positively predicted by concentration (accuracy = 0.040; variation = 0.040) and negatively by frustration (0.017; 0.056), which supported the predictions. The outcome isSTEM was also predicted by off-task (0.020; 0.039) with an uncertain direction and negatively by gaming (0.020; 0.039; Table 3; Figure 1). These two behaviors were relatively new in research on online learning, and no prediction direction was hypothesized (cf. Section 1.5).

To summarize, the results of LR and ORz revealed that female and male students had the same prediction patterns, except for the negative effect of gaming for only female students. The RFPS results revealed more gender differences than LR and ORz results (Tables 2–3).

3.2.3. Action-level data

Both LR and RFPS obtained theoretically non-plausible and visually unidentifiable results using the action-level data for both female and male students (Tables 2–3). Similar results were obtained using the action-level data for all students by LR and RFPS (Table 1). As such, only the results obtained by ORz were interpreted in this section.

For female students, the results using ORz revealed that both boredom (beta = -0.016; beta-squared < 0.0005) and gaming (-0.057; 0.003) negatively predicted isSTEM, and being off-task (0.011; < 0.0005) positively predicted isSTEM. A note to make is that the effect sizes (i.e., beta-squared) were very small even though the regression coefficient (i.e., betas) were significant.

Male students had quite a different prediction pattern from female students. The only exception was the negative predictive capacity of gaming, which was the same for both genders (males' gaming: beta = -0.038; beta-squared = 0.001). Additionally, male students' STEM choice (isSTEM) could be positively predicted by concentration (0.009; < 0.0005) and negatively by frustration (-0.009; < 0.0005), which were plausible results based on the literature.

3.3. DIFFERENT ALGORITHMS

3.3.1. Student-level data

For the student-level data, LR and ORz obtained the same results in the predictive directions of the regression coefficients, and RFPS obtained different "important predictors" over different student samples (Tables 1–3). The predictive direction of the important predictors identified by RFPS could be partially identified by the summary plots of SHAP values (Figure 1) but still could not be completely certain.

Overall performance measures were used to examine the three algorithms over the three student samples (Table 4). LR performed better than ORz for the all-student sample: the LR model was significant (LLR p = 0.043), and the ORz model was not (p (F-statistic) = 0.077). In addition, the LR model had a larger effect size (pseudo R-squared = 0.019) than the ORz model (adj. R-squared = 0.009). However, for the female and males student samples, LR was not better than ORz because both LR and ORz models were not significant despite LR having larger effect sizes (female: 0.028; male: 0.023) than ORz (0.004; 0.003).

LR also performed slightly better than RFPS because LR's AUCs for the all-student and female-student samples (0.641 and 0.535) were higher than RFPS's AUCs (0.540 and 0.492). However, for male students, RFPS (0.641) performed better than LR (0.571).

Table 2: Analysis results for female students.

algorithm \ predictors		boredom	concentration	confusion	frustration	off-task	gaming
		Student-level data					
	VIF	8.612	2.440	1.325	1.127	5.551	1.977
LR (individual -feature model)	log odds	-5.716	-2.223	-13.187	-10.292	-6.381	-10.177
	std err	0.644	0.250	1.546	1.257	0.767	1.752
	p > |z|	0.000	0.000	0.000	0.000	0.000	0.000
	odds ratio	0.003	0.108	0.000	0.000	0.002	0.000
LR (all-feature model)	log odds	-19.914	4.210	-3.132	2.504	5.725	-4.320
	std err	11.678	2.979	5.428	3.338	4.936	2.071
	p > |z|	0.088	0.158	0.564	0.453	0.246	0.037
	odds ratio	0.000	67.367	0.044	12.226	306.493	0.013
ORz (individual -feature model)	beta	-0.037	0.033	-0.060	0.032	-0.005	-0.070
	std err	0.064	0.064	0.064	0.064	0.064	0.064
	p > |t|	0.559	0.603	0.349	0.616	0.942	0.274
	beta-squared	0.001	0.001	0.004	0.001	0.000	0.005
ORz (all-feature model)	beta	-0.312	-0.002	-0.041	0.032	0.184	-0.198
	std err	0.186	0.099	0.073	0.067	0.150	0.089
	p > |t|	0.095	0.982	0.580	0.637	0.221	0.028
	beta-squared	0.097	0.000	0.002	0.001	0.034	0.039
RFPS	accuracy	-0.007	-0.007	-0.016	-0.029	0.016	-0.032
	variation	0.016	0.016	0.029	0.024	0.050	0.020
	important	no	no	no	no	yes	no
	impact direction	?	?	?	?	positive?	?
		Action-level data					
	VIF	1.766	1.288	1.068	1.007	1.361	1.129
LR (individual -feature model)	log odds	-5.536	-2.179	-2.150	-1.833	-5.293	-2.654
	std err	0.031	0.011	0.028	0.022	0.046	0.028
	p > |z|	0.000	0.000	0.000	0.000	0.000	0.000
	odds ratio	0.004	0.113	0.117	0.160	0.005	0.070
LR (all-feature model)	log odds	-1.975	-1.385	-0.070	-0.069	0.119	-0.785
	std err	0.058	0.019	0.029	0.023	0.038	0.028
	p > |z|	0.000	0.000	0.017	0.003	0.002	0.000
	odds ratio	0.139	0.250	0.932	0.934	1.126	0.456
ORz (individual -feature model)	beta	0.008	-0.003	0.000	0.005	0.017	-0.055
	std err	0.003	0.003	0.003	0.003	0.003	0.003
	p > |t|	0.002	0.246	0.879	0.071	0.000	0.000
	beta-squared	0.000	0.000	0.000	0.000	0.000	0.003
ORz (all-feature model)	beta	-0.016	-0.004	-0.003	0.004	0.011	-0.057
	std err	0.004	0.003	0.003	0.003	0.003	0.003
	p > |t|	0.000	0.256	0.301	0.201	0.001	0.000
	beta-squared	0.000	0.000	0.000	0.000	0.000	0.003
RFPS	accuracy	0.063	0.056	0.034	0.021	0.016	0.006
	variation	0.002	0.001	0.001	0.001	0.002	0.001
	important	yes	yes	yes	yes	yes	yes
	impact direction	?	?	?	?	?	?

Note. The notes are the same as those in Table 1.

Table 3: Analysis results for male students.

algorithm \ predictors		boredom	concentration	confusion	frustration	off-task	gaming
			Student-level data				
	VIF	9.982	2.711	1.457	1.101	4.511	2.289
LR	log odds	-4.534	-1.760	-9.734	-9.314	-4.854	-6.117
(individual	std err	0.609	0.234	1.420	1.257	0.704	1.078
-feature	p > |z|	0.000	0.000	0.000	0.000	0.000	0.000
model)	odds ratio	0.011	0.172	0.000	0.000	0.008	0.002
LR	log odds	-6.031	1.944	1.744	-6.021	-0.228	-2.691
(all-feature	std err	8.722	2.397	4.188	3.571	3.811	1.669
model)	p > |z|	0.489	0.417	0.677	0.092	0.952	0.107
	odds ratio	0.002	6.986	5.718	0.002	0.796	0.068
ORz	beta	0.027	0.051	0.020	-0.116	0.023	-0.087
(individual	std err	0.065	0.065	0.065	0.065	0.065	0.065
-feature	p > |t|	0.678	0.436	0.764	0.074	0.721	0.184
model)	beta-squared	0.001	0.003	0.000	0.013	0.001	0.007
ORz	beta	0.044	0.113	0.016	-0.105	-0.033	-0.110
(all-feature	std err	0.205	0.107	0.078	0.068	0.138	0.098
model)	p > |t|	0.830	0.289	0.842	0.124	0.810	0.262
	beta-squared	0.002	0.013	0.000	0.011	0.001	0.012
RFPS	accuracy	-0.003	0.040	-0.033	0.017	0.020	0.020
	variation	0.013	0.040	0.030	0.056	0.039	0.039
	important	no	yes	no	yes	yes	yes
	impact direction	?	positive?	?	negative?	?	negative?
			Action-level data				
	VIF	1.786	1.266	1.071	1.008	1.372	1.150
LR	log odds	-4.388	-1.707	-1.617	-1.550	-3.510	-1.767
(individual	std err	0.028	0.010	0.025	0.021	0.035	0.019
-feature	p > |z|	0.000	0.000	0.000	0.000	0.000	0.000
model)	odds ratio	0.012	0.182	0.199	0.212	0.030	0.171
LR	log odds	-1.578	-1.060	0.005	-0.166	0.000	-0.460
(all-feature	std err	0.056	0.018	0.027	0.022	0.035	0.021
model)	p > |z|	0.000	0.000	0.855	0.000	0.677	0.000
	odds ratio	0.206	0.347	1.005	0.847	1.015	0.632
ORz	beta	0.008	0.006	0.007	-0.008	0.007	-0.037
(individual	std err	0.003	0.003	0.003	0.003	0.003	0.003
-feature	p > |t|	0.003	0.047	0.017	0.003	0.016	0.000
model)	beta-squared	0.000	0.000	0.000	0.000	0.000	0.001
ORz	beta	-0.002	0.009	0.004	-0.009	0.000	-0.038
(all-feature	std err	0.004	0.003	0.003	0.003	0.003	0.003
model)	p > |t|	0.604	0.006	0.199	0.001	0.999	0.000
	beta-squared	0.000	0.000	0.000	0.000	0.000	0.001
RFPS	accuracy	0.0124	0.0924	0.0214	0.0377	0.0589	0.0805
	variation	0.0014	0.002	0.0014	0.0015	0.0046	0.0009
	important	yes	yes	yes	yes	yes	yes
	impact direction	?	?	?	?	?	?

Note. The notes are the same as those in Table 1.

Table 4: Overall performance of the three algorithms for different datasets.

		All students		Female students		Male students	
Algorithm	Data level	student	action	student	action	student	action
LR	LL	-299.220	-161190.000	-116.840	-63440.000	-127.750	-71538.000
	LL-null	304.920	-160430.000	-120.200	-63183.000	-130.740	-71166.000
	LLR p	0.043	1.000	0.242	1.000	0.307	1.000
	pseudo R^2	0.019	-0.005	0.028	-0.004	0.023	-0.005
	AUC	0.641	0.531	0.535	0.535	0.571	0.531
ORz	F-statistic	1.912	127.900	1.154	70.680	1.117	33.850
	p (F-statistic)	0.077	0.000	0.332	0.000	0.353	0.000
	adj. R^2	0.009	0.002	0.004	0.003	0.003	0.002
RFPS	AUC	0.540	0.476	0.492	0.492	0.641	0.475

Note. The orange cells indicate statistically significant results at $p < 0.050$. The value "0.000" refers to "< 0.0005". LL = log likelihood; LLR = log-likelihood ratio.

3.3.2. Action-level data

In terms of regression coefficients, ORz tended to generate theoretically plausible results and could sensitively detect effective predictors for different student samples (Tables 1–3). Both LR and RFPS identified many significant predictors. However, LR generated non-plausible results, and RFPS found uncertain ones.

In terms of overall algorithm performance, ORz performed better than LR because the LR models for all the three student samples were not significant (all LLR ps = 1.000) and the effect sizes (pseudo R-squared) became negative (-0.005; -0.004; -0.005), which showed that the LR models did not fit the empirical data. On the other hand, the ORz models were significant (all *p*s (F-statistic) < 0.0005), and their effect sizes were positive though small (0.002; 0.003; 0.002). LR performed better than RFPS because LR models had higher AUCs (0.531; 0.535; 0.531) than did the RFPS models (0.476; 0.492; 0.475).

4. Discussion

4.1. Methodological Issues

4.1.1. Two Approaches to the Concern of Multicollinearity

Regression-related algorithms using multiple predictors (or features) need to examine multicollinearity. A salient indication of multicollinearity is sign changes of regression coefficient estimates from individual-feature to all-feature models, which normally occur when regression models include multiple highly correlated predictors. This study uses six predictors of similar constructs (i.e., affective states and behaviors), which inevitably increases the necessity to address the concern of multicollinearity. This study handles multicollinearity using two approaches: the methodological approach and the theoretical or conceptual approach.

The methodological approach. This approach sees including multiple highly correlated predictors into a regression analysis as an undesirable procedure, which will generate distorted and thus non-trustworthy regression coefficient estimates. Based on this approach, the aim is to reduce the problem of multicollinearity or provide evidence showing few problems of multicollinearity. For the former, to reduce the problem of multicollinearity, researchers can use statistical measures such as penalizing highly correlated predictors in ridge and Lasso regressions (Bowles, 2015), combining similar predictors to one factor by factor analysis or using z-scores (Aiken & West, 1991; as ORz used in this study). For the latter, researchers can provide evidence showing how severe is the problem of multicollinearity. For example, this study presents and compares the results obtained by individual-feature models and all-feature models. Sign changes do occur from individual to all-feature models, which indicates multicollinearity. However, the low correlations among predictors and the low VIF for each predictor (Hair et al., 2006) reflect a low degree of the problem of multicollinearity.

The theoretical approach. In interpreting the results of regression analysis, the regression coefficient estimates for a particular predictor is the pure relationship between the outcome and the predictor controlling for, partial out, or relative to the other predictors in the regression model. This means that the results obtained by individual-feature models may be disguised predictive effects without including essential control variables in a regression model or without considering its relativity to other related, essential factors in the world. From this approach, judging the plausibility of the results obtained by individual-feature and all-feature models should depend on multiple criteria, as used in this study: the low (or below-criterion) correlations among predictors, the low (or below-criterion) VIF for each predictor, and the predictions based on a pre-determined theoretical framework (i.e., the PAS). The three criteria suggest that the results obtained from all-feature models tend to be more plausible than those from individual-feature models.

4.1.2. ORz as the best analysis method

This study used three algorithms or data analysis methods (LR, ORz, and RFPS) to identify effective predictors for STEM choice. As indicated in Section 2.3.2, the three algorithms were appropriate for the present data and aims of this study because they are typical methods for predicting or classifying dependent variables. LR suited the present student-level data structure and the aim to predict whether go to STEM or not; ORz might suit the present action-level structure and the aim to predict STEM choice; random forest is a typical non-parametric regression tree algorithm with feature selection functions, which can assess the degrees of the importance of each feature in determining STEM choice. This triangulation among the three algorithms (Hussein, 2015) may help find suitable algorithms for educationally meaningful findings on the present novel datasets at both student and action levels from an intelligent tutoring system, ASSISTments.

Combining the results of Sections 3.1–3 about LR and ORz, ORz tended to be a conservative but valid analysis method, which generated theoretically plausible predictions and significant overall model performance. LR identified non-plausible predictors that were contrary to the literature, and LR's overall model performance showed a bad fit to the data. This study was actually a binary classification task. LR did perform best on the student-level data but became worse on the action-level data. ORz violated some of its assumptions for this task (Allison, 2012) but performed excellently on the action-level data and generated the same regression coefficient patterns as LR on the student-level data. This finding appears to be in accordance with empirical research perspectives that ORz is a suitable choice for most predictive tasks, even if its

assumptions were violated (Pohlmann & Leitner, 2003). Future research needs to validate these findings using different datasets.

RFPS over-identified important predictors in indicating that all predictors were important on the action-level data. Most signs of the prediction were not visually identifiable using the summary plots of SHAP values in RFPS. Given the unreliability of using the SHAP-value summary plot to judge the directions of prediction, RFPS needs to develop further measures for certain predictive directions in order to fully solve the issue of the black box in the random forest algorithm. Another concern is that random forest is a proper algorithm for a dataset with a large number of features, but fewer cases than features (Strobl, Malley, & Tutz, 2009) and can be exempt from the problem of multicollinearity even with many features. Perhaps the task of this study contains only six features and many cases (especially on the action-level data) and is not suitable for using random forest. The ASSISTments dataset contains many more variables than those used in this study. Random forest may be more suitable for a study using all the variables in the ASSISTments dataset, which is an issue that could be addressed by future research.

To summarize, ORz generally performed better than LR and RFPS did in terms of plausible regression coefficients and certain prediction directions on both student- and action-level data over the three student samples. Given the merits of ORz, the following discussion only focuses on the results obtained by ORz. However, the generally small effect sizes of the regression coefficients and the overall model performances in the models of this study (Tables 1–4) suggest taking a conservative approach to interpreting the results.

4.2. EFFECTS OF THE PREDICTORS (RQ1)

4.2.1. Gaming as the top stable, negative predictor

Gaming the system is the most stable predictor of STEM choice in the negative direction at both student and action levels among the six affective states and behavioral constructs investigated in this study, a result consistent with San Pedro et al.'s (2014) and Pardos et al.'s (2014) studies using student-level data. Gaming the system is a behavior during online learning that includes continuously or quickly guessing solutions, requesting hints, or exploiting the functions in the system irrelevant to learning (Baker et al. 2010; Ocumpaugh et al., 2015). The result may also suggest that when considering affective states and behavior together (placing them all in one regression model), affective behavior (instead of affective states) will capture all the predictive capacity in predicting STEM choice, a result supporting the PAS at affective behavior.

Gaming (the system) is new for traditional educational research on mathematical problem solving. Linking gaming to educational research on similar issues may further elaborate on the term. During mathematical problem solving, students need to experience the process of contemplating (Mason, Burton, & Stacey, 1996). The present use of the term 'gaming the system' may be a proxy for the concepts of 'lacking contemplation,' 'hyperactivity,' 'impulsiveness,' 'lack of discipline,' or 'lack of self-regulation' in education and psychology research. Self-regulation or executive functioning is a higher-order cognitive process for inhibitory control, planning, and flexible goal-directed behaviors (Bernier, Carlson, & Whipple, 2010). Future research may need to validate in greater depth whether or not gaming the system is "lack of self-regulation" behavior in online learning.

Another reason for the negative role of gaming in predicting STEM choice may be that the tasks in the present ASSISTments dataset are difficult enough and invite gaming behaviors for low mathematics achievers, which in turn can negatively predict STEM choice. As indicated in a related study, unsuccessful problem-solvers are likely to game the system on difficult tasks

(Baker & Rossi, 2013). Future research needs to control for task difficulty in investigating related topics.

4.2.2. Boredom as the second negative predictor only at the action level

Boredom is the second predictor of STEM choice in the negative direction but only at the action level, not at the student level, which partially fits the PAS hypothesis. The results suggest that boredom may be more of an action-level affective state than a student-level one. As defined in ASSISTments and evidenced in related studies, boredom is an aversive state that disengages students and relates to low knowledge, skills (Eastwood et al., 2012; San Pedro et al., 2013b), and poor learning (Baker et al., 2010).

According to the educational literature on affective states during mathematical problem solving, boredom may occur at the start of the process of mathematical problem solving and correlate with interest and task attraction (Mason et al., 1996). Students' STEM choice is largely determined by their affect toward STEM including interest and curiosity, identity, and values, and next by teaching activities and context such as parental encouragement and pressure, teachers' pedagogies, and inside- and outside-school learning experiences (Cerinsek et al., 2013; Maltese & Tai, 2010). Diverse novel, interesting problem-solving designs (e.g., games) need to be incorporated into online mathematical problem-solving platforms in order to reduce students' boredom.

4.2.3. Other predictors

The other three affective states in ASSISTments (i.e., concentration, confusion, and frustration) do not significantly predict STEM choice. The results do not fully support the major PAS hypothesis, which assumes that concentration should positively affect STEM choice, and confusion and frustration should negatively predict STEM choice. Past studies using the ASSISTments data suggest that STEM choice is (A) positively predicted by concentration given its capacity to predict mathematical knowledge (San Pedro et al., 2013b) and STEM vocational self-efficacy (Ocumpaugh et al., 2016), (B) negatively predicted by confusion given its capacity to predict STEM vocational interest (Ocumpaugh et al., 2016) and course grades (Lee et al., 2011), and (C) uncertain in its capacity to be predicted by confusion and frustration. This is because frustration has an uncertain relationship with knowledge or task difficulty (San Pedro et al., 2013b), and short-period confusion and frustration relate positively to learning gains, but lengthy-period confusion and frustration negatively relate to learning gains (Liu et al., 2013). The uncertain relationship between confusion, frustration, and outcomes is suggested by the minor PAS hypothesis that occasional negative affect in response to negative feedback can partially explain long-term success (Lyubomirsky et al. 2005). In this sense, the present non-significant results are reasonable because this study does not consider in-depth short- and long-period affects. Future research needs to take into account the time factor in student experiences of the three affective states (i.e., concentration, confusion, and frustration).

Off-task behavior also fails to predict STEM choice. Being off-task is negative in its meaning, and it is hard to hypothesize its role in predicting STEM choice due to the diversity of off-task behaviors (e.g., staring into space, interacting with peers, and playing with objects; Ocumpaugh et al., 2015).

The non-significant results for the three affective states, however, are reasonable because the significance has been captured by the most relatively important features or predictors (i.e., gaming and boredom) in the all-feature regression model. In terms of educational literature, the result appears to be reasonable if we consider STEM choice as a complex decision determined

by diverse student personal, social, and cultural factors (Chiu, 2017). In addition to personal cognitive, affective, and behavioral aspects of STEM learning and problem solving, students' pursuit of advanced STEM studies and careers may be determined by sociocultural factors. For example, students are likely to choose STEM if they perceive STEM as special, beneficial, practical, influential, and conducive to future career development, meeting their ideal job-related reputation or expectations (Gazley et al., 2014; Hsu, Roth, Marshall, & Guenette, 2009). Combining diverse online and off-line personal and sociocultural data may provide a clearer picture of effective predictors for STEM choice.

4.3. GENDER DIFFERENCES (RQ2)

4.3.1. Few gender differences at the student level

For student-level constructs, female students' gaming negatively predicts their STEM choice, but there is no effective predictor for male students. Educational research indicates that girls use more self-regulated learning skills (e.g., record keeping, monitoring, goal-setting, planning, and environmental structuring) than boys (Zimmerman & Martinez-Pons, 1990). In addition to the tendency to use self-regulated learning strategies, compared with boys, girls are more reluctant to compete with others, to take action because of extrinsic motivation, and to respond to the environment strongly (OECD, 2015). Because girls are less likely to respond strongly to the environment and more self-regulated, girls' gaming the system may be more a sign of being less likely to choose STEM than it is for boys.

This speculation, however, needs to be examined by future research. It may be particularly important to better understand the relationship between online gaming behavior and self-regulation. As suggested, "gaming the system" behavior may be correlated with self-regulation in a negative direction, in which individuals attempt to succeed in problem solving without focusing on learning the intended curricula but on irrelevant tasks such as intentionally rapid guessing, making mistakes, and requesting hints (Baker, Corbett et al., 2013). Based on these understandings, researchers could investigate gender differences in how gaming behavior predicts their STEM choices.

4.3.2. Many gender differences at the action level

For action-level constructs, gaming is the only common significant predictor of STEM choice for both female and male students. The main gender differences are that boredom negatively predicts females' STEM choice, and being off-task positively predicts females' STEM choice; by contrast, frustration negatively predicts males' STEM choice, and concentration positively predicts males' STEM choice. The results imply that there are more gender differences at the action level than gender similarities (Hyde, 2005). Females and males may have different patterns of affective states and behaviors in approaching mathematics problem solving, which may, in turn, play a role in their future STEM choice. The results are consistent with the stable research findings that there are gender differences in STEM-related affects (e.g., Carli et al., 2016; OECD, 2014).

The findings of this study may be used to provide insights into appropriate designs for both genders. For example, females may need to feel interested (not bored) in solving mathematical problems. Females may also need to take time off from online learning tasks when they need to ponder or handle other (e.g., social) matters not directly related to learning tasks. Females generally have more interest in social communication and others' feelings in solving game-based mathematical problems (Ke, 2008), which implies that off-task social behavior may

dominate females' learning with little harm or even positive support to learning. A related study indicates that the negative relationships between off-task behavior and mathematics achievement are stronger for boys than for girls (Peterson, & Fennema, 1985), which also partially suggest that being off-task appears to be more harmful to boys than girls. If the speculation is a wise guess, then female students especially need a mathematics learning platform that provides some interesting elements and allows for seemingly irrelevant social behaviors during online mathematical problem solving.

For males, educators may need to manage degrees of task difficulty (Gee, 2005a; San Pedro et al., 2013b) and notice male students' sensitivity to failure in mathematical problem solving, which may be a major source of frustration. The interaction between task difficulty and frustration and the way to manage this interaction are complex problems, which appear to be an issue for boys. Concentration is a positive predictor for males. The results are consistent with past research findings that concentration relates to STEM vocational self-efficacy (Ocumpaugh, San Pedro, Lai, Baker, & Borgen, 2016) and mathematical knowledge (San Pedro et al., 2013b). This also leads to an interesting comparison with females' being off-task as a positive predictor. Females pay more attention to social affairs and males to tasks in game-based mathematical problem solving (Ke, 2008), which invites future research to investigate this likely gender difference further.

4.4. CONTRIBUTIONS, LIMITATIONS, AND SUGGESTIONS FOR FUTURE RESEARCH

4.4.1. Contributions

ASSISTments provides valuable big data on student action-level data, which is rarely researched in traditional education. This study offers a pioneering approach to such research and contributes to two aspects in particular.

Firstly, it uses both student-level and action-level data in mathematical problem solving to predict future STEM choice.

Secondly, gender differences are investigated. For educational practice, the differential teaching for addressing gender differences in affective states and behaviors during online mathematical problem solving, as suggested by the present findings, may be a key to encouraging both genders to pursue STEM advanced studies and careers, especially for female students, who are persistently underrepresented in STEM (Else-Quest et al., 2010; Koller, Baumert, & Schnabel, 2001).

4.4.2. Limitations and future research

Despite the novel dataset, new topics, and diverse data analysis methods used in this study, this study has the following limitations for future researchers to consider.

STEM choice relates to high STEM or mathematics ability (Nugent et al., 2015; San Pedro et al., 2014). It, therefore, can infer that affective states or behaviors linking to higher STEM achievement may link to STEM choices such as a low degree of boredom (Tze, Daniels, & Klassen, 2016) and a high degree of motivation (including confidence, interest, value, control, and goal; Pintrich, 2003). Future research may need to include student knowledge in the proposed model.

Research has indicated that the relationships between affective states or behaviors and problem-solving scores may be moderated by problem types. For example, there is a negative relationship between boredom and online problem-solving scores on original problems but a

positive relationship on scaffolding problems (Pardos et al., 2014). This study does not include problem types as a moderator, which can be addressed by future research.

Student selection bias may be an issue. The data is collected through an online tutoring platform (i.e., ASSISTments). If the platform is used as homework, then family computer availability and skills may influence student performance. If the platform is used at school, a control-experimental design may be the best choice to draw a cause-and-effect relationship, which may partially resolve the problem of selection bias.

The problem of multicollinearity should be further addressed. One solution would be to use factor analysis to reduce measure numbers (i.e., combining correlated measures into factors). Another solution would be to use regression algorithms penalizing highly correlated predictors (e.g., Ridge and Lasso linear regressions; Bowles, 2015). Given that there were only six predictors in the regression models, however, there appeared to be no need to perform other complicated linear regressions that are not typical in educational research. Further, multilevel analysis may be needed for the present data structure. However, the disadvantages of using multilevel analysis are its model complexity and time-consuming computation. ORz appears to be an effective method but needs to be examined further for its validity in handling the present type of dataset, which had multiple levels of data, a dichotomous outcome, and a large number of observations.

Gender difference is a complex issue and may be an outcome from interactions between multiple biological, psychological, and social factors (Halpern, Wai, & Saw, 2005). There are still debates between gender differences and similarities (Hyde, 2005). Any research results or claims relating to gender differences should be explained and used with caution. Gender differences in affective states and behaviors during mathematical problem solving may vary by culture (Ho et al., 2000), by age, in different time periods, for different problem types, and on online and offline platforms. The results obtained in this study need to be examined further with data from other cultures, cohorts, and platforms and on different problem-solving tasks.

This study uses diverse measures to assess the overall performances of the three algorithms (LR, ORz, and RFPS; Section 3.3). However, there is a lack of systematic literature review, research design, data analysis, and discussion to generate robust findings across the three algorithms that can be applied to guiding future research into selecting proper algorithms for these particular types of data. This topic can be addressed by future research on data analysis algorithms.

Some findings of this study may not be robust. For example, the predictive directions of features obtained by the random forest plus related feature selection algorithm are not reliably or identifiable by visualization. The effect sizes of the features' effects are very small even for significant effects, which may be due to the large sample sizes in this study. This is especially true for the results obtained by using the action-level data. The low effect size, however, is reasonable because it is challenging to predict a student's STEM choice using that student's affective states and behaviors during one mathematical problem. Nonetheless, these findings may provide insight for understanding the relationships between students' online learning behavior, adaptive teaching, and career development.

4.5. CONCLUSION

Understanding which factors contribute to STEM choice and how educational designs can promote this choice remains a challenge. Based on the triangulation between educational literature, three student samples, and three data analysis methods, this study uses data on

students' affective states and behaviors from an online mathematics learning platform (i.e., ASSISTments) to predict STEM choice and provides the following major findings.

1. Gaming the system at both student and action levels stably predicts STEM choice in a negative direction. Whether "gaming the system" is a sign of "lack of self-regulation" in educational and psychological literature needs to be clarified in terms of its predictive capacity for STEM choice.

2. At the action level of problem solving, in addition to less gaming, female students are more likely to study STEM if they show less boredom and have more off-task behaviors (perhaps because females may engage in socially or other off-task behaviors that support their learning; Section 4.3.2; Ke, 2008; Peterson, & Fennema, 1985). Male students are more likely to study STEM if they exhibit more concentration and less frustration. Differential intervention for both genders could be designed for both off-line and online learning platforms in order to encourage both genders to pursue advanced studies and careers in STEM.

ACKNOWLEDGMENTS

I thank the members of the ASSISTments Data Mining Competition Team 2017, including Ryan Baker and Thanaporn "March" Patikorn, for their generous provision of the datasets and support for my inquiries. This work was partially supported by the Ministry of Science and Technology, Taiwan (MOST 108-2629-H-004-001; MOST 108-2511-H-004-002). The funder only provides financial support and does not substantially influence the entire research process, from study design to submission. The author is fully responsible for the content of the paper.

REFERENCES

AIKEN, L. S., & WEST, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

ALLISON, P. D. (2012). *Logistic regression using SAS: Theory and application* (2nd ed.). Cary, NC: SAS Institute.

BAKER, R. S. J. D., CORBETT, A. T., ROLL, I., KOEDINGER, K. R., ALEVEN, V., COCEA, M., HERSHKOVITZ, A., DE CARAVALHO, A. M. J. B., MITROVIC, A., & MATHEWS, M. (2013). Modeling and studying gaming the system with educational data mining. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 97-115). New York: Springer.

BAKER, R. S., & INVENTADO, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), *Learning analytics: From research to practice* (pp. 61-75). New York, NY: Springer.

BAKER, R. S. J. D., & ROSSI, L. M. (2013). Assessing the disengaged behavior of learners. In R. Sottilare, A. Graesser, X. Hu, & H. Holden (Eds.), *Design recommendations for intelligent tutoring systems: Vol. 1. Learner modeling* (pp. 155-165). Orlando, FL: U.S. Army Research Lab.

BAKER, R. S., D'MELLO, S. K., RODRIGO, M. M. T., & GRAESSER, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–

affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies, 68,* 223-241.

BANERJEE, P. A. (2016). A longitudinal evaluation of the impact of STEM enrichment and enhancement activities in improving educational outcomes: Research protocol. *International Journal of Educational Research, 76,* 1-11.

BARKATSAS, A. T., KASIMATIS, K., & GIALAMAS, V. (2009). Learning secondary mathematics with technology: Exploring the complex interrelationship between students' attitudes, engagement, gender and achievement. *Computers & Education, 52,* 562-570.

BECKER, D. (2019). *Machine learning explainability course home page***.** Retrieved from https://www.kaggle.com/dansbecker/permutation-importance

BERNIER, A., CARLSON, S. M., & WHIPPLE, N. (2010). From external regulation to self-regulation: Early parenting precursors of young children's executive functioning. *Child Development, 81,* 326-339.

BOTELHO, A. F., BAKER, R. S., & HEFFERNAN, N. T. (2017). Improving sensor-free affect detection using deep learning. *Proceedings of the 18th International Conference on Artificial Intelligence in Education* (pp. 40-51). Springer.

BOWLES, M. (2015). *Machine learning in Python: Essential techniques for predictive analysis*. Indianapolis, IN: John Wiley & Sons.

CARLI, L. L., ALAWA, L., LEE, Y., ZHAO, B., & KIM, E. (2016). Stereotypes about gender and science: Women≠ scientists. *Psychology of Women Quarterly, 40,* 244-260.

CERINSEK, G., HRIBAR, T., GLODEZ, N., & DOLINSEK, S. (2013). Which are my future career priorities and what influenced my choice of studying science, technology, engineering or mathematics? Some insights on educational choice—case of Slovenia. *International Journal of Science Education, 35,* 2999-3025.

CHIU, M.-S. (2007). Mathematics as mother/basis of science in affect: Analysis of TIMSS 2003 data. In J. H. Woo, H. C. Lew, K. S. Park, & D. Y. Seo (Eds.), *Proceedings of the 31st Conference of the International Group for the Psychology of Mathematics Education, 2,* 145-152. Seoul: PME.

CHIU, M.-S. (2012). The internal/external frame of reference model, big-fish-little-pond effect, and combined model for mathematics and science. *Journal of Educational Psychology, 104,* 87-107.

CHIU, M.-S. (2017). High school student rationales for studying advanced science: Analysis of their psychological and cultural capitals. *Journal of Advances in Education Research, 2, 171-182.*

CLIFFORD, M. M. (1988). Failure tolerance and academic risk-taking in ten-to twelve-year-old students. *British Journal of Educational Psychology, 58,* 15-27.

COHEN, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159.

DE WITTE, K., HAELERMANS, C., & ROGGE, N. (2015). The effectiveness of a computer-assisted math learning program. *Journal of Computer Assisted Learning, 31,* 314-329.

D'MELLO, S., PICARD, R. W., & GRAESSER, A. (2007). Toward an affect-sensitive AutoTutor. *IEEE Intelligent Systems, 22*(4), 53-61.

EASTWOOD, J. D., FRISCHEN, A., FENSKE, M. J., & SMILEK, D. (2012). The unengaged mind: Defining boredom in terms of attention. *Perspectives on Psychological Science, 7,* 482-495.

ELSE-QUEST, N. M., HYDE, J. S., & LINN, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136*, 103-127.

FAWCETT, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*, 861-874.

GAZLEY, J. L., REMICH, R., NAFFZIGER-HIRSCH, M. E., KELLER, J., CAMPBELL, P. B., & MCGEE, R. (2014). Beyond preparation: Identity, cultural capital, and readiness for graduate school in the biomedical sciences. *Journal of Research in Science Teaching, 51*, 1021-1048.

GEE, J. P. (2005a). Good video games and good learning. *Phi Kappa Phi Forum, 82*(2), 34-37.

GEE, J. P. (2005b). Learning by design: Good video games as learning machines. *E-learning and Digital Media, 2*(1), 5-16.

GLYNN, S. M., TAASOOBSHIRAZI, G., & BRICKMAN, P. (2007). Nonscience majors learning science: A theoretical model of motivation. *Journal of Research in Science Teaching, 44*, 1088-1107.

GUISO, L., MONTE, F., SAPIENZA, P., & ZINGALES, L. (2008). Culture, gender, and mathematics. *Science, 320*, 1164-1165.

HAIR, J. F., JR., BLACK, W. C., BABIN, B. J., ANDERSON, R. E., & TATHAM, R. L. (2006). *Multivariate data analysis* (6th ed.). Upper Saddle River, NJ: Prentice-Hall.

HEFFERNAN, N. T., & 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 in Education, 24*(4), 470-497.

HELLEVIK, O. (2009). Linear versus logistic regression when the dependent variable is a dichotomy. *Quality & Quantity, 43*, 59-74.

HO, H., SENTURK, D., LAM, A. G., ZIMMER, J. M., HONG, S., OKAMOTO, Y., CHIU, S., NAKAZAWA, Y., & WANG, C. (2000). The affective and cognitive dimensions of math anxiety: A cross-national study. *Journal for Research in Mathematics Education, 31*, 362-379.

HOQUE, M. E., MCDUFF, D. J., & PICARD, R. W. (2012). Exploring temporal patterns in classifying frustrated and delighted smiles. *IEEE Transactions on Affective Computing, 3*(3), 323-334.

HSU, P. L., ROTH, W. M., MARSHALL, A., & GUENETTE, F. (2009). To be or not to be? Discursive resources for (dis-)identifying with science-related careers. *Journal of Research in Science Teaching, 46*, 1114-1136.

HUSSEIN, A. (2015). The use of triangulation in social sciences research: Can qualitative and quantitative methods be combined?. *Journal of Comparative Social Work, 4*(1), 1-12.

HYDE, J. S. (2005). The gender similarities hypothesis. *American Psychologist, 60*, 581-592.

KAI, S., ALMEDA, M. V., BAKER, R. S., HEFFERNAN, C., & HEFFERNAN, N. (2018). Decision tree modeling of wheel-spinning and productive persistence in Skill Builders. *Journal of Educational Data Mining, 10*(1), 36-71.

KOEHLER, M. S. (1990). Classrooms, teachers, and gender differences in mathematics. In E. Fennema & G. C. Leder (Eds.), *Mathematics and gender* (pp. 128-148). New York: Columbia University, Teachers College.

KOLLER, O., BAUMERT, J., & SCHNABEL, K. (2001). Does interest matter? The relationship between academic interest and achievement in mathematics. *Journal for Research in Mathematics Education, 32*, 448-470.

LEE, D. M. C., RODRIGO, M. M. T., BAKER, R. S., SUGAY, J. O., & CORONEL, A. (2011). Exploring the relationship between novice programmer confusion and achievement. *Proceedings of the 4th bi-annual International Conference on Affective Computing and Intelligent Interaction* (pp. 175-184). Berlin, Heidelberg: Springer.

LIU, Z., PATARANUTAPORN, V., OCUMPAUGH, J., BAKER, R.S.J.D. (2013) Sequences of frustration and confusion, and Learning. *Proceedings of the 6th International Conference on Educational Data Mining*, 114-120.

LYUBOMIRSKY, S., KING, L., & DIENER, E. (2005). The benefits of frequent positive affect: Does happiness lead to success? *Psychological Bulletin, 131*, 803-855.

MALTESE, A. V., & TAI, R. H. (2010). Eyeballs in the fridge: Sources of early interest in science. *International Journal of Science Education, 32*, 669-685.

MASON, J., BURTON, L., & STACEY, K. (1996). *Thinking mathematically*. Essex: Addison-Wesley.

MCLEOD, D. B. (1992). Research on affect in mathematics education: A reconceptualisation. In D. A. Grouws (Ed.), *Handbook of Research on Mathematics Teaching and Learning: a Project of the National Council of Teachers of Mathematics* (pp. 575-596). New York: Macmillan.

MCLEOD, D. B. (1994). Research on affect and mathematics learning in the JRME: 1970 to the present. *Journal for Research in Mathematics Education, 25*, 637-647.

MEECE, J. L. WIGFIELD, A., & ECCLES, J. S. (1990). Predictors of math anxiety and its influence on young adolescents' course enrollment intentions and performance in mathematics. *Journal of Educational Psychology, 82*, 60-70.

NAKAMURA, J., & CSIKSZENTMIHALYI, M. (2002). The concept of flow. In C. R. Snyder & S. J. Lopez (Eds.), *Handbook of Positive Psychology* (pp. 89-105). New York, NY: Oxford University Press.

NUGENT, G., BARKER, B., WELCH, G., GRANDGENETT, N., WU, C., & NELSON, C. (2015). A model of factors contributing to STEM learning and career orientation. *International Journal of Science Education, 37*, 1067-1088.

OCUMPAUGH, J., BAKER, R. S., & RODRIGO, M. M. T. (2015). *Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual*. New York, NY: Teachers College, Columbia University. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.

OCUMPAUGH, J., SAN PEDRO, M. O., LAI, H. Y., BAKER, R. S., & 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.

ORGANIZATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. (2014). *PISA 2012 results: What students know and can do – student performance in mathematics, reading and science* (Volume I, Revised edition, February 2014). Paris, France: Author. Retrieved from http://dx.doi.org/10.1787/9789264201118-en

ORGANIZATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. (2015). *The ABC of gender equality in education: Aptitude, behavior, confidence*. Paris: OECD Publishing.

PAJARES, F., & MILLER, M. D. (1994). Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis. *Journal of Educational Psychology, 86*, 193-203.

PARDOS, Z. A., BAKER, R. S., SAN PEDRO, M., GOWDA, S. M., & GOWDA, S. M. (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*, 107-128.

PETERSON, P. L., & FENNEMA, E. (1985). Effective teaching, student engagement in classroom activities, and sex-related differences in learning mathematics. *American Educational Research Journal, 22*, 309-335.

PINTRICH, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology, 95*, 667-686.

POHLMANN, J. T., & LEITNER, D. W. (2003). A comparison of ordinary least squares and logistic regression. *Ohio Journal of Science, 103*, 118-125.

SAN PEDRO, M. O. Z., BAKER, R. S., BOWERS, A. J., & HEFFERNAN, N. T. (2013a). Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. *Proceedings of the 6th International Conference on Educational Data Mining* (pp. 177–184).

SAN PEDRO, M. O. Z., BAKER, R., GOWDA, S. M., & HEFFERNAN, N. T. (2013b). Towards an understanding of affect and knowledge from student interaction with an intelligent tutoring system. In *Proceedings of the 16th International Conference on Artificial Intelligence and Education*, 9–13 July, Exeter, UK, 41–50.

SAN PEDRO, M. O. Z., OCUMPAUGH, J., BAKER, R., & HEFFERNAN, N. T. (2014). Predicting STEM and non-STEM college major enrollment from middle school interaction with mathematics educational software. *Proceedings of the 7th International Conference on Educational Data Mining,* 276-279. Memphis, TN: International Educational Data Mining Society.

SHUTE, V. J., D'MELLO, S., BAKER, R., CHO, K., BOSCH, N., OCUMPAUGH, J., VENTURA, M., & ALMEDA, V. (2015). Modeling how incoming knowledge, persistence, affective states, and in-game progress influence student learning from an educational game. *Computers & Education, 86*, 224-235.

STROBL, C., MALLEY, J., & TUTZ, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods, 14*(4), 323-348.

TEMPELAAR, D. T., RIENTIES, B., & GIESBERS, B. (2015). In search for the most informative data for feedback generation: Learning Analytics in a data-rich context. *Computers in Human Behavior, 47*, 157-167.

TZE, V. M., DANIELS, L. M., & KLASSEN, R. M. (2016). Evaluating the relationship between boredom and academic outcomes: A meta-analysis. *Educational Psychology Review, 28*, 119-144.

YÜKSEL-ŞAHIN, F. (2008). Mathematics anxiety among 4th and 5th grade Turkish elementary school students. *International Electronic Journal of Mathematics Education, 3*, 179-192.

ZELDIN, A. L., & PAJARES, F. (2000). Against the odds: Self-efficacy beliefs of women in mathematical, scientific, and technological careers. *American Educational Research Journal, 37*, 215-246.

ZHU, Z. (2007). Gender differences in mathematical problem solving patterns: A review of literature. *International Education Journal, 8,* 187-203.

ZIMMERMAN, B. J., & MARTINEZ-PONS, M. (1990). Student differences in self-regulated learning: Relating grade, sex, and giftedness to self-efficacy and strategy use. *Journal of Educational Psychology*, *82*, 51-59.