Using Machine Learning Explainability Methods to Personalize Interventions for Students

Paul Hur
University of Illinois
Urbana–Champaign
khur4@illinois.edu

HaeJin Lee
University of Illinois
Urbana–Champaign
haejin2@illinois.edu

Suma Bhat
University of Illinois
Urbana–Champaign
spbhat2@illinois.edu

Nigel Bosch
University of Illinois
Urbana–Champaign
pnb@illinois.edu

ABSTRACT

Machine learning is a powerful method for predicting the outcomes of interactions with educational software, such as the grade a student is likely to receive. However, a predicted outcome alone provides little insight regarding how a student's experience should be personalized based on that outcome. In this paper, we explore a generalizable approach for resolving this issue by personalizing learning using explanations of predictions generated via machine learning explainability methods. We tested the approach in a self-guided, self-paced online learning system for college-level introductory statistics topics that provided personalized interventions for encouraging self-regulated learning behaviors. The system used explanations generated by SHAP (SHapley Additive exPlanations) to recommend specific actions for students to take based on features that most negatively influenced predicted learning outcomes; an "expert system" comparison condition provided recommendations based on predefined rules. A randomized controlled trial of 73 participants (37 expert-system condition, 36 explanation condition) revealed similar learning and topic-choosing behavior between conditions, suggesting that XAI-informed interventions facilitated student statistics learning to a similar degree as expert-system interventions.

Keywords

Machine Learning Explainability, Online Learning,
Self-regulated Learning, Educational Interventions

1. INTRODUCTION

Personalization promotes learning by providing meaningful, timely, and relevant support that is tailored and paced to an individual's needs and preferences [4, 32]. Thus, many intelligent tutoring systems (ITSs) have integrated personalization aspects that can automatically suggest which materials to study [9], reorient attentional states [15], and construct personalized feedback [20]. Such interventions are often driven by predictions using past learners' data to build machine learning models, whose underlying mechanisms can be difficult to interpret. Yet, understanding the reasons behind a prediction is essential for educational software that needs to respond not only to *what* is likely (i.e., predicted) to happen, but also *why* it is likely.

Explainable artificial intelligence (XAI) methods [8, 17] have been developed to circumvent the opaque nature of complex machine learning models, which may thus enable a new generation of educational software with increased user trust and perceived usefulness [12, 13]. In this paper, we create personalized interventions driven by explanations, rather than by predictions, for the purpose of adapting students' behaviors in a computer-based learning environment. We focus on encouraging self-regulated learning (SRL) behaviors in particular [33]. SRL is especially important in online and computer-based learning contexts, where teachers are often less available (versus classroom learning contexts) to guide the learning process. However, many students need assistance with these SRL decisions [37, 45, 36], and thus stand to benefit from computer-based learning environments that fill the gaps in SRL skills by suggesting appropriate activities to students.

We present work from a randomized controlled trial for which we developed an online, computer-based education platform for college-level introductory statistics topics. We explored how machine learning model predictions, coupled with explanations, can personalize interventions to support SRL reviewing behaviors. Our study provides a rigorous comparison of an XAI-driven intervention against an active expert-system intervention consisting of predefined rules based on the amount of time spent studying each topic and the expected order of topics in the curriculum.

The XAI-driven interface adaptations in this work raise several research questions (RQs) related to the effect interventions have on learning and the effects adaptations have on behaviors. These RQs have implications for computer-based education (and for broader understanding of how simple

XAI-driven interventions affect student behaviors).

RQ1: What are the effects on learning and self-regulated learning behaviors when students receive XAI-informed interventions vs. expert-system interventions?

Hypothesis: We expected that the participants in the XAI-informed intervention group will learn more than those receiving expert-system interventions due to XAI-informed intervention group studying topics directly related to improving their predicted learning outcome. Furthermore, we expected that XAI-informed interventions would lead to more frequent reviewing SRL behaviors. Although both the XAI-informed and expert-system condition interventions in our study were created with the goal of supporting SRL reviewing behaviors (re-taking quizzes and re-reading texts), we expected that XAI-informed interventions would better highlight topic areas that needed the most studying based on learning outcome predictions.

RQ2: Do XAI-informed interventions lead to different topic choosing behaviors compared to the expert-system intervention?

Hypothesis: XAI-informed interventions may impart topic choosing strategies based on predicted knowledge gaps, which we expected would lead to lower proportions of students following the default intended order of topics. Furthermore, we expected that students would follow interventions in the expert-system condition more frequently because these interventions often recommended a top-to-bottom reading order of topics that may align with students' natural inclinations.

2. RELATED WORK

Here, we highlight work on self-regulated learning in online and computer-based education environments and interface adaptations informed by XAI.

2.1 Supporting Self-Regulated Learning

Self-regulated learning (SRL) refers to the metacognitive, motivational, and emotional processes behind acquiring information or skills [48, 33]. SRL has been identified as an important skill for succeeding in postsecondary education [27, 31]. Developing SRL skills is difficult, however; students struggle to differentiate effectiveness between learning strategies [50], and may not be aware of how to develop SRL skills [5]. There are three general groups of strategies identified in major SRL models: preparation, performance, and regulation [33, 10, 38, 44, 50, 38, 44] (though there is some variation, including SRL strategies that occur after/between learning sessions [49]). It is regulation behaviors (e.g., revisiting materials or re-taking quizzes to prepare for a final test) our study interventions target, since reviewing behaviors can be supported by recommending review of specific learning material during test preparation.

Within the past two decades, a substantial amount of work has been carried out encouraging SRL in online learning environments, such as MOOCs [25, 23], where SRL skills may be especially important since learners are required to learn autonomously [2, 42]. Researchers have developed computer-based education environments to support SRL skills, such as MetaTutor [3], Betty's Brain [24], and Cognitive Tutor [40], which aid SRL skills via adaptive pedagogical agents or by automatically personalizing the presentation of information. These systems, along with other research in online contexts [14, 34], demonstrate the feasibility of utilizing data recorded in log files to examine SRL behaviors through modeling SRL behaviors and predicting student outcomes.

2.2 Adaptations using XAI

In this paper, we focus on a particular XAI method called SHAP (SHapley Additive exPlanations) [26]. SHAP is well-suited to driving interface adaptations because it provides, for every prediction, an indication of how much each feature (i.e., predictor variable) influenced the decision made by a machine learning model. SHAP values capture directionality (e.g., the value of feature X_1 for this prediction contributed positively vs. negatively to the prediction) as well as magnitude, via a game-theoretic approach [22]. Hence, an interface can adapt to the needs of users based on feature values and the effects those values have on predictions (i.e., the SHAP values), provided that the features themselves are interpretable [6].

Within XAI research, there has been less focus on XAI systems that leverage machine learning explainability for adaptation for education purposes. Conati et al. used XAI for integrating explanation functionality for adaptive hints in an Adaptive CSP (ACSP) [12], and found that explanations increase students' trust, perceived usefulness, and intention to use the hints again. In another study by Mu [30], researchers used XAI to develop suitable interventions for wheel-spinning students with simulated data and hypothetical interventions predicted for a previous study [30]. The work in our paper significantly extends this previous work [30] by examining one possible application of XAI-driven interventions (i.e., supporting SRL behaviors) via a randomized controlled trial. We also explore how XAI approaches such as SHAP can help education researchers discover sensible interventions for any learning behavior (e.g., suggest different things to different students in a plausible way)—in our case, SRL reviewing behaviors.

3. METHODS

Next, we discuss our online learning system and SRL interventions, the machine learning model for predicting student learning outcome and interpretation via SHAP (SHapley Additive exPlanations) values, and the experiment setup.

3.1 Self-guided Online Learning System

We developed a self-guided, self-paced online learning system which displays both learning content and interventions as students navigated through the interface, agnostic of content type (images, text, videos, etc.). The system also collected logs including some general interaction behaviors such as web page visits, time spent on each page, and more specific study-related data such as automatically assessed quiz scores, pretest scores, and posttest scores.

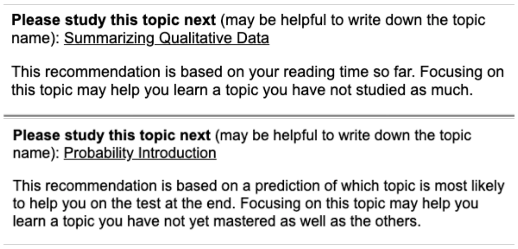
We focused on introductory statistics because it is an important yet difficult-to-learn subject for many college degrees [46, 39, 41]. We developed a small curriculum of 12 introductory statistics topics in consultation with university statistics instructors and educational websites (e.g., course pages). Each topic consisted of a reading (text tutorials and accompanying figures) and a corresponding 3-question, multiple choice mini quiz. These materials could be accessed from the main interface of the topics menu page (Fig. 2). The curriculum also included two variations of a 12-question multiple choice question test (a pretest and a posttest), with each question of the tests corresponding directly to one of the 12 topics. Both tests asked about the same core concepts and differed only with slight variations in questions, such as the specific values used. We designed the final curriculum to take a total of 90 minutes to complete, including the pretest, 12 readings, 12 quizzes, and the posttest.

3.2 Expert-system Intervention

We designed an expert-system version of the self-regulated learning intervention for the system (Fig. 1, top image) with a simple yet precise message of a topic suggestion based on reading time. However, we decided that the expert-system intervention should first suggest an unseen topic over topics with little study time, since learning outcomes improve when students at least touch on all material in time-limited scenarios [28]. We anticipated that the statistics topics on the system required careful reading in order to fully learn and perform well, as the study's statistics topics have been cited to be prone to misconception [43], and difficult to teach [11]. Thus, we expected that time spent on readings was closely connected to posttest scores. If a student knew that they had spent a lower amount of time reading one of the topics relative to others, they may self-reflect and be more likely to prioritize reviewing that topic over others they have already studied more thoroughly.

We implemented the expert-system intervention in the online learning system by displaying the intervention message when the student reached the 30 minute mark in the self-guided study session (Fig. 2), then again at 40 minutes, and finally, at 50 minutes. We chose these time points to provide sufficient data collection before the first intervention to enable an accurate prediction of student outcome, and repeated the intervention at 10-minute intervals to give the students additional suggestions. At the 60 minute mark, students were automatically taken to the posttest. Since the study session was self-guided, it was left up to the discretion of the student to read, review, or skip topics, and spend as little or as much time—up to 60 minutes—as they wished to complete the study session.

3.3 Piloting and Training Data Collection



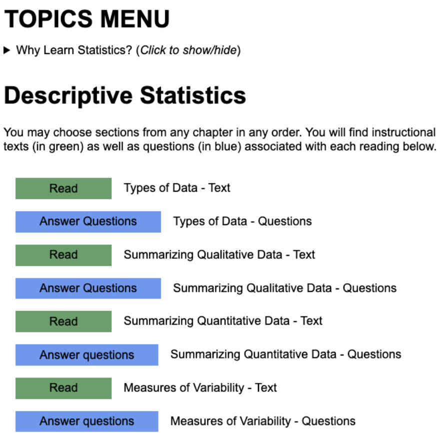
Please study this topic next (may be helpful to write down the topic name): Summarizing Qualitative Data

This recommendation is based on your reading time so far. Focusing on this topic may help you learn a topic you have not studied as much.

Please study this topic next (may be helpful to write down the topic name): Probability Introduction

This recommendation is based on a prediction of which topic is most likely to help you on the test at the end. Focusing on this topic may help you learn a topic you have not yet mastered as well as the others.

Figure 1: Examples of the expert-system (top) vs. XAI-informed (bottom) intervention messages.



TOPICS MENU

▶ Why Learn Statistics? (*Click to show/hide*)

Descriptive Statistics

You may choose sections from any chapter in any order. You will find instructional texts (in green) as well as questions (in blue) associated with each reading below.

Read Types of Data - Text

Answer Questions Types of Data - Questions

Read Summarizing Qualitative Data - Text

Answer Questions Summarizing Qualitative Data - Questions

Read Summarizing Quantitative Data - Text

Answer questions Summarizing Quantitative Data - Questions

Read Measures of Variability - Text

Answer questions Measures of Variability - Questions

Figure 2: A portion of the topics menu from a self-paced learning session.

We recruited student participants via student mailing lists and digital bulletin boards, seeking students with minimal college statistics experience (0 or 1 college-level statistics courses) in order to avoid ceiling effects from participants with extensive preexisting knowledge of the material. The study session was fully online. The study included a demographics survey, a pretest, a self-guided learning session (12 readings and 12 quizzes), and a final posttest. Participants were compensated $15 USD. Based on participant feedback from semi-structured interviews (compensated an additional $5), we made various minor changes, such as clarifying the topics menu page instructions, adjusting names to reduce cultural specificity, noting topics from the topics menu were related to the pretest and posttest questions, and including a proceed to posttest confirmation page.

After making the final changes to our system, we recruited a total of 58 participants for the first round of data collection. The goal of the first round of data collection was to collect training data for the machine learning model to predict the posttest score, which is described in the next section.

3.4 XAI-informed Intervention

Table 1: Example subset of SHAP values from a posttest score prediction for one student, indicating that the student's current time spent on topic 8 (i.e., 0 seconds) has the most negative impact on their predicted posttest outcome.

Feature name	Feature value	SHAP value
Pretest score	50%	2.454
Quiz 1 score	67%	0.206
Quiz 2 score	100%	0.221
...	...	...
Topic 6 reading time	658 seconds	3.734
Topic 7 reading time	0 seconds	-3.860
Topic 8 reading time	0 seconds	-6.187

With the training data collected with 58 participants, we trained a random forest regressor using pretest score, quiz score (12 topics), and reading time (12 topics) features to predict posttest score. We trained the final model on all data with 100 trees and a maximum tree depth of 4 (the only hyperparameter tuned). We used the tree explainer (`shap.TreeExplainer`) in the Python `shap` library [26] to interpret model behavior of the posttest score predictions in terms of SHAP values. The feature with the most negative SHAP value represented the feature which contributed most to lowering a student's posttest score.

For the XAI-informed intervention message (Figure 1, right image) used for the XAI-informed condition, the topic (quiz or reading) with the most negative SHAP value was selected to be recommended to the student. The intervention text also communicated to the student that the recommendation was based on a system prediction of the student's posttest score to help the student better understand the reason for suggesting the particular topic.

We examined what happens when a student follows the XAI-informed intervention recommendation of the feature with the most negative SHAP value. Table 1 shows an excerpt of results from a SHAP analysis of a student during a study session. At the time of this particular prediction, *Topic 8 reading time* had the most negative SHAP value with -6.187—substantially lower than the next feature, *Topic 7 reading time* at -3.860. This indicates that, based on the model's posttest prediction, topic 8 reading time is negatively impacting the predicted posttest score of this student by 6.187 points out of 100 possible points on the posttest, and should be recommended to the student to study.

Figure 3 shows how the SHAP value of one example feature, *Topic 6 reading time*, changed with each additional 20 seconds of studying time for one student. The SHAP value trended in the positive direction as reading time increased until a point between around 140 seconds of total reading time when the SHAP value plateaued at ≈ 4. The model appears to have learned that short studying times do not yield learning, and that very long studying times (relative to the brief topics used in this experiment) do not help past a certain point, thus yielding an approximately sigmoid-shaped curve. In this example, what was previously negatively influencing the predicted posttest score is now predicted to contribute around +4 points toward the final posttest score. Using the intuition from this example, our XAI-informed intervention recommends a topic to review to the student such that each student receives a personalized recommendation of the most helpful topic for improving their posttest score.

3.5 Expert-system vs. XAI-informed Intervention Experiment

We carried out a randomized controlled trial to compare the expert-system intervention and the XAI-informed intervention. We recruited 73 participants with minimal college statistics experience (0 or 1 college-level statistics courses) via campus mailing lists and—with the aid of university research support—targeted emails to undergraduate students with no statistics course on their academic course record. We also recruited students from research subject pools and introductory psychology undergraduate courses.

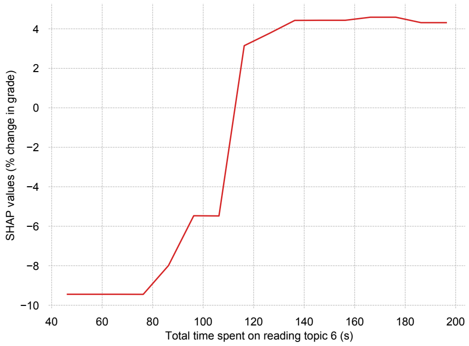


Figure 3: Changes in student's total time spent on reading topic 6 and resulting SHAP values (in % change in grade)

We randomly assigned students to conditions, with 37 students and 36 students assigned to the expert-system and XAI-informed conditions, respectively. The study session structure was identical to the training data study sessions: a video-call meeting followed by the student independently working through the study consisting of a demographics survey, a pretest, a self-guided learning session (12 readings and 12 quizzes), and a final posttest. Students were told that the estimated total completion time was around 90 minutes, and were compensated $15 USD. However, for the experiment, in the case that a participant wanted to skip out of the self-guided learning session early and proceed to the posttest, they would receive an intervention on the confirmation page and be offered the opportunity to return to the learning session. We included this final intervention to ensure that every student saw at least one intervention message regardless of whether they reached the 30 minute mark.

4. RESULTS

In this section, we report participant demographics, various learning outcome comparisons, and finer-grained analyses of participants' topic ordering learning behaviors.

4.1 Demographics Information

Among the 73 participants from the randomized control trial, 73% identified as female, 26% as male, and 1% as non-binary. Students had a mean age of 19.58 ($SD = 1.71$) years old, with a minimum age of 18 and a maximum of 27. Over 35 college majors were represented by our participant population. Finally, 55% identified as White, 27% Asian, 12% Hispanic or Latina/o, 3% Black/African American, and 1% Native American.

4.2 Learning Behaviors and Outcomes

Table 2 summarizes the differences in pretest and posttest scores between the expert-system and XAI-informed groups. The mean improvement from pretest to posttest score was 18.03 (out of 100) for the expert-system condition, and 10.88 for the XAI-informed conditions, suggesting—contrary to RQ1 expectations—that students in the expert-system condition may have learned more. However, the difference in improvement between between the two conditions was not significant, $t(71) = 1.924$, $p = .058$. We also calculated the Bayes factor (BF) using JASP [19]; BF represents how likely the null or alternate hypothesis model is through a Bayesian approach [18]. Established guidelines [18] suggest that $BF = 1$–3 provides anecdotal evidence and $BF = 3$–10 provides substantial evidence. Through this metric, there is anecdotal evidence that the expert-system intervention group had greater grade improvement, $BF = 2.24$. Table 2 also shows that both groups' mean scores were not likely to have been influenced by ceiling effects from prior statistics knowledge; furthermore, no student in either condition achieved a perfect score (100) on the pretest.

Table 2: Comparison of pretest and posttest scores between expert-system and XAI-informed conditions

Group	Count	Mean score	Std. dev.
Expert-system (pretest)	37	47.1%	16.10
XAI-informed (pretest)	36	50.2%	19.87
Expert-system (posttest)	37	65.1%	16.30
XAI-informed (posttest)	36	61.1%	20.70

4.3 Model Evaluation and XAI-informed Predictions

We performed 5-fold cross-validation with the model training data to estimate model accuracy and obtained a mean R^2 value of .262 (SD = .067), a mean RMSE of 15.17 (on a 0–100 posttest grade scale, SD = 2.20), and a mean Pearson's r of .576. Mean R^2 value was somewhat variable across cross-validation folds; however, the trained model worked relatively well overall when considering the small size of our training data.

We analyzed the predictions made by our model and the activity logs of the XAI-informed intervention group participants. We found that the top three most often recommended topics were *Probability Introduction*, *Introduction to Regression*, and *Calculating Probability*. These represent the most frequently recommended topics that had the most negative SHAP values at the time of displaying the intervention, across the 30 minute, 40 minute, 50 minute mark, and on the posttest confirmation page. Two of the three most-recommended topics were related to probability. Probability is widely recognized as a difficult topic to learn for students due to misconceptions about the subject [7, 16, 21, 43], and our findings here support this assertion.

4.4 Self-Regulated Learning Behaviors

In order to evaluate the effects of the interventions on self-regulated learning behaviors, we defined two metrics which are shown in Table 3: attempts at quizzes already taken and rereading texts that had already been read. The differences in the number of quiz retakes were not significant, $t(71)$ = -1.618, p = .110, but there was anecdotal evidence that the XAI-intervention group did more quiz retakes, BF = 1.354. Similarly, the number of text reviews was not significantly different, $t(71)$ = -1.186, p = .240, but there was anecdotal evidence of the XAI-intervention group having higher number of text reviews, $BF = 2.263$. The mean number of interventions seen prior to the posttest was 3.16 for the XAI-informed intervention group, and 2.75 for the expert-system intervention group. However, there were several participants who only saw a single intervention: 10 from the XAI-informed intervention group and 3 from the expert-system intervention group.

The results in Table 3 suggest minimal reviewing behaviors in both conditions, though that may be expected given that learning session included enough content that students could spend most or all of their time on new topics.

Table 3: Comparison of metrics for SRL reviewing behaviors.

Group	Quiz retakes	Texts reread
Expert-system	1.054	11.514
XAI	1.583	15.444

4.5 Learning Order Analysis

We carried out an analysis to examine orders in which students in each condition studied the 12 topics. We analyzed the degree to which students deviated from the baseline learning topic order by calculating the proportion of topic component (reading or quiz) selection actions that did not follow the direction of the default learning order presented on the topics menu page (Fig. 2). For example, if the student studied the first three topics in order and then studied the sixth topic, the learning order deviation value for readings would be .333. These learning order deviation values were calculated for the learning periods before and after students saw the first intervention. This analysis was done for participants who studied any amount of material after the first intervention n(XAI) = 26, n(expert-system) = 34.

The results in Table 4 show that students in both conditions deviated from the typical top-to-bottom topic significantly more frequently after intervention: XAI-informed, $t(25)$ = 4.262, p < .001; expert-system, $t(33)$ = 3.240, p = .003. These differences before and after the intervention were expected, per RQ1, especially for the XAI-informed intervention condition since it is more likely to recommend topics out of order according to students' individual needs.

Table 4: Proportion of actions in which students deviated from a typical top-to-bottom topic order during their selections of what topic to pursue next.

	Topic order deviation	
	Expert-system	**XAI**
Before 1st intervention	.552	.539
After 1st intervention	.709	.727
Difference before/after	**.157**	**.188**

5. DISCUSSION

Here, we discuss the main findings and implications and also discuss generalization of our approach, limitations of our study, and possible future work.

5.1 Learning and SRL Behaviors (RQ1)

We hypothesized that the students in the XAI-informed condition would have greater learning gains when compared to those receiving the expert-system intervention because XAI interventions would give suggestions to review the most critical topics for improving posttest score rather than the expert-condition suggestions based on time. However, the findings did not support our hypothesis. Learning gain was not significantly different between the two conditions. The overall average pretest score being 48.6% and average posttest score 63.1%, and thus students may have benefited from studying almost *any* topic. In such cases, an intervention to encourage specific SRL behaviors would not be needed until the student has spent much longer studying.

Additionally, our machine learning model was trained from a relatively small amount of training data of 58 students, which may have contributed noise to the predictions, and consequently, less effective XAI-informed interventions which recommended unhelpful topics for improving the posttest score. However, both the expert-system and XAI-informed groups had significant, notable improvements in pretest to posttest scores, showing that our curriculum (and perhaps both interventions) was effective for teaching the statistics topics.

We hypothesized that the XAI-informed interventions would have lead to more frequent SRL reviewing behaviors due to bringing to light more directed learning strategies motivated by improving one's posttest score, and therefore, have lead to more regular and frequent self-reflection to identify apparent gaps in learning. The findings were inconclusive for answering our hypothesis since there were few instances of SRL reviewing behaviors (Table 3). While there were no statistical differences between the mean quiz and text reviewing behaviors, Bayes factor values show that there was anecdotal evidence of the XAI-informed group having both slightly higher mean rates of retaking quizzes and rereading texts. This suggests that there may indeed be intervention effects that could emerge more clearly in large datasets and longer learning sessions.

5.2 Topic Choosing Behaviors (RQ2)

We expected that participants receiving the expert-system interventions would have been more receptive to following the interventions' suggestions compared to those receiving the XAI-informed interventions. The expert-system interventions recommended topics based on the lowest reading time spent, or to suggest the next unstudied topic in the expected order. This would have aligned with some students' natural inclinations of studying through the topics in the default order as presented on the topics menu page (Figure 2). Furthermore, we expected the XAI-informed intervention group to strategize their topic choosing behavior without the influence of a necessarily top-to-bottom order suggestion, and choose more autonomously, based on strategies of identifying gaps in knowledge (discussed in section 5.1).

The results show that there were increased effects on topic choosing deviation behaviors in the two conditions. XAI-informed and expert-system conditions both had increased proportions of topic selection behaviors deviating from the top-to-bottom order after the first intervention. The results suggest the possibility that both interventions facilitated self-directed learning behaviors of students to make topic choosing behaviors most beneficial for their learning. When understanding the results of RQ1 on learning performance and topic suggestions together, the findings from our paper may suggest that the XAI-informed interventions facilitated students' statistics learning to a similar degree to a method which prioritized unseen topics over mastery of a smaller amount of material. Additionally, average pretest scores were quite low for both conditions (Table 2), which likely made any amount of scaffolding helpful.

5.3 Generalization to Other Domains

In our study, we used the possible reasons from quizzes and readings for the posttest score predictions to personalize interventions to scaffold SRL behaviors through encouraging the review of specific topics. However, the same XAI approach could be applied to cases where the goal is to help a student improve almost any predicted outcome. For example, one could predict student dropout from online learning based on relevant student factors (constructs related to study skills, learning material interaction patterns, stress, motivation, etc.) [35], combined with complex or uninterpretable factors, and implement XAI-informed alerts via interfaces or targeted emails for the student throughout the course period to take actions reducing the likelihood of dropout based on the most impactful of the interpretable factors. Other applications could follow a similar approach, such as helping students resolve confusion during attempting algebra assignments [1], preventing learner disengagement from reading texts [29], or improving student performance in interactive online question pools [47].

5.4 Future Work and Conclusion

Future work could expand on this research by increasing the number of participants in both the training data and experiment as we were limited by our relatively small pilot and experiment sample sizes. It may also be possible to reduce idling behavior instances by using pre-screening surveys to gauge their motivation for learning the material, or better controlling the experiment through an in-person lab setting where participants may be observed. It would also be informative to explore our approach to other types of study material, such as more advanced statistics topics or mathematics, and explore ways to support other SRL behaviors such as planning or goal-setting.

In this study, we leveraged machine learning to predict future student outcomes, explained the predictions via an XAI method, and implemented personalized system interventions. Specifically, we explored supporting SRL behaviors in an online learning environment for learning college-level introductory statistics topics through personalized interventions. Despite limited differences in learning gain, SRL reviewing behaviors, and topic choosing behaviors, our findings suggest that XAI-informed interventions facilitate learning to a similar benefit as expert-system interventions. We expect that the approach examined in this experiment could be generalized across other applications, and could serve as one reference for designing system implementation informed by XAI methods.

6. ACKNOWLEDGEMENTS

This research was supported by a grant from the Technology Innovation in Educational Research and Design (TIER-ED) initiative at the University of Illinois Urbana–Champaign.

7. REFERENCES

[1] S. M. R. Abidi, M. Hussain, Y. Xu, and W. Zhang. Prediction of confusion attempting algebra homework in an intelligent tutoring system through machine learning techniques for educational sustainable development. *Sustainability*, 11(1):105, 2019.

[2] M. E. Alonso-Mencía, C. Alario-Hoyos, J. Maldonado-Mahauad, I. Estévez-Ayres, M. Pérez-Sanagustín, and C. Delgado Kloos. Self-regulated learning in moocs: lessons learned from a literature review. *Educational Review*, 72(3):319–345, 2020.

[3] R. Azevedo, A. Johnson, A. Chauncey, and C. Burkett. Self-regulated learning with metatutor: Advancing the science of learning with metacognitive tools. In I. M. Khine, Myint Sweand Saleh, editor, *New science of learning*, pages 225–247. Springer, New York, NY, 2010.

[4] M. L. Bernacki, M. J. Greene, and N. G. Lobczowski. A systematic review of research on personalized learning: Personalized by whom, to what, how, and for what purpose (s)? *Educational Psychology Review*, 33(4):1675–1715, 2021.

[5] R. A. Bjork, J. Dunlosky, and N. Kornell. Self-regulated learning: Beliefs, techniques, and illusions. *Annual review of psychology*, 64:417–444, 2013.

[6] N. Bosch. AutoML feature engineering for student modeling yields high accuracy, but limited interpretability. *Journal of Educational Data Mining*, 13(2):55–79, 2021.

[7] K. Carolyn and S. Kirk. A new approach to learning probability in the first statistics course. *Journal of Statistics Education*, 9(3), 2001.

[8] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

[9] N. Chakraborty, S. Roy, W. L. Leite, M. K. S. Faradonbeh, and G. Michailidis. The effects of a personalized recommendation system on students' high-stakes achievement scores: A field experiment. *International Educational Data Mining Society*, 2021.

[10] A. Cicchinelli, E. Veas, A. Pardo, V. Pammer-Schindler, A. Fessl, C. Barreiros, and S. Lindstädt. Finding traces of self-regulated learning in activity streams. In *Proceedings of the 8th international conference on learning analytics and knowledge*, pages 191–200, 2018.

[11] G. W. Cobb and D. S. Moore. Mathematics, statistics, and teaching. *The American mathematical monthly*, 104(9):801–823, 1997.

[12] C. Conati, O. Barral, V. Putnam, and L. Rieger. Toward personalized xai: A case study in intelligent tutoring systems. *Artificial Intelligence*, 298:103503, 2021.

[13] C. Conati, K. Porayska-Pomsta, and M. Mavrikis. Ai in education needs interpretable machine learning: Lessons from open learner modelling. *arXiv preprint arXiv:1807.00154*, 2018.

[14] S. Crossley, M. Dascalu, D. S. McNamara, R. Baker, and S. Trausan-Matu. Predicting success in massive open online courses (moocs) using cohesion network analysis. Philadelphia, PA: International Society of the Learning Sciences., 2017.

[15] S. D'Mello, A. Olney, C. Williams, and P. Hays. Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of human-computer studies*, 70(5):377–398, 2012.

[16] J. Garfield and A. Ahlgren. Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for research in Mathematics Education*, 19(1):44–63, 1988.

[17] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[18] A. F. Jarosz and J. Wiley. What are the odds? a practical guide to computing and reporting bayes factors. *The Journal of Problem Solving*, 7(1):2, 2014.

[19] JASP Team. JASP (Version 0.16.1)[Computer software], 2022.

[20] E. Kochmar, D. D. Vu, R. Belfer, V. Gupta, I. V. Serban, and J. Pineau. Automated personalized feedback improves learning gains in an intelligent tutoring system. In *International Conference on Artificial Intelligence in Education*, pages 140–146. Springer, 2020.

[21] C. Konold. Issues in assessing conceptual understanding in probability and statistics. *Journal of statistics education*, 3(1), 1995.

[22] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler. Problems with Shapley-value-based explanations as feature importance measures. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5491–5500. PMLR, Nov. 2020.

[23] D. Lee, S. L. Watson, and W. R. Watson. Systematic literature review on self-regulated learning in massive open online courses. *Australasian Journal of Educational Technology*, 35(1), 2019.

[24] K. Leelawong and G. Biswas. Designing learning by teaching agents: The betty's brain system. *International Journal of Artificial Intelligence in Education*, 18(3):181–208, 2008.

[25] A. Littlejohn, N. Hood, C. Milligan, and P. Mustain. Learning in moocs: Motivations and self-regulated learning in moocs. *The internet and higher education*, 29:40–48, 2016.

[26] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, 2020.

[27] C. Mega, L. Ronconi, and R. De Beni. What makes a good student? how emotions, self-regulated learning, and motivation contribute to academic achievement. *Journal of educational psychology*, 106(1):121, 2014.

[28] P. Michlík and M. Bieliková. Exercises recommending for limited time learning. *Procedia Computer Science*, 1(2):2821–2828, 2010.

[29] C. Mills, N. Bosch, A. Graesser, and S. D'Mello. To quit or not to quit: predicting future behavioral disengagement from reading patterns. In *International Conference on Intelligent Tutoring Systems*, pages

19–28. Springer, 2014.

[30] T. Mu, A. Jetten, and E. Brunskill. Towards suggesting actionable interventions for wheel-spinning students. In *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)*, pages 183–193. International Educational Data Mining Society, July 2020.

[31] W. C. Naumann, D. Bandalos, and T. B. Gutkin. Identifying variables that predict college success for first-generation college students. *Journal of College Admission*, (181):4, 2003.

[32] U. S. D. of Education. Transforming american education: Learning powered by technology. 2010.

[33] E. Panadero. A review of self-regulated learning: Six models and four directions for research. *Frontiers in psychology*, 8:422, 2017.

[34] A. Pardo, F. Han, and R. A. Ellis. Combining university student self-regulated learning indicators and engagement with online learning events to predict academic performance. *IEEE Transactions on Learning Technologies*, 10(1):82–92, 2016.

[35] J.-H. Park. Factors related to learner dropout in online learning. *Online Submission*, 2007.

[36] M. Pedrotti and N. Nistor. How students fail to self-regulate their online learning experience. In *European Conference on Technology Enhanced Learning*, pages 377–385. Springer, 2019.

[37] N. E. Perry, L. Hutchinson, and C. Thauberger. Talking about teaching self-regulated learning: Scaffolding student teachers' development and use of practices that promote self-regulated learning. *International Journal of Educational Research*, 47(2):97–108, 2008.

[38] M. Puustinen and L. Pulkkinen. Models of self-regulated learning: A review. *Scandinavian journal of educational research*, 45(3):269–286, 2001.

[39] J. B. Ramsey. Why do students find statistics so difficult. *Proccedings of the 52th Session of the ISI. Helsinki*, pages 10–18, 1999.

[40] S. Ritter, J. R. Anderson, K. R. Koedinger, and A. Corbett. Cognitive tutor: Applied research in mathematics education. *Psychonomic bulletin & review*, 14(2):249–255, 2007.

[41] R. D. Snee. What's missing in statistical education? *The american statistician*, 47(2):149–154, 1993.

[42] Y.-h. Tsai, C.-h. Lin, J.-c. Hong, and K.-h. Tai. The effects of metacognition on online learning interest and continuance to learn with moocs. *Computers & Education*, 121:18–29, 2018.

[43] A. Tversky and D. Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232, 1973.

[44] A. Uzir, D. Gašević, W. Matcha, J. Jovanović, A. Pardo, L.-A. Lim, S. Gentili, et al. Discovering time management strategies in learning processes using process mining techniques. In *European Conference on Technology Enhanced Learning*, pages 555–569. Springer, 2019.

[45] M. V. Veenman. Learning to self-monitor and self-regulate. In *Handbook of research on learning and instruction*, pages 249–273. Routledge, 2016.

[46] D. G. Watts. Why is introductory statistics difficult to learn? and what can we do to make it easier? *The American Statistician*, 45(4):290–291, 1991.

[47] H. Wei, H. Li, M. Xia, Y. Wang, and H. Qu. Predicting student performance in interactive online question pools using mouse interaction features. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, pages 645–654, 2020.

[48] B. J. Zimmerman. Self-regulated learning and academic achievement: An overview. *Educational psychologist*, 25(1):3–17, 1990.

[49] B. J. Zimmerman. Becoming a self-regulated learner: An overview. *Theory Into Practice*, 41(2):64–70, 2002.

[50] B. J. Zimmerman and M. M. Pons. Development of a structured interview for assessing student use of self-regulated learning strategies. *American educational research journal*, 23(4):614–628, 1986.