MATHstream and UpGrade: Using Rapid, Large-Scale Experimentation for Data-Driven Improvements in a Digital Learning Tool

Tyree Cowell, April Murphy, Saumya Mehta, Unekwu-Ojo Shaibu, Rae Bastoni, Stephen E. Fancsali, & Steve Ritter

A/B Testing
Digital Learning
Digital Tools
learning engineering
Mathematics Education
MATHstream
Rapid Cycle Experimentation

MATHstream is an innovative digital learning tool from Carnegie Learning, a leading provider of K-12 EdTech solutions, used by over 30,000 middle and high school students in the 2023/24 school year. MATHstream provides supplemental math instruction through an engaging video platform. Throughout the development of this product, our team has been able to leverage UpGrade, a free

and open-source A/B testing platform designed specifically with educational use cases in mind, in our approach to learning engineering. Integrating MATHstream with UpGrade allows us to conduct large-scale field experiments directly in the platform, facilitating rapid deployment and data collection for different learning experiences, to better understand what works best to improve learning outcomes.

Introduction

MATHstream is an innovative digital learning tool from Carnegie Learning, a leading provider of K-12 EdTech solutions, currently used by over 30,000 middle and high school students in the 2023/24 school year. MATHstream provides supplemental math instruction through an engaging video platform. The platform drives engagement by leveraging “rock star math teachers” with large followings on various social media platforms, interspersing assessment items throughout instructional videos, and applying gamified elements such as coins and badges. Throughout the development of this product, our team has been able to leverage UpGrade (Ritter et al., 2020, 2022), a free and open-source A/B testing platform designed specifically with educational use cases in mind. Integrating MATHstream with UpGrade allows us to engage in the cyclical process of Learning Engineering including creation, implementation and investigation (Kessler, et. al, 2022). To this end, we conduct large-scale field experiments directly in the platform, facilitating rapid deployment and data collection of different learning experiences, to better understand what works best to improve learning outcomes.

In what follows, we consider two types of field experiments we’ve conducted on MATHstream using UpGrade, with detailed examples of each. The first type of experiment - what we call “content experiments” - allow us to improve specific pieces of content (including video content, assessment items, and hints, among others) and rigorously evaluate the impact of proposed changes on learning outcomes. The second type of experiments - what we call “feature experiments” - allow us to test new product or instructional features with a randomized sub-population of MATHstream users to ascertain the effects of those features on learning and/or user experience.

Over the past two years, our team has been able to launch over 10 different experiments in support of rapid, iterative processes to find out what works best for learners using MATHstream, reaching over 20,000 students. These experiments often involve cutting edge technology and techniques such as generative AI video, the use of large language models

(LLMs) to rewrite text content, and the introduction of new instructional features based on feedback from students themselves. We explore these examples to demonstrate how rapid, iterative experimentation using UpGrade can help make evidence-based product decisions and improve the learning experience for thousands of students.

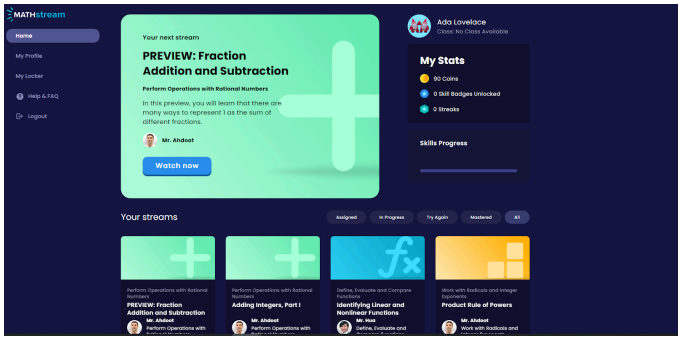
The Platforms: MATHstream and UpGrade

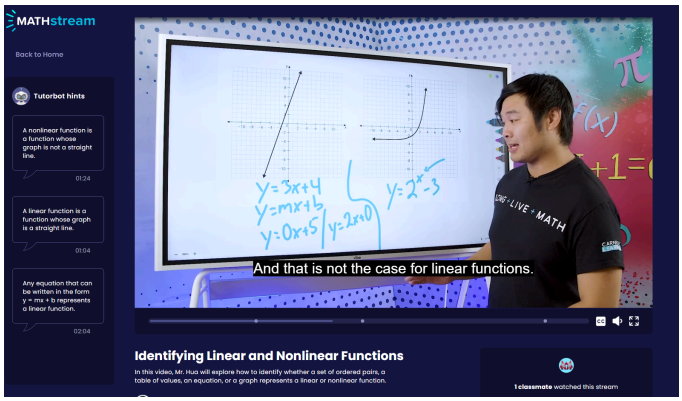
MATHstream: A digital learning tool

MATHstream provides supplemental math instruction through an engaging video platform. It is currently used by over 30,000 middle school students in the 2023/24 school year. The initial vision for MATHstream was to help fill learning gaps left by COVID-19 (Moscovitz & Evans, 2022) and the current teacher shortage in the United States (Sutcher et. al, 2019), but the platform has evolved into a resource widely used to support both supplemental and core math instruction. The platform drives engagement by leveraging “rock star” math teachers with large followings on various social media platforms, interspersing assessment items throughout instructional videos, and applying gamified elements such as coins and badges.

Figure 1

Sample screenshots of the MATHstream platform, including the landing page for students (top) and the stream player (bottom)

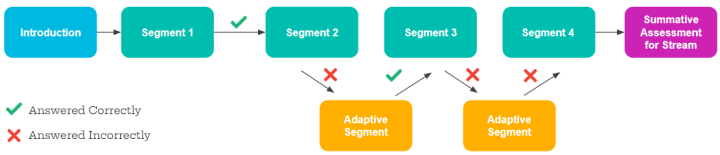




A typical flow for a student using MATHstream is as follows: After logging in and selecting an assigned stream to watch, the student watches a few minutes of engaging, instructional video content before encountering an in-stream assessment item. If the student answers that question correctly, they return to the main video segment and continue watching. If a student answers incorrectly, they may be directed to an "adaptive segment" - a brief video segment providing a step-by-step walkthrough of the incorrectly answered question. After watching the adaptive segment, students will typically get another chance to answer a similar question before returning to the main video. Students may encounter adaptive segments 2-3 times during a stream before an end-of-stream checkpoint, which typically consist of 5 summative assessment questions. Student performance on all these assessment items are used to determine their level of proficiency in the content of the stream, which is then made available to their teacher via our detailed teacher reports system. Figure 2 below demonstrates a usage flow similar to the one described above.

Figure 2

A typical workflow for a student watching a single stream on MATHstream



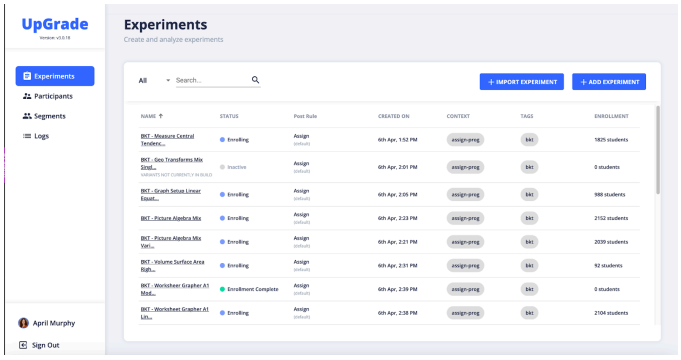
UpGrade: an A/B testing platform

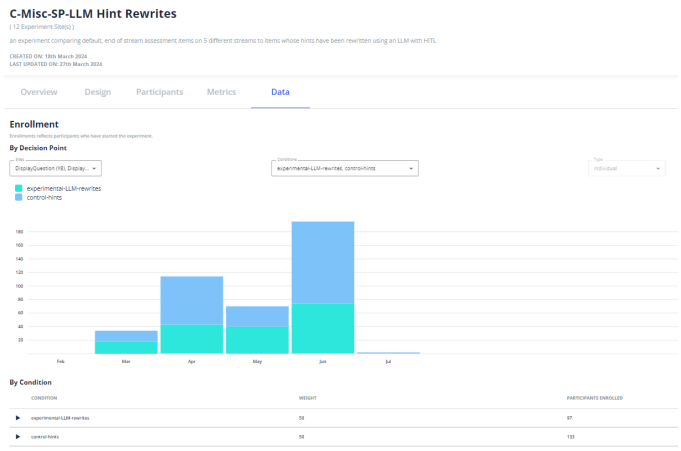
UpGrade is a free and open-source A/B testing and experimentation platform for educational software, developed by Carnegie Learning to address challenges in conducting large-scale randomized field tests in classroom environments (Ritter et. al, 2020, 2022). UpGrade allows for group-level random assignment (for example, at the class, instructor, or school level), which can help researchers maintain consistency of educational experiences when embedding field tests in software (Ritter, Murphy, & Fancsali, 2020). UpGrade handles coordinating experimental activities that are linked, and we consider how to manage anomalies that can arise, such as how to handle condition assignment and consistency if students are using adaptive or self-paced software and they reach instructional materials

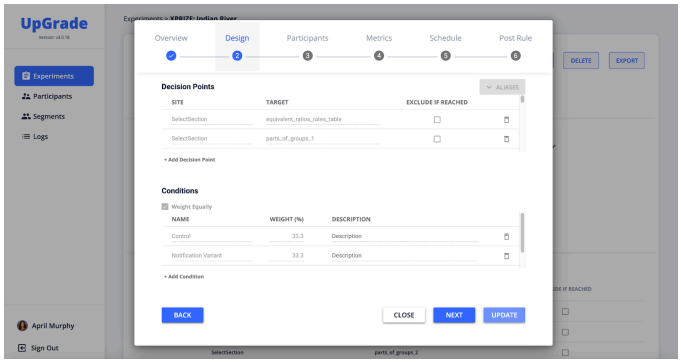
asynchronously. In addition to simple weighted randomization, UpGrade can facilitate factorial designs, within-subject experiments, and feature flags, as well as use stratified random sampling for ensuring condition balance among subgroup populations. Advanced algorithms for experimentation such as multi-armed bandits are currently in development for the platform.

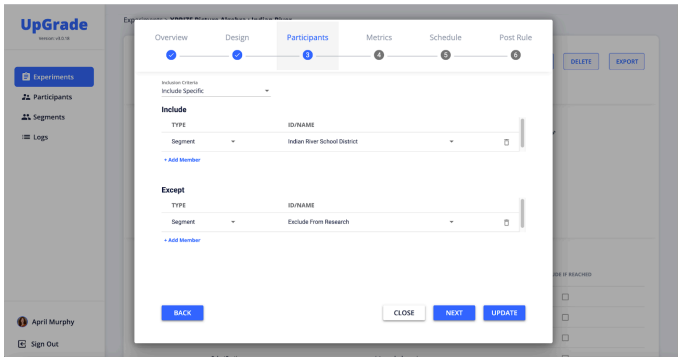
Figure 3

Sample screenshots of the UpGrade platform including (first) the UpGrade landing page listing all active experiments, (second) a results tab for a given experiment demonstrating enrollment numbers across months in each condition, (third) a screen from the set up workflow where the experimental design is implemented including decision points and conditions, and (fourth) another page from the set up flow where researchers can indicate which users to include in and which to exclude from the experiment









Content Experiments

What we call “content experiments” are a more traditional form of Learning Engineering research (Goodell, Kessler & Schatz, 2023) where instructional designers, researchers, and content experts create multiple versions of a piece of instruction or assessment (often with a different instructional approach) and serve both versions to students in an A/B test to evaluate which approach is more effective. With the UpGrade integration, we can serve variant versions of nearly any type of content in MATHstream, including main video segments, adaptive video segments, in-stream assessment items, end of stream checkpoint items, hints within assessment items, feedback provided after completion of assessment items, and more. To date, we have launched 7 different A/B tests experimenting with different versions of content, impacting over 2,500 students.

Using AI-generative video to provide feedback to students

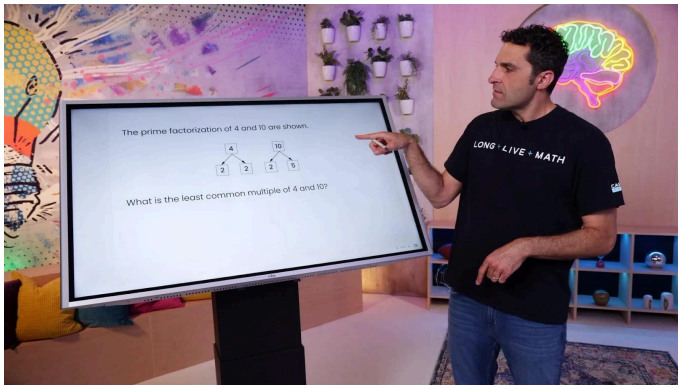
The most common form these experiments take is using AI-generative video technology to create alternate versions of adaptive segments. We imagine a future where each adaptive segment is generated in real time and responds to the specific error made by the student. Before that can happen, we must evaluate the AI video technology itself and ensure it does not have a detrimental impact on learning. To date, Carnegie Learning has worked with several vendors in the AI-generative video space (including HourOne.ai, HeyGen, and Elai.io)

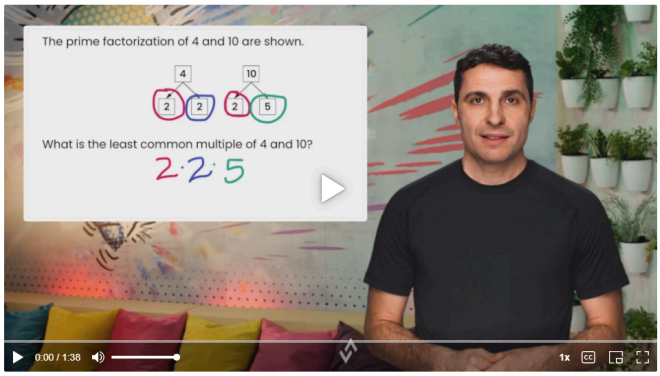
each with different processes and results. In general, these vendors have some out-of-the-box avatar options, but in order to create AI versions of our existing “rock star” math teachers like Robert Ahdoot, the founder of Yay Math! and a popular online math educator, we underwent a longer process in the studio of recording Robert from different angles and even speaking in different languages. Once the customized avatar has been built, we can use the interfaces of each vendor to generate videos with it like we would for any other out-of-the-box avatar.

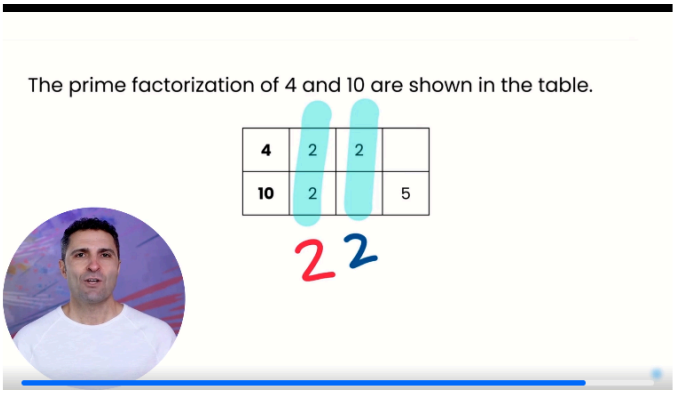
As a preliminary test of the impact of AI-video on learning, we generated AI versions of adaptive segments for several streams and compared them to the original, “human” content in an A/B experiment. We have since iterated on this style of experiment with different vendors, different content, and different avatars. Early versions of this experiment indicated that there was no significant difference in performance between students in the control vs. experimental groups, suggesting that AI video is not detrimental to learning. Supporting evidence from smaller-scale, quantitative studies have shown us that, as the technology improves, any impact on engagement and likeability is also decreasing. These findings allow us to make conscientious iterations on our implementation of AI generative videos.

Figure 4

Screenshots demonstrating the evolution of our AI generative video on a single stream teaching Least Common Multiples. Includes (first) the original, human instructor, Robert Ahdoot, (second) our first attempt at creating AI generative video using HourOne, (third) our most recent iteration of this technology using HeyGen, and (fourth) a version of this most recent iteration with a smaller likeness of the AI instructor.







Experimenting with playback speeds

A similar experiment on adaptive segments investigated the effect of playback speeds on student learning. Based on feedback from our users we learned that 1) some of the teachers spoke very quickly and students, especially ESL students, wished to slow down the video and 2) some students wished to speed up the video on their second watch-through in order to reach sections of the topic they were struggling with faster. This experiment used 3 conditions (normal speed, 1.25 speed, and 0.75 speed) to evaluate what a playback speed feature would mean for learning. We found no significant difference in performance between students in the normal speed and slow speed conditions, but found that students in the fast

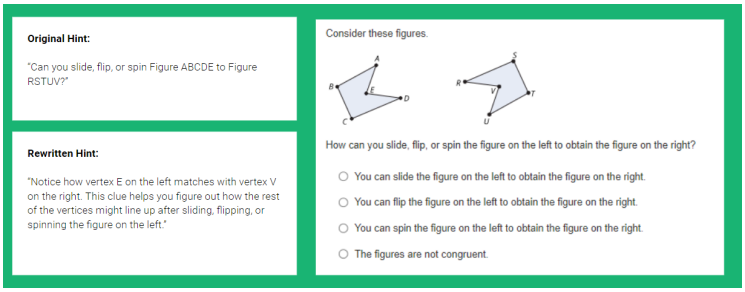
speed condition performed significantly worse on related assessment items. These findings directly support a product decision to provide students with the means to slow down a video, but not speed one up.

Using LLMs to rewrite text content

Aside from experimenting with video content, we have also begun to use UpGrade to evaluate the impact of rewriting text content in MATHstream such as hints and feedback. Learning Science research has shown that providing hints, particularly on-demand hints, in computer-assisted learning can reduce student cognitive load and lead to increased learning gains (Kessler, Stein, & Schunn, 2015; Razzaq & Heffernan, 2010). During a review of our assessment content, we identified assessment items that had lower than usual correctness rates. Our internal content experts determined that these items may benefit from improved hint content to make them more problem-specific, more verbose, and more action-oriented. In order to improve the efficiency of these rewrites, our content authoring team has leveraged large language models (LLMs) to perform the first round of rewrites. One ongoing experiment involves rewrites of hints for end-of-stream items in 5 different streams. In MATHstream, students can request up to 3 progressive hints for any given problem. The use of these hints does not affect their score on the problem, nor any measurement of their mastery of the stream. Figure 5 below provides an example of one of these rewrites. We are still collecting data on this experiment and hope to have results soon.

Figure 5

Example of rewritten content from our "Hint Rewrite" experiment. Each hint on the left is associated with the assessment item on the right. The hint on the top is the original content, while the hint on the bottom is the content that was rewritten with the help of an LLM.



Feature Experiments

In addition to the more "traditional" types of experiments described above, we have also been able to use UpGrade to launch new product features to a subset of users in order to evaluate their impact on learning and engagement. This functionality essentially works like a feature flag. When a student reaches a decision point (e.g. by logging in or launching a stream), that student is assigned to either the control condition, where the feature flag is off, or the experimental condition, where the feature flag is on. One of the benefits of these types of experiments is that because they are not tied to a particular piece of content, participant numbers are typically much higher. Of the 5 feature experiments we've run to date, two of them had over 20,000 unique participants each. The ability to quickly collect data at this

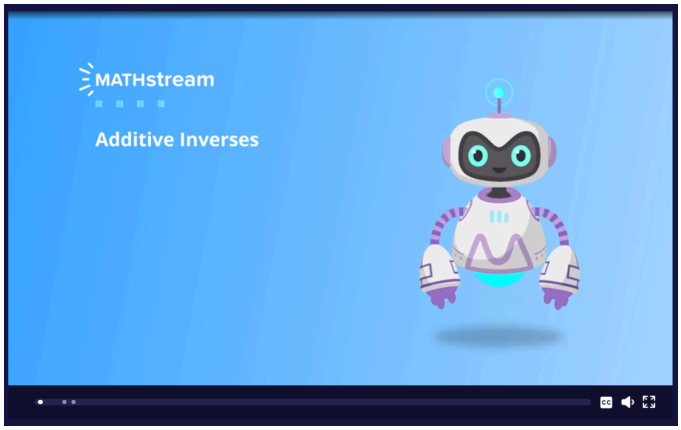
scale allows for rapid iteration cycles when making fast-paced product decisions about feature development.

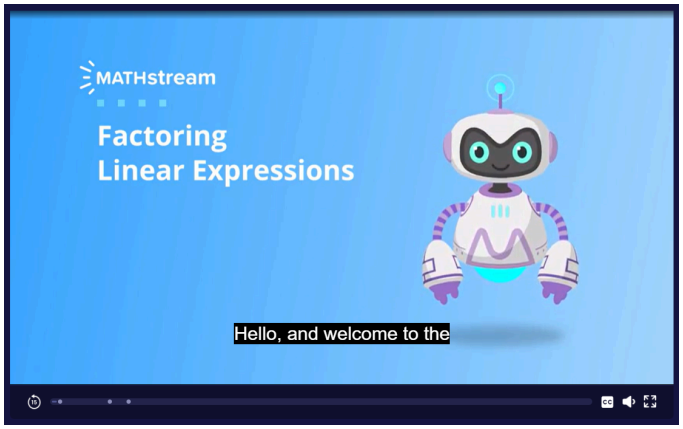
Rewind feature experiment

One such example of this experimentation process at work is the development of the rewind feature. In the original version of MATHstream, students were not able to rewind the streams - the idea being that it should be similar to a livestream experience. However, we received overwhelming feedback from students and teachers asking for the ability to rewind. In order to evaluate any impact on learning, we first released the rewind feature to half the user base (randomly assigned using UpGrade). While we saw some non-significant improvements in performance from students who had access to the rewind feature, we generally found no significant differences between the control and the experimental groups. Because this strongly suggests that access to the rewind feature was not detrimental to learning, and because it was our most requested product feature at the time, we felt comfortable releasing the feature to our entire user population.

Figure 6.

Screenshots demonstrating (top) the control condition with no rewind button in bottom left of the video player to (bottom) the experimental condition including the rewind button in the bottom left of the video player.



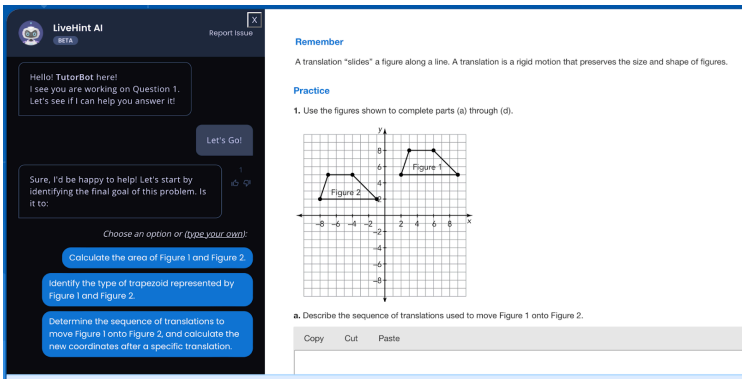


Future experimentation: LiveHint AI

Looking forward to the following school year, one feature we plan to release to certain research partners is an interactive chatbot tool called LiveHint AI. This feature was developed leveraging Carnegie Learning's 25 years of research on how students learn best. It is the first generative AI math tutor to not only understand how students approach problem solving, but also predict common mistakes they make (Fisher, et. al, 2023). Our vision is that as an alternative to requesting one of the pre-determined, progressive hints currently provided by MATHstream, students can instead open a chat window to get more specific answers to any questions they may have. Understanding that the use of generative AI in education is a topic in active discussion among some schools and families, we will make sure of UpGrade's ability to include and exclude participants at the school, classroom, and/or student level. This way, we can opt in our research partner schools but still exclude individual students within those schools who choose to opt out.

Figure 7

Screenshots of a demo of LiveHint AI, demonstrating how the bot may respond to a student asking for help on a particular assessment item.



Conclusion

Integrating the UpGrade A/B testing platform into MATHstream, a supplemental, digital learning tool has allowed us to perform rapid experimentation on instructional content and approaches, as well as on product features, to measure their impact on student learning. This integrated experimentation not only allows us to get real data from real users in real-world environments, but also allows us to make informed product decisions that provide real benefits to our users. The integration of these two technologies allow us to make progress on two of the three high-level "opportunities" of Learning Engineering described by Baker, Boser & Snow (2022). UpGrade allows us to engage in "Better Learning Engineering" by extending experimentation infrastructure, which in turn allows us to create "Better Learning Technologies" in MATHstream. This process, like all Learning Engineering, is inherently iterative and ongoing (Goodell, Kessler & Schatz, 2023), leading to multiple rounds of improvement throughout the product lifecycle as we collect data about what works best for learning. It is our goal as members of the Learning Engineering community to move away from relying on our intuitions to drive improvements. This infrastructure allows us to empirically test our intuitions and make the best, data-driven decisions possible for our students.

Acknowledgments

We would like to acknowledge the LEVI (Learning Engineering Virtual Institute) grant for funding the research detailed in these proceedings. Additionally, we would like to acknowledge the engineers responsible for creating the infrastructure necessary for this work to take place, namely Patrick McMahon, Katie Ladd, Robert Lowman, Kyle Lauffer, and Jorge Reyes. We would like to acknowledge team members who have worked in a project management or oversight role, namely Jamie Sterling and Mark Swartz. We would also like to acknowledge Robert Ahdoot, one of our rockstar MATHstreamers, for allowing us to use his likeness in our generative AI efforts and for being an enthusiastic partner.

References

Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning Engineering: A View on Where the Field Is at, Where It's Going, and the Research Needed. *Technology, Mind, and Behavior, 3*(1: Spring 2022).

Fancsali, S. E., Murphy, A., & Ritter, S. (2022). "Closing the Loop" in Educational Data Science with an Open Source Architecture for Large-Scale Field Trials. International Educational Data Mining Society.

Fisher, J., Almoubayyed, H., Fancsali, S. E., Ritter, S., De Ley, L., & Lee, Z. (2023). Building an Instructional Design–Backed, GPT-Driven AI Tutor for Math Homework Support. In *AI for Education: Bridging Innovation and Responsibility at the 38th AAAI Annual Conference on A*I.

Goodell, J., Kessler, A., & Schatz, S. (2023). Learning Engineering at a Glance. *Journal of Military Learning*. Army University Press.

Kessler, A., Craig, S. D., Goodell, J., Kurzweil, D., & Greenwald, S. (2022). Learning Engineering is A Process. In J Goodell & J. Kolodner (Eds.), *Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond.* (pp. 29-46). Routledge.

Kessler, A. M., Stein, M. K., & Schunn, C. D. (2015). Cognitive demand of model tracing tutor tasks: Conceptualizing and predicting how deeply students engage. *Technology, Knowledge and Learning, 20(3)*, 317-337.

Moscoviz, L., & Evans, D. K. (2022). Learning loss and student dropouts during the covid-19 pandemic: A review of the evidence two years after schools shut down.

Razzaq, L., & Heffernan, N. T. (2010). Hints: is it better to give or wait to be asked?. In *Intelligent Tutoring Systems: 10th International Conference, ITS 2010*, Pittsburgh, PA, USA, June 14-18, 2010, Proceedings, Part I 10 (pp. 349-358). Springer Berlin Heidelberg.

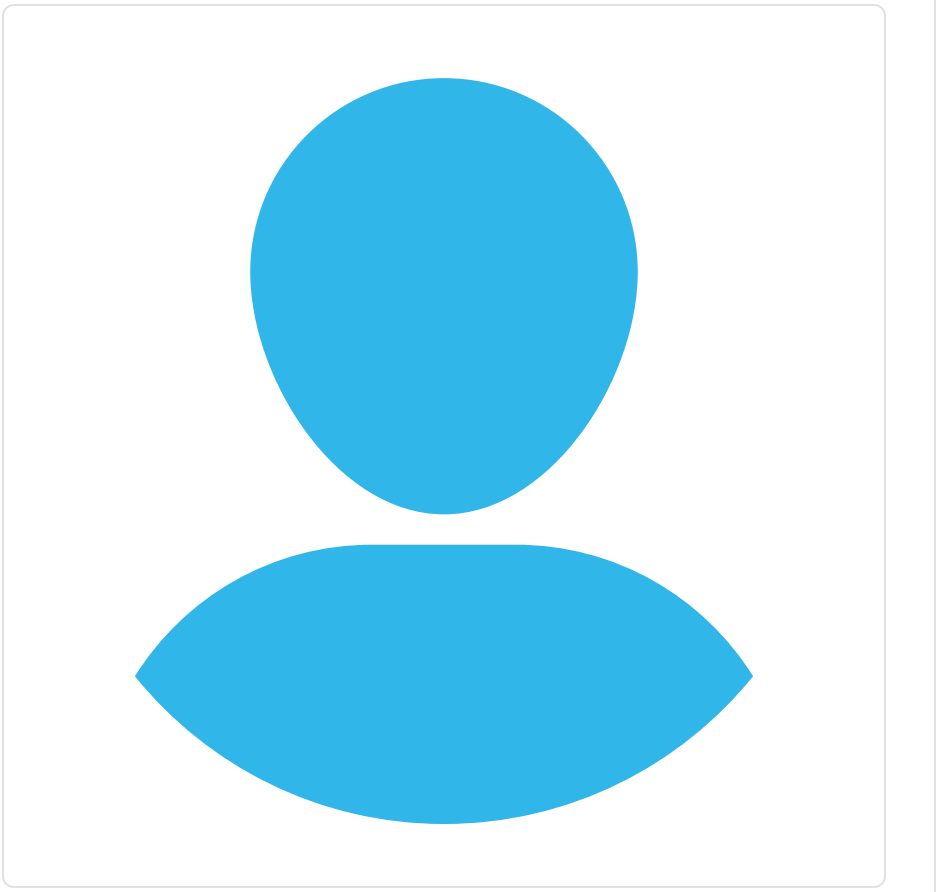
Ritter, S., Murphy, A., Fancsali, S. E., Fitkariwala, V., Patel, N., & Lomas, J. D. (2020). UpGrade: An open source tool to support A/B testing in educational software. In *Proceedings of the First Workshop on Educational A/B Testing at Scale (at Learning@ Scale 2020)*.

Ritter, S., Murphy, A., & Fancsali, S. (2020). Managing group random assignments in UpGrade. In *Proceedings of the First Workshop on Educational A/B Testing at Scale (at Learning@ Scale 2020*).

Ritter, S., Murphy, A., & Fancsali, S. (2022). Curriculum-embedded experimentation. In *Proceedings of the Third Workshop on A/B Testing and Platform-Enabled Research (at Learning@ Scale 2022)*.

Sutcher, L., Darling-Hammond, L., & Carver-Thomas, D. (2019). Understanding teacher shortages: An analysis of teacher supply and demand in the United States. *Education policy analysis archives, 27(35)*.

Tyree Cowell



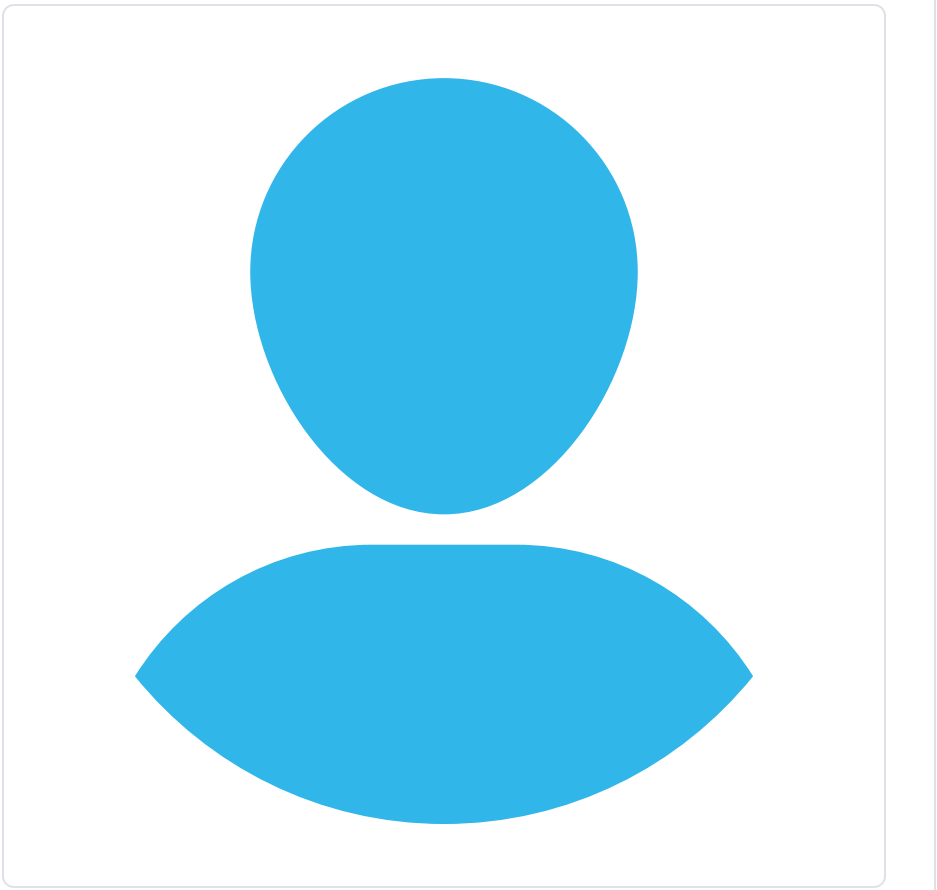
April Murphy

Saumya Mehta

Unekwu-Ojo Shaibu

Rae Bastoni

Stephen E. Fancsali
Carnegie Learning



Steve Ritter