Platform-based Adaptive Experimental Research in Education

Lessons Learned from The Digital Learning Challenge

Ilya Musabirov
University of Toronto
Toronto, Canada
ilya@musabirov.info

Mohi Reza
University of Toronto
Toronto, Canada
mohireza@cs.toronto.edu

Haochen Song
University of Toronto
Toronto, Canada
fred.song@mail.utoronto.ca

Steven Moore
Carnegie Mellon University
Pittsburgh, USA
StevenJamesMoore@gmail.com

Pan Chen
University of Toronto
Toronto, Canada
pan.chen@utoronto.ca

Harsh Kumar
University of Toronto
Toronto, Canada
harsh@cs.toronto.edu

Tong Li
University of Toronto
Toronto, Canada
tongli@cs.toronto.edu

John Stamper
Carnegie Mellon University
Pittsburgh, USA
john@stamper.org

Norman Bier
Carnegie Mellon University
Pittsburgh, USA
nbier@cmu.edu

Anna Rafferty
Carleton College
Northfield, USA
arafferty@carleton.edu

Thomas Price
North Carolina State University
Raleigh, USA
twprice@ncsu.edu

Nina Deliu
MEMOTEF
Sapienza University of Rome
Rome, Italy
nina.deliu@mrc-bsu.cam.ac.uk

Audrey Durand
Université Laval
Quebec City, Canada
audrey.durand@ift.ulaval.ca

Michael Liut
University of Toronto Mississauga
Mississauga, Canada
University of Toronto
Toronto, Canada
michael.liut@utoronto.ca

Joseph Jay Williams
University of Toronto
Toronto, Canada
williams@cs.utoronto.ca

Abstract

Adaptive Experimentation is one of the most promising approaches to support complex decision-making in learning experience design and delivery. This paper reports on our experience with a real-world, multi-experimental evaluation of an adaptive experimentation platform within the XPRIZE Digital Learning Challenge framework, and summarizes data-driven lessons learned and best practices for Adaptive Experimentation in education. We outline key scenarios of the applicability of platform-supported experiments and reflect on lessons learned from this two-year project, focusing on implications relevant to platform developers, researchers, practitioners, and policy stakeholders to integrate Adaptive Experiments in real-world courses.

LAK 2025, Dublin, Ireland

ACM ISBN 979-8-4007-0701-8/25/03
https://doi.org/10.1145/3706468.3706471

CCS Concepts

• **Human-centered computing** → *Interaction design process and methods*; • **Information systems** → *Decision support systems*; • **Applied computing** → **Education**.

Keywords

adaptive experiments, posterior sampling, experimentation platforms, educational technology, human-computer interaction

ACM Reference Format:
Ilya Musabirov, Mohi Reza, Haochen Song, Steven Moore, Pan Chen, Harsh Kumar, Tong Li, John Stamper, Norman Bier, Anna Rafferty, Thomas Price, Nina Deliu, Audrey Durand, Michael Liut, and Joseph Jay Williams. 2025. Platform-based Adaptive Experimental Research in Education: Lessons Learned from The Digital Learning Challenge. In *LAK25: The 15th International Learning Analytics and Knowledge Conference (LAK 2025), March 03–07, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3706468.3706471

1 Introduction

Online courseware platforms provide instructors, learning designers, and researchers with extensive opportunities to enhance learning experiences and improve outcomes [35]. However, this introduces additional complexity, as there are now many decision points

with different potential improvements at each step of the student's learning experience [22].

Although major technology companies rely on systematic A/B testing, backed by years of research and evidence-based best practices [23], educational settings present distinct challenges that extend beyond traditional online experimentation [36]. These challenges include managing limited sample sizes and complex populations, engaging stakeholders in pedagogically meaningful experiments [39], and balancing scientific discovery with practical student impact [35].

Instructors often worry about the fairness of experiments, concerned that students might end up in a less effective learning condition. Moreover, when an experiment shows differences in outcomes between conditions, educators may hesitate to replicate it, making it difficult to understand what led to its success or how underrepresented students were affected. However, it is both practical and scientifically important to verify that an intervention works or to test new, potentially better ideas. Rather than viewing experimentation as a yes-or-no decision, we should consider it a balance between two strategies: "exploring" (gathering data to discover the best condition) and "exploiting" (choosing the condition believed to work best based on current data). This balance lies at the heart of adaptive experimentation.

Adaptive experiments differ from traditional A/B tests by dynamically updating the assignment policies throughout the experiment. This approach enables conditional assignment strategies, flexible experimental grouping, and personalized interventions while balancing multiple objectives, such as optimizing student outcomes and advancing scientific discovery. Numerous applications demonstrate the effectiveness of adaptive experiments across fields [4], including advertising [15], public health [26, 32], clinical trials [8], and policy evaluation [2].

In this paper, we report on the data-informed lessons learned, which are centered on a multi-experiment field deployment evaluating an adaptive experimentation platform within the two-year XPRIZE Digital Learning Challenge[1] (*the Challenge*). This deployment involved the rapid conduction and systematic replication of five experiments within 30 days across multiple courses and institutions. Through application scenarios, we demonstrate two key ideas of adaptive experimentation: *evidence-driven outcome optimization* and *context-sensitive personalization*. We explore how stakeholder-centric design can empower instructors, learning designers, and researchers to create and manage experiments for continuous learning improvement. We emphasize the use of data-informed simulations, customized visualizations, and strategies to address the small sample sizes typical of education as crucial tools for stakeholder-centric design..

While the existing literature primarily approaches adaptive experiments from an optimization perspective, emphasizing theoretical guarantees and algorithmic properties, we adopt a pragmatic approach focused on designing experiments grounded in real-world settings to enhance practical applicability while maintaining a high level of scientific rigor. By providing recommendations and tools, this research contributes to addressing key education-specific design constraints:

- Accounting for **realistic effect size estimates** for educational interventions, informed by recent meta-analyses [24, 25] that suggest more conservative effects than traditional social science expectations;
- Addressing **sample size constraints** inherent to classroom-based research;
- Supporting **stakeholder agency and decision support** for researchers, instructors, and learning designers, acknowledging their distinct yet overlapping interests.

This research explores how adaptive experiments can improve learning analytics and engineering workflows, bridging the gap between analytics and action. Our approach is implemented through the Experiments As a Service Infrastructure (EASI), a platform for adaptive experimentation [35]. While most tools are limited to specific platforms, EASI provides flexible random assignment methods and broad cross-platform compatibility. EASI builds on integrations with major learning platforms (e.g., edX, Coursera, Moodle, Canvas, ASSISTments, OLI) and any LTI-compliant Learning Management Systems. EASI offers access to a library of machine learning algorithms and statistical methods for adaptive experiment design and real-time analysis. This work extends our earlier Practitioner Report [30] presented at LAK'24.

2 The Approach

First and foremost, at the core of our design process, we formed an interdisciplinary group of research and software deployment experts representing the fields of human-computer interaction, statistics and machine learning, learning science and engineering, as well as educational practitioners, all with previous experience in field randomized controlled trials in education. In addition, this expert group (in short, "experts") included developers of learning infrastructure from the Open Learning Initiative (OLI) at Carnegie Mellon University and developers of adaptive experimentation infrastructure. Three team members were also course instructors with extensive experience using the course delivery platforms and learning management systems within which we deployed our experiments.

During *the Challenge*, we aimed to demonstrate the adaptive approach to experimentation and our platform capabilities in the rapid multi-replication of educational interventions for different student demographics. The scope of expert work was not limited just by *the Challenge* participation. Firstly, the experts reflected on previous adaptive experimentation experiences of the Intelligent Adaptive Interventions Group at the University of Toronto in the computer science education and mental health research domains, finding potentially promising patterns of adaptive experiments from previous work [26, 31, 35]. Secondly, during *the Challenge*, in addition to guiding the team's work, the experts outlined key adaptivity scenarios, demonstrating the design space of learning interventions expanded by adaptive experimentation. These scenarios, teamwork protocols, and empirical data served as the foundation for a series of statistical simulations, expanding from concrete experiments to scenarios and hypothetical outcomes, representative of a wider class of educational interventions. Lastly, the experts reflected on challenges, lessons learned, and the broader applicability in the

[1] https://www.xprize.org/challenge/digitallearning

educational technology and learning analytics research and practitioner community.

2.1 Iterative Refinement Through Field Deployments

Participation in *the Challenge* included one pilot study and five replication studies in different courses within a 30-day time frame. These courses were offered across different institutions, subject domains, and academic levels. Our settings included research and teaching, public and private, and both two- and four-year institutions. Within these settings, our experiments were deployed in a range of courses representing diverse domains, with undergraduate and graduate learners, and in both general education and for-majors contexts. At Bethune-Cookman University, an HBCU in Florida, we integrated our study into an undergraduate Anatomy & Physiology course for allied health and pre-med majors. At Georgia State University (GSU), a public Predominantly Black Institution (PBI) with both four- and two-year campuses, we embedded our studies in each campus's respective gen-ed Introduction to Statistics course. At Carnegie Mellon University in Pennsylvania, we deployed in Interactive Data Science, a STEM course with a mixed undergraduate/graduate population, and Tools for Online Learning, a graduate-level class in Learning Engineering.

To support deployment in a diverse set of courses, we chose intervention cases designed as loosely coupled with the course content and EASI, enabling researchers and instructors to rapidly replicate them in any new course using existing course content. These interventions were created for a common formal education context in which students independently work through an online textbook containing short passages and videos with content knowledge. Students engage in various activities at the end of each textbook section to promote learning. The first intervention aimed to provide students with retrieval practice prompts tied to course activities, using accuracy on the following problem as a proximal outcome (algorithm reward). The second intervention focused on the motivational domain, encouraging students to participate in optional course activities through self-/peer-based frameworks and tracking engagement outcomes.

All deployments during *the Challenge* were conducted on courses using the Open Learning Initiative (OLI) platform [3], which is designed to support robust experimentation at scale in collaboration with institutions already using the OLI courseware. Instructors of these courses had previously used OLI for one to three semesters before EASI integration.

2.2 Simulations for Supporting Experimental Design and Decision-Making

Statistical simulations are an essential tool in the design of randomized controlled trials (RCTs). When creating classic randomized experiments using the frequentist approach, one of the key considerations is the acceptable margin of error. This involves fixing an acceptable false positive error rate in advance and analyzing the potential probability of a false negative error in relation to sample size and effect size [6, 27]. In the most straightforward cases of common statistical tests and models, these calculations can be done analytically and are implemented in statistical packages, online "calculators" for experiments, and more advanced software that assists with statistical experimental designs [10]. More complex models, experimental designs, and metrics require the use of "what-if" statistical simulation analysis based on the Monte Carlo approach: creating multiple simulated experiments with known design parameters and data-generation processes and analyzing the range of hypothetical outcomes across thousands of experiments to make design decisions [29].

In this paper, we extend this approach, using statistical simulations to expand the understanding of concrete cases and adaptive experimental designs to a wide range of practically relevant scenarios and constraints typical of educational interventions. Based on this, we ground expert recommendations in data-informed "what-if" analysis, focusing on opportunities, challenges, and up-scaling considerations for adaptive experiments in education.

2.3 Adaptive Experiments Definitions

Decision Point A step in a learner's journey in online courseware when a choice is made between alternative conditions informed by data, policy algor ithms, and other relevant considerations.

Condition A specific treatment, arm, or variation assigned to a learner during the experiment.

Reward A measurable outcome resulting from the interaction of a learner with one of the conditions.

Contextual Variable A factor or characteristic that provides relevant context for the experiment, for example, the learner's prior knowledge estimate or time before the deadline.

Multi-armed Bandit (MAB) A machine learning approach to decision problems, including adaptive experimentation, where conditions are selected sequentially over time to optimize the target performance metric [19][2].

3 Application Scenarios for Adaptive Experiments

In this section, we use two example experiments adapted from *the Challenge* to illustrate the core ideas and features of adaptive experiments. First, we present an example that demonstrates how these experiments *adapt to better outcomes* based on evidence gathered, highlight key metrics that provide insight into this process, and discuss potential enhancements enabled by adaptive experimentation. Next, we discuss another experiment showcasing the potential for *personalization to subgroups*. These scenarios illustrate how adaptive experiments can help explore and replicate the impact of novel variations of existing interventions in practice, targeting diverse outcomes of interest such as assessments and participation in learning activities. We use each application scenario to demonstrate helpful decision affordances, and example analyses providing insight into adaptive experiments.

[2]Throughout this paper, we use the Thompson Sampling [5] MAB algorithm in our examples and refer the reader to [31] for a detailed case study on its use in adaptive experimentation. EASI implements a wider range of MAB algorithms [28]

3.1 Adapting to Better Outcomes: Encouraging Self-Reflection

Learning designer Yi wants to evaluate whether prompting students to think critically about their responses and the learning process improves immediate learning outcomes. She chooses a specific **decision point**: a particular asynchronous activity in the course where students are prompted to reflect. Students can begin the activity at different times. To experiment with different methods, she specifies two conditions:

- **Condition 1**: No self-explanation prompt is shown to the student.
- **Condition 2**: A self-explanation prompt is provided, which asks: *Can you explain why you chose your answer? Imagine that you were explaining it to another student or your instructor.*

She selects the correctness of the student's answer to a multiple-choice question (relevant to the activity) as the **reward** for her experiment. This metric measures the impact of reflection prompts on learning outcomes. She configures the **algorithm updates** to occur after every interaction with the exercise.

Decision Affordances and Example Analysis: Adaptive Experiment Monitoring. Traditional experiments assign conditions uniformly, while adaptive experiments adjust assignment probabilities based on accumulated evidence. Over time, as more student data accumulate, the approach shifts from a standard experimental split (that is, 50/50), preferring better empirically performing conditions based on available evidence. To better understand this process, we zoom in on hypothetical steps of our *Encouraging Self-Reflection experiment* with the help of metrics and visualizations used for adaptive experiment monitoring.

Decision Affordances. Table 1 showcases instruments that are useful for both understanding and monitoring probabilistic adaptive experiments. The current state of knowledge of an adaptive experiment algorithm can be summarized using the posterior probability distribution for each condition (column "Probability Distributions" in the table). These distributions can provide researchers and analysts with deep insight into the learning dynamics of the algorithm and serve as a foundation for more user-centered metrics.

One such metric is the set of *selection probabilities* (or assignment probabilities) for each condition (column "Selection Probability" in the table). These probabilities indicate what condition is more likely to be assigned in the next step. The more consistent evidence the algorithm gathers in relative favor of a particular condition, the more likely it will be selected in the future. Selection probabilities summarize the relative effectiveness of conditions at the current step.

Example Analysis. Table 1 summarizes the number of successes and failures and the changing selection probability for each condition at each step. It also illustrates how evolving knowledge about the performance of each condition and the associated uncertainty is represented in the adaptive algorithm using probability distributions. We used the Thompson Sampling [5] adaptive learning algorithm in our examples and refer the reader to [31] for a detailed case study on its use in adaptive experimentation.

Step 1 represents an early stage of the experiment. The algorithm selected Condition 1 (no self-explanation prompt) for the first participant, and we received a negative reward from them. Taking into account the prior information (often set equal to one success and one failure for each condition if no additional information is available), the selection probability for Condition 2 in the next step would be approximately twice as high as for Condition 1.

By *Step 2*, more data has been gathered, leading to improved efficiency estimates for both conditions. This is reflected in narrower probability distributions, which summarize the algorithm's updated confidence in each condition's effectiveness. Selection probabilities continue to adjust dynamically, favoring the condition with better observed performance.

At *Step 3*, the algorithm's accumulated evidence more strongly supports Condition 2 as the more effective option. The probability distributions for each condition become narrower, indicating reduced uncertainty. Additionally, the slightly narrower shape of the distribution for Condition 2 at this step indicates that we have more information about its effectiveness as a result of allocating more participants to it. This process illustrates how adaptive experiments use evidence dynamically to refine decision-making and optimize outcomes.

3.2 Personalizing Interventions: Improving Participation in Optional Learning Activities

Instructor Steve is concerned that some students participate less in asynchronous optional generative activities, which ask students to create a revision question based on the learned material. He aims to evaluate different ways to motivate diverse groups of students to participate more actively in these activities.

Based on prior research, Steve finds that brief motivational messages before an activity can increase student engagement and participation. He decides to test two motivational approaches:

- **Condition 1 (Self-Focused)**: *Creating a question is a way to help you learn better by revising the content. Think of it like flashcards that help you review. We are not going to share the results with other students.*
- **Condition 2 (Peer-Focused)**: *By creating a question, you are helping to improve this course, contributing to the learning of your peers, and assessing your own understanding of the content.*

Steve hypothesizes that students with lower self-efficacy or lower demonstrated knowledge might respond better to the **Condition 1, Self-Focused** message, as it places less pressure on them. In contrast, students with higher self-efficacy or greater demonstrated knowledge may be more motivated by the **Condition 2, Peer-Focused** message, as it emphasizes contributing to the group by demonstrating their skill.

Condition	Successes	Failures	Selection Probability	Probability Distributions
Step 1				
1	0	1	33%	
2	0	0	67%	
Step 2				
1	4	6	20%	
2	6	4	80%	
Step 3				
1	12	18	2%	
2	24	12	98%	

Table 1: Condition outcomes, selection probabilities, and evolving probability distributions across steps in a hypothetical experiment.

To test these hypotheses, Steve incorporates the **contextual variable** of whether a student correctly answered a multiple-choice test question (Higher or Lower Prior Accuracy), using this as a proxy for their demonstrated level of knowledge.

He chooses the fact that a student submitted an answer to an optional activity as the **reward** for his experiment. Steve configures the **algorithm updates** to occur after every asynchronous interaction with the exercise.

3.2.1 Decision Affordances and Example Analysis: Adaptive Experiment Outcomes. In this section, we use our *Improving Participation in Optional Learning Activities* application scenario to discuss how collected data can be used for statistical analysis. This scenario also illustrates the potential of contextual interventions. As before, we base this discussion **on simulated data**, with the details of the experiment design adapted from real deployments during *the Challenge*.

Decision Affordances. To analyze the outcomes of the experiment, we rely on two key groups of decision affordances.

The first is hypothesis testing and alternative frequentist and Bayesian approaches to analyze experimental outcomes. Although traditional statistical methods (e.g., Wald Z test or Bayesian probability of superiority) can be applied, adaptively collected data present unique challenges. These challenges are beyond the scope of this paper, and we refer the reader to [9, 33, 38] for a deeper discussion.

The second is the use of data visualizations to summarize and interpret experimental results. Common approaches include bar charts with error bars to summarize outcomes (Figure 1). While widely used, recent research highlights the limitations of bar charts in accurately conveying effects and supporting decision-making [16, 42]. To improve decision support, we encourage analysts and tool developers to explore best practices, such as alternative ways of presentation and visualization of results [1, 18, 21].

Beyond summarizing the efficiency of each condition, it is essential to assess how well the experiment fulfilled the primary promise of adaptive experimentation: reallocating to better-performing conditions based on accumulated evidence. For this purpose, we use frequency-based framing in visualizations [14, 20] (Figure 2).

Example Analysis. Applying the Wald Z test to simulated data for *Improving Participation in Optional Learning Activities* application scenario, we could conclude (Figure 1A) that the Peer-Focused condition (Condition 2) outperforms the Self-Focused condition (Condition 1) ($z = 4.55$, $p < 0.0001$, Cohen's $\omega = 0.12$)[3].

However, a closer analysis (Figures 1B and 1C) reveals that, for a statistical minority (the 20% of students with Lower Prior Accuracy), the Peer-Focused condition performs worse than the Self-Focused condition, with a larger effect size ($z = 3.31$, Cohen's $\omega = 0.2$). The benefit of the Self-Focused message for Lower Prior Accuracy students is obscured by the majority group (80% of students with Higher Prior Accuracy), where the Peer-Focused condition outperforms the Self-Focused condition ($z = 6.77$, Cohen's $\omega = 0.2$). This majority effect drives the overall average positive effect, even though assigning the Peer-Focused message universally would harm students with Lower Prior Accuracy.

Figure 2 illustrates how the contextual adaptive experiment successfully personalized the treatment using accumulated knowledge. It allocated more participants in each group (Higher Prior Accuracy and Lower Prior Accuracy) to the conditions most effective for their respective groups.

[3]Wald Z-test for two independent samples.

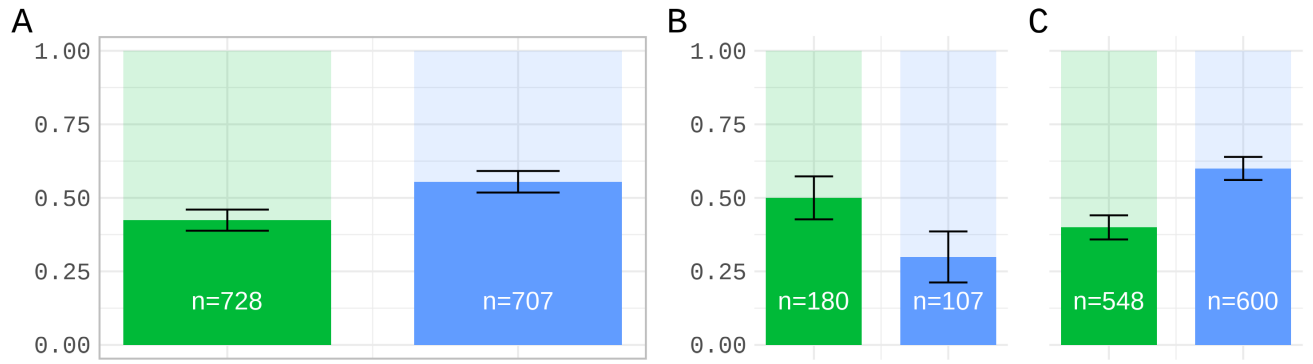


Figure 1: Simulated experiment results. Bar colors denote Condition 1, Self-Focused and Condition 2, Peer-Focused. Panel A represents the combined results for all participants. Panels B and C represent results for the groups with Higher Prior Accuracy and Lower Prior Accuracy on a previous task, respectively.

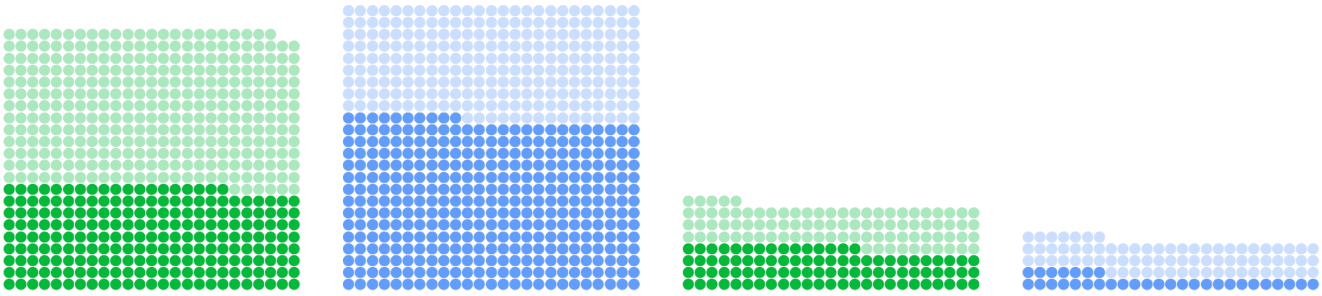


Figure 2: Simulated experiment results. Allocation and outcomes for the majority (Higher Prior Accuracy, left panel) and the minority (Lower Prior Accuracy, right panel) groups. Condition 1, Self-Focused assignments are represented by a green bullet • for success and by a lighter green • for failure. Similarly, Condition 2, Peer-Focused assignments are represented by a blue bullet • for success and by a lighter blue • for failure.

3.3 Summary: Promise of Adapting Experiments Proportional to Uncertainty

The features demonstrated in this section enable researchers, instructors, and learning designers to conduct a wide range of adaptive experiments aimed at course improvement. These experiments can address micro-level objectives, such as refining specific actions or enhancing a particular course element, or macro-level goals, such as improving overall course outcomes. Dynamic assignment of students to better-performing conditions reduces the decision-making burden on course teams while improving student outcomes.

In similar scenarios, the potential for adaptivity and personalization must be clearly communicated. Experimental platforms should support users in exploring and visualizing these capabilities early in the design process to enable informed decisions about interventions, particularly from an equity perspective.

4 Supporting Experiment Design Decisions

One of the key challenges in adaptive experimentation is providing stakeholders with the tools and affordances needed to understand and design adaptive interventions. In traditional RCTs, power analysis plays a primary role in guiding decision-making, such as determining the required sample size during the design phase [10].

However, adaptive experiments introduce a trade-off between participant benefit and statistical inference [7, 9, 33, 34], complicating statistical analysis compared to traditional A/B tests. As a result, stakeholders must consider the potential benefits and constraints of transitioning from traditional RCTs to adaptive experiments during the design process.

Additionally, there is a need to establish metrics and design strategies that can help stakeholders make more informed and efficient decisions.

4.1 Data-Informed Simulation for Intervention Design Decisions

During *the Challenge*, we prototyped modules aimed at letting instructors, designers, and researchers take data from one experiment and apply it to specify alternative scenarios for what the effects of conditions might be in different replications. Here, we demonstrate this approach step-by-step[4].

Decision Affordances. We need to provide stakeholders with complementary tools that clarify the various aspects and metrics influencing the design of an adaptive experiment.

Multiple **decision metrics** are critical for stakeholders when adopting adaptive experiments. These include the correctness rate and average reward, which respectively measure the performance of experimental and optimization treatments of the Multi-armed Bandit problem.

- **Average Reward** represents the overall outcome of an experiment by combining results across conditions[5]. Consider a scenario with binary rewards. In an experiment of 100 participants distributed between Conditions A and B, 60 participants are allocated to Condition A, and 40 are allocated to Condition B. Of these, 45 participants in Condition A have successful outcomes, and 20 in Condition B have successful outcomes. The average reward for the experiment is $\frac{45+20}{60+40} = 0.65$.
- **Correctness Rate** measures how often the best condition is assigned to a participant. During the experiment, an adaptive algorithm explores all conditions to estimate their effects. This exploratory phase results in some participants being assigned to suboptimal conditions, lowering the correctness rate.

In addition, we need tools informed by recent research in **data visualization for decision making** to help stakeholders make more informed design decisions [13, 16, 18, 21, 42].

- **Raincloud Plots** [1] (Figure 3c) offer a robust visualization method that combines multiple perspectives, including density plots, boxplots, and raw data points, to represent the shape of underlying probability distributions. Fitzgerald and Tipton [13] advocated their use for meta-analyses of multiple experiments, and we apply this tool to summarize simulated experiments.
- **Hypothetical Outcome Plots** [18] (Figure 4) address challenges in helping stakeholders interpret observed effects, particularly when statistical expertise varies. Kale et al. [18] suggest demonstrating a number of outcomes of simulated experiments, sharing characteristics with the experiment under consideration (hypothetical outcomes). Determining the number of static outcomes needed to build stakeholder intuition remains an open question. For this purpose, Kale et al. [18] suggest using animation where possible.

Example Analysis. To understand the potential outcomes of the experiment during the design stage, we first need to define the effect sizes of interest. There are multiple approaches to do this [27]. Even if there is no prior information about a particular intervention, looking at the realistic ranges of effect sizes for relevant educational interventions [24] and taking into account types and "closeness" of outcomes to the interventions and other design characteristics [24, 25, 41] is critical to make informed decisions. For adaptive interventions, it is also important to consider proximal outcomes as rewards to enable the intervention to adapt quickly.

In this example, we use a range that roughly corresponds to Cohen's d from 0.05 to 0.3, reflecting more realistic estimates for small to large sizes of educational interventions [24]. We also use a range of sample sizes of interest, 20-1000, representing typical classroom enrollment in *the Challenge* (see Table 3).

Figure 3 illustrates how an Adaptive Experiment compares to a traditional RCT by simulating hypothetical experiments across the specified ranges of effect sizes and sample sizes.

The correctness rate (Figure 3a) assumes that every time a suboptimal condition is chosen, a student is underserved during a particular interaction, treating all errors with equal weight regardless of the actual difference in effectiveness between conditions. In contrast, the average reward (Figure 3b) provides stakeholders with insight into potential performance across the entire classroom. Notably, the average reward is lower than the correctness rate, reflecting the reality that achieving a 100% outcome (e.g., perfect correctness in answering test questions for all participants) is unlikely, even with 100% assignment to the better condition.

However, these outcomes represent *average* results from a series of simulated experiments, which could mislead stakeholders if treated as guarantees. Figure 3c uses raincloud plots to provide a much-needed closer look at the hypothetical outcomes for a sample size of interest, usually dictated by course enrollment. In this case, we selected a classroom size of 100 to explore the potential performance of interventions across different effect sizes. Despite reasonable average results even for small effect sizes, a substantial proportion of experiments could perform worse in reallocating to better conditions than traditional uniform random allocation[6].

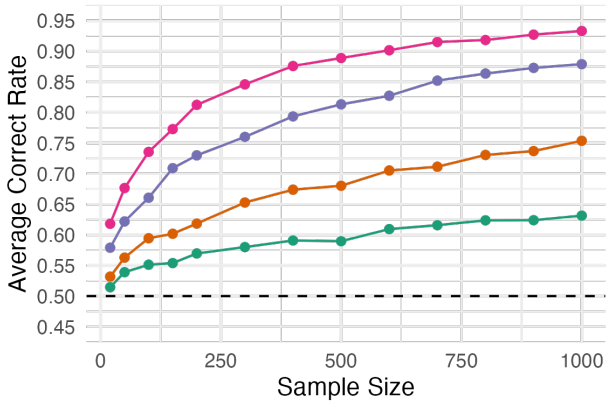
In practice, depending on the primary objective of the adaptive experiment, this might not necessarily constitute a problem. If the focus is on maintaining reasonably good performance (as represented by the average reward), it is evident that most fluctuations occur when effect sizes and/or sample sizes are small. However, small effect sizes often imply that there is no substantial difference between conditions, so mistakes may not be costly in terms of overall benefit.

Conversely, adaptive experiments demonstrate clearer advantages when larger differences between conditions exist. For example,

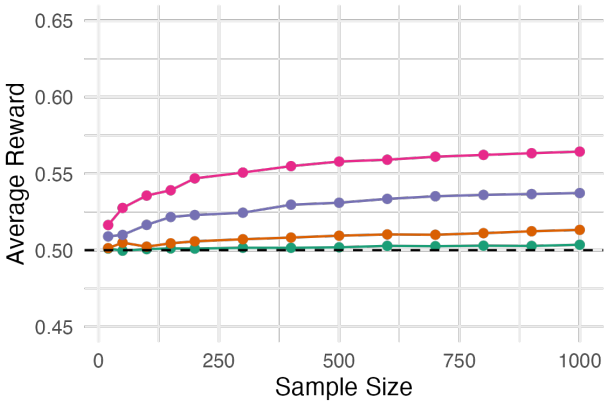
[4] *Technical note*: Simulation setup should closely replicate the specific design and algorithms of the intended deployment. In this example, we simulate the application scenario from 3.1 with effect sizes typical of educational interventions. We use one MAB algorithm, Thompson Sampling [5], and compare it with uniform random allocation, the baseline approach for traditional randomized controlled trials (RCTs). Simulations are based on a stochastic process with fixed reward distributions for different conditions. We simulate a case with binary rewards modeled as Bernoulli distributions with $p_i \in (0, 1)$ representing the default success probability for condition i. While ground-truth success probabilities are controlled in simulation, they are unobservable to experimenters in practice. The two-condition experiment uses effect sizes converted from Cohen's d to p'_1, p'_2 centered around 0.5. Thompson Sampling updates posterior success probabilities using Bayesian inference with a conjugate Beta distribution, leveraging the prior (Beta(1,1) in this case) and observed data for each condition.

[5] *Technical note*: In Machine Learning literature, the focus is usually on minimizing *regret*.

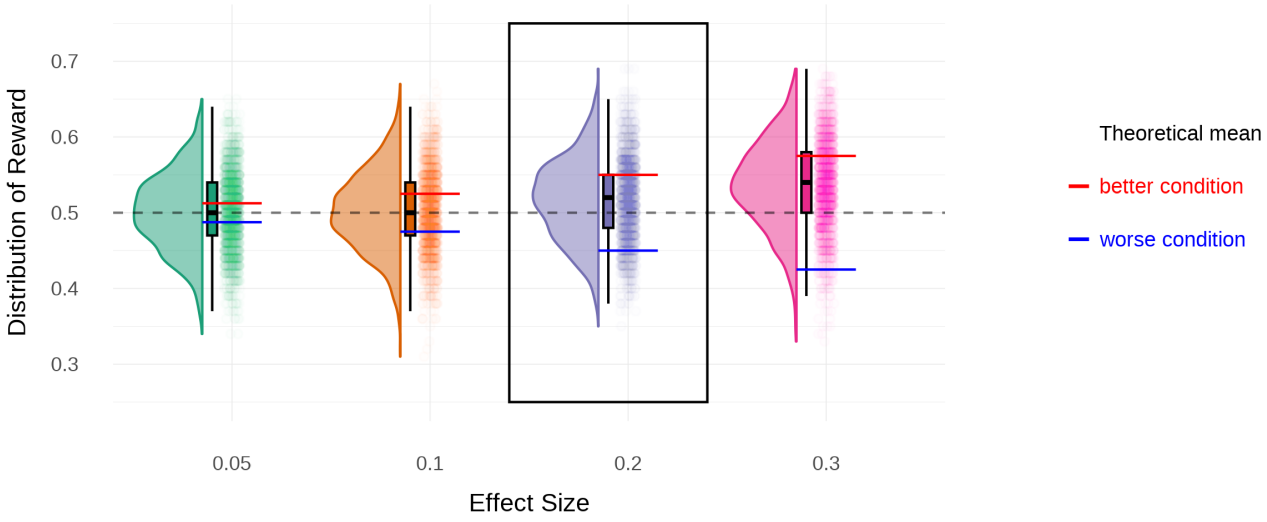
[6] *Technical note*: These results are specific for the particular algorithm and setup we used in the example. For a broader discussion of sources of biases in Multi-armed Bandits we refer the reader to [37].



(a) Correct rate across different sample and effect sizes



(b) Average reward across different sample and effect sizes



(c) Zoomed in reward distribution with a sample size of 100 at different effect sizes

Figure 3: Average rate of correct assignments and average reward in relation to sample size and effect size. The black dashed line represents the performance of the corresponding metric under traditional equal allocation RCT for comparison. Effect sizes are shown in the following colors: 0.05, 0.1, 0.2, 0.3. Results are based on replications of the hypothetical experiment.

the highlighted box in Figure 3c corresponds to a Cohen's d of 0.2. In approximately 25% of cases, the performance approaches the average reward of the optimal condition. Moreover, in over 50% of cases, adaptive experiment outperform traditional RCT.

Therefore, it is crucial to communicate both the correctness rate and average reward perspectives to stakeholders during experimental design.

While this information may provide stakeholders with enough context to make a decision, hypothetical outcome plots can offer additional insights by visualizing multiple simulated "classrooms"–instances of the experiment. Figure 4 presents results from five hypothetical adaptive experiments, each with an effect size of 0.2 and a sample size of 200. Although the experimental setup is identical across all five cases, we observe that four experiments quickly identified the optimal condition (Condition 2). However, in one case, the experiment required more time to converge to the optimal arm. This delayed convergence typically occurs when the exploration phase encounters early negative outcomes, which can mislead the policy learning process [38].

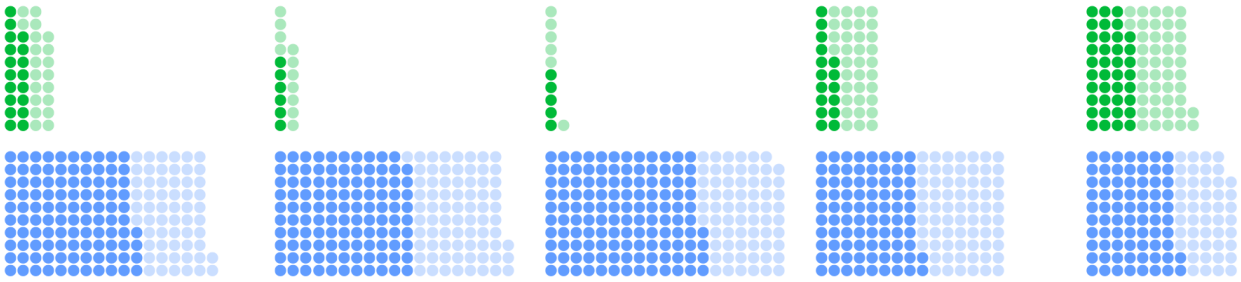


Figure 4: Five hypothetical outcomes for the same experimental design parameters (sample size 200 and effect size 0.2). Condition 1 assignments are represented by a green bullet • for success and by a lighter green • for failure. Similarly, Condition 2 assignments are represented by a blue bullet • for success and by a lighter blue • for failure.

Table 2: Illustrative examples of research, learning design, and classroom improvement interventions.

Example #	Intervention	Level
E1	Motivational prompts to pursue more practice connected to classroom events	Classroom/Section level
E2	Alternative explanations of a course concept	Classroom/Section level or Course level
E3	Adding self-explanation prompts after course videos	Course level or General research level

4.2 Summary: Simulations for Stakeholder-Centered Adaptive Intervention Design

We suggest presenting the results from research-centered simulations summarized via raincloud and hypothetical outcome plots to support stakeholder understanding and decision-making. The approach helps researchers simulate the collection and analysis of data from thousands of repeated runs of an experiment under different scenarios of what the effects could be and what might mediate these effects, such as student characteristics. This allows researchers and instructors to specify precisely and explore different kinds of effects they could discover in future replications of their experiment and to understand what impact the particular adaptive experimental design can achieve compared to traditional approaches. Another related requirement demonstrated in the example is to provide a set of custom data visualizations and data analysis workflows tied to the experimental design. This avoids potential issues arising from the application of unsuitable or suboptimal analytical methods. It allows us to understand not only what we have learned – causal effects, but also how we did – the impact of adaptation and personalization on students. Small sample sizes are the most problematic area for all effect sizes considered, as expected. Sadly, they are also typical for educational settings. We will discuss some ways to alleviate limitations imposed by them in section 5.

5 Overcoming Constraints of Classroom Experiments

In the previous section, we focused on analyzing the performance of adaptive experiments across classrooms of different sample sizes. To bring the benefits of adaptive experimentation to a wide range of real-world contexts, such experiments need to be applied not just to ideal settings with large samples but also to smaller courses where both traditional and adaptive experiments are not guaranteed to deliver optimal results. While external researchers may want to run experiments in ideal situations, a crucial design constraint emerges when we account for what other stakeholders such as learning designers want, as they may be more keen on bringing the practical benefits of the experiments to their students, and have more on more agency in continuous improvement and decision-making.

From our analysis of course sizes on OLI (see Table 3), *90%* of the courses on the plaftorm have fewer than or 113 students. In such course contexts, different stakeholders may be interested in adopting interventions that can potentially have a meaningful impact in single-classroom settings. In Table 2, we list some examples of interventions that can be initiated with different stakeholders (researchers, learning designers, instructors) on different levels, setting constraints for experimental design. Although there are no ways to overcome the limit of the information that can be extracted from small samples, the Bayesian nature of adaptive experiments suggests some promising ways to help alleviate these constraints and support instructors' and learning designers' agency in experimentation and evidence-based decision-making. In this section, we present two strategies that we found to be helpful in bringing adaptive experiments to small courses – using *prior information* and *sharing experiments* between courses.

5.1 Strategies for Overcoming Constraints for Smaller Classroms

Starting Adaptive Experiment with Prior Information. While our examples **E1** and **E2** (see Table 2) might look for solutions, working

Table 3: Summary Statistics of Typical Classroom Sizes for OLI.

Mean	Median	P0	P25	P50	P75	P80	P90	P100
54	26	11	19	26	41	51	113	2564

for a particular course section, we can incorporate prior information into the experiment. For example, we can assume that prompts working in other relevant courses might, but not necessarily will, work the same way in ours and scale that assumption according to our degree of belief and classroom size, balancing reliance on existing evidence and the instructor's judgment. In this case, our course section builds on existing evidence, but converges to its own solution by the end of the experiment. On the platform level, we can support the search for relevant prior information, for example, by allowing us to easily build on results from previous course iterations or other sections while incorporating subjective beliefs of how our section is similar or different or adding our alternative options while building on proven defaults.

Shared Experiment Between Course Sections to Courses. Alternatively, we can run an experiment (e.g., **E2**, **E3**) by pulling data across multiple sections and courses, taking advantage of a larger sample and faster convergence to better options. This option requires more learning platform-supported coordination but might be critical to empower collaboration.

6 Discussion and Conclusion

Discussion. As adaptive experiments combine experimental and optimization perspectives [33], planning them should rely on more elaborate analyses derived from the decision-making perspectives of educational experimentation stakeholders: researchers, instructors, learning designers, and students [36]. Although the set of metrics for analysis might be relatively simple, experimental platform tools should focus on accounting for domain constraints and provide actionable decision-making affordances for stakeholders.

In addition to providing tools for ex-ante power analysis, which has become standard in experimentation [10] and remains critical for adaptive experiments [33], we need tools specific to adaptive experimentation that demonstrate decision consequences of particular patterns in early design stages using statistical data-informed simulations grounded in field experiences and expert design work. This will help stakeholders support a pragmatic, learning impact-centered [40] view on adaptive experiments embedded within complex real-world educational settings [36, 41].

In designing these tools, we should take a responsible and critical approach, going beyond providing average rates and considering how to communicate risks, suggesting adaptive designs only in cases where analysis shows they may be warranted.

Finally, we emphasize that integrating adaptive experimentation within broader learning analytics and engineering workflows should be a key direction for future work. Observational data from existing systems can help identify intervention points, explore potential contexts for personalization, and evaluate the alignment between short-term measures used in adaptive experiments and longer-term learning outcomes. In turn, adaptive experimentation can enhance the causal grounding of learning analytics [12] and help bridge the analytics-action gap [11, 17, 40].

Conclusion. In this paper, based on our experience designing and using the EASI platform across multiple educational settings and design iterations, centered around the XPRIZE Digital Learning Challenge, we highlight the potential of adaptive experimentation to improve learning experiences and outcomes, taking into account the complexities of continuous improvement decision support in modern learning environments.

We present data-informed lessons learned, best practices, and key application scenarios for adaptive experimentation, emphasizing the importance of stakeholder-centered design. Furthermore, we provide practical recommendations for addressing the challenges of implementing adaptive experiments in real educational settings, particularly in smaller classrooms, which are crucial for providing researchers, instructors, and learning designers with agency in evidence-based continuous improvement of their courses.

Acknowledgments

We would like to recognize the support of many individuals who helped our work during the XPRIZE Digital Learning Challenge representing: the University of Toronto (Jiakai Shi, and Kobi Choy), CMU Open Learning Initiative (Raphael Gachuhi, Tanvi Domadia, Gene Hastings, and Hal Turner), the XPRIZE DLC Team (Dr. Monique Golden-Harrison, Shashi Rai, and Devin Krotman), as well as instructors and students of the courses who participated in our trials. We also express our gratitude to the people who helped guide us and shape this work: Vsevolod Suschevskiy, Dr. Nathan Taback, Dr. Ashton Anderson, Dr. Alena Suvorova, Dr. Pavel Okopnyi, and Yi Wang.

This work was partially supported by the XPRIZE and IES, NSF grant #2209819 to Dr. John Stamper, Norman Bier and Dr. Joseph Jay Williams, NIHM grant #R34MH124960, the Office of Naval Research grant numbers #N00014-21-1-2576 and #N00014-18-1-2755, and the Natural Sciences and Engineering Research Council of Canada (NSERC) grant #RGPIN-2019-06968 awarded to Dr. Joseph Jay Williams. Additionally, we acknowledge the financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC) grant #RGPIN-2024-04348 awarded to Dr. Michael Liut.

References

[1] Micah Allen, Davide Poggiali, Kirstie Whitaker, Tom Rhys Marshall, Jordy Van Langen, and Rogier A. Kievit. 2021. Raincloud plots: a multi-platform tool for robust data visualization. *Wellcome Open Research* 4 (Jan. 2021), 63. https://doi.org/10.12688/wellcomeopenres.15191.2

[2] Susan Athey, Undral Byambadalai, Vitor Hadad, Sanath Kumar Krishnamurthy, Weiwen Leung, and Joseph Jay Williams. 2022. Contextual Bandits in a Survey Experiment on Charitable Giving: Within-Experiment Outcomes versus Policy Learning. http://arxiv.org/abs/2211.12004 arXiv:2211.12004 [cs, econ, stat].

[3] Norman Bier, John Stamper, Steven Moore, Darren Siegel, and Ariel Anbar. 2023. OLI Torus: a next-generation, open platform for adaptive courseware development, delivery, and research. In *Companion Proceedings of the 13th International Learning Analytics and Knowledge Conference.* https://stevenjamesmoore.com/assets/papers/lak23_prac_bier.pdf

[4] Djallel Bouneffouf, Irina Rish, and Charu Aggarwal. 2020. Survey on Applications of Multi-Armed and Contextual Bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC).* 1–8. https://doi.org/10.1109/CEC48606.2020.9185782

[5] Olivier Chapelle and Lihong Li. 2011. An Empirical Evaluation of Thompson Sampling. In *Proceedings of the 24th International Conference on Neural Information*

Processing Systems (NIPS'11). Curran Associates Inc., Red Hook, NY, USA, 2249–2257.

[6] Jacob Cohen. 1992. Statistical Power Analysis. *Current Directions in Psychological Science* 1, 3 (June 1992), 98–101. https://doi.org/10.1111/1467-8721.ep10768783

[7] Nina Deliu. 2024. Multinomial Thompson sampling for rating scales and prior considerations for calibrating uncertainty. *Statistical Methods & Applications* 33, 2 (April 2024), 439–469. https://doi.org/10.1007/s10260-023-00732-y

[8] Nina Deliu, Rajenki Das, Angelique May, Joseph Newman, Jo Steele, Melissa Duckworth, Rowena J. Jones, Martin R. Wilkins, Mark R. Toshner, and Sofia S. Villar. 2024. StratosPHere 2: study protocol for a response-adaptive randomised placebo-controlled phase II trial to evaluate hydroxychloroquine and phenylbutyrate in pulmonary arterial hypertension caused by mutations in BMPR2. *Trials* 25, 1 (Oct. 2024), 680. https://doi.org/10.1186/s13063-024-08485-z

[9] Nina Deliu, Joseph J. Williams, and Sofia S. Villar. 2021. Efficient Inference Without Trading-off Regret in Bandits: An Allocation Probability Test for Thompson Sampling. https://doi.org/10.48550/arXiv.2111.00137 arXiv:2111.00137 [stat].

[10] Alexander Eiselmayer, Chat Wacharamanotham, Michel Beaudouin-Lafon, and Wendy E. Mackay. 2019. *Touchstone2*: An Interactive Environment for Exploring Trade-offs in HCI Experiment Design. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland UK, 1–11. https://doi.org/10.1145/3290605.3300447

[11] Stephen Fancsali, April Murphy, and Steven Ritter. 2022. "Closing the Loop" in Educational Data Science with an Open Source Architecture for Large-Scale Field Trials. In *Proceedings of the 15th International Conference on Educational Data Mining*, Antonija Mitrovic and Nigel Bosch (Eds.). International Educational Data Mining Society, Durham, United Kingdom, 834–838. https://doi.org/10.5281/zenodo.6852930

[12] Rebecca Ferguson and Doug Clow. 2017. Where is the evidence?: a call to action for learning analytics. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. ACM, Vancouver British Columbia Canada, 56–65. https://doi.org/10.1145/3027385.3027396

[13] Kaitlyn G. Fitzgerald and Elizabeth Tipton. 2022. The Meta-Analytic Rain Cloud Plot: A New Approach to Visualizing Clearinghouse Data. *Journal of Research on Educational Effectiveness* 15, 4 (Oct. 2022), 848–875. https://doi.org/10.1080/19345747.2022.2031366

[14] Gerd Gigerenzer and Ulrich Hoffrage. 1995. How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review* 102, 4 (Oct. 1995), 684–704. https://doi.org/10.1037/0033-295X.102.4.684

[15] Benjamin Han and Carl Arndt. 2021. Budget Allocation as a Multi-Agent System of Contextual & Continuous Bandits. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. ACM, Virtual Event Singapore, 2937–2945. https://doi.org/10.1145/3447548.3467124

[16] Jake M. Hofman, Daniel G. Goldstein, and Jessica Hullman. 2020. How Visualizing Inferential Uncertainty Can Mislead Readers About Treatment Effects in Scientific Results. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–12. https://doi.org/10.1145/3313831.3376454

[17] Rasmus Leth Jørnø and Karsten Gynther. 2018. What Constitutes an 'Actionable Insight' in Learning Analytics? *Journal of Learning Analytics* 5, 3 (Dec. 2018). https://doi.org/10.18608/jla.2018.53.13

[18] Alex Kale, Francis Nguyen, Matthew Kay, and Jessica Hullman. 2018. Hypothetical outcome plots help untrained observers judge trends in ambiguous data. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 892–902. Publisher: IEEE.

[19] Michael N. Katehakis and Arthur F. Veinott. 1987. The Multi-Armed Bandit Problem: Decomposition and Computation. *Mathematics of Operations Research* 12, 2 (1987), 262–268. https://www.jstor.org/stable/3689689

[20] Matthew Kay, Tara Kola, Jessica R. Hullman, and Sean A. Munson. 2016. When (ish) is My Bus?: User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 5092–5103. https://doi.org/10.1145/2858036.2858558

[21] Yea-Seul Kim, Jake M Hofman, and Daniel G Goldstein. 2022. Putting scientific results in perspective: Improving the communication of standardized effect sizes. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–14. https://doi.org/10.1145/3491102.3502053

[22] Kenneth R. Koedinger, Julie L. Booth, and David Klahr. 2013. Instructional Complexity and the Science to Constrain It. *Science* 342, 6161 (Nov. 2013), 935–937. https://doi.org/10.1126/science.1238056

[23] Ron Kohavi, Diane Tang, and Ya Xu. 2020. *Trustworthy online controlled experiments: a practical guide to A/B testing*. Cambridge University Press, Cambridge, United Kingdom ; New York, NY.

[24] Matthew A. Kraft. 2020. Interpreting Effect Sizes of Education Interventions. *Educational Researcher* 49, 4 (May 2020), 241–253. https://doi.org/10.3102/0013189X20912798 Publisher: American Educational Research Association.

[25] Matthew A. Kraft. 2023. The Effect-Size Benchmark That Matters Most: Education Interventions Often Fail. *Educational Researcher* 52, 3 (April 2023), 183–187. https://doi.org/10.3102/0013189X231155154 Publisher: American Educational Research Association.

[26] Harsh Kumar, Tong Li, Jiakai Shi, Ilya Musabirov, Rachel Kornfield, Jonah Meyerhoff, Ananya Bhattacharjee, Chris Karr, Theresa Nguyen, David Mohr, Anna Rafferty, Sofia Villar, Nina Deliu, and Joseph Jay Williams. 2024. Using Adaptive Bandit Experiments to Increase and Investigate Engagement in Mental Health. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 22906–22912. https://doi.org/10.1609/aaai.v38i21.30328 Number: 21.

[27] Daniël Lakens. 2022. Sample Size Justification. *Collabra: Psychology* 8, 1 (March 2022), 33267. https://doi.org/10.1525/collabra.33267

[28] Tong Li, Jacob Nogas, Haochen Song, Harsh Kumar, Audrey Durand, Anna Rafferty, Nina Deliu, Sofia S. Villar, and Joseph J. Williams. 2022. Algorithms for Adaptive Experiments that Trade-off Statistical Analysis with Reward: Combining Uniform Random Assignment and Reward Maximization. https://doi.org/10.48550/arXiv.2112.08507 arXiv:2112.08507 [cs].

[29] Thomas M.Carsey and Jeffrey J.Harden. 2014. *Monte Carlo Simulation and Resampling Methods for Social Science*. SAGE Publications, Inc.

[30] Ilya Musabirov, Mohi Reza, Steven Moore, Pan Chen, Harsh Kumar, Tong Li, Haochen Song, Jiakai Shi, Koby Choy, Thomas Price, John Stamper, Norman Bier, Nina Deliu, Sofia Villar, Anna Rafferty, Audrey Durand, and Joseph Williams. 2024. Platform-based Adaptive Experimental Research in Education: Lessons Learned from Digital Learning Challenge. In *Companion Proceedings 14th International Conference on Learning Analytics & Knowledge (LAK 24)*. Kyoto.

[31] Ilya Musabirov, Angela Zavaleta Bernuy, Pan Chen, Michael Liut, and Joseph Williams. 2024. Opportunities for Adaptive Experiments to Enable Continuous Improvement in Computer Science Education. In *Proceedings of the 26th Western Canadian Conference on Computing Education (WCCCE '24)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3660650.3660659

[32] Mashfiqui Rabbi, Min Hane Aung, and Tanzeem Choudhury. 2017. Towards Health Recommendation Systems: An Approach for Providing Automated Personalized Health Feedback from Mobile Data. In *Mobile Health*, James M. Rehg, Susan A. Murphy, and Santosh Kumar (Eds.). Springer International Publishing, Cham, 519–542. https://doi.org/10.1007/978-3-319-51394-2_26

[33] Anna Rafferty, Huiji Ying, and Joseph Williams. 2019. Statistical consequences of using multi-armed bandits to conduct adaptive educational experiments. *Journal of Educational Data Mining* 11, 1 (2019), 47–79. https://jedm.educationaldatamining.org/index.php/JEDM/article/view/357

[34] Anna N. Rafferty, Huiji Ying, and Joseph Jay Williams. 2018. Bandit Assignment for Educational Experiments: Benefits to Students Versus Statistical Power. In *Artificial Intelligence in Education*. Vol. 10948. Springer International Publishing, Cham, 286–290. https://doi.org/10.1007/978-3-319-93846-2_53 Series Title: Lecture Notes in Computer Science.

[35] Mohi Reza, Juho Kim, Ananya Bhattacharjee, Anna N. Rafferty, and Joseph Jay Williams. 2021. The MOOClet Framework: Unifying Experimentation, Dynamic Improvement, and Personalization in Online Courses. In *Proceedings of the Eighth ACM Conference on Learning@ Scale*. 15–26.

[36] Steve Ritter, Neil Heffernan, Joseph Jay Williams, Derek Lomas, Klinton Bicknell, Jeremy Roschelle, Ben Motz, Danielle McNamara, Richard Baraniuk, and Debshila Basu Mallick. 2023. Fourth Annual Workshop on A/B Testing and Platform-Enabled Learning Research. In *Proceedings of the Tenth ACM Conference on Learning@ Scale*. 254–256. https://dl.acm.org/doi/abs/10.1145/3573051.3593397

[37] Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. 2019. Are sample means in multi-armed bandits positively or negatively biased? *Advances in Neural Information Processing Systems* 32 (2019). https://proceedings.neurips.cc/paper/2019/hash/65b1e92c585fd4c2159d5f33b5030ff2-Abstract.html

[38] Joseph Jay Williams, Jacob Nogas, Nina Deliu, Hammad Shaikh, Sofia S. Villar, Audrey Durand, and Anna Rafferty. 2021. Challenges in Statistical Analysis of Data Collected by a Bandit Algorithm: An Empirical Exploration in Applications to Adaptively Randomized Experiments. http://arxiv.org/abs/2103.12198

[39] Joseph Jay Williams, Anna N. Rafferty, Dustin Tingley, Andrew Ang, Walter S. Lasecki, and Juho Kim. 2018. Enhancing Online Problems Through Instructor-Centered Tools for Randomized Experiments. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–12. https://doi.org/10.1145/3173574.3173781

[40] Alyssa F. Wise, Simon Knight, and Xavier Ochoa. 2021. What Makes Learning Analytics Research Matter. *Journal of Learning Analytics* 8, 3 (Dec. 2021), 1–9. https://doi.org/10.18608/jla.2021.7647

[41] Betsy Wolf and Erica Harbatkin. 2023. Making Sense of Effect Sizes: Systematic Differences in Intervention Effect Sizes by Outcome Measure Type. *Journal of Research on Educational Effectiveness* 16, 1 (Jan. 2023), 134–161. https://doi.org/10.1080/19345747.2022.2071364

[42] Sam Zhang, Patrick R. Heck, Michelle N. Meyer, Christopher F. Chabris, Daniel G. Goldstein, and Jake M. Hofman. 2023. An illusion of predictability in scientific results: Even experts confuse inferential uncertainty and outcome variability. *Proceedings of the National Academy of Sciences* 120, 33 (Aug. 2023), e2302491120. https://doi.org/10.1073/pnas.2302491120