

“Closing the Loop” in Educational Data Science with an Open Source Architecture for Large-Scale Field Trials

Stephen E. Fancsali
April Murphy
Steve Ritter
Carnegie Learning, Inc.
{sfancsali, amurphy, sritter}
@carnegielearning.com

ABSTRACT

Ten years after the announcement of the “rise of the super experiment” at Educational Data Mining 2012, challenges to implementing “internet scale” educational experiments often persist for educational technology providers, especially when they seek to test substantive instructional interventions. Studies that deploy and test interventions, when informed by data-driven modeling, are often described as “close the loop” studies. Studies that close the loop attempt to link improvements in statistical and machine learning models of learning to real-world learning outcomes. After first considering challenges to internet scale experiments, we review several educational data science/mining studies that close the loop between data-driven modeling and learning outcomes. Next, we describe UpGrade, an open source architecture that, when integrated with educational technologies, helps overcome challenges to large-scale field trials (or internet scale experiments) that close the loop between data-driven work and real-world learning outcomes. In addition to describing preliminary randomized experiments that have been conducted and will soon be conducted using the architecture in two educational technology platforms, we end with a “call for contributors and integrators.” UpGrade contributors and integrators will be researchers and developers who seek to drive continuous, data-driven improvements in real-world settings where learning with technology occurs.

Keywords

A/B testing, closing the loop, educational technology, large-scale field trials, experimentation, open source software.

1. INTRODUCTION

At Educational Data Mining 2012, Stamper et al. [18] described “the rise of the super experiment” and the Super Experiment Framework (SEF), which conceptualizes data-driven educational experimentation at the lab scale, school scale, and internet scale. Lab scale experiments may have sample sizes in the range of 1-100; school scale experiments range in sample size between tens of learners and thousands of learners, and internet scale experiments may have sample sizes ranging from thousands of learners to

S. Fancsali, A. Murphy, and S. Ritter. “closing the loop” in educational data science with an open source architecture for large-scale field trials. In A. Mitrovic and N. Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 834–838, Durham, United Kingdom, July 2022. International Educational Data Mining Society.

© 2022 Copyright is held by the author(s). This work is distributed under the Creative Commons Attribution NonCommercial NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license.
<https://doi.org/10.5281/zenodo.6852930>

millions of learners. Experiments at each scale have benefits and drawbacks, and there are important ways in which experiments at each scale can inform the design and implementation of experiments at the other scales. The authors’ presentation of the SEF concludes by laying out several key challenges for internet scale experiments. We briefly present their four posited challenges and review several (types of) educational data science/mining (EDS/EDM) studies that “close the loop” before describing the UpGrade open source architecture for conducting experiments in educational technologies at any of the SEF’s scales and how UpGrade helps educational technology developers address these challenges.

2. CHALLENGES FOR INTERNET SCALE EXPERIMENTS

Stamper et al. [18] lay out four key challenges for internet scale educational experiments. The first challenge is attracting a large user-base to the learning platform on which one would like to conduct such experiments. While an important issue, we leave the much broader discussion of large-scale adoption of educational technologies and software for learning for another day and assume that a researcher seeking to “close the loop” already is satisfied that they have a sufficiently large and diverse user-base to answer their research questions.

The second challenge they suggest, in the context of an educational game deployed via the Brainpop.com platform, is “instrumenting software for generating data logs that measure player performance, learning, and engagement” [18]. This is another broad, general challenge for developing software for learning, whether an educational game, intelligent tutoring system (ITS), or other type of learning software. Well-instrumented learning software provides insights about learners and their learning process, behaviors, engagement, and related facets of their learning experience that are the purview of nearly all work published at Educational Data Mining and related venues. We assume that readers and platform developers already recognize the importance of generating meaningful data from their learning platforms if they seek to run studies that close the loop or similar internet scale educational experiments.

Features of UpGrade target the third and fourth challenges raised by Stamper et al. [18]. The third challenge they pose is “the configuration of the software to allow for experimental designs” [18]. Being able to run experiments within a piece of software requires that different variants or instances of elements within an application’s design space can be instantiated and deployed in software, abstracted in such a way that the software can deliver

users to particular experiences based on their condition assignment. The software must also have some way of assigning users to particular experimental conditions. While variants within a target application's design space must still be created, UpGrade serves as an enabling technology to make experimental management (e.g., handling complexities of random assignment and tracking of users' condition assignments) less onerous for technology developers. One way of viewing UpGrade is as an enabling technology that makes A/B testing or experimentation a relatively simple matter of instrumenting target applications by implementing appropriate application programming interface (API) "hooks" to UpGrade rather than having to implement complex experimental management logic within the target application itself. Just as learning software developers increasingly understand the importance of instrumenting their software for high quality learning data collection, UpGrade may serve to further the goal that basic instrumentation in learning software will allow for rigorous learning engineering, involving experimentation and testing of new content, software features, instructional approaches, and other factors of interest to both researchers and technology developers.

The fourth challenge is two-fold, as Stamper et al. [18] note that "researchers increasingly face the challenge of making use of tens of thousands of subjects in an efficient manner" while also pointing out that some experiments may create inconsistent user experiences. While the authors view inconsistent user experiences as a potential catalyst for reducing overall participation, we take inconsistent user experiences as a more fundamental potential barrier to delivering internet scale experiments. Educational technology providers might reasonably refuse to deploy experiments at all if there are potential scenarios in which unsatisfying, inconsistent, or disengaging user experiences are likely to result.

Enabling technologies for large-scale educational experiments like UpGrade ought to deliver flexible options and capabilities to researchers and developers to deploy complex, substantive experiments and experimental designs in ways that still maintain high quality, consistent learning experiences. By illustrating practical examples of how studies that "close the loop" based on EDS/EDM insights might noticeably affect (or not) the learning experiences of real K-12 students and teachers, we motivate how UpGrade helps to meet many of the challenges raised by internet scale experiments.

3. "CLOSING THE LOOP"

We consider two kinds of potential "close the loop" studies drawn from literature in EDS/EDM and ITS research to provide a simple illustration of the types of considerations educational technology providers might make in delivering internet scale field trials or experiments to users in settings like K-12 classrooms. One case is intended to illustrate a situation in which a relatively simple, user-level random assignment study is unlikely to raise any concerns about consistency of learner experience while the other illustrates some potential concerns about consistency that might arise. In the section that follows, we describe important features and affordances of the UpGrade architecture in more detail to show how it enables researchers and educational technology developers to deploy experiments that address these concerns about consistency as well as provide other options for delivering high quality experiments for learning engineering.

The overarching goal of "closing the loop" with experiments is to evaluate whether observed improvements in (usually statistical) outcomes like prediction accuracy of particular models translate

into improved learning outcomes for students in a target system. Learning outcomes of interest in target systems might include efficiency of practice, time to skill mastery, or gains in performance from a pre-test measure to a post-test measure, among others.

Our example "close the loop" studies are related to how data-driven modeling of learner performance informs the specification and parameterization of so-called knowledge component (KC; or skill) [10] cognitive models frequently used within ITSs like Cognitive Tutor/MATHia [14] or in tutors built with tools like the Cognitive Tutor Authoring Tools [2].

A bevy of research in EDS/EDM and related areas (e.g., [3, 15, 19]) consider different approaches to fitting the parameters of KC models deployed in ITSs, typically within the four-parameter framework of Bayesian Knowledge Tracing [6], holding the set of KCs that appear to a learner during the learning experience fixed.

Contrast (1) studies that close the loop by contrasting *two or more sets of parameter estimates for the same KC model* with (2) studies that contrast two or more *different specifications of a KC model*, of which there are several examples in the literature (e.g., [11]). The second type of experiments are typically supported by semi-automated, data-driven techniques (e.g., Learning Factors Analysis [5]), various types of task analysis (e.g., [4]), or more recent multi-method approaches for "design loop adaptivity" [1, 8]. These techniques are used to re-evaluate the underlying KC model that drives the ITS's adaptive learning and determine how the specification of the KCs themselves might better represent what a student is learning (not just the performance parameters related to their learning) as they practice in the ITS.

In the first "parameter estimation" experiment in an ITS like MATHia, one or more experimental conditions and a control condition each have the same "skillometer" or dashboard display for students to see their progress toward KC mastery. Similarly, the same KCs or skills reported to teachers in their classroom analytics. The only difference in a parameter estimation experiment is likely to be exceedingly subtle, in that there are different parameters for subsets of KCs in each condition. In the latter "specification" experiment, the control condition and one or more experimental conditions vary in what KCs constitute the set of KCs used to drive adaptation for the topic as well as *what is displayed to students and their teachers*.

In the hypothetical parameter estimation study, individual random assignment is likely to be a reasonable choice for the researcher running the experiment. If the ITS, for example, implements some form of mastery learning (e.g., [17]), then students will be accustomed to receiving different amounts of practice on KCs they encounter within different topics in the ITS. Different parameterizations for different students in the same classroom are *not likely to lead to drastically different perceptions of the learning experience* for students. Nevertheless, if the experiment is successful, one or more parameterizations may lead to more efficient practice or provide students with additional practice that they need to achieve mastery. Teachers' experience of using analytics and reports are likely to be nearly indistinguishable across the different parameterizations.

Though perhaps still subtle, KC model specification studies are more likely to create inconsistencies in K-12 learners' and instructors' experiences in an ITS if they were to be deployed within, for example, the same classroom, or even to all of the classrooms of the same teacher. Changes in the KC model specification may also be accompanied by design differences in the

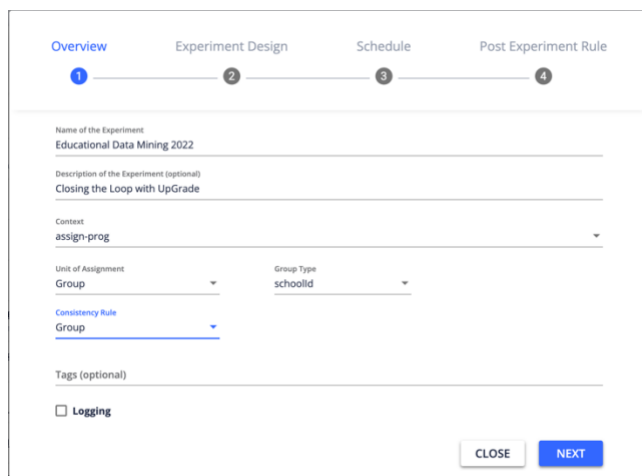
tasks presented to students (e.g., as in [8, 18]), presenting further opportunities for noticeable differences in learner experiences to emerge.

Generally, the contrast we seek to illustrate is between any type of relatively “stealthy” experiment with more “conspicuous” and/or visually salient experiments. More stealthy examples, like the hypothetical experiment that only manipulates parameter estimates, are those in which differences in the learner experiences are subtle and likely to be unnoticed between conditions. Experiments are likely to be more conspicuous when they seek to test substantive differences in learning experiences and may have more easily-discernable variations, as in the “specification” experiment.

More conspicuous differences might reasonably be evaluated between instructional methods, encouraging students to adopt different problem-solving strategies for the same topic, and contrast along any number of a wide variety of instructional decisions content designers must make [9]. Many such differences between experimental conditions are easily perceived by students and teachers. If condition assignment is not thoughtfully considered (e.g., by considering a group-level assignment), students may realize that they are receiving a different experience compared to other learners in their classrooms, leading to unanticipated changes in learning patterns. Teachers might easily become overwhelmed by the need to support different learning experiences for the same topic, keep track of differences in analytics and reporting, and other potential inconsistencies. In what follows, we describe how the UpGrade platform can help educational technology developers and researchers not only deploy internet scale educational experiments but also do so in ways that handle many of the challenges that arise in such deployments in real world settings in which learning takes place.

4. UPGRADE

UpGrade¹ [16] is an open source software architecture (available via GitHub²) intended to lower barriers to learning engineering and enable internet scale experiments (also sometimes referred to as “A/B tests” or randomized field trials) that test substantive changes to learning experiences in settings like K-12 classrooms while preserving consistent, high-quality learning experiences for end-users.



The screenshot shows the 'Overview' step of a four-step process for creating a new experiment. The steps are: Overview (1), Experiment Design (2), Schedule (3), and Post Experiment Rule (4). The 'Overview' step is currently active. The form includes the following fields:

- Name of the Experiment: Educational Data Mining 2022
- Description of the Experiment (optional): Closing the Loop with UpGrade
- Context: assign-prog
- Unit of Assignment: Group
- Group Type: schoolid
- Consistency Rule: Group
- Tags (optional):
- Logging:

At the bottom right, there are 'CLOSE' and 'NEXT' buttons.

Figure 1. Screenshot of the “overview” of an experiment in the UpGrade user interface for creating a new experiment

We consider several of the major affordances provided by UpGrade, several of which are illustrated in the screenshot of Figure 1 that shows the preliminary “overview” of an UpGrade experiment as it is being specified. Group random assignment, consistency rules, post-experiment rules, and user segmentation are features of experiments that help to ensure that consistent learning experiences are delivered during experiments. Researchers determine how best to set these options as a part of designing experiments in UpGrade.

4.1 Group Random Assignment

While UpGrade has logic for assigning experimental conditions on an individual-student or user basis, the ability to randomly assign conditions by group (e.g., at the level of classrooms within a K-12 school) enables researchers to conduct internet scale experiments in educational software products that are both deployed at scale and used in authentic classroom contexts. In educational settings, it may be undesirable for students within the same group (e.g., by class, teacher, school district, or some other grouping) to be assigned different conditions within the same experiment, particularly if such conditions involve conspicuous or salient visual changes (e.g., different “skillometers” displays and teacher analytics in the KC model specification experiment illustrated in §3) or substantively different models of instruction. UpGrade manages coherence of learning experiences by group as well as anomalies that may arise in group membership, enabling researchers to specify how an experiment should behave if a student switches classes or is in multiple classes simultaneously.

4.2 Consistency Management

The second way in which UpGrade helps deliver consistent experiences is via associating deployed experiments with consistency rules that govern how users are treated for inclusion/exclusion in an experiment who have already encountered pieces of instructional content or other design features that are included in experiments. This is particularly useful when instructional content is delivered via adaptive software in which self-paced progress is often a crucial feature of the learning experience; in such software students may reach the content of interest at different times. If a student in a class encounters the content of interest earlier (or later) than their fellow students, consistency management can specify whether condition assignment binds more strongly to group membership or individual students.

4.3 Post-Experiment Rules

A “post-experiment rule” is a parameter that manages delivery of experimental conditions once an experiment has stopped running, but students may still interact with the target content in the educational application. Researchers may wish to have a “winning” condition be delivered to subsequent students who encounter the content of interest, or may wish to maintain the condition assignment weighting even if the experiment has ended. For example, if a study using UpGrade runs from September-November of a school year, but a student goes back to review content in preparation for an end-of-semester exam in December, a post-experiment rule can specify whether that student should receive the same experimental condition they were originally assigned, or whether they are permitted to experience a default or other condition.

¹ <https://www.upgradeplatform.org/>

² <https://github.com/CarnegieLearningWeb/UpGrade>

4.4 User Segmentation

Another challenge when conducting internet-scale experiments in real-world classrooms is that researchers may not necessarily want or need to target all members of a population. Segmentation and the ability to pre-define include/exclude lists empowers researchers to target experimental interventions to groups of interest, at the level of interest: e.g., middle schools, 7th grade, specific districts, or even geographic regions. Similarly, districts or schools can fully “opt-out” and join a global exclude list without impacting research at scale conducted by an educational technology company.

While learning platforms often do not automatically collect data about individual learner demographics, when such data are available (or when such data are available at an aggregated level like that of particular schools), user segmentation is likely to also play an important role in better understanding what works for particular sub-populations of students, helping researchers close the loop in ways that promote inclusion and equity across diverse populations of learners.

4.5 Monitoring Metrics

UpGrade enables researchers running experiments to monitor their progress with respect to “enrolled” users (i.e., those users who have encountered relevant content and been assigned to a condition) as well as those who have been excluded. In addition, APIs are available for target learning applications to send specific metrics (e.g., time to complete a particular piece of content) of interest to UpGrade’s monitoring dashboards for real-time progress monitoring of experimental progress. The screenshot in Figure 2 illustrates a particular case of an experiment with two conditions in the UpGrade platform, displaying enrollment data over time and by experimental condition.

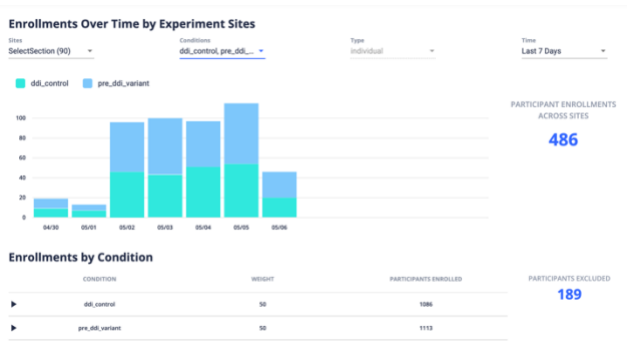


Figure 2. Screenshot of monitoring dashboard in UpGrade, showing enrollment metrics over time and by condition, for two experimental conditions, entitled “ddi_control” and “pre_ddi_variant,” where “ddi” stands for “data driven improvement” of particular content in Carnegie Learning’s MATHia platform.

4.6 Support for Diverse Experimental Designs

UpGrade currently supports relatively simple experimental designs, including weighted random assignment to two or more conditions (see “weight” column near the bottom of the screenshot in Figure 2) as well as within-subject, factorial, and partial-factorial designs. UpGrade developers have a roadmap for implementing additional designs including multi-arm bandits and stepped-wedge designs in the near future. More sophisticated designs and those

incorporating adaptive experimentation could be valuable contributions from the EDM and allied communities. This leads to our call for contributors and integrators.

5. CALL FOR CONTRIBUTORS & INTEGRATORS

Our goal in the present paper has been to introduce UpGrade to the EDM community. We hearken back to the “rise of the super experiment” and seek to build awareness of challenges often raised by internet scale, experimental “close the loop” studies and how UpGrade presents a solution that can be integrated within an existing or emerging educational technology product to help overcome those challenges. Efforts at the intersection of EDS/EDM and the emerging field of learning engineering rigorously seek to establish causal links between data-driven insights that inform improvements in educational technologies and practical learning outcomes resulting from the use of these technologies. We applaud efforts like E-TRIALS [12], MOOClets [13], and Terracotta³ that aim to provide similar support for rigorous experimental or A/B testing of learning innovations within particular contexts (e.g., E-TRIALS within ASSISTments [7], TerraCotta within the Canvas Learning Management System). UpGrade can be integrated into existing or new learning applications and technologies to similarly drive rigorous data-driven improvements to learning platforms.

Educational technology developers must still (as ever) address challenges to attracting large and diverse user-bases, instrumenting their technology to capture rich learning data, and appropriately abstracting features and content in their systems so that different learning experiences can be delivered to learners (Challenges #1-2 and part of the third challenge described by [18]). However, UpGrade removes many of the barriers imposed by the challenges of large-scale experimental management in real-world learning settings. The educational technology developer need only implement software instrumentation that calls UpGrade’s API to determine which alternative learning experience (if any) ought to be delivered to a particular user, given characteristics of that user that the target system “knows” about (e.g., via communication with a rostering system) such as the class or school in which the user is learning. The target system for experimentation can also implement UpGrade’s API to provide metrics for monitoring experiments as they proceed. UpGrade handles complex logic of managing condition assignments and dealing with real-world complexities that inevitably arise in settings like K-12 schools.

As an open source platform, developers can contribute new functionality and features to the codebase for the future benefit of all integrators and researchers using platforms that integrate with UpGrade. For example, code might be contributed to build connections to deliver A/B tests in different LMSs and implement appropriate APIs to have metrics for monitoring delivered to UpGrade. Support for new types of experimental designs and algorithms for adaptive experimentation are also a natural place for future development. We welcome such new contributions from the EDM community as well as from the broader educational technology and learning engineering communities.

UpGrade has already been used to deliver experiments to tens of thousands of learners in a math game similar to that targeted in the paper that introduces the SEF [18]. A number of experiments are currently deployed in Carnegie Learning’s MATHia, and new experiments will be deployed in the coming months using

³ <https://terracotta.education/>

UpGrade. These experimental field trials will close the loop between data-driven improvements to facets of learning experiences like KC models as well as personalization and motivational features, and we look forward to presenting those results in the near future. We encourage educational technology developers to consider integrating UpGrade into their platforms to enable rigorous, iterative learning engineering improvements and platform-enabled educational research.

6. ACKNOWLEDGMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305N210045 to Carnegie Learning, Inc. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. The first and third authors are also supported by the National Science Foundation under award The Learner Data Institute (Award #1934745). The opinions, findings, and results are solely the authors' and do not reflect those of the National Science Foundation.

7. REFERENCES

- [1] Aleven, V., McLaughlin, E.A., Glenn, R.A., Koedinger, K.R. 2017. Instruction based on adaptive learning technologies. In *Handbook of Research on Learning and Instruction*, 2nd Edition. Routledge, New York, 522-560.
- [2] Aleven, V., Sewall, J., McLaren, B. M., & Koedinger, K. R. 2006. Rapid authoring of intelligent tutors for real-world and experimental use. In *Proceedings of the 6th IEEE International Conference on Advanced Learning Technologies*. ICALT 2006. IEEE Computer Society, Los Alamitos, 847-851.
- [3] Baker, R.S., Corbett, A.T., Gowda, S.M., Wagner, A.Z., MacLaren, B.M., Kauffman, L.R., Mitchell, A.P., Giguere, S. 2010. Contextual Slip and Prediction of Student Performance After Use of an Intelligent Tutor. In *Proceedings of the 18th Annual Conference on User Modeling, Adaptation, and Personalization*, 52-63.
- [4] Baker, R. S., Corbett, A.T., and Koedinger, K.R. 2007. The difficulty factors approach to the design of lessons in intelligent tutor curricula. *Int. J. Artif. Intell. Educ.* 17(4), 341-369.
- [5] Cen, H., Koedinger, K.R., and Junker, B. 2006. Learning factors analysis: A general method for cognitive model evaluation and improvement. In *Intelligent Tutoring Systems 2006*. ITS 2006. Springer-Verlag, Berlin, 164-175.
- [6] Corbett, A.T., Anderson, J.R. 1995. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Model. User-Adap.* 4, 253-278.
- [7] Heffernan, N.T., and Heffernan, C.L. 2014. The ASSISTments Ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *Int. J. Artif. Intell. Educ.* 24, 470-497.
- [8] Huang Y., Aleven V., McLaughlin E., and Koedinger K. 2020. A general multi-method approach to design-loop adaptivity in intelligent tutoring systems. In *Artificial Intelligence in Education 2020*. AIED 2020. LNCS, vol 12164. Springer, Cham, 124-129.
- [9] Koedinger, K.R., Booth, J.L., and Klahr, D. 2013. Instructional complexity and the science to constrain it. *Science* 342(6161), 935-937.
- [10] Koedinger, K.R., Corbett, A.T., and Perfetti, C. 2012. The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science* 36(5), 757-798.
- [11] Liu, R., and Koedinger, K.R. 2017. Closing the loop: Automated data-driven cognitive model discoveries lead to improved instruction and learning gains. *Journal of Educational Data Mining* 9(1), 25-41.
- [12] Ostrow, K., and Emberling, R. 2020. E-TRIALS: A web-based application for educational experimentation at scale. In *Proceedings of the First Workshop on Educational A/B Testing at Scale*. EdTech Books.
- [13] Reza, M., Kim, J., Bhattacharjee, A., Rafferty, A.N., & Williams, J.J. 2021. The MOOClet Framework: Unifying experimentation, dynamic improvement, & personalization in online courses. In *Proceedings of the 8th ACM Conference on Learning at Scale*. L@S 2021. ACM, New York, NY, 15-26.
- [14] Ritter, S., Anderson, J.R., Koedinger, K.R., and Corbett, A.T. 2007. Cognitive Tutor: applied research in mathematics education. *Psychon. B. Rev.* 14, 249-255.
- [15] Ritter, S., Harris, T.K., Nixon, T., Dickison, D., Murray, R.C., Towle, B. 2009. Reducing the knowledge tracing space. In *Proceedings of the 2nd International Conference on Educational Data Mining* (Cordoba, Spain, 2009), 151-160.
- [16] Ritter, S., Murphy, A., Fancsali, S. E., Fitkariwala, V., Patel, N., and 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*. EdTech Books.
- [17] Ritter, S., Yudelson, M., Fancsali, S.E., and Berman, S.R. 2016. How mastery learning works at scale. In *Proceedings of the 3rd ACM Conference on Learning at Scale* (April 25 - 26, 2016, Edinburgh, UK). L@S 2016. ACM, New York, NY, 71-79.
- [18] Stamper, J.C., Lomas, D., Ching, D., Ritter, S., Koedinger, K.R., Steinhart, J. 2012. The rise of the super experiment. In *Proceedings of the 5th International Conference on Educational Data Mining* (Chania, Greece, June 19-21, 2012). EDM 2013. International Educational Data Mining Society, 196-199.
- [19] Yudelson, M., Koedinger, K., Gordon, G. 2013. Individualized Bayesian Knowledge Tracing models. In *Proceedings of 16th International Conference on Artificial Intelligence in Education* (Memphis, TN). AIED 2013. LNCS vol. 7926. Springer-Verlag, Berlin / Heidelberg, 171-180.