User-Centered Design to Democratize Research Experiments on Digital Learning Platforms

Marshall An
Carnegie Mellon University
Pittsburgh, USA
haokanga@andrew.cmu.edu

Jessica Fortunato
Carnegie Mellon University
Pittsburgh, USA
jfortun2@andrew.cmu.edu

Ilya Musabirov
University of Toronto
Toronto, Canada
imusabirov@cs.toronto.edu

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

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

Joseph Jay Williams
University of Toronto
Toronto, Canada
williams@cs.toronto.edu

John Stamper
Carnegie Mellon University
Pittsburgh, USA
jstamper@andrew.cmu.edu

ABSTRACT

The expansion of online learning and the growing integration of AI in educational settings create novel opportunities for impactful research and enhanced learning experiences. However, the existing disparities in technical expertise and access to engineering support among researchers highlight the urgent need to democratize online experiments to ensure broad community participation. This paper begins with a critical evaluation of existing research platforms, examining their utility from the perspectives of both researchers and engineers. Building on insights from current platforms, we introduce a user-centered design within the Open Learning Initiative (OLI) Torus, specifically aimed at lowering the barriers to online educational research. We detail the design, development, and integration of features tailored for researchers lacking in technical expertise or dedicated engineering resources. Transitioning from traditional, resource-intensive experimental setups to a more accessible methodology, our work addresses the enduring technical and logistical challenges that have traditionally impeded progress in online educational research. Targeted at the diverse Learning @ Scale community, this work speaks to researchers contending with logistical complexities, instructors looking to incorporate research into their online teaching, and EdTech professionals seeking to support research endeavors. By democratizing access to educational research and encouraging broader participation, we strive to support the expansion of effective learning practices in the age of AI.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **HCI design and evaluation methods**; • **Applied computing** → **Learning management systems**; **Computer-assisted instruction**; **Interactive learning environments**; • **General and reference** → **Empirical studies**; **Experimentation**.

KEYWORDS

A/B Testing, Learning Engineering, Educational Technology, Digital Experimentation, Randomized Control Trials, Online Field Experiments, Adaptive Experimentation, Massive Open Online Courses, Learning Management System (LMS)

1 INTRODUCTION

1.1 Growing Need for Online Educational Research Experiments

The swift expansion of online learning has significantly transformed the educational landscape [6]. Accelerated by global events like the COVID-19 pandemic, the shift towards digital learning platforms has become widespread [5]. This shift has naturally prepared a foundation for researchers to conduct studies, highlighting an increasing need to facilitate experiments within these online environments. The enthusiasm within the learning science community for the innovative and responsible application of AI to enhance learning scalability further amplifies the importance of integrating research into online technology-enhanced learning.

1.2 Embedding Research Experiments in Online Teaching

Online learning platforms present a unique opportunity to embed in-vivo research experiments within authentic teaching contexts. Unlike lab-based experiments, in-vivo studies within naturalistic learning settings offer empirical evidence on the impact of situational factors on learning outcomes [16], yielding insights that are directly applicable to educational practices [20]. This research modality, however, necessitates platform features that cater specifically to the nuanced requirements of educational researchers.

1.3 Standards for Excellence in Education Research (SEER)

To elevate the rigor and impact of educational research, the Institute of Education Sciences (IES) has established the Standards for

Excellence in Education Research (SEER) [11]. These standards underscore the importance of transparency, actionability, and equity in research methodologies. The recent SEER guidelines serve as a beacon for conducting meaningful and impactful educational studies, highlighting the necessity for platform features that support these standards [9].

1.4 An Emerging Community of Platform-Enabled Research

An increasing number of researchers, instructors, and engineers recognize the significance of facilitating online educational research. *SEERNet* (https://seernet.org/), funded by IES, exemplifies collaborative efforts to leverage digital learning platforms for equitable and rigorous research [25]. Furthermore, the Learning at Scale community has been at the forefront of fostering collaboration by hosting *annual workshops on A/B Testing and Platform-Enabled Learning Research* since 2020 [27–30]. Despite the growth of this interest group, the community remains relatively small compared to the potential audience that could benefit from platform-enabled research. This paper aims to catalyze broader engagement and participation, facilitating a more inclusive dialogue on online educational experiments.

1.5 Navigating Research on Learning Platforms

Learning platforms often lack features specifically designed for research, necessitating considerable manual effort in setting up and managing experiments. This paper examines the traditional overhead associated with online experiments and reviews existing research platforms to assess their utility and limitations in supporting educational experiments. Drawing from these insights, we have designed and developed user-centered features within our platform to streamline the research process and reduce operational overhead. We discuss the outcomes of our prototyping and user studies, providing key takeaways that have guided the further refinement and development of these features.

1.6 Target Audience and Paper Contributions

This paper is written for a diverse audience within the Learning @ Scale community, including:

- Researchers, especially those contending with the logistical complexities of running educational research experiments.
- Instructors, especially those using online learning platforms for teaching and interested in embedding research but deterred by the lack of research support features.
- Software engineers and UI/UX designers of learning platforms, especially those seeking to support research experiments on their platforms.

We contribute to the field by providing:

- A discussion on embedding research within online teaching contexts and the associated overhead.
- A timely review and critical reflection on existing research experiment platforms, offering insights from both the researcher's and the engineer's perspectives.
- A report on the design, development, and integration of user-centered features, informed by user studies, tailored for research atop an established learning platform.

2 EXAMINING REPRESENTATIVE TASKS IN CONDUCTING ONLINE EXPERIMENTS

Rather than abstracting from concrete user needs to produce a general specification of the system and its interface—as traditional requirements analysis often does by examining abstract, partial task elements—we adopt the task-centered user interface design process [18]. This approach focuses on designing around *real, complete, and representative tasks* that users must accomplish. In this section, we use an in-depth user narrative to present these tasks required to integrate a randomized experiment into an online course. This narrative, based on the authors' direct experiences with setting up experiments and supporting researchers across various platforms, aims to resonate with a broad audience familiar with these research challenges.

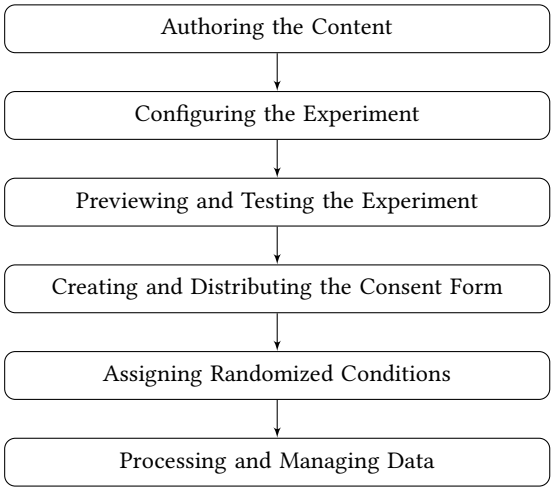

Figure 1: Representative tasks required for conducting online experiments on a digital learning platform.

2.1 Scenario: Integrating an Experiment into an Online Course

Consider the case of an aspiring researcher, who has been serving as a faculty member at a community college. This individual is proficient in utilizing an online learning platform for teaching purposes and views this digital environment as a fertile ground for research aimed at enhancing teaching methods through data analysis. Despite their proficiency in developing and delivering online courses, the educator is constrained to rely on the platform's existing features, due to the lack of dedicated engineering support. This researcher will need to complete a series of representative tasks, as depicted in Figure 1. In the following subsections, we will describe how these tasks are traditionally performed and highlight typical obstacles that researchers encounter.

For illustration in this section, we will reference screenshots created with Canvas LMS, a widely-used, open-source learning

management system, with which the authors have no affiliation. Consider a typical unit of an online curriculum as illustrated in Figure 2, into which our protagonist intends to incorporate a research component. The instructor opts to employ a "between-subject design with test form counterbalancing" to mitigate the variability in assessment difficulty (refer to Figure 3). This particular study design is chosen for our discussion to highlight the administrative hurdles that arise when a single condition impacts multiple learning components (e.g., pre-test, treatment, and post-test), a complexity that is similarly encountered in other research methodologies, such as crossover studies.

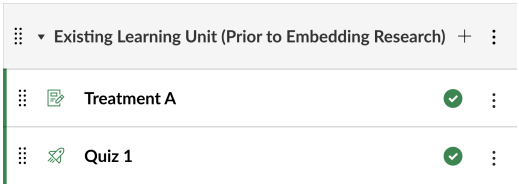

Figure 2: A unit in an existing online curriculum (before embedding a research experiment) showing a learning activity followed by a graded quiz. "Treatment A" here represents the business-as-usual content typically delivered in this unit.

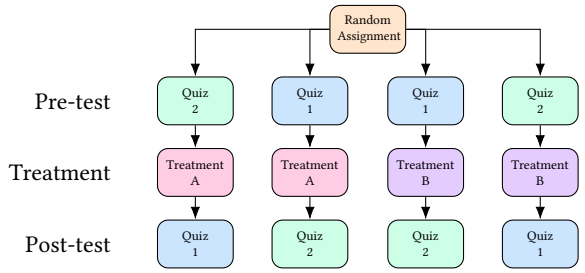

Figure 3: Between-subjects design with test form counterbalancing. This design involves two treatments (Treatment A vs. Treatment B) and two assessments (Quiz 1 vs. Quiz 2), with each participant having an equal chance of being assigned to either treatment and receiving either Quiz 1 or Quiz 2 as the pre-test, followed by the other as the post-test. This methodological approach addresses potential biases in testing, ensuring that differences in learning outcomes between the pre-test and post-test are due to the instructional conditions rather than the assessments' difficulty.

2.2 Authoring the Content

It is possible to make use of Canvas's existing functionalities to orchestrate the planned research study. The process begins with the creation of content, where the researcher develops the pre-test, treatment, and post-test for each condition, as illustrated in Figure 4. A notable aspect of this setup is the deliberate redundancy in content creation, with each quiz being replicated four times and each intervention twice. The decision to duplicate Quiz 1 and Quiz 2 may raise questions—specifically, why not assign the entire cohort to identical singular instances of Quiz 1 and Quiz 2 followed by directing each participant to their specific quiz sequence? This deliberate choice prioritizes mitigating potential errors and participant burden, as a precaution to ensure the integrity of the study's outcomes. While the Canvas' "Question Bank" feature aids in the quiz question development process by allowing for the creation of each question once and its subsequent reuse in multiple instances [13, 14], remember that modifications might often be necessary after the initial content authoring and deployment, e.g., to address issues discovered during testing. Due to the duplication, researchers must meticulously ensure consistent changes across all instances. Canvas cautions users about nuanced concerns regarding updates to deployed questions, emphasizing the need for careful double-checking to avoid unintended consequences [13, 14].

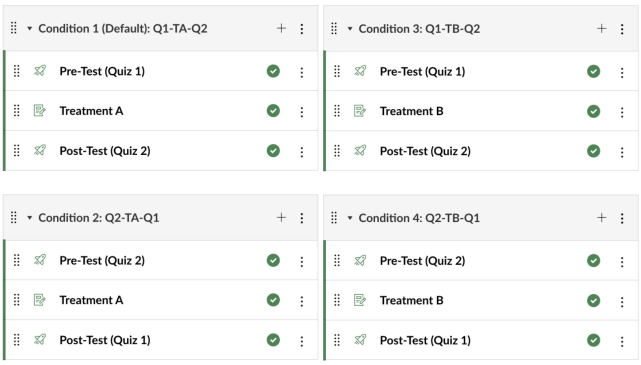

Figure 4: An example of implementing between-subjects design with test form counterbalancing in Canvas.

2.3 Configuring the Experiment

Embedding experiments often necessitate more complex configurations than standard teaching practices. For instance, quizzes in research settings require intricate configurations to ensure the validity of the results. Researchers might configure the pre-test to withhold correct answers and detailed feedback until participants complete all activities in the study; the post-test, being the last activity, may immediately display correctness and feedback. The key point here is not to suggest that withholding correctness or feedback in pre-tests is universally optimal for all learning science experiments, but rather to emphasize that pre- and post-tests commonly require different configurations, demanding manual effort from researchers for setup and verification.

As illustrated in Figure 5, although Canvas offers advanced configuration capabilities for quizzes, such as withholding correct answers until a specified date, the labeling of these options is primarily tailored for instructors. This lack of intuitive clarity regarding their applicability in research contexts may lead to their underutilization among researchers.

2.4 Previewing and Testing the Experiment

The ability to accurately preview and test experiments from a student's perspective poses a significant challenge. Many learning

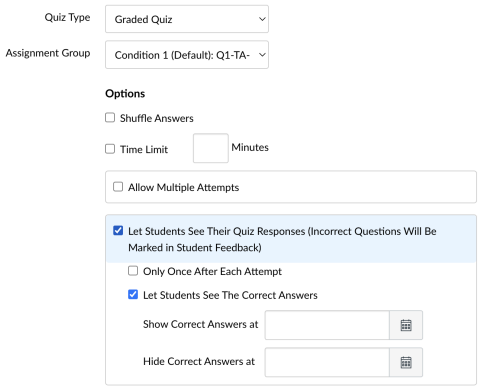

Figure 5: Canvas LMS allows users to disable the quiz option "Let Students See Their Quiz Responses (Incorrect Questions Will Be Marked in Student Feedback)". This feature, while versatile, can be specifically employed to conceal correctness and feedback in pre-tests.

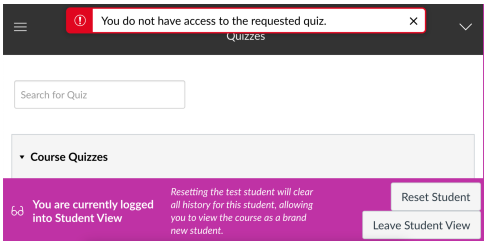

Figure 6: Limitations of Canvas "Student View" arise when content is restricted to specific individuals. The dummy account used in "Student View" cannot be assigned to access content designated only for specific individuals, which hinders accurate previewing.

platforms, including Canvas and Blackboard Learn, provide features that enable instructors to simulate a student's view of a course. However, these platforms can fall short in providing preview functionalities that *accurately mirror* the student's experience in research conditions, as illustrated in Figure 6. Consequently, this necessitates the employment of dummy accounts for manual testing, a method that is not only inefficient but also risks data contamination. An alternative strategy, cloning courses for testing purposes, avoids contaminating actual course data but requires substantial manual effort. Making changes to the experiment's configurations or content may require a repetition of the entire testing process, exacerbating the already significant burden of experiment testing.

2.5 Creating and Distributing the Consent Form

Following content creation, the research experiment moves to the configuration phase. The initial step in conducting ethical research with human participants is to secure informed consent. The instructor creates a digital consent form outlining the study's purpose, procedures, and participant rights. This form is distributed to students, and responses are collected.

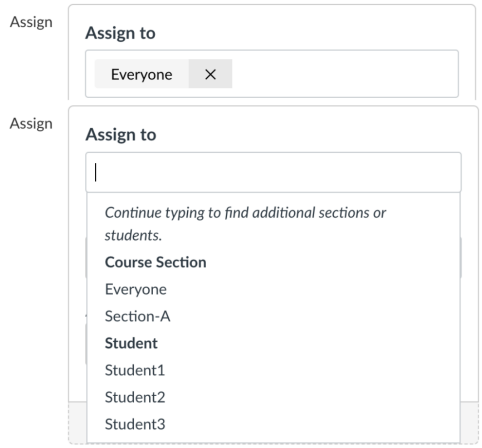

Figure 7: Utilizing the "Assign to" feature in Canvas to manually assign quizzes or tasks to specific students is possible but also cumbersome. To assign a quiz or task to specific students, a user needs to first click the "Remove" icon next to the Everyone label, then start typing and select the names of individual students in the "Assign to" field [12]. This procedure, while straightforward in theory, can be susceptible to human error. For example, quizzes are by default assigned to everyone, and Canvas does not issue a warning if "Everyone" is not removed from the "Assign to" field before assigning to specific individuals. This is because the "Assign to" feature is designed primarily for logistical tasks in teaching, such as granting individual deadline extensions, rather than tailored for research condition assignments. The operational complexity of manual condition assignment increases with the class size; for instance, in a class of 40 participants, manual condition assignment would require at least 40 * 3 (pre-test, treatment, and post-test) = 120 operations.

2.6 Assigning Randomized Conditions

After collecting consent forms, the instructor first separates students who opted out into a separate group, ensuring they receive the default condition. Next, they employ true randomization to assign participants who opted in to their respective conditions. The instructor utilizes Canvas to assign them accordingly. As Figure 7 illustrates, utilizing the Canvas "Assign to" feature to allocate students to their designated conditions is feasible but fraught with potential for error and inefficiency.

2.7 Processing and Managing Data

Data analysis follows experiment completion. However, this stage presents complexities. Test data, such as data generated by dummy accounts, must be meticulously separated from participant data. Further complicating aggregation are opt-out student removal, anonymization, and data protection compliance. Additionally, in-vivo experiments often necessitate dual data processing streams: one for research and one for course grade. These streams, though similar, hold nuanced and critical differences in data management and processing.

2.8 Reflecting on the Traditional Workflow

While the traditional online experiment workflow serves its purpose, it presents a significant administrative burden. Despite researchers' commitment to meticulousness, the numerous manual steps outlined earlier inherently elevate the risk of human error. Recognizing the laborious and error-prone nature of the traditional workflow, we turned our attention to existing efforts aimed at streamlining the process and mitigating error, as detailed in the next section.

3 APPROACHES TO A/B TESTING IN THE TECHNOLOGY INDUSTRY

To maintain a competitive edge and respond effectively to evolving market conditions, the technology industry has increasingly adopted agile methodologies to accelerate production improvements and enhance value delivery [4]. Continuous innovation and improvement are recognized as the cornerstones of these efforts [7]. One effective strategy for achieving continuous innovation and improvement is A/B testing, which swiftly evaluates alternatives to determine and adopt the more effective option promptly. To lay the groundwork for research-centered features, we have examined the primary approaches to A/B testing commonly employed within the technology industry. This section will describe the key differences and potential adaptations when applying these approaches in educational research settings.

A research team at Amazon.com, Inc. has identified two primary methods of A/B testing in the technology industry: redirecting users to different links and utilizing front-end real-time rendering within the same link [33]. Each approach offers distinct advantages and challenges, especially when viewed through the lens of educational research.

- **Redirection Approach:** This method allows for significant variations in instructional design and sequence. It is particularly useful for educational research that demands broad-scope modifications. However, as noted by the Amazon research team, this approach may lead to asset duplication and the risk of learners encountering outdated or incorrect URLs [33].
- **Real-Time Rendering Approach:** Suitable for more granular modifications, such as tweaks to individual pages or learning objects, this method pre-deploys all conditions within a single page, hidden behind conditional statements. It simplifies the process of identifying and transitioning to the winning condition for future learners, eliminating the need for course redeployment or link redistribution.

While the redirection approach offers depth and flexibility for wide-ranging experiments, real-time rendering provides efficiency and ease of management for smaller-scale modifications. The choice between these approaches depends on the specific needs of the research experiment, balancing the scope of variation against operational simplicity.

4 SYSTEM REVIEWS

Next, we will report a critical reflection on existing systems designed to facilitate platform-enabled research. The selection of these platforms for examination is based on meeting one or both of the following criteria: their consistent participation and presentation of updates at previous *annual workshops on A/B Testing and Platform-Enabled Learning Research*; and/or their being funded by IES as part of *SEERNet*, which demonstrates adherence to and contribution towards the Department of Education's SEER Principles for educational research.

4.1 System Review #1: UpGrade

Developed by Carnegie Learning in partnership with PlayPower Labs, UpGrade is a free, open-source A/B testing framework designed to facilitate randomized experiments in educational software [31]. Based on our engineering team's efforts to deploy an instance of UpGrade and integrate UpGrade with our learning platform, we present an analysis of UpGrade's contributions and limitations.

4.1.1 Key Contributions of UpGrade. UpGrade introduces several innovative features tailored to the unique demands of educational research, including:

- **Centralized Experiment Management:** UpGrade offers a unified platform enabling researchers to design experiments, automate participant randomization, and monitor progress with customizable metrics.
- **Individual and Group Consistency:** UpGrade allows researchers to enable "individual consistency" for maintaining consistent experimental conditions across multiple learning units for individuals. Alternatively, researchers may choose "group consistency", which will ensure that entire groups such as classes or schools experience uniform experimental conditions, which is particularly beneficial in classroom settings to avoid disruption.
- **Post Experiment Rule:** UpGrade provides options for defining participant experiences post-experiment, including continuation in the assigned condition, reverting to a default state, or transitioning to the most effective condition.

4.1.2 Challenges in Adopting UpGrade. Despite its strengths, several barriers hinder the adoption of UpGrade:

- **Technical Complexity in User Experience:** The interface introduces unnecessary technical complexity, using technical jargon (e.g., "App Context", "Payloads") that may deter non-technical researchers.
- **Overexposure of Advanced Features:** The availability of advanced features for niche scenarios can overwhelm users, complicating the interface for those in need of basic experiment setups.
- **Separate Web Service Integration:** Operating as an independent web service, UpGrade requires users to alternate between it and their primary educational tools, complicating the experiment setup process.
- **Lack of Fine-Grained Access Control:** The absence of detailed Role-Based Access Control (RBAC) poses compliance challenges in environments with multiple research teams, compounded by the impracticality of each researcher team managing a separate UpGrade instance.

4.1.3 Overall Assessment. UpGrade contributes to platform-enabled research through its provision of a platform for centralized experiment management. It highlights the importance of considering the complexities inherent in educational research environments. The

platform distinguishes itself with features such as group experiment consistency and post-experiment flexibility, showcasing its capabilities. However, the technical requirements and the necessity for integration with UpGrade as a separate web service present ongoing challenges. The platform's rigorous approach to managing the nuances of educational research, particularly in handling edge cases, is noteworthy. Meanwhile, simplifying the interface by concealing advanced settings would greatly assist researchers in search of more straightforward configurations. Implementing the principle of gradual exposure, as elaborated in the Discussion section), could markedly facilitate UpGrade's adoption, extending its reach beyond the initial cohort of technologically adept users.

4.2 System Review #2: The AdapComp/MOOClet Framework

The AdapComp (Adaptive Component) Framework, also known as the MOOClet Framework, introduces a novel framework that transforms user interface components into *Adaptive Components* that are dynamic, mutable, adaptable, and personalized [26]. Designed to be context-independent, AdapComp's infrastructure and technical specifications can support a diverse range of settings, from education to healthcare to website design. In educational research, in particular, AdapComp enables the dynamic rendering of adaptive learning resources, significantly enriching the possibilities for A/B testing and personalized learning experiences.

To facilitate understanding among a broader audience, including researchers without extensive technical expertise, we will intentionally describe the operation of the AdapComp in a simple, less technical manner. Here is how an Adaptive Component functions:

(1) A learning platform sends a request to AdapComp; this request essentially asks, *"Given this student and the context, which version of the learning materials should be assigned for this Adaptive Component?"*
(2) AdapComp determines the appropriate version for the user, and sends a response back to the learning platform.
(3) The learning platform then displays the chosen version of the content to the student.

To achieve this functionality, the AdapComp infrastructure consists of three core elements:

- **Version Set:** Different versions of learning content from which AdapComp can select, i.e., *"Which versions to choose from?"*.
- **Policy Set:** Rules or algorithms that determine which version to choose, i.e., *"How to choose a version?"*.
- **Learn Data Store:** The repository of data related to learning interactions and outcomes, serving as the input for deciding on the version assignment, i.e., *"What data to inform the choice?"*

Drawing on our prior successful integration of AdapComp with our learning platform, we present a nuanced analysis of both the framework's strengths and areas for potential improvement.

4.2.1 Key Contributions of AdapComp.

- **Dynamic Content Rendering:** At the heart of AdapComp is its real-time rendering capability. When a student accesses a learning resource, the client platform requests content via API calls to the AdapComp infrastructure relaying relevant learning context. The infrastructure then assigns and returns the specific content for the client platform to render.
- **Advanced Condition Assignment Policies:** Beyond simple randomization, AdapComp supports an array of advanced condition assignment policies, with adaptive multi-armed bandit algorithms being a key highlight. In educational experiments, researchers and instructors face an exploration-exploitation tradeoff: exploration involves gathering more data to become increasingly certain about the effectiveness of an intervention, while exploitation refers to promptly using gathered evidence to benefit from effective interventions [22]. Striking the right balance is crucial for enhancing educational practices in a fair and empirically grounded manner. AdapComp, equipped with adaptive algorithms, can automatically phase out less effective conditions in a timely manner. A case study presented in [23] demonstrates how these capabilities can support the rapid application of gathered evidence, ensuring that a larger portion of participants (compared to random assignment) receive more effective conditions while maintaining the statistical power required for research.
- **Personalization of Learning Experiences:** A single winning condition may not adequately address the diverse needs of all learners. As highlighted by the expertise reversal effect [15], there is a critical need to tailor instructional techniques as learners' knowledge develops. AdapComp supports this personalization through its flexible policies for condition assignment, enabling the system to deliver instructional materials tailored to individual progress and skill levels. For example, when integrated with a Learn Data Store that captures necessary learning data, AdapComp is capable of personalizing feedback for students in accordance with Shute's guidelines on delivering formative feedback [35]: directive feedback that guides the learning process is provided to those with less mastery, while facilitative feedback, which poses challenges and promotes exploration, is offered to more advanced learners.

4.2.2 Challenges in Adopting AdapComp. Despite its innovative approach, adopting AdapComp presents several challenges:

- **Integration Complexity:** The necessity for external infrastructure integration adds a layer of complexity to utilizing AdapComp. This includes the technical challenge of making multiple, sequenced API calls to retrieve condition-specific content for students.
- **Content Authoring and Preview Difficulties:** AdapComp's external content hosting deviates from native authoring tools in many learning platforms. This can obstruct seamless content creation and previewing processes, detracting from the design experiences educators and researchers are accustomed to.
- **Data Management Hurdles:** The external hosting of educational content and data logging requirements pose significant challenges in ensuring data integrity and compliance with privacy standards. The operational overhead associated with managing this data could exceed the capacity of many research teams, potentially restricting the framework's adoption.

4.2.3 Overall Assessment. AdapComp's forward-thinking features, such as real-time content rendering and sophisticated condition assignment, offer vast potential for conducting educational research

and personalizing learning. However, the integration and data management challenges require strategic solutions to unlock its full potential as an impactful tool in digital learning environments.

4.3 System Review #3: E-TRIALS

E-TRIALS (an EdTech Research Infrastructure to Advance Learning Science) is a specialized testbed built atop the ASSISTments platform to facilitate streamlined educational research in mathematics. Initiated by Krichevsky et al. [17] and further developed by McCarthy [19], E-TRIALS equips researchers with a conducive environment for executing randomized educational studies. E-TRIALS streamlines the execution of student-level randomized experiments by tapping into ASSISTments' rich repository of mathematics problems and its substantial user engagement, thereby facilitating the entire research cycle from participant recruitment to study design implementation. Also, E-TRIALS provides an intuitive interface complemented by a preview panel that enables researchers to visually plan and modify their study designs in real time, enhancing the research setup experience for users of varying technical backgrounds. In summary, E-TRIALS presents a user-centered and efficient platform tailored for education research. Leveraging the ASSISTments ecosystem, E-TRIALS streamlines the process of setting up and conducting experiments. The foundation of E-TRIALS' success is deeply rooted in its strategic alignment with the core principles established at the inception of ASSISTments. Distinct from platforms primarily focused on curriculum delivery, ASSISTments was designed as a "platform for learning sciences" [8]. This research-oriented ethos forms the cornerstone of ASSISTments' notable achievements and its extensive adoption within the educational data mining research community.

4.4 System Review #4: Terracotta

Terracotta (Tool for Education Research with RAndomized COnTrolled TriAls) emerges as an innovative bridge connecting research endeavors with everyday educational practices [21]. As a platform-specific plugin developed for Canvas LMS, Terracotta leverages the platform's existing infrastructure to simplify the experimental setup process for educational researchers. Through an intuitive interface, researchers can specify their experimental configurations, which Terracotta then translates into the requisite operational components within Canvas to facilitate the study. This integration allows for the seamless incorporation of various research elements, such as consent forms and alternative content modules, directly within the Canvas LMS. By doing so, Terracotta ensures minimal disruption and enhances the user experience for both researchers and study participants, fostering a more integrated and efficient research environment.

4.5 Executive Summary of System Reviews

Table 1 summarizes the comparative analysis of four distinct educational research platforms: UpGrade, AdapComp, E-TRIALS, and Terracotta, focusing on their platform specificity, barriers to adoption, and their contributions to the domain of educational research. The examination of prior works and system reviews reveals a fundamental trade-off: platform-agnostic systems, while flexible, introduce considerable complexity and integration challenges. In contrast, platform-specific systems simplify usage and integration but restrict the scope of application to particular learning environments. This insight points to the need for a solution that blends ease of integration with the flexibility to meet the varied demands of educational research, without introducing complex technical barriers that might deter researchers.

5 PROTOTYPING AND USER STUDY

After examining industry practices and existing systems, we extended our existing learning platform by integrating a prototype component designed to support research experiments. This addition enables the collection of empirical data through user studies.

5.1 A Brief Overview of Our Existing Platform: Open Learning Initiative (OLI) Torus

The Open Learning Initiative (OLI) at Carnegie Mellon University offers a comprehensive suite of online courses and resources that meets the needs of both instructors and learners in blended and fully online educational settings. OLI distinguishes itself by its emphasis on immediate feedback for students and an instructor dashboard for tracking learning outcomes. OLI is extensively used, with over 5 million independent learners and more than 750,000 enrollments from hundreds of academic institutions. Additionally, over 6,000 educators have created instructor accounts and 300 courses have been authored. Notably, OLI supports research by providing access to over 700 researchers for both primary and secondary analysis of its datasets. Decades of dedication to leading the development of evidence-based, research-driven adaptive courseware have enhanced learning outcomes and established a robust testbed for learning research. The insights gained from collaborations with numerous researchers and educators, coupled with the support of a dedicated engineering team, have positioned us at the forefront of platform-enabled educational research. Torus is the next generation implementation of the OLI's adaptive learning platform.

5.2 Prototype Design

The prototype extends the existing content authoring and release functionality of our learning platform by incorporating features specifically tailored for researchers to configure experiments. We have integrated the UpGrade platform for its research condition assignment and management capacity. To streamline the workflow, we minimized direct user interaction with the UpGrade interface. Therefore, our prototype provides a feature that allows users to complete their configuration and then export a configuration file. This file can be directly imported into UpGrade, connecting the two platforms and enabling condition assignment. The entire process for setting up and launching an A/B experiment through this prototype requires 18 distinct steps.

5.3 User Study Design

The user study utilizes a think-aloud protocol, wherein participants configure a minimal research experiment, specifically a randomized controlled trial with two conditions. During the think-aloud sessions, participants interact with the system and verbalize their thought processes, providing real-time insights into their user experience. The think-aloud activity is followed by a semi-structured

Table 1: Comparison of Educational Research Platforms

Feature	UpGrade	AdapComp	E-TRIALS	Terracotta
Platform Specificity	Agnostic	Agnostic	Specific (ASSISTments)	Specific (Canvas LMS)
Barriers to Adoption	High	High	Low	Low
Key Features	Centralized management for experiments	Real-time rendering, Advanced condition assignment policies	Streamlined experiment setup and execution, Intuitive interface	Seamless experiment embedding, Intuitive interface

interview and a System Usability Scale (SUS) survey. The SUS, a ten-item Likert scale questionnaire, provides a global view of subjective usability assessment [3], which is widely used and recognized as a highly robust and versatile tool in usability evaluation [2]. Each session lasted for 1 hour.

5.4 Results and Takeaways

We conducted a user study with 6 participants, including researchers, content developers, and instructors.

- **Quantitative Measure:** The SUS score obtained from the user study provides a quantitative measure of the usability of our prototype, which recorded a SUS score of 42. According to the synthesis by Bangor et al., a score of 42 is qualitatively described as "poor," indicating significant usability challenges faced by users [2]. This quantitative assessment is consistent with the qualitative feedback from the study's participants, who reported various difficulties and frustrations, as summarized below.
- **Qualitative Feedback:** The user studies identified several usability issues with the prototype. Firstly, the workflow for setting up experiments was perceived as complex and time-consuming. Additionally, the interface did not intuitively guide users through this workflow, often necessitating navigation across multiple pages, which led to confusion and frustration. Despite efforts to minimize direct user interaction with the UpGrade platform, the integration with UpGrade still appeared overly complex, and users found that direct exposure to UpGrade was not particularly helpful. There was also a notable demand for onboarding training to help users fully utilize the platform's features, as many found the process challenging to navigate by themselves.

As demonstrated by Bangor et al., user experience design inherently involves an iterative process. The introduction of new, untested features typically results in lower initial SUS scores; and these scores generally improve as user feedback is incorporated and modifications are implemented in subsequent iterations [2]. The results of this user study underscore the necessity for further iterations of the new prototype, incorporating user feedback to enhance the system. Key improvements needed include a simplified workflow, clearer navigation, and more effective onboarding processes, all of which are essential for enabling users to effectively adopt and utilize platform-enabled research solutions.

6 OUR PROPOSED APPROACH: STREAMLINING RESEARCH EXPERIMENTS ON LEARNING PLATFORMS

Reflecting on the industry practices, existing research platforms and our user studies, we propose the following approach:

6.1 Laying the Groundwork with a Platform-Specific Strategy

To strike a balance between flexibility and accessibility, we advocate for *initially* adopting a platform-specific strategy. This strategy involves adding experimental research functionalities—such as A/B testing—directly atop existing learning platforms. This approach leverages the existing features and interfaces of these platforms that users are familiar with, which provides the following benefits:

- **Enhancing User Experience:** Adding educational research functionalities atop existing learning platforms eliminates the need for researchers to navigate multiple systems for setting up, validating, monitoring, and analyzing experiments. Such integration reduces administrative overhead and cognitive burdens.
- **Reducing Technical Barriers:** A unified approach mitigates the technical complexities associated with external integration. Recognizing that many in the educational research community may not have access to dedicated engineering support, embedding functionalities within a learning platform reduces reliance on external technical assistance. It makes advanced research tools more accessible to a wider range of users, including those with limited technical expertise.
- **Simplifying Data Management:** Integrating research functionalities within a single platform offers a unified solution for data management, addressing challenges related to data storage, transfer, and privacy across multiple platforms. Centralizing all functions within a single platform simplifies data management, reducing the potential for data breaches and easing adherence to data protection regulations. This also eliminates the inherent complexity of combining data from separate platforms, particularly when these platforms are developed with varied technical stacks and logging standards.

Recommending the platform-specific strategy is not to discredit the value of platform-agnostic solutions such as UpGrade and AdapComp. Meanwhile, for platform-agnostic solutions, we argue that there should be a clear separation of concerns between distinct user personas. Even if a separate service is used, direct exposure to the external service should be minimized, or ideally eliminated, so that most researchers can set up and manage their experiments within a single learning platform with which they are more familiar.

6.2 Forward-Looking of an Interoperable Platform-Enabled Research Standard

Our intention is not to dismiss the need of a research platform that is both accessible and interoperable. Rather, we aim to build upon our existing learning platform to provide a practical example that contributes to the formulation of universal standards for

tools used in educational research. Establishing an interoperable infrastructure for streamlined research experiments in educational settings necessitates a standardized set of API protocols universally accepted across learning platforms.

To illustrate the potential and practicality of such standards, we can look to existing success stories within the educational community. A prominent example is the widespread adoption of Learning Tools Interoperability (LTI) by 1EdTech, which has demonstrated the benefits and challenges of establishing interoperable standards.

The efficacy of LTI as an educational technology specification that facilitates interoperability among diverse learning tools across different platforms can be ascribed to two principal factors. First, LTI captures and automates essential functionalities like seamless enrollment and grade passback, thereby alleviating the manual workload typically borne by educators [34]. Secondly, the technical barrier for adopting LTI is relatively low, supported by a community that develops and maintains libraries in multiple programming languages.

However, it is equally important to reflect upon the setbacks encountered by the LTI initiative. The ambition to incorporate extensive functionalities in LTI versions 1.2 and 2.0 resulted in significant complexity, which substantially hindered their adoption, leading to their deprecation [10]. This highlights the critical balance between comprehensive functionalities and user-friendliness. Therefore, we caution against prematurely adding complex features that could introduce unnecessary technical complexities and administrative burdens.

7 KEY FEATURES TAILORED FOR RESEARCH

In this section, we detail the design, development, and integration of the research capacity on top of our existing learning platform, informed by the prototype and insights gathered from the previous user study. User experience design is inherently iterative; accordingly, our team conducted three additional internal reviews and iterations of the interface prototyping, leading to a stable version ready for beta release. We will present the core features designed to streamline the process of embedding in-vivo educational experiments within existing curricula, ensuring minimal disruption to the educators' workflow while providing powerful tools for researchers.

Figure 8: This annotated screenshot showcases the interface for editing decision points in condition assignment. It illustrates an example decision point defining four conditions for a between-subjects design with test form counterbalancing.

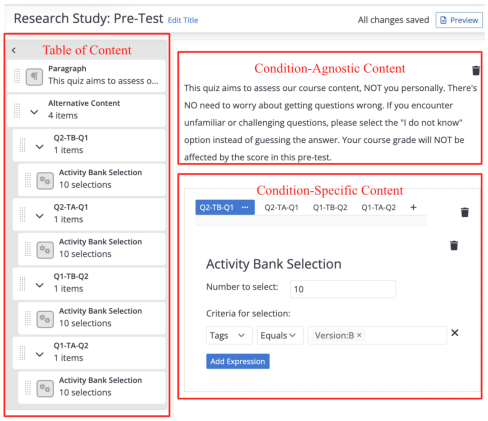

Figure 9: This screenshot presents the content authoring interface designed for authoring content variants for research experiments. This example contains four tabs, each corresponding to a distinct condition, governed by a central decision point. On the left, a Table of Contents panel enables researchers to review the page's structure, facilitating the verification of condition parity directly from a single interface and eliminating the necessity to navigate through multiple pages for comparison. Additionally, the interface accommodates both condition-specific and condition-agnostic content.

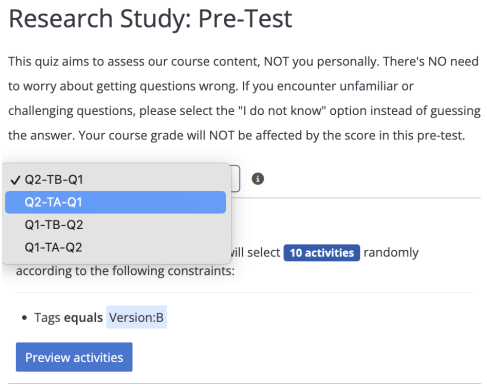

Figure 10: This screenshot shows the content preview feature, which enables the preview of content variants. Similar to previewing uniform content, this feature provides a familiar experience, allowing researchers to select and preview different conditions from one interface.

7.1 Streamlined Condition Assignment

Inspired by UpGrade, we introduced a "decision point" feature (refer to Figure 8) to simplify condition assignment. This feature enhances control across multiple learning components, such as pre-tests, treatments, and post-tests, eliminating the need for the repetitive and laborious task of duplicating condition assignments in each learning component to maintain consistency. The subsequent section on content authoring will detail the use of decision points in controlling the assignment to different content variants.

7.2 Streamlined Content Authoring for Experiments

To address the need for seamless integration of research experiments into educational materials, we have developed an enhanced content authoring process. This feature, as illustrated in Figure 9, enables instructors and researchers to author content variations for experiments within the familiar environment of the existing content authoring interface. The creation of content variations is designed to be intuitive, requiring only slight modifications from the standard authoring procedure, such as the incorporation of alternative content configurations within the same interface. By maintaining a close resemblance to the standard authoring practices, this process provides a non-intrusive user experience, promoting swift adaptation and enhancing the overall efficiency of content development for experimental purposes.

7.3 Streamlined Experiment Preview

As Figure 10 illustrates, to reduce overhead in experiment testing, our platform enables researchers to validate configurations and content without the need to create test student accounts or publish experiments prematurely. This feature is built upon existing course preview functionalities, offering a consolidated view for easy comparison and validation of experiment variants.

8 DISCUSSION

8.1 Adopting the Progressive Disclosure Design Principle

In the design of digital learning platforms, a continuous tension exists between accommodating basic features for novice users and incorporating advanced features for expert users. Drawing from industry best practices, we observe that leading IT companies manage this tension by employing the principle of progressive disclosure, which relegates advanced or rarely used features to a secondary interface layer, making an interface easier to learn and less error-prone [24]. For instance, Google Cloud Platform (GCP), as illustrated by Figure 11, employs the design principle of progressive disclosure to fulfill multiple objectives: it minimizes clutter and cognitive load for users who primarily require basic functionality, and it provides advanced users with access to advanced options without overwhelming novice users. We advocate for the adoption of progressive disclosure in the user interface design for research experiments. A minimally viable set of features should be provided in the default workflow, with advanced features becoming visible only when a user actively opts to access them.

8.2 Creating Video Tutorials to Onboard Users

Recognizing the indispensable need for effective onboarding training, as highlighted by the user study, we plan to create tutorials in the format of demonstration videos, given that research indicates video tutorials outperform traditional paper-based guides in supporting procedural knowledge development by software users [36]. These how-to videos will be designed in accordance with demonstration-based training (DBT) principles [32], integrating dynamic examples of task performance with instructional features.

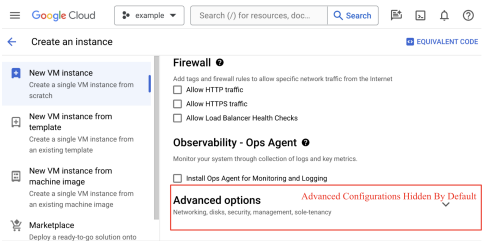

Figure 11: Google Cloud Platform (GCP) employs the progressive disclosure principle in their web console.

This approach, grounded in Bandura's theory of observational learning [1], has been shown to significantly enhance the development of procedural knowledge and motivation for software users [37]. Furthermore, creating these video tutorials offers additional benefits. First, the act of recording and articulating the procedure serves as a sanity check for the engineering team, confirming that the demonstrated procedures are not overly complex and are manageable for the intended users. Second, while maintaining video tutorials requires effort due to the need for updates when the interface changes, each iteration of the video tutorials provides a tangible piece of evidence that chronicles the evolution of platform functionality and usability over time. This archival value offers a clear, visual history that informs future updates and facilitates knowledge sharing within the community.

8.3 Limitations and Future Work

The completed development of researcher-centered features within the learning platform has established a solid foundation, serving as a tangible proof of concept. However, we acknowledge that these features still require further refinement. These features will undergo continuous, iterative development and rigorous assessment to align precisely with user needs and expectations. Our goal is to achieve a SUS score of approximately 70, which corresponds to an adjective rating of "good." Future plans include conducting additional rounds of user studies. We plan to share our findings and subsequent improvements in future academic publications.

9 CONCLUSION

This paper contributes to ongoing efforts in platform-enabled learning research, with a particular focus on democratizing online educational research. We aim to lower barriers to adoption, especially for researchers who lack technical expertise or dedicated engineering resources. We present a user-centered design that integrates industry best practices with insights drawn from existing platforms, aiming to simplify the technical and logistical complexities often encountered in digital learning experiments. While the iterative nature of our platform development calls for ongoing evaluation and refinement, this work provides a valuable perspective on enhancing the usability of research tools in educational technology. By sharing our developments and promoting open collaboration, we strive to foster further dialogue and innovation within the community, contributing to the broader goal of seamlessly integrating research into educational settings.

ACKNOWLEDGMENTS

This material is based upon work partially supported by the National Science Foundation under Grant No. 2209819. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Albert Bandura et al. 1986. Social foundations of thought and action. *Englewood Cliffs, NJ* 1986, 23-28 (1986), 2.

[2] Aaron Bangor, Philip T Kortum, and James T Miller. 2008. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction* 24, 6 (2008), 574–594.

[3] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[4] Amadeu Silveira Campanelli and Fernando Silva Parreiras. 2015. Agile methods tailoring–A systematic literature review. *Journal of Systems and Software* 110 (2015), 85–100.

[5] Shivangi Dhawan. 2020. Online learning: A panacea in the time of COVID-19 crisis. *Journal of educational technology systems* 49, 1 (2020), 5–22.

[6] Rizky Firmansyah, Dhika Putri, Mochammad Wicaksono, Sheila Putri, Ahmad Widianto, and Mohd Palil. 2021. Educational transformation: An evaluation of online learning due to COVID-19. *International Journal of Emerging Technologies in Learning (iJET)* 16, 7 (2021), 61–76.

[7] Brian Fitzgerald and Klaas-Jan Stol. 2014. Continuous software engineering and beyond: trends and challenges. In *Proceedings of the 1st International Workshop on rapid continuous software engineering*. 1–9.

[8] Neil T Heffernan and Cristina Lindquist Heffernan. 2014. The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education* 24 (2014), 470–497.

[9] Carolyn J Hill, Lauren Scher, Joshua Haimson, and Kelly Granito. 2023. Conducting Implementation Research in Impact Studies of Education Interventions: A Guide for Researchers. Toolkit. NCEE 2023-005. *National Center for Education Evaluation and Regional Assistance* (2023).

[10] IMS Global Learning Consortium. 2021. Security Update and Deprecation Schedule for Early Versions of LTI. https://www.1edtech.org/lti-security-announcement-and-deprecation-schedule Accessed on June 3, 2024.

[11] Institute of Education Sciences. 2022. Standards for Excellence in Education Research - About. https://ies.ed.gov/seer/ Accessed on February 11, 2024.

[12] Instructure Community. 2024. How do I assign an assignment to an individual student? https://community.canvaslms.com/t5/Instructor-Guide/How-do-I-assign-an-assignment-to-an-individual-student/ta-p/717 Accessed on February 15, 2024.

[13] Instructure Community. 2024. How do I create a quiz by finding questions in a question bank? https://community.canvaslms.com/t5/Instructor-Guide/How-do-I-create-a-quiz-by-finding-questions-in-a-question-bank/ta-p/1034 Accessed on February 15, 2024.

[14] Instructure Community. 2024. How do I create a quiz with a question group linked to a question bank? https://community.canvaslms.com/t5/Instructor-Guide/How-do-I-create-a-quiz-with-a-question-group-linked-to-a/ta-p/1033 Accessed on February 15, 2024.

[15] Slava Kalyuga. 2009. The expertise reversal effect. In *Managing cognitive load in adaptive multimedia learning*. IGI Global, 58–80.

[16] Kenneth R Koedinger, Julie L Booth, and David Klahr. 2013. Instructional complexity and the science to constrain it. *Science* 342, 6161 (2013), 935–937.

[17] Nicholas Krichevsky, Kamryn Spinelli, Neil Heffernan, Korinn Ostrow, and Mr Ryan Emberling. 2020. *E-TRIALS*. Ph. D. Dissertation. Doctoral dissertation, Worcester Polytechnic Institute.

[18] Clayton Lewis and John Rieman. 1993. Task-centered user interface design. *A practical introduction* (1993).

[19] Tim McCarthy. 2021. Continuing the Development of E-TRIALS.

[20] Benjamin A Motz, Paulo F Carvalho, Joshua R de Leeuw, and Robert L Goldstone. 2018. Embedding experiments: Staking causal inference in authentic educational contexts. *Journal of Learning Analytics* 5, 2 (2018), 47–59.

[21] Benjamin A. Motz, Öykü Üner, Harmony E. Jankowski, Marcus A. Christie, Kim Burgas, Diego Del Blanco Orobitg, and Mark A. McDaniel. 2023. Terracotta: A tool for conducting experimental research on student learning. *Behavior Research Methods* (July 2023). https://doi.org/10.3758/s13428-023-02164-8

[22] Ilya Musabirov, Mohi Reza, Steven Moore, Pan Chen, Harsh Kumar, Li Tong, Fred Haochen Song, Jiakai Shi, Koby Choy, Thomas Price, et al. 2024. Platform-based Adaptive Experimental Research in Education: Lessons Learned from Digital Learning Challenge. In *The Fourteenth International Conference on Learning Analytics & Knowledge (LAK24): Learning Analytics in the Age of Artificial Intelligence*. Association for Computing Machinery, Inc., 37–40.

[23] Ilya Musabirov, Angela Zavaleta-Bernuy, Pan Chen, Michael Liut, and Joseph Jay Williams. 2023. Opportunities for Adaptive Experiments to Enable Continuous Improvement that Trades-off Instructor and Researcher Incentives. *arXiv preprint arXiv:2310.12324* (2023).

[24] Jakob Nielsen. 2006. Progressive disclosure. https://www.nngroup.com/articles/progressive-disclosure/ Accessed on February 16, 2024.

[25] Stefani Pautz Stephenson, Rebecca Banks, and Deblina Pakhira. 2022. *Practitioners at the Center: Catalyzing Research on Problems of Practice in Realistic Settings*. Technical Report. Digital Promise.

[26] 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* (Virtual Event, Germany) *(L@S '21)*. Association for Computing Machinery, New York, NY, USA, 15–26. https://doi.org/10.1145/3430895.3460128

[27] Steve Ritter, Neil Heffernan, Joseph Jay Williams, Derek Lomas, and Klinton Bicknell. 2021. Second Workshop on Educational A/B Testing at Scale. In *Proceedings of the Eighth ACM Conference on Learning @ Scale* (Virtual Event, Germany) *(L@S '21)*. Association for Computing Machinery, New York, NY, USA, 1–3. https://doi.org/10.1145/3430895.3460876

[28] Steve Ritter, Neil Heffernan, Joseph Jay Williams, Derek Lomas, Klinton Bicknell, Jeremy Roschelle, Ben Motz, Danielle McNamara, Richard Baraniuk, Debshila Basu Mallick, Rene Kizilcec, Ryan Baker, Stephen Fancsali, and April Murphy. 2023. Fourth Annual Workshop on A/B Testing and Platform-Enabled Learning Research. In *Proceedings of the Tenth ACM Conference on Learning @ Scale* (Copenhagen, Denmark) *(L@S '23)*. Association for Computing Machinery, New York, NY, USA, 254–256. https://doi.org/10.1145/3573051.3593397

[29] Steven Ritter, Neil Heffernan, Joseph Jay Williams, Derek Lomas, Ben Motz, Debshila Basu Mallick, Klinton Bicknell, Danielle McNamara, Rene F. Kizilcec, Jeremy Roschelle, Richard Baraniuk, and Ryan Baker. 2022. Third Annual Workshop on A/B Testing and Platform-Enabled Learning Research. In *Proceedings of the Ninth ACM Conference on Learning @ Scale* (New York City, NY, USA) *(L@S '22)*. Association for Computing Machinery, New York, NY, USA, 252–254. https://doi.org/10.1145/3491140.3528288

[30] Steven Ritter, Neil Heffernan, Joseph Jay Williams, Burr Settles, Phillip Grimaldi, and Derek Lomas. 2020. Workshop Proposal: Educational A/B Testing at Scale. In *Proceedings of the Seventh ACM Conference on Learning @ Scale* (Virtual Event, USA) *(L@S '20)*. Association for Computing Machinery, New York, NY, USA, 219–220. https://doi.org/10.1145/3386527.3405933

[31] Steven Ritter, April Murphy, Stephen E Fancsali, Vivek Fitkariwala, Nirmal Patel, and J Derek Lomas. 2020. UpGrade: An open source tool to support A/B testing in educational software. In *Proceedings of the First Workshop on Educational A/B Testing at Scale (at Learning@ Scale 2020)*.

[32] Michael A Rosen, Eduardo Salas, Davin Pavlas, Randy Jensen, Dan Fu, and Donald Lampton. 2010. Demonstration-Based Training: A Review of Instructional Features. *Human factors* 52, 5 (2010), 596–609.

[33] Sree Sankaranarayanan, Chellie Harrison, Sarita Kumari, Kim Larson, Robert Lemiesz, Nicolas Mesa, Ryan Mitts, Esha Verma, and Dawn Zimmaro. 2023. Building an infrastructure for A/B experiments at scale: The challenges, opportunities, and lessons for the learning analytics community. In *LAK 2023*. https://www.amazon.science/publications/building-an-infrastructure-for-a-b-experiments-at-scale-the-challenges-opportunities-and-lessons-for-the-learning-analytics-community

[34] Charles Severance, Ted Hanss, and Josepth Hardin. 2010. Ims learning tools interoperability: Enabling a mash-up approach to teaching and learning tools. *Technology, Instruction, Cognition and Learning* 7, 3-4 (2010), 245–262.

[35] Valerie J Shute. 2008. Focus on formative feedback. *Review of educational research* 78, 1 (2008), 153–189.

[36] Hans Van der Meij and Jan Van Der Meij. 2014. A comparison of paper-based and video tutorials for software learning. *Computers & education* 78 (2014), 150–159.

[37] Hans Van Der Meij and Jan Van Der Meij. 2016. Demonstration-based training (DBT) in the design of a video tutorial for software training. 44, 6 (2016), 527–542. https://doi.org/10.1007/s11251-016-9394-9