The Utilization of Artificial Intelligence in Evaluating Student Outcomes over Online Learning of Mathematics

Eesha Sinha, Janani Sekar

Eesha Sinha, Menlo Atherton High School, Atherton, California
Janani Sekar, Inspirit AI, Palo Alto, California

Student Authors

Eesha Sinha, High School

ABSTRACT

Online math learning is rapidly increasing in popularity for its ability to engage students (Olajumoke et al 2025); however, there is no consensus on its efficacy and most important factors (Hickman et al 2022). This study utilizes data from the ASSISTments online learning platform to explore the effectiveness of online math learning through classification models, such as Logistic Regression models and a FeedForward Neural Network using Scikit-learn's Multi-Layer Perceptron (MLP) Classifier (E-TRIALS, 2015). We hypothesized that a Neural Network would be the most effective in predicting student outcomes and finding patterns to optimize online learning because of its ability to process complex data. The dataset included 30 features about the learning platform setting and student decision-making at that stage; our models focus on a subset of these features to learn. Example features include "skill ID," which labels different math skills such as exponents or rounding, and "answer type," which categorizes the type of math, such as algebra. The results suggested the best model was a Neural Network with two hidden layers, using nine features including tutor mode, hint count, attempt count, and response time. We report an accuracy of .9408, precision score of .9292, recall score of .9882, F1 score of .9578, and ROC/AUC of .9142. Our model suggests that attempt count and hint count are the most influential factors. Based on these findings, platforms that include multiple attempts, real-time feedback, or step-by-step guidance may be able to optimize online math learning.

Introduction

The U.S. education system currently faces major challenges. A 2023 survey of K-12 public school teachers found that more than 80% of the respondents believed the education system has declined compared to five years ago (PEW Research Center, 2023). This suggests that teachers face challenges to meet students' needs. At the same time, as society becomes more dependent on technology, schools are increasingly incorporating electronic devices into classrooms to help support learning (Schindler et al 2017). These shifts signify the need to discover or improve methods to foster education. One promising approach is online learning, through platforms like ASSISTments. ASSISTments is a free online learning platform where students and teachers get immediate feedback as students progress. Teachers can use this information to find where students are stuck and tailor future lessons, making learning more efficient and personalized (Heffernan et al 2014). However, fully understanding which features of online learning are most influential on student learning remains complex (Prabowo et al 2022).

This research aims to assess the effectiveness of online learning through a machine learning analysis of data from the 2015 ASSISTments Skill Builder Data. The primary objective is to identify which features are most closely linked to successful learning. This will offer insights to guide the design of more effective educational tools.

Data is analyzed by comparing in-game behaviors with performance outcomes using classification models. Key performance metrics, such as accuracy, AUC, precision, recall, and f1 score, found through logistic regression and neural networks, are used to measure the models' accuracy. Results are verified through these metrics as well as comparisons to previous results, such as those outlined in the paper *Prediction of Confusion Attempting Algebra Homework in an Intelligent Tutoring System through Machine Learning Techniques for Educational Sustainable Development* by Abidi and others.

Background

Previous studies have established the effectiveness of online learning, even showing that active engagement with digital educational games enhances students' motivation to learn. Research indicates a positive relationship between digital educational games and students' motivation, mediated through their learning engagement (Li et al. 2024). In Jeon et al 2023, the authors applied a multi-layer perceptron (MLP) model on the Jo Wilder dataset. Jo Wilder is an online educational game that helps students learn history skills, this dataset was part of a Kaggle competition. They achieved an F1 score of .83 and an accuracy of .74, outperforming other models. Their analysis underscored the importance of data preprocessing and large datasets to improve predictive performance (Jeon et al. 2023). Abidi and others analyzed the ASSISTments 2009-2010 skill builder dataset, similar to the dataset we are using, and applied a random forest model. They achieved an accuracy of .86 and focused on predicting confused students' outcomes.

However, other studies contrast the findings above. In one experiment, learning math online through iPads and digital games failed to make a difference between pre and post-assessments (Carr, 2012). Additionally, a study conducted on the impact of Khan Academy on public schools in Brazil found that online learning improved attitudes toward math but did not significantly increase test scores (Ferman et al 2019). The study builds on previous studies by looking at specific features of the online learning platform and determining the most influential factors in learning math.

The 2015 ASSISTments Skill Builder Data was developed by the E-TRAILS platform to research and explore the science of learning (ASSISTments 2019). In our study, we try to predict which questions students answer correctly, based on metadata about the question, such as its topic, difficulty, and the amount of time the student spent on the question. The binary target variable — *correct* — indicates if the student completed the question, and this variable is measured throughout the game to evaluate learning. The ASSISTments platform is designed such that when a student gets a question correct, they move on to the next question. This dataset consists of students using the skill builder tool of ASSISTments, where they are assigned certain math skills by teachers. They practice problems within that skill until they master it before progressing to other skills (Heffernan et al. 2014).

This paper aims to contribute to the growing understanding of online math learning and what factors of platforms have the most effect. We found that “attempt count” and “hint count” are the most influential factors. This signifies that online math platforms should include multiple attempts and real time hints for the user to optimize learning.

Dataset

The dataset used in this study is from the ASSISTments online learning platform, sourced from E-TRIALS (ASSISTments, 2015). The data comes from 19,917 students from all over the United States and they are mostly in 7th or 8th grade. This dataset provides important insights into how students interact with and learn from online educational resources by capturing specific student interactions, such as problem attempts, hint count, response times, problem type, and more. There are 708,631 interactions across 100 math skills, such as percentages and slope.

After transferring the dataset into Google Colab, we cleaned the data by dropping any rows that were missing the target variable (“correct”). Then we converted the response time feature from milliseconds to seconds. Additionally, we encoded categorical features like "tutor_mode," "answer_type," "type," and "skill_id" into numerical values using LabelEncoder and selected a subset of features, including both numerical and categorical columns, and defined the target variable as "correct." Lastly, we standardized the numerical features and randomly split the data into training and testing sets, with 80% used to train and 20% for testing.

Feature	**Description**	**Mean**	**Median**
tutor_mode	Mode of tutoring (tutor, test, pre-test, post-test)	N/A	N/A
answer_type	Type of answer (multiple choice, short answer)	N/A	N/A
type	Type of problem	N/A	N/A

skill_id	ID of skill associated with the problem	N/A	N/A
attempt_count	Number of attempts made by the student on the problem	1.502	1
response_time_sec	Time in seconds it took the student to respond	47.992	20.791
hint_count	Number of hints requested by the student	0.441	0
hint_total	Total number of hints available for the problem	2.343	3
opportunity	Number of opportunities the student has to practice this skill	79.017	12

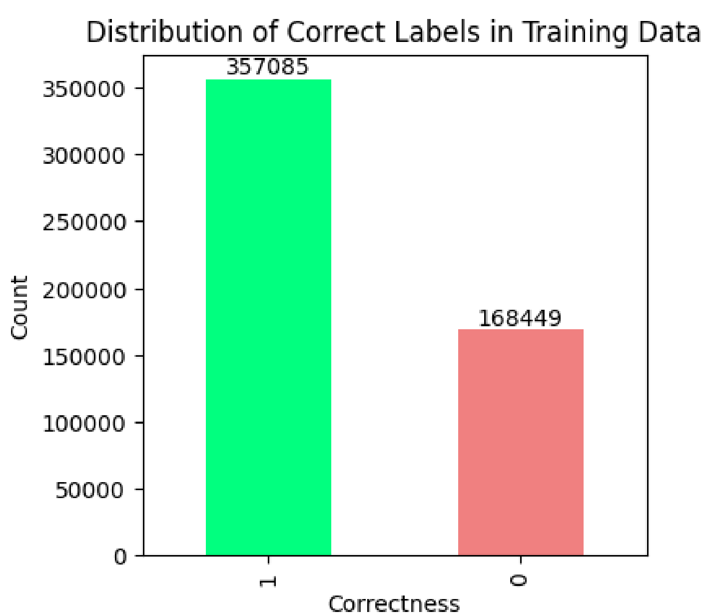

Figure 1. Shows the distribution of correct and incorrect labels in the training data

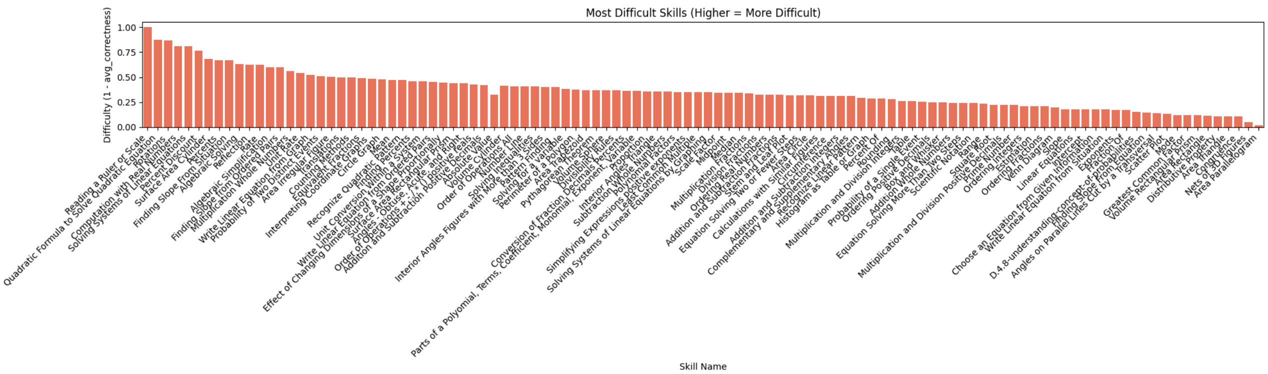

Figure 2. Shows most difficult skills calculated by subtracting the average difficulty by 1, therefore the higher the bar the more difficult

Figure 3. Shows the count of each math skill tested in the dataset

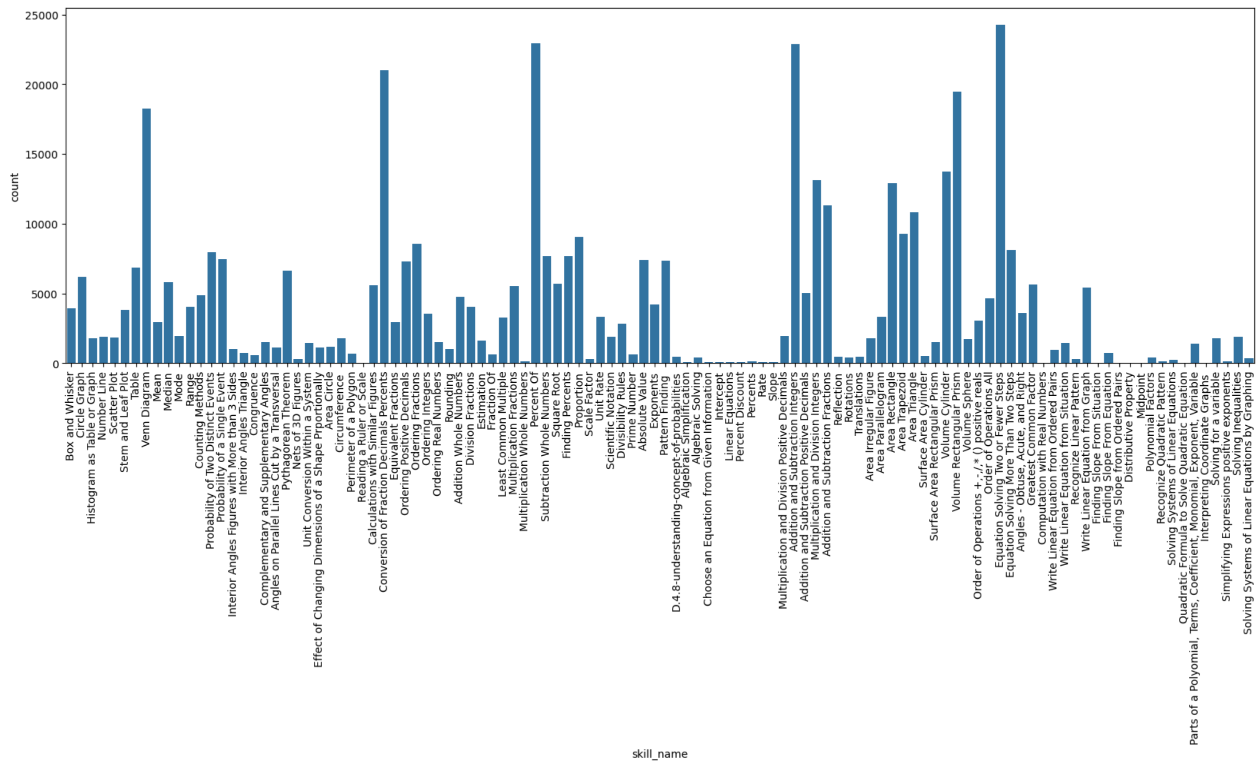

Figure 4. Shows the distribution of correct and incorrect labels across each math skill

<u>Methodology/Models</u>

The primary features of the dataset that we focused on are “tutor mode”, “answer type”, “type”, “skill id”, “attempt count”, “response time”, “hint count”, “hint total”, and “opportunity.” These inputs were used to predict whether a student answered a question correctly. Since this is a classification exercise, we applied logistic regression and neural networks to predict student performance based on these features. To assess the effectiveness of our models, we used several metrics: accuracy, precision, recall, F1 score, and ROC/AUC. All training and testing were conducted on Google Colab.

We first created several basic logistic regression models, experimenting with different combinations of features to identify the most influential features. Then, we added L1 and L2 regularization to the models to remove unhelpful features and compared those results to the baseline models.

We then also used a neural network based on the features named before. The network consisted of two hidden layers with 64 and 32 neurons, each using ReLU activation functions. The output layer used a sigmoid activation function to produce a probabilistic prediction of correct vs. incorrect. To try and obtain better results, we started hyperparameter tuning. We adjusted the number of hidden layers to five with 128, 64, 32, 16, and 8 neurons. We also changed the features to just "attempt count", "response time", and "tutor mode" to the two layer neural network.

To see which features affected the accuracy the most, we applied permutation importance on the neural network models.

Results and Discussion

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
LR (all selected features)	0.9129	0.8896	0.9954	0.9395	0.8668
Neural Network w/ 2 layers (all selected features)	0.9408	0.9292	0.9882	0.9578	0.9142
LR L1 Regularization (all selected features)	0.8879	0.8496	0.9969	0.9173	0.8114
LR L2 Regularization (all selected features)	0.8777	0.8493	0.9969	0.9172	0.8111
LR (tutor_mode, attempt_count, response_time_sec)	0.8629	0.8344	0.9958	0.9080	0.7885
Neural Network w/ 2 layers (tutor_mode, attempt_count, response_time_count)	0.8932	0.8691	0.9922	0.9266	0.8377
Neural Network w/ 5 layers (all selected features)	0.9405	0.9268	0.9906	0.9577	0.9124

The results between the logistic regression models show that a basic logistic regression model with all the selected features achieved the highest metrics, with an accuracy of .9129, precision of .8896, recall of .9954, F1 score of .9395, and ROC/AUC of .8668. However, the neural network with two hidden layers with all the selected features performed the best. It attained an accuracy of .9408, precision of .9292, recall .9882, F1 score .9578, and ROC/AUC of .9142, but the neural network with five layers had almost identical results. Throughout all the models, the recall was very high, almost one, meaning the models successfully predicted almost all the instances when a student answered a question correctly. The best model was the neural network with two hidden layers that used all the selected features. The most influential features that contributed to the accuracy of this model were attempt_count and hint_count (see Figure 7).

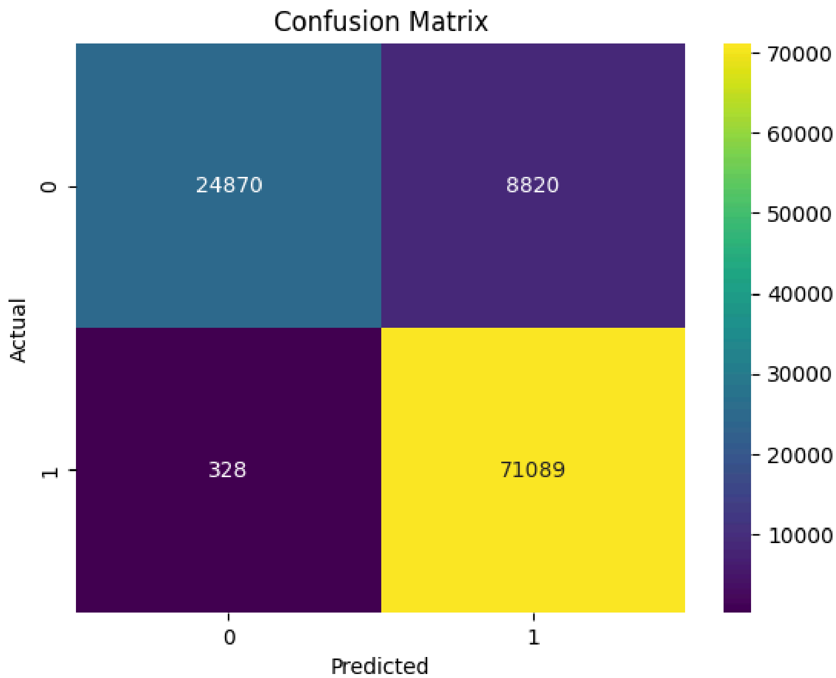

Figure 5. The confusion matrix of the basic logistic regression model using all the selected features

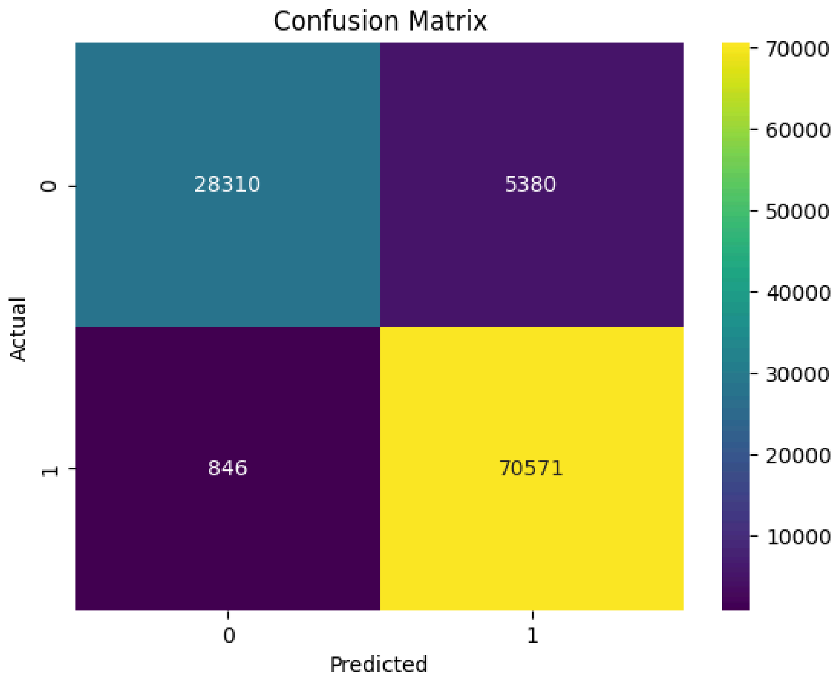

Figure 6. The confusion matrix of the neural network with two hidden layers using all the selected features

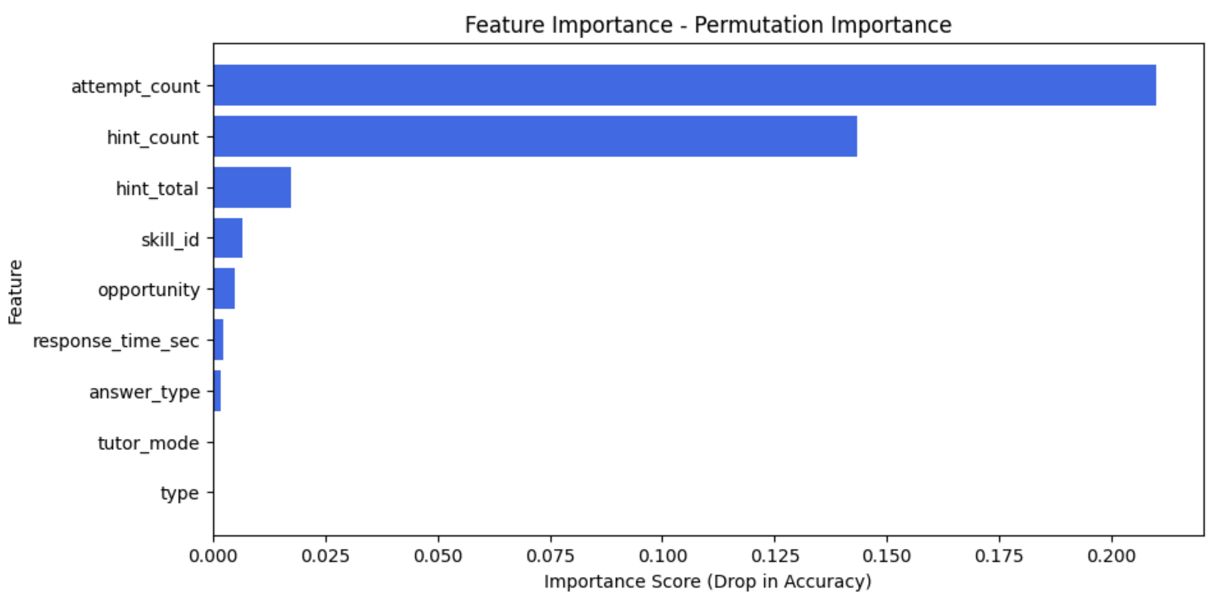

Figure 7. The permutation importance of the features of the best-performing neural network shows attempt_count and hint_count as being the most influential. The longer the bar, the more importance it has.

Discussion and Conclusions

In this project, we created a classification model that predicts student performance on different math skills on the ASSISTments online learning platform. The ASSISTments 2015 Skill Builder dataset has been used in previous research. One of the limitations of this project is we could only load a fixed dataset due to the limited RAM available, so we could not use a larger dataset. Additionally, the dataset did not include features such as gender, specific grade, race, or information about the teacher. Another interesting factor we could explore in the future is the learning environment, such as at school or home. Moreover, we were unable to analyze the visual or auditory elements of the ASSISTments platform, limiting our understanding of the diverse ways students learn (e.g., auditory or visual learners). Despite these limitations, our project highlights the potential of machine learning in exploring the efficiency of the online learning of math and uncovering key factors that optimize student learning.

The key takeaway from this project is the most significant features are attempt_count and hint_count, indicating learning platforms should include multiple attempts and hints to enhance online math learning.

Acknowledgments

I would like to thank Janani Sekar for her invaluable guidance, support, and encouragement throughout this research project.

References

Li, Y., Chen, D., & Deng, X. (2024). The impact of digital educational games on student's motivation for learning: The mediating effect of learning engagement and the moderating effect of the digital environment. *PloS one*, *19*(1), e0294350.

Schindler, L. A., Burkholder, G. J., Morad, O. A., & Marsh, C. (2017). Computer-based technology and student engagement: a critical review of the literature. *International journal of educational technology in higher education*, *14*, 1-28.

Pandey, S., & Karypis, G. (2019). A self-attentive model for knowledge tracing. *arXiv preprint arXiv:1907.06837.*

Williams, J. J., Ostrow, K., Xiong, X., Glassman, E., Kim, J., Maldonado, S. G., ... & Heffernan, N. (2015, March). Using and designing platforms for in vivo educational experiments. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale* (pp. 409-412).

Heffernan, N.T., Heffernan, C.L. The ASSISTments Ecosystem: Building a Platform that Brings Scientists and Teachers Together for Minimally Invasive Research on Human Learning and Teaching. Int J Artif Intell Educ 24, 470–497 (2014). https://doi.org/10.1007/s40593-014-0024-x

Prabowo, H., Ikhsan, R. B., & Yuniarty, Y. (2022, November). Student performance in online learning higher education: A preliminary research. In *Frontiers in Education* (Vol. 7, p. 916721). Frontiers Media SA.

Wang, Y., Kai, S., & Baker, R. S. (2020). Early detection of wheel-spinning in ASSISTments. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I 21* (pp. 574-585). Springer International Publishing.

PEW RESEARCH CENTER 2023 SURVEY OF TEACHERS FINAL TOPLINE OCTOBER 17-NOVEMBER 14, 2023. (2023). https://www.pewresearch.org/social-trends/wp-content/uploads/sites/3/2024/04/ST_24.04.04_teacher-survey_topline.pdf

Olajumoke, S., & Spangenberg, E. (2025). Assessing the role of online mathematics tools in enhancing student learning and engagement. Educational Technology Quarterly. https://doi.org/10.55056/etq.874

Luo, T., Hickman, C., & Muljana, P. (2022). Examining the impact of online math games on student performance and attitudes. Old Dominion University Digital Commons. https://digitalcommons.odu.edu/cgi/viewcontent.cgi?article=1228&context=stemps_fac_pubs

E-TRIALS | ASSISTments. (2019). Assistments.org. https://new.assistments.org/e-trials

Abidi, S. M. R., Hussain, M., Xu, Y., & Zhang, W. (2019). Prediction of Confusion Attempting Algebra Homework in an Intelligent Tutoring System through Machine Learning Techniques for Educational Sustainable Development. Sustainability, 11(1), 105. https://doi.org/10.3390/su11010105

Li, Y., Chen, D., & Deng, X. (2024). The impact of digital educational games on student's motivation for learning: The mediating effect of learning engagement and the moderating effect of the digital environment. PloS one, 19(1), e0294350.

Carr, J. M. (2012). Does math achievement h'APP'en when iPads and game-based learning are incorporated into fifth-grade mathematics instruction?. Journal of information technology education: Research, 11(1), 269-286.

Ferman, B., Finamor, L., & Lima, L. (2019). Are Public Schools Ready to Integrate Math Classes with Khan Academy?

Jeon, H., He, H., Wang, A., & Spooner, S. (2023). Modeling Student Performance in Game-Based Learning Environments. arXiv preprint arXiv:2309.13429.

E-TRIALS (n.d.). Publications Using Public Datasets. Etrialstestbed.org. https://www.etrialstestbed.org/resources/featured-studies/dataset-papers