Gaze-Enhanced Student Modeling for Game-based Learning

Andrew Emerson
Department of Computer Science
North Carolina State University
ajemerso@ncsu.edu

Robert Sawyer
Department of Computer Science
North Carolina State University
rssawyer@ncsu.edu

Roger Azevedo
Department of Psychology
North Carolina State University
razeved@ncsu.edu

James Lester
Department of Computer Science
North Carolina State University
lester@ncsu.edu

ABSTRACT

Recent advances in eye-tracking technologies have introduced the opportunity to incorporate gaze into student modeling. Creating student models that leverage gaze information holds significant promise for game-based learning environments. This paper introduces a gaze-enhanced student modeling framework that incorporates student eye tracking to dynamically predict students' performance in a game-based learning environment for microbiology education, CRYSTAL ISLAND. The gaze-enhanced student modeling framework was investigated in a study comparing a gaze-enhanced student model with a baseline student model that does not utilize student eye-tracking. Results of a study conducted with 65 college students interacting with the CRYSTAL ISLAND game-based learning environment indicate that the gaze-enhanced student model significantly outperforms the baseline model in dynamically predicting student problem-solving performance. The findings suggest that incorporating gaze into student modeling can contribute to a new generation of student models for game-based learning environments.

KEYWORDS

Student modeling; Gaze; Game-based learning

ACM Reference format:
Andrew Emerson, Robert Sawyer, Roger Azevedo, and James Lester. 2018. Gaze-Enhanced Student Modeling for Game-based Learning. In *Proceedings of the Twenty-Sixth ACM Conference on User Modeling, Adaptation and Personalization, Singapore, Singapore, July 2018,* (UMAP'18), 10 pages. https://doi.org/10.1145/3209219.3209238.

1 INTRODUCTION

Student modeling plays a central role in user-adaptive environments for learning [12]. Student models offer an explicit representation of a learning environment's representation of student characteristics. Student modeling techniques have been devised to infer a broad range of student characteristics, including student knowledge [3, 9, 25, 35], plans and goals [1, 25, 30-32], and affective states [4, 7, 17, 40, 43, 46]. Despite these advances, with only a few notable exceptions [11, 21, 28, 45], limited work has investigated the potential of leveraging information about student gaze to improve the accuracy of student modeling.

Gaze offers a potentially rich source of information about student learning. Temporal patterns in eye movements, such as variations in fixations and saccades, may indicate the attentional foci of student interactions with a learning environment. Recent work such as the investigation of how gaze may signal the presence of cognitive processes [37] suggests that gaze can inform student modeling. While emerging work has begun to examine how learning can be inferred from gaze data [6, 24, 34] and how to infer mind wandering from gaze data [29], there has been limited investigation of how gaze can enhance student modeling to improve predictive accuracy for student problem-solving performance during learning interactions.

Game-based learning environments offer a promising context for investigating gaze-enhanced student modeling. Over the past decade game-based learning environments have emerged as a vehicle for creating engaging learning experiences through game mechanics [5, 14, 18, 27, 39, 47]. Immersive game-based learning environments, such as the CRYSTAL ISLAND learning environment for microbiology education [39], feature rich story worlds, an expansive cast of characters, and a large set of digital artifacts that students interact with during learning episodes. Because immersive game-based learning environments feature 3D worlds that students navigate during problem solving, these environments may elicit fine-grained gaze behaviors that provide significant diagnostic value for student modeling, while providing a manifestation of intent-related cognitive processes [19, 33, 44].

This paper reports on an investigation of gaze-enhanced student modeling for game-based learning. With the goal of creating student models augmented with gaze data to improve the predictive accuracy of student models, in a study with 65 students we evaluate two student modeling approaches: a gaze-enhanced student model that leverages gaze data collected from a version of CRYSTAL ISLAND that was instrumented with eye-tracking, and a baseline student model that does not use gaze data. We compare both with respect to their predictive accuracy on student problem-

UMAP '18, July 8–11, 2018, Singapore, Singapore

ACM ISBN 978-1-4503-5589-6/18/07…$15.00
https://doi.org/10.1145/3209219.3209238

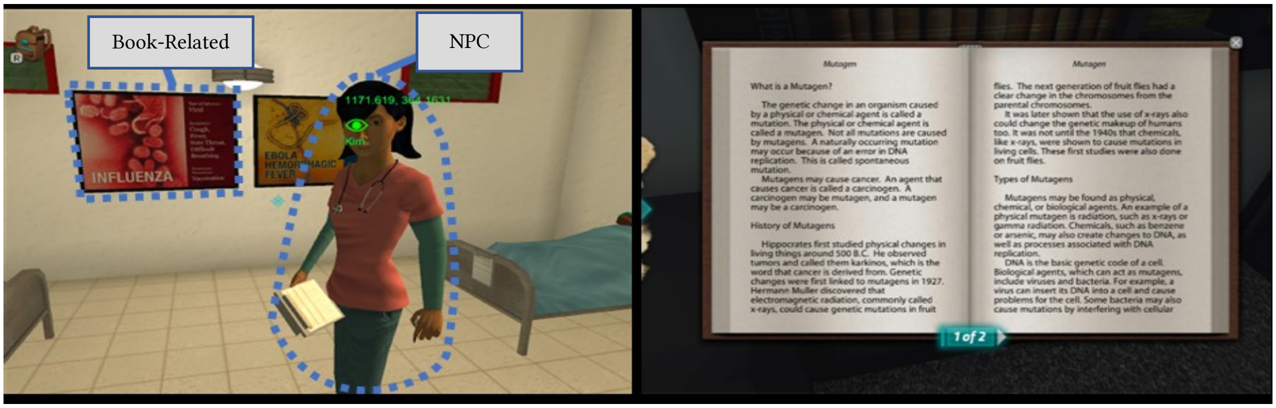

Figure 1. The CRYSTAL ISLAND learning environment with gaze entity categories (left) and book related content (right).

solving performance during the game-based problem-solving episodes. In developing student models that can dynamically predict problem-solving performance, we investigate the capability of gaze-enhanced student models to predict student problem-solving effectiveness and efficiency throughout learning interactions, depicting the fruitfulness of the learning experience.

2 RELATED WORK

Gaze holds considerable promise for guiding user-adaptive interactions. Because movements of the human eye may indicate attention, engagement, and motivation [2, 13, 36], they may signal cognitive states such as foci of attention [10] as well as mind wandering, and thus offer significant potential for designing "attention-adaptive" learning environments [19-21, 28, 29]. Gaze may also be used to recognize off-task behavior and disengagement. For example, Mills et al. used reading patterns to predict a severe form of disengagement, quitting [30]. Gaze may also be able to play an important role in user-adaptive systems that consider learners' metacognitive processes [11, 22, 23, 27, 42] and assess or predict learning [6, 24]. While we only explore gaze, prior work has explored affect and gameplay interactions for this problem [3, 17, 42, 43, 45].

Gaze has successfully been applied to student goal recognition in game-based learning environments [33]. Deep learning, and more specifically sequence-based recurrent neural networks using eye-tracking features, has emerged as a powerful student modeling technique for goal recognition. This line of investigation has shown that using multimodal data, including eye tracking, can yield high predictive accuracy for student goal recognition.

Building on these developments, the work reported here explores how gaze data can be used to improve student modeling. By exploiting gaze transition data streams-in contrast to saccade patterns previously explored, which offer a potentially complementary source of gaze information-for improving student modeling, we explore how gaze-enhanced student models can improve the accuracy of dynamically predicting student problem-solving effectiveness and student problem-solving efficiency. This work can inform future research that builds internal student models and use gaze as an indicator of cognitive states.

3 GAME-BASED LEARNING ENVIRONMENT TESTBED

To investigate gaze-enhanced student modeling, we conducted a study with college students interacting with the CRYSTAL ISLAND game-based learning environment for microbiology education [39] (Figure 1). When students interact with CRYSTAL ISLAND, they embark on a mission to solve a mysterious illness outbreak by collecting evidence and testing hypotheses. Students first arrive at the central research camp and travel to buildings such as the infirmary, dining hall, virtual laboratory, and living quarters. At each location, students interact with non-player characters (NPCs). Students speak with a variety of NPCs to collect evidence and obtain guidance. Two of the NPCs are domain experts in bacteria and viruses, allowing the students to gather information through dialogue. Students can read text resources distributed throughout the island including books, articles, and posters, which they use to learn about the potential diseases causing the outbreak. All student actions including navigation, dialogue, and text resource interactions are recorded in log files.

4 METHODS AND DATA

To evaluate the potential contribution of gaze information to student modeling, we compare two approaches. First, we instrumented CRYSTAL ISLAND with eye-tracking and introduced a real-time gaze-driven entity tracker (described in Section 4.4) to monitor the in-game objects of students' focus on a moment-to-moment basis. We then created a gaze-enhanced student model that uses student gaze data stream information together with students' goal orientation [15], gameplay time, and prior knowledge (as assessed with a pre-test) to predict student problem-solving performance. Second, we created a baseline student model that uses all information in the gaze-enhanced student model but does not have access to any gaze information. We compared the performance of the two student models on predictions of student problem-solving effectiveness and student problem-solving efficiency.

4.1 Participants and Experimental Set-Up

During the study, 65 college students interacted with the CRYSTAL ISLAND game-based learning environment in a controlled setting. Three students were removed because they were missing key pieces of data, which resulted in 62 students (*M* = 20.0 years old, *SD* = 1.57). In this group, 42 subjects (68%) were female. Each of the students played CRYSTAL ISLAND until successfully solving the mystery, with game times ranging from a minimum of 39.7 minutes to a maximum of 170.3 minutes (*M* = 81.3, *SD* = 22.8). Before playing the game, each student completed the 12-question Achievement Goal Questionnaire to measure their goal orientation [16]. The students also took a pre-test preceding their interaction with CRYSTAL ISLAND to assess their prior knowledge and a post-test following gameplay.

4.2 Goal Orientation

Because gameplay in CRYSTAL ISLAND features goal-oriented problem-solving activities, we used students' goal orientation to inform the student model. To measure goal orientation, prior to gameplay each student was asked to complete the Achievement Goal Questionnaire (AGQ) to measure their goal orientation [16], which is represented with a 2x2 matrix (Table 1).

A student may exhibit traits from all four categories. Each competency is scored on a scale of 1 to 5. The self-reported measure includes four subscales with averages for Mastery-Approach (*M* = 3.98, *SD* = 0.53), Mastery-Avoidance (*M* = 3.42, *SD* = 0.83), Performance-Approach (*M* = 3.47, *SD* = 0.80), and Performance-Avoidance (*M* = 3.43, *SD* = 0.85). We standardize each of these categories to a unit normal distribution to allow comparison among the four subscales of the AGQ.

Table 1. 2x2 Framework for goal orientation construct.

		Definition	
		Mastery	*Performance*
Valence	*Approach (towards success)*	Mastery-Approach Goal	Performance-Approach Goal
	Avoidance (of failure)	Mastery-Avoidance Goal	Performance-Avoidance Goal

4.3 Problem-Solving Performance

We designed the gaze-enhanced student model and baseline student model to predict real-time student problem-solving performance. Rather than designing student models to predict end-of-session metrics such as post-test scores, we sought to devise models that dynamically predict student in-game problem-solving performance as this family of models could yield actionable information to inform the user-adaptive tailoring of gameplay and scaffolding. To measure students' in-game problem-solving performance in CRYSTAL ISLAND, the actions they take, and the timing of those actions are used to compute their problem-solving performance scores at a given moment.

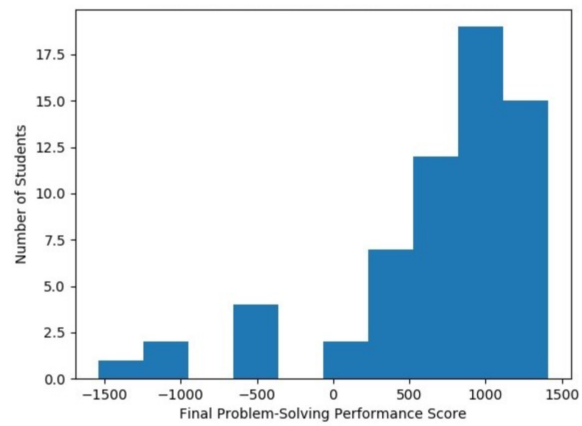

Figure 2. Problem-solving performance distribution.

Performance scores increase or decrease as students make progress (or fail to make progress) in solving the science mystery. The game-based problem-solving performance metric is parameterized on action types and elapsed time (Table 2). For example, a correct solution of the CRYSTAL ISLAND science mystery yields (+500) points, whereas scans of incorrect objects are penalized up to 35 points if the student is also scanning for the wrong contaminant. As such, there is usually an increase in the

Table 2. Student game-based problem-solving performance metric.

Action	Points (pts)
Overall Mystery Solution	
Correct Solution	500 pts
Solution Efficiency	(7500 / elapsed mins) pts
Incorrect Solution Attempt	-100 pts
In-game Quiz Questions	
First Attempt Correct	25 pts
Second Attempt Correct	10 pts
Second Attempt Incorrect	-10 pts
Object Contaminant Testing	
Correct Object and Correct Contaminant	200 pts
Incorrect Object and Correct Contaminant	15 pts
Correct Object and Incorrect Contaminant	-15 pts
Incorrect Object and Incorrect Contaminant	-35 pts
Character Interactions	
Talk to Kim	(25 / elapsed mins) pts
Talk to Teresa	(50 / elapsed mins) pts
Talk to Ford	(125 / elapsed mins) pts
Talk to Robert	(125 / elapsed mins) pts
Talk to Quentin	(125 / elapsed mins) pts
Total Maximum Points	**1665 pts**

score at the conclusion of gameplay if the student correctly solves the mystery. The metric thereby provides a real-time assessment of two facets of student problem solving: problem-solving effectiveness and problem-solving efficiency. It should also be noted that while the problem-solving performance measure was developed to assess student problem solving, it has also been used as a factor in measuring in-game student engagement with the problem-solving scenario [39].

In this student gameplay data, the final problem-solving scores have a mean of 679.9 and standard deviation of 608.80. We note that the score can be a negative value; the minimum score achieved was a -1542.73, and the maximum score was 1413.17.

Figure 2 displays a histogram of the problem-solving performance scores. Previous work has shown problem-solving performance to be significantly correlated with learning [38] and a marginally significant correlation between problem-solving performance and normalized learning gain was observed in this study ($r(60) = 0.271$ $p = 0.033$).

4.4 Eye Tracking and Gaze-based Entity Tracking

To provide eye gaze data stream information to the gaze-enhanced student model, students' gaze was tracked during gameplay using the SMI RED 250 eye tracker, which was mounted on a desktop (Figure 3). As students interacted with CRYSTAL ISLAND, the system pinpointed the coordinates on the screen at each timestamp where the student was gazing to log gaze data. Eye movements were tracked at 120 Hz, and following established conventions [38], a "fixation" is operationalized as engaging in a sustained gaze for a minimal threshold of 250 milliseconds.

The gaze-enhanced student model utilizes a gaze-driven entity tracking system that we have incorporated into the CRYSTAL ISLAND game-based learning environment. The gaze-driven entity tracker automatically detects which in-game objects the student is fixating on by analyzing the angle and gaze point on the screen by using ray casting to detect the intersection with specific in-game objects. It operates in real-time to generate a gaze data stream from synchronized sequences of data representing each fixation event, the in-game object that is the subject of the fixation, and the duration of the fixation. The gaze-enhanced student model uses these three elements to compute proportions for multiple categories of gaze objects, and it uses the fixation duration to compute total fixation time per student.

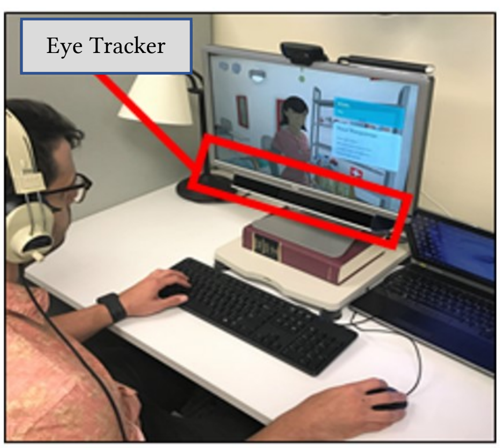

Figure 3. Gaze-instrumented CRYSTAL ISLAND environment.

5 RESULTS

We compared the predictive accuracies of the gaze-enhanced student model with the baseline student model to explore the potential additive value of gaze. The gaze-enhanced student model used gaze data in addition to students' goal orientation, gameplay time, and prior knowledge, while the baseline model had access to the same information sources as the gaze-enhanced student model except for gaze. The gaze-enhanced student model used 15 features (9 gaze proportions and 6 baseline features), while the baseline model only had access to the 6 baseline features: 4 goal-orientation subscales, pre-test score, and gameplay duration.

We conducted two sets of evaluations to compare the performance of the two student models. To begin, we compare the baseline model and gaze-enhanced model using data generated from full gameplay sessions (Baseline Full-Gameplay Model and Gaze-Enhanced Full-Gameplay Model). We then compare the baseline model and gaze-enhanced model on interval-based models that make dynamic predictions throughout gameplay from incrementally available data (Baseline Interval-Based Model and Gaze-Enhanced Interval-Based Model). In these comparisons, we evaluate the dynamic predictive capabilities of these models by using incrementally available cumulative gameplay to predict the problem-solving performance after discrete time intervals. Specifically, we investigate the use of percentage gameplay versus constant time-elapsed intervals.

The evaluation uses two machine learning frameworks for student modeling: an L2-regularized linear regression model (Ridge Regression) and an ensemble partitioning method (Random Forests). We compare the use of Ridge Regression and Random Forests for each feature set and each gameplay accumulation setup. Ridge Regression provides a simple, interpretable, and regularized method for predicting a continuous outcome from continuous variables. Random Forests are better suited to avoid overfitting due to their ensemble nature, which is a concern when presented with relatively small datasets. These models were chosen because of their ability to prevent overfitting. To evaluate each model, we use leave-one-out cross-validation (LOOCV). In this context, we are leaving one student out of each iteration to create the test set. When examining each model's performance, we evaluate the model test set R^2. We also examine the mean absolute error (MAE) for each model. R^2 is a way to determine the model's relative fit, and it provides a quantified evaluation of explained variance (using held-out data) versus total variance in the data. MAE is advantageous with respect to interpretability because it determines the average error of the model predictions.

5.1 Baseline Full-Gameplay Models

Baseline models include 6 features. Out of these, 5 are known beforehand (4 AGQ values, pretest score). The remaining value is the total time the student spent playing the game. This can be an important factor in predicting problem-solving performance because depending on students' pre-test score and goal orientation, a longer game time could indicate a lack of self-regulatory skills, spending a disproportionate amount of time acquiring knowledge, or using less efficient learning strategies. For Ridge Regression, we set up and evaluate the models by standardizing the input features and reporting the most significant features using the standard error of coefficients to calculate the t-statistic under the null hypothesis that the coefficient is equal to zero. We also report both the R^2 and MAE.

In addition to using Ridge Regression (RR), we also use a Random Forest (RF), where we use MAE as the criterion for deciding the quality split. We report leave one out cross validation R^2 and MAE for this model as well. The "feature importance" of each feature is also presented for each RF model. This accounts for the sum of the decision tree splits that include the given feature in proportion to the number of students that it splits. Baseline student model results are shown in Tables 3, 4, and 5.

Table 3. Overall baseline model LOOCV accuracy results.

	R^2	MAE
Baseline RR	0.225	372.248
Baseline RF	0.140	409.517

Table 4. Baseline Ridge Regression model.

Feature	Coefficient (B)	STD Error (B)	β	p-value
Constant	391.165	621.592	3.88e-16	1.0
Mastery-Approach	232.162	139.094	0.146	0.885
Performance-Approach	-101.717	126.289	-0.141	0.890
Mastery-Avoidance	110.104	86.213	0.139	0.891
Performance-Avoidance	73.270	116.963	0.0834	0.935
Game Time	-24.757	1.001	-0.5839	0.564
Pre-Test Score	25.624	25.599	0.0614	0.952

Ridge Regression was conducted on the baseline data to determine significant attributes. The model coefficients are significantly different from a null model, $F(6,55) = 15.221$ ($p < 0.001$) with an R^2 of 0.398 and an adjusted R^2 of 0.332.

We note that we are able to achieve an R^2 of 0.225 for the Ridge Regression model with the baseline feature set. With this performance, it is clear that the most significant feature is gameplay duration (Game Time). There are two plausible explanations for this result. A longer gameplay duration could indicate there is a lack of understanding of the game content or less efficient problem-solving strategies, which would negatively affect their problem-solving performance. Alternatively, the student may be exploring the environment or performing other off-task behaviors unrelated to solving the science mystery.

Table 5. Random Forest results for baseline feature set.

Feature	Feature Importance
Mastery-Approach	0.0828
Performance-Approach	0.0928
Mastery-Avoidance	0.136
Performance-Avoidance	0.0828
Game Time	0.422
Pre-Test Score	0.184

5.2 Gaze-Enhanced Full-Gameplay Models

Next, we investigated gaze-enhanced student models that extend the baseline models with gaze data streams. We augmented the baseline model with category-based gaze patterns representing the sequence of categories of in-game objects that were the focus of students' fixations. Specifically, gaze-enhanced student models use gaze pertaining to in-game objects in the following categories: non-player characters (NPCs), Travel/Game Items, Food-Related, Lab-Related, Diagnosis-Related, Book-Related, Concept-Matrix-Related, Miscellaneous, and Fixations-per-second. "Book Related" refers to the material that the students read throughout the game in order to gather information; "Concept-Matrix-Related" refers to the in-game testing the students complete after reading scientific content; "Travel/Game Items" refers to objects within the game that are related to transitioning between locations; "Lab Related", "Diagnosis Related", and "Food Related" are objects within the game that are relevant to solving the mystery (e.g., pieces of evidence towards a hypothesis, equipment to test hypotheses). "Miscellaneous" in this context encompasses game-related objects that are not associated with game content, such as the heads-up display, settings menu, and achievement panel. The "Fixations per second" feature quantifies the student's general fixation pattern. These categories were chosen to group specific game world objects identified by the gaze-driven entity tracker into higher-level game-based learning objects.

Table 6. Overall gaze-enhanced model LOOCV accuracy.

	R^2	MAE
Gaze RR	0.361	343.051
Gaze RF	0.453	326.968
PCA RR	0.212	373.339
PCA RF	0.389	330.102

We perform dimensionality reduction on these features to determine a transformed, reduced set of features that helps remove subjective bias of the categories, noise from a high dimensionality relative to the size of the data, and multicollinearity among the features. We used principal component analysis (PCA) with 5 components and then used these new orthogonal features in Ridge Regression and Random Forest gaze-enhanced student models for comparison. Tables 6, 7, and 8, display the results from the overall gaze-enhanced models and the PCA reduced models.

As performed previously, Ridge Regression was performed on the gaze-enhanced data to determine if there were significant attributes. The model coefficients are significantly different from a null model, $F(15,46) = 9.096$ ($p < 0.001$) with an R^2 of 0.662 and an adjusted R^2 of 0.553.

Table 7. Gaze-enhanced Ridge Regression model.

Feature	Coefficient (B)	STD Error (B)	β	p-value
Constant	337.381	1.86e4	4.44e-16	1.0
NPCs	-101.452	2.19e4	-0.206	0.858
Travel/Game Items	-274.562	1.87e4	-0.0937	0.968
Food-Related	9.996	1.97e4	0.0795	0.942
Lab-Related	-394.812	1.87e4	-0.341	0.839
Diagnosis-Related	-137.246	1.87e4	-0.203	0.872
Miscellaneous	170.129	2.00e4	-0.0853	0.956
Book-Related	839.123	1.86e4	0.135	0.955
Concept-Matrix-Related	216.927	1.84e4	0.0933	0.947
Fixations/Second	-36.119	6.96e2	-0.0299	0.976
Mastery-Approach	221.771	1.49e2	0.120	0.906
Performance-Approach	-92.090	1.34e2	-0.0384	0.970
Mastery-Avoidance	87.125	9.84e1	0.0271	0.979
Performance-Avoidance	76.264	1.29e2	-7.82e-3	0.994
Game Time	-24.790	1.00	-0.598	0.556
Pre-Test Score	26.564	2.70e1	0.152	0.880

The gaze-enhanced full-gameplay models were the best performing models. With respect to contribution to problem-solving performance prediction, we note that "Lab Related," "Diagnosis Related," and "Book Related" are of interest. The RF regressor model found these to be strong elements in predicting student performance according to feature importance. The RR model also found these to be important features (although not significant), as well as "Travel/Game Items." The negative impact of "Lab Related" fixations could be due to the fact that students are awarded points based on whether their testing is correct. Thus, the longer a student spends fixated on testing equipment in-game, could indicate they are performing additional, incorrect tests.

The negative impact of fixation proportion of "Lab Related" is reinforced by the fact the Ridge Regression model found a negative coefficient, while the Random Forest Regressor found this feature to be the most important feature. "Book Related" fixations are important since book material within the environment is a primary source of how the student acquires content learning. Perhaps the longer students are fixating on these content items, the more relevant scientific information they acquire, which might enable them to solve the science mystery more efficiently. The negative impact of "Travel/Game Items" on problem-solving performance could indicate that a student performed more off-task behavior, as travel and game-related objects have little relevance to solving the mystery.

Table 8. Random Forest results for gaze-enhanced models.

Feature	Feature Importance
NPCs	0.0278
Travel/Game Items	0.0409
Food Related	0.0144
Lab Related	0.207
Diagnosis Related	0.124
Miscellaneous	0.110
Book Related	0.120
Concept Matrix Related	0.0232
Fixations/Sec	0.0367
Mastery-Approach	0.0268
Performance-Approach	0.0262
Mastery-Avoidance	0.0161
Performance-Avoidance	0.00927
Game Time	0.125
Pre-Test Score	0.0935

Another explanation could be that the students have reached an impasse with respect to understanding the content, and they could be navigating widely to try out many different ideas. Previous work has shown that large amounts of time spent in irrelevant locations in the game may indicate off-task behavior, as there are no relevant materials in these locations [41]. Another interesting result is the fact that the duration of "Game Time" is considered to be a strong predictor of problem-solving performance. This could be attributed to the fact that students who are spending longer times in the environment are either (1) not fully engaged in the game and take longer to complete it or (2) do not fully grasp the material as quickly as other students and must devote more time to gain this level of understanding. It is also worth noting that problem-solving performance considers efficiency, which is based on game duration.

Another interesting result from the Ridge Regression model evaluation is the fact that "Mastery-Approach" is weighted so highly. This poses an important question. Are students who exhibit this goal orientation potentially better performers in this type of environment? While difficult to determine, we note that each of the other goal orientation categories were weighted much lower in magnitude. Mastery-Approach students might strive to comprehend the material in its entirety and not leave the learning environment without gaining a firm understanding of these materials. In CRYSTAL ISLAND, an example of this phenomenon would be a student striving to understand a particular disease or microbiology concept at a deep level.

5.3 Interval-Based Models

The evaluations of Full-Gameplay student models shed light on the additive diagnostic value of gaze for problem-solving performance in *toto*, and we are particularly interested in investigating the additive diagnostic value of gaze for student models that are to operate dynamically and only have access to gaze data and gameplay data that have been produced before the current moment in the gameplay. A user-adaptive learning environment could exploit such models at runtime to make in-game adaptations dynamically. The objective of informing runtime adaptations motivates the exploration of interval-based student models that predict students' problem-solving performance in a cumulative fashion. We seek to design student models that can accurately predict students' problem-solving performance scores in cumulative intervals using the same feature sets as the non-interval-based (i.e., Full-Gameplay) models.

The baseline feature set containing goal orientation values are static throughout gameplay since these are calculated from a pre-game survey. However, the gaze features will be dynamic throughout gameplay. For example, if an interval of the gameplay contains predominantly "Book Related" objects, this proportion will be very high relative to the other gaze attributes for this period. Table 9 shows that the gaze-enhanced models showed a predictive accuracy improvement of 62.3% over the baseline based on R^2 using the RF model.

We can distinguish two alternative approaches to interval-based models. First, a percentage of total time played up to each point (i.e., 10%, 20%, ... 100%) could be adopted, or second, a constant-time approach (i.e., 1 minute, 2 minutes, ... total minutes played) could be used. We chose the constant-time approach as it supports the use of a more standard time period for each student. If we chose the percentage-based approach, then each student would likely have very different amounts of gameplay in each segment. It should be noted that using a constant time approach also presents the challenge of having a different number of time segments per student, which we address below. A constant time approach is more realistic for real-time prediction, and the results shown below in Table 9 use 1-minute intervals. Thus, predictions were made each minute during the student's gameplay, using all student actions and gaze behavior up to the current prediction.

The best performing models for using the cumulative interval-based gameplay models were the gaze-enhanced models. The Random Forest regressors performed well, but we note that as expected the general performance of these models was not as high as the Full-Game models. This could be because students generally proceed through different phases in the CRYSTAL ISLAND game. For example, the tutorial will feature different objects than latter parts of the game, thus creating a difference in gaze proportions between intervals. Early predictions may therefore be more inaccurate than later gameplay predictions, and because of the dynamic nature of student problem-solving in game-based learning, the models will improve over time as they observe additional gameplay for each student.

Table 9. Summary results for interval-based models.

	R^2	MAE
Gaze RR	0.00599	202.613
Gaze RF	0.227	161.617

5.4 Convergence of Interval-Based Models

A desirable characteristic of dynamic models is continual improvement as they observe more gameplay. As a student progresses through the game, this would manifest as the student model achieving increases in predictive accuracy, which could enable user-adaptive learning environment to better adapt to the student's needs as indicated by changes to predicted problem-solving scores. Below, in Figure 4, we can see that the best models from the cumulative interval-based gameplay evaluation improve as they observe additional gameplay.

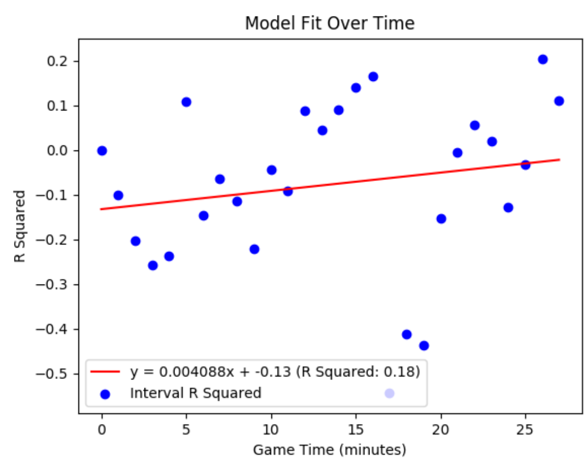

Figure 4. Random Forest model accuracy over time.

It should be noted that for purposes of evaluation, we use the predictions up to the time of the shortest student gameplay to guarantee that we have a prediction value for each student. As we can see, the R^2 value starts off very unstable, and then appears to approach a positive, improving value. We believe the reason for the less accurate early fit is that much of the time early in the game is unrelated to problem-solving performance, e.g., becoming familiar with the environment.

In addition, there are different stages of the game for which different types of objects may be present. Thus, significant features the model finds may not be present at certain early stages of the game. After an initial period, the model fit improves continually, suggesting that the model is useful in predicting student problem-solving performance over time as it observes more student gameplay and gaze behavior.

6 DISCUSSION

The results of this evaluation demonstrate that gaze-enhanced student models achieve higher predictive accuracy on student problem-solving performance than baseline student models. After evaluating the baseline student models on the entire gameplay data for each student, we find that we significantly improve these models by augmenting them with student gaze patterns. Specifically, the gaze-enhanced student models use gaze proportions with respect to categories of objects within the game. Notably, the strongest predictor categories of objects within the game were "Book Related," "Lab Related," and "Diagnosis Related" materials. "Travel/Game Items" were also informative, which could be due to off-task behavior [41]. For "Book Related" objects, this could be generalized to learning science content within the game. Many game-based learning environments feature learning content in one form or another, and CRYSTAL ISLAND uses books, articles, and posters (reading material) to convey relevant microbiology and science content to students. If students are fixated upon these types of objects for a larger proportion of time, it could indicate that they are spending additional time using sophisticated cognitive strategies (e.g., making inferences) to understand material deeply, or it could mean that they are trying to absorb a large amount of science related material using a combination of accurate metacognitive monitoring and cognitive strategies. Either way, it would seem "Book Related" material has a positive effect on their problem-solving performance. Examining more fine-grained gaze information for this category, e.g., saccades while reading, offers a promising direction for future work, as it has been shown that various patterns in reading can affect performance and learning [45].

For "Lab Related" and "Diagnosis Related" objects, we note that these can be generalized to be a form of applying scientific reasoning processes associated with generating hypotheses, collecting relevant evidence, and confirming or contradicting hypotheses through testing evidence. In terms of this specific learning environment, a higher proportion of time could indicate that a student is spending more time testing items they believe are related to solving the mysterious outbreak. At a high level, this would appear to have a positive effect on one's performance. However, there are ways a student could "game the system" by exhibiting "guess-and-check" problem-solving behavior to simply test every single item to determine the correct answer, without learning the material or engaging in scientific reasoning [39]. Thus, the prediction model seems to be able to identify students pursuing a "guess-and-check" strategy.

It is possible that students who spend more time testing items are not grasping concepts as well (e.g., differences between viruses and bacteria) and are perhaps testing irrelevant evidence. In fact, it would be logical for a student who efficiently determines the solution to the science mystery to have a very low proportion of their time fixating upon these test related objects, since they would use the test related areas of the environment less after confirming their hypothesis. This is perhaps related to self-regulated learning in that students who employ certain strategies may have a particular learner profile [8].

The fixation object categories were chosen to create generalizable, common categories that are broadly applicable. Data-driven methods of aggregating specific fixation items into higher-level categories, such as sparse autoencoders, should be explored to increase the generalizability of fixation models to other game-based learning environments. To address the concern of generalizability, dimensionality reduction (PCA) was used to calculate the components that maximize the total variance of the data. We chose 5 components, and the performance of the PCA components within the same prediction models outperformed the baseline. However, because a side effect of dimensionality reduction is losing interpretability of the original features, more interpretable aggregation techniques should be investigated.

7 CONCLUSION

With rapid improvements in eye-tracking technologies, incorporating gaze data streams into student modeling offers considerable promise for creating more robust student models, which in turn can yield user-adaptive learning environments that are more effective and engaging. To explore the potential of gaze for student modeling, we created gaze-enhanced student models that use student gaze fixation transitions to predict student problem-solving performance in a game-based learning environment. We conducted a comparative evaluation of the predictive accuracy of the gaze-enhanced student models relative to baseline models and found that the gaze-enhanced student models significantly outperformed baseline student models that did not have access to gaze data streams. The results also demonstrate that gaze-enhanced student models more accurately predict student problem-solving performance in both static contexts (when full-session gameplay data is made available) and in dynamic contexts (when predictions must be made incrementally in an interval-based fashion).

Gaze-based student modeling can inform dynamic user-adaptations, and it will be important to investigate this in future work. Two additional lines of investigation that are also promising are to explore sequence-based models that directly represent the temporal dimension of fixations, and multimodal student models that further extend gaze-enhanced models with additional modalities such as affect and posture to yield even higher predictive accuracies and support more effective user-adaptation.

Acknowledgements

We would like to thank our group's collaborators at the Center for Educational Informatics and the SMART Lab at North Carolina State University. This study was supported by funding from the Social Sciences and Humanities Research Council of Canada. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Social Sciences and Humanities Research Council of Canada.

REFERENCES

[1] Alvarez, N., Sanchez-Ruiz, A., Cavazza, M., Shigematsu, M. and Prendinger, H. 2015. Narrative Balance Management in an Intelligent Biosafety Training Application for Improving User Performance. *International Journal of Artificial Intelligence in Education*. 25, 1, 35–59.

[2] Baker, R. 2007. Modeling and Understanding Students' Off-Task Behavior in Intelligent Tutoring Systems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.*, 1059.

[3] Baker, R., Corbett, A. and Aleven, V. 2008. More Accurate Student Modeling through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. *Intelligent Tutoring Systems.*, 406–415.

[4] Baker, R., K. D'Mello, S., Mercedes T. Rodrigo, M. and Graesser, A. 2010. Better to be Frustrated than Bored: The Incidence, Persistence, and Impact of Learners' Cognitive Affective States during Interactions with three Different Computer-Based Learning Environments. *International Journal of Human-Computer Studies*. 68, 4, 223–241.

[5] Bergey, B., Ketelhut, D.J., Liang, S., Natarajan, U. and Karakus, M. 2015. Scientific Inquiry Self-Efficacy and Computer Game Self-Efficacy as Predictors and Outcomes of Middle School Boys' and Girls' Performance in a Science Assessment in a Virtual Environment. *Journal of Science Education and Technology*. 24, 5, 696–708.

[6] Bondareva, D., Conati, C., Feyzi-Behnagh, R., Harley, J.M., Azevedo, R. and Bouchet, F. 2013. Inferring Learning from Gaze Data during Interaction with an Environment to Support Self-Regulated Learning. *Proceedings of the 16th International Conference on Artificial Intelligence in Education*. 229–238.

[7] Botelho, A., Baker, R. and Heffernan, N. 2017. Improving Sensor-Free Affect Detection Using Deep Learning. *Proceedings of the International Conference on Artificial Intelligence in Education*. 40–51.

[8] Bouchet, F., Harley, J.M., Trevors, G. and Azevedo, R. 2013. Clustering and Profiling Students According to their Interactions with an Intelligent Tutoring System Fostering Self-Regulated Learning. *Journal of Educational Data Mining*. 5, 1, 104–146.

[9] Conati, C., Gertner, A. and VanLehn, K. 2002. Using Bayesian Networks to Manage Uncertainty in Student Modeling. *User Modeling and User-Adapted Interaction*. 12, 4, 371–417.

[10] Conati, C., Jaques, N. and Muir, M. 2013. Understanding attention to adaptive hints in educational games: an eye-tracking study. *International Journal of Artificial Intelligence in Education*. 23, 1–4, 136–161.

[11] Conati, C. and Merten, C. 2007. Eye-Tracking for User Modeling in Exploratory Learning Environments: an Empirical Evaluation. *Knowledge-Based Systems*. 20, 6, 557–574.

[12] Desmarais, M. and Baker, R. 2012. A Review of Recent Advances in Learner and Skill Modeling in Intelligent Learning Environments. *User Modeling and User-Adapted Interaction*. 22, 1, 9–38.

[13] Deubel, H. and Schneider, W. 1996. Saccade Target Selection and Object Recognition: Evidence for a Common Attentional Mechanism. *Vision Research*. 36, 12, 1827–1837.

[14] Easterday, M., Aleven, V., Scheines, R. and Carver, S. 2017. Using Tutors to Improve Educational Games: A Cognitive Game for Policy Argument. *Journal of the Learning Sciences*. 26, 2, 226–256.

[15] Elliot, A. and McGregor, H. 2001. A 2 × 2 achievement goal framework. *Journal of Personality and Social Psychology*. 80, 3, 501–519.

[16] Elliot, A.J. and Murayama, K. 2008. On the Measurement of Achievement Goals: Critique, Illustration, and Application. *Journal of Educational Psychology*. 100, 3, 613–628.

[17] Grafsgaard, J., Wiggins, J., Boyer, K., Wiebe, E. and Lester, J. 2014. Predicting Learning and Affect from Multimodal Data Streams in Task-Oriented Tutorial Dialogue. *Proceedings of the 7th International Conference on Educational Data Mining.*, 122–129.

[18] Habgood, J. and Ainsworth, S. 2011. Motivating Children to Learn Effectively: Exploring the Value of Intrinsic Integration in Educational Games. *The Journal of the Learning Sciences*. 20, 2, 169–206.

[19] Huang, C.-M., Andrist, S., Sauppé, A. and Mutlu, B. 2015. Using gaze patterns to predict task intent in collaboration. *Frontiers in Psychology*. 6, 1–12.

[20] Hutt, S., Hardey, J., Bixler, R., Stewart, A., Risko, E. and D'Mello, S. 2017. Gaze-based Detection of Mind Wandering during Lecture Viewing. *10th International Conference on Educational Data Mining.* , 226–231.

[21] Hutt, S., Mills, C., Bosch, N., Krasich, K., Brockmole, J. and D'Mello, S. 2017. Out of the Fr-"Eye"-ing Pan. *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*. 94–103.

[22] Hutt, S., Mills, C., White, S., Donnelly, P. and D 'Mello, S. 2016. The Eyes Have It: Gaze-based Detection of Mind Wandering during Learning with an Intelligent Tutoring System. *Proceedings of the 9th International Conference on Educational Data Mining. International Educational Data Mining Society*. 86–93.

[23] Jaques, N., Conati, C., Harley, J. and Azevedo, R. 2014. Predicting Affect from Gaze Data During Interaction with an Intelligent Tutoring System. *Intelligent Tutoring Systems*. 29–38.

[24] Kardan, S. and Conati, C. 2012. Exploring Gaze Data for Determining User Learning with an Interactive Simulation. *Proceedings of 20th International Conference on User Modeling, Adaptation, and Personalization*. 126–138.

[25] Khajah, M., V. Lindsey, R. and Mozer, M. 2016. How deep is knowledge tracing? *Proceedings of the Ninth International Conference on Educational Data Mining*. 94–101.

[26] Lee, S.J., Liu, Y. and Popovi, Z. 2014. Learning Individual Behavior in an Educational Game : A Data-Driven Approach. *Proceedings of the 7th International Conference on Educational Data Mining*. 114–121.

[27] M. McLaren, B., M. Adams, D., Mayer, R. and Forlizzi, J. 2017. A Computer-Based Game that Promotes Mathematics Learning More than a Conventional Approach. *International Journal of Game-Based Learning*. 7, 1, 36–56.

[28] Merten, C. and Conati, C. 2006. Eye-Tracking to Model and Adapt to User Meta-cognition in Intelligent Learning Environments. *Proceedings of the 11th International Conference on Intelligent User Interfaces.* 39.

[29] Mills, C., Bixler, R., Wang, X. and D'Mello, S. 2016. Automatic Gaze-Based Detection of Mind Wandering during Narrative Film Comprehension. *Proceedings of the 9th International Conference on Educational Data Mining.* 30–37.

[30] Mills, C., Bosch, N., Graesser, A. and D'Mello, S. 2014. To Quit or Not to Quit: Predicting Future Behavioral Disengagement from Reading Patterns. *Proceedings of the 12th International Conference on Intelligent Tutoring Systems.* 19–28.

[31] Min, W., Ha, E.Y., Rowe, J., Mott, B. and Lester, J. 2014. Deep Learning-Based Goal Recognition in Open-Ended Digital Games. *Proceedings of the Tenth Annual Conference on Artificial Intelligence and Interactive Digital Entertainment.* 37–43.

[32] Min, W., Mott, B., Rowe, J., Liu, B. and Lester, J. 2016. Player Goal Recognition in Open-World Digital Games with Long Short-Term Memory Networks. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence.* 2590–2596.

[33] Min, W., Mott, B., Rowe, J., Taylor, R., Wiebe, E., Boyer, K. and Lester, J. 2017. Multimodal Goal Recognition in Open-World Digital Games. *Proceedings of the Thirteenth Annual AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment.* 80–86.

[34] Peterson, J., Pardos, Z., Rau, M., Swigart, A., Gerber, C. and McKinsey, J. 2015. Understanding Student Success in Chemistry Using Gaze Tracking and Pupillometry. *Proceedings of the 17th International Conference on Artificial Intelligence in Education.* 883.

[35] Piech, C., Spencer, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. and Sohl-Dickstein, J. 2015. Deep Knowledge Tracing. *Advances in Neural Information Processing Systems.* 505–513.

[36] Randall, J., Oswald, F. and Beier, M. 2014. Mind-Wandering, Cognition, and Performance: A Theory-Driven Meta-Analysis of Attention Regulation. *Psychological Bulletin.* 140, 6, 1411–1431.

[37] Raptis, G., Katsini, C., Belk, M., Fidas, C., Samaras, G. and Avouris, N. 2017. Using Eye Gaze Data and Visual Activities to Infer Human Cognitive Styles: Method and Feasibility Studies. *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization.* 164–173.

[38] Rayner, K. 1998. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin.* 124, 3, 372–422.

[39] Rowe, J., Shores, L., Mott, B. and Lester, J. 2011. Integrating Learning, problem Solving, and Engagement in Narrative-Centered Learning Environments. *International Journal of Artificial Intelligence in Education.* 21, 1–2, 115–133.

[40] Sabourin, J. and Lester, J. 2014. Affect and Engagement in Game-Based Learning Environments. *IEEE Transactions on Affective Computing.* 5, 1, 45–56.

[41] Sabourin, J., Rowe, J., Mott, B. and Lester, J. 2011. When Off-Task is On-Task: The Affective Role of Off-Task Behavior in Narrative-Centered Learning Environments. *Proceedings of the Fifteenth International Conference on Artificial Intelligence in Education.*, 534–536.

[42] Sabourin, J.L. and Lester, J.C. 2014. Affect and Engagement in Game-Based Learning Environments. *IEEE Transactions on Affective Computing.* 5, 1, 45–56.

[43] Sawyer, R., Smith, A., Rowe, J., Azevedo, R. and Lester, J. 2017. Enhancing Student Models in Game-based Learning with Facial Expression Recognition. *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, 1–10.

[44] Slanzi, G., Balazs, J.A. and Velásquez, J.D. 2017. Combining eye tracking, pupil dilation and EEG analysis for predicting web users click intention. *Information Fusion.* 35, 51–57.

[45] Taub, M., Mudrick, N., Azevedo, R., Millar, G., Rowe, J. and Lester, J. 2017. Using multi-channel data with multi-level modeling to assess in-game performance during gameplay with CRYSTAL ISLAND. *Computers in Human Behavior.* 76, 641–655.

[46] Vail, A., Grafsgaard, J., Boyer, K.E., Wiebe, E. and Lester, J. 2016. Gender Differences in Facial Expressions of Affect During Learning. *Proceedings of the Twenty-Fourth Conference on User Modeling, Adaptation, and Personalization.* 65–73.

[47] Wiburg, K., Chamberlin, B., Valdez, A., Trujillo, K. and Stanford, T. 2016. Impact of Math Snacks Games on Students' Conceptual Understanding. *Journal of Computers in Mathematics and Science Teaching.* 35, 2, 173–193.