Tomorrow’s EdTech Today: Establishing a Learning Platform as a Collaborative Research Tool for Sound Science

Korinn S. Ostrow
Learning Sciences & Technologies
Worcester Polytechnic Institute
Worcester, MA 01609
ksostrow@wpi.edu

Korinn Ostrow is a Ph.D. candidate in Learning Sciences & Technologies at Worcester Polytechnic Institute. Her concentrations include applied educational statistics and cognitive psychology, with research interests in learning interventions, experimental methods at scale, learning analytics within adaptive technologies, and enhancing student motivation and engagement. She expects to graduate in 2018. Recent publications include “The Future of Adaptive Learning: Does the Crowd Hold the Key?” in the International Journal of Artificial Intelligence in Education, and “The Assessment of Learning Infrastructure (ALI): The Theory, Practice, and Scalability of Automated Assessment” in the Proceedings of the 6th International Conference on Learning Analytics and Knowledge.

Neil T. Heffernan
Computer Science, Learning Sciences & Technologies
Worcester Polytechnic Institute
Worcester, MA 01609
nth@wpi.edu

Neil Heffernan is a Professor of Computer Science and the Director of the Learning Sciences and Technologies Graduate Program at Worcester Polytechnic Institute. He is best known for creating ASSISTments. He has used the platform to conduct and publish two-dozen randomized controlled experiments, and now strives to expand the platform as a tool for others to do the same. In addition, he has published three dozen papers in predictive analysis, using large educational datasets to predict student performance on standardized state tests, affective states like boredom and frustration, and even college admission years later. He cares deeply about helping others learn about personalized learning in a methodically rigorous way.

Joseph Jay Williams
Vice Provost for Advances in Learning Research Group
Harvard University
Cambridge, MA 02138
joseph_jay_williams@harvard.edu

Joseph Jay Williams is a Research Fellow in the Vice Provost for Advances in Learning Research Group at Harvard, where he conducts human-computer interaction and statistical machine learning research on digital education. This bridges his postdoctoral work at Stanford's Graduate School of Education conducting randomized experiments in MOOCs and Khan Academy to motivate learners through psychological interventions. He received his Ph.D. from U.C. Berkeley, where he focused on computational cognitive science and developed the Subsumptive Constraints account of why generating self-explanations enhances learning. He also developed the MOOClet Framework for designing intelligent digital lessons that personalize learning through randomized comparisons of crowdsourced content. Recent publications include "Generating Explanations at Scale with Learnersourcing and Machine Learning" at ACM Learning at Scale, and "Revising Learner Misconceptions without Feedback: Prompting for Reflection on Anomalous Facts" at Computer-Human Interaction.

Structured Abstract

Background/Context

Large-scale randomized controlled experiments conducted in authentic learning environments are commonly high stakes, carrying extensive costs and requiring lengthy commitments for “all or nothing” results amidst many potential obstacles. Educational technologies harbor an untapped potential to provide researchers with access to extensive and diverse subject pools of students interacting with educational materials in authentic ways. These systems log extensive data on student performance that can be used to identify and leverage best practices in education and guide systemic policy change. Tomorrow’s educational technologies should be built upon rigorous standards set forth by the research revolution budding today.
Purpose/Objective/Research Question/Focus of Study
The present work serves as a call to the community to infuse popular learning platforms with the capacity to support collaborative research at scale.

Research Design
This article defines how educational technologies can be leveraged for use in collaborative research environments by highlighting the research revolution of ASSISTments (www.ASSISTments.org), a popular online learning platform with a focus on mathematics education. A framework described as the ‘cycle of perpetual evolution’ is presented, and research exemplifying progression through this framework is discussed in support of the many benefits that stem from infusing EdTech with collaborative research. Through a recent NSF grant (SI2-SSE&SSI: 1440753), researchers from around the world can leverage ASSISTments’ content and user population by designing and implementing randomized controlled experiments within the ASSISTments TestBed (www.ASSISTmentsTestBed.org). Findings from these studies help to define best practices within technology driven learning, while simultaneously allowing for augmentation of the system’s content, delivery, and infrastructure.

Conclusions/Recommendations
Supplementing educational technologies with environments for sound, collaborative science can result in a broad range of benefits for students, researchers, platforms, and educational practice and policy. This article outlines the successful uptake of research efforts by ASSISTments in hopes of advocating a research revolution for other educational technologies.

Executive Summary
Educational psychologists, researchers, and practitioners have grown accustomed to the complex and time consuming nature of studying effective classroom practices. The gold standard in determining causality, the randomized controlled experiment (RCE), can be difficult to conduct in real-world classrooms. While a handful of traditional classroom RCEs have led to significant implications for educational practice and policy, these high stakes explorations often include thousands of students and span multiple years but still fall short of identifying learning interventions that reliably enhance student achievement. The present work investigates the use of educational technologies to simplify the process of conducting RCEs within authentic learning environments, making research at scale more feasible and more accessible to researchers. Infusing popular learning platforms with the capacity to support collaborative research environments has the potential to lower the stakes by drastically reducing costs, promoting validated universal measures of achievement, and assisting researchers through the process of designing, implementing, and analyzing RCEs conducted at scale within real-world classrooms. Supplementing educational technologies with environments for sound, collaborative science can result in a broad range of benefits for students, researchers, platforms, and educational practice and policy.

Educational technologies already have great promise for extending the accessibility of educational materials and improving learning outcomes across diverse populations. At scale, data collected from these platforms can be leveraged in dynamic ways that may reveal revolutionary insights about learning. Entire fields of research are growing alongside educational technologies in hopes of better understanding how these tools and their data can be used to improve education (e.g., Learning Analytics, Educational Data Mining). However, despite significant growth in researcher interest, few technologies that are currently available to teachers and students allow real-time hypothesis testing. Examining the causal effects of specific learning interventions through “Big Experimentation” would allow researchers to begin answering three questions that truly drive personalized education: “What works best? For whom? When?” By determining the interventions that work best for particular students and the optimal time to deliver those interventions, controlled experimentation conducted within educational technologies could revolutionize the future of education.

The present work highlights the research revolution of ASSISTments (www.ASSISTments.org), a popular online learning platform that was designed with the flexibility to house RCEs and has supported numerous publications since its inception in 2002. ASSISTments is used by hundreds of teachers to assign classwork and homework to over 50,000 students around the world, with over 10 million problems solved in the
2013-2014 school year. The platform, offered as a free service of Worcester Polytechnic Institute (WPI), is an increasingly powerful tool that provides students with assistance while offering teachers assessment.

A recent NSF grant (SI2-SSE&SSI: 1440753) has helped to launch a formal infrastructure that allows external researchers to leverage ASSISTments as a shared scientific tool. This supplementary infrastructure, the ASSISTments TestBed (www.ASSISTments TestBed.org), allows researchers to design minimally invasive RCEs within easily accessible and highly used educational content, and receive organized reports detailing student performance that streamline the analysis of learning interventions. There are currently over 130 RCEs running in the ASSISTments TestBed that are helping researchers to identify evidence-based instructional improvements while inspiring a ‘cycle of perpetual evolution’ for the platform and its users. In this cycle, a simple hypothesis acts as the seed for an expanse of research that germinates via related ideas, eventually pushing the limits of the system until infrastructure improvements must be made to accommodate further questions. The cycle begins when researchers form novel hypotheses that compare manipulations within the platform to best (known) practices (either comparable traditional classroom practices or previous versions of the platform’s material). Early results inspire collaborative idea expansion through replications and extensions of studies that serve to enhance system content and content delivery, while improving student learning and advancing the state of knowledge in the field through peer reviewed publication. New hypotheses form and grow as results are observed, naturally evolving until pushing the boundaries of the platform’s infrastructure. In response, scientifically validated infrastructure improvements can be tailored to research demand, allowing researchers to start the cycle anew with novel hypotheses.

With a focus on disseminating the ASSISTments TestBed and enhancing its validity as a collaborative tool for sound research, this work also briefly defines how the cycle of perpetual evolution will bring about a number of significant infrastructure changes for ASSISTments in the near future. The ASSISTments team is currently focusing on expanding the platform’s capacity to support learnersourced feedback by allowing students to ‘show their work’ and provide explanations for their peers through a tool called PeerASSIST. The successful implementation of learnersourcing will give way to another goal for the future of ASSISTments: establishing an automated process for selecting optimal feedback using contextual k-armed bandits. This approach will quickly filter ineffective content to minimize potential detriment to students, and will provide opportunities for the dynamic versioning of materials and for truly adaptive learning personalization (i.e., algorithmically establishing “What works best? For whom? When?”). The ASSISTments team expects that these goals will strengthen the platform and inspire new avenues for scientific inquiry.

Systemic change does not stem from a small number of large-scale RCEs funded by government grants, but instead from a revolution in thought surrounding the value of technology based learning applications. Infusing pre-existing learning technologies with the capacity to support RCEs is the first step in kick starting this revolution. Research infused platforms have the potential to drive inquiry for a diverse community of researchers through the low-cost, rapid iteration of valid, generalizable, and noninvasive investigations within authentic learning environments. Systems like ASSISTments can provide researchers with access to an extensive and diverse subject pool, an automated fine-grained logging of educational data, validated measures of student learning and affect, and automated data reporting and analysis to tackle the high stakes nature of typical education research. With similar research-focused platforms in the field, it would also be possible for researchers to compare learning interventions across platforms to better measure the reliability and generalizability of results. Tomorrow’s educational technology demands a revolution in today’s approaches to research at scale: pave the way for sound collaborative science and the rest will follow.

Introduction

Educational psychologists, researchers, and practitioners have grown accustomed to the complex and time consuming nature of studying effective classroom practices. When studying learning interventions, seasoned experts turn to the gold standard in determining causality: the randomized controlled experiment (RCE). Yet despite a recent call encouraging the use of RCEs within authentic learning environments (IES, 2013), and despite the nearly infinite array of complexities to be examined within the context of instruction
(Koedinger, Booth, & Klahr, 2013), RCEs can be difficult to conduct in real-world classrooms (National Research Council, 2002). Common complications include IRB restrictions, lengthy and invasive pre- and post-tests, curriculum restrictions for the design of strict controls, and large sample populations required to detect significantly reliable results. Further, experimental designs must be carefully vetted prior to implementation in an attempt to account for as much variance as possible. Thorough organization is also necessary when recording and maintaining anonymized student data. With so many moving parts, traditional classroom RCEs leave numerous windows for error and bias. Even when reporting findings, publication bias and the cherry picking of results can lead to the inability for replication, contributing to a growing crisis of faith in RCEs spanning numerous scientific fields (Ioannidis, 2005; Achenbach, 2015; Open Science Collaboration, 2015). Additionally, while a handful of traditional classroom RCEs have led to significant implications for educational practice and policy, most lack the statistical power necessary to observe reliable improvements in student achievement because they are restricted by class- or school-level randomization (i.e., all students within a particular class or school fall within the same experimental condition, resulting in drastically reduced sample sizes). High stakes explorations at scale (i.e., stressful ‘make or break’ longitudinal studies costing millions of dollars) often include thousands of students and span multiple years but still fall short of identifying learning interventions that reliably enhance student achievement.

While it is crucial that high standards exist for educational research, the present work investigates the use of educational technologies to simplify the process of conducting RCEs within authentic learning environments, making research at scale more feasible and more accessible to researchers. Infusing popular learning platforms with the capacity to support collaborative research environments has the potential to lower the stakes by drastically reducing costs, promoting validated universal measures of achievement, and assisting researchers through the process of designing, implementing, and analyzing RCEs conducted at scale within real-world classrooms. Supplementing educational technologies with environments for sound, collaborative science can result in a broad range of benefits for students, researchers, platforms, and educational practice and policy.

The Growth of Educational Technologies

Educational technologies offer the novel opportunity to drive best practices in K-12 education by testing what works in authentic learning environments while simultaneously simplifying the process of educational research. Technology is gaining acceptance in the modern classroom, with intelligent tutoring systems, computer-aided testing platforms, and adaptive learning applications offering new and unique approaches to learning, beckoning a transition from ‘teaching’ based practices to ‘learning’ based practices (Bush & Mott, 2009), and producing exponential growth in the availability of educational data. Educational technologies commonly include immediate feedback, adaptive assistance, elements that enhance student motivation and engagement, and assessment tools for teachers and administrators that help to drive data-driven classroom practices. As such, the National Education Technology Plan predicted that these platforms would play a key role in personalizing educational interventions (U.S. Dept. of Ed., 2010). However, less focus has been devoted to one of the primary forces driving successful personalization: the use of adaptive learning technologies to conduct educational research.

These platforms and applications already have great promise for extending the accessibility of educational materials and improving learning outcomes across diverse populations. At scale, the data collected from these technologies can be leveraged in dynamic ways that may reveal revolutionary insights about learning. Entire fields of research are growing alongside educational technologies in hopes of better understanding how these tools and their data can be used to improve education (e.g., Learning Analytics, Educational Data Mining). However, despite significant growth in researcher interest, few platforms currently available to teachers and students allow for real-time hypothesis testing. In lieu of in vivo experimentation, researchers often turn to logged data to model student performance, make predictions regarding learning, and determine the effectiveness of system features (Koedinger, Baker, et al., 2010). “Big Data” in education has grown synonymous with solutions that enhance educational practices, platforms, and theories. Still, a critical link is missing: causality. Examining the causal effects of specific learning interventions through “Big Experimentation” would allow researchers to begin answering three questions to truly drive personalized education: “What works best? For whom? When?” By determining the
interventions that work best for particular students and the optimal time to deliver those interventions, controlled experimentation conducted within these platforms has the potential to revolutionize the future of education.

The ASSISTments Platform

Despite expanse in the availability of adaptive learning technologies in recent years, popular platforms have been very slow to mobilize, support, and leverage randomized controlled experimentation (Williams, Ostrow, et al., 2015; Williams, Maldonado, et al., 2015). ASSISTments is an online learning platform that was designed with the flexibility to house RCEs and has supported the publication of more than two dozen peer-reviewed articles on learning since its inception in 2002 (Heffernan & Heffernan, 2014). The platform, offered as a free service of Worcester Polytechnic Institute (WPI), is an increasingly powerful tool that provides students with assistance while offering teachers assessment. Over $14 million in grant funding from the IES and the NSF has supported twelve years of co-development with teachers and researchers to establish a unique tool for educational research at scale. Historically, the primary investigators of these studies have had close connections to WPI (e.g., graduate students or other researchers working closely with the ASSISTments Team). However, a recent NSF grant (SI2-SSE&SSI: 1440753) has helped to launch a formal infrastructure that allows external researchers to use ASSISTments as a shared scientific tool. This supplementary infrastructure is called the ASSISTments TestBed (www.ASSISTmentsTestBed.org). While other systems have the potential to provide many of the same classroom benefits as ASSISTments, none promote an infrastructure allowing educational researchers to design and implement content-based experimentation, and to do so with ease.

Doubling its user population each year for almost a decade, ASSISTments is used by hundreds of teachers and over 50,000 students around the world, with over 10 million problems solved in the 2013-2014 school year. Although most content pertains to middle school mathematics, teachers from alternative domains like history, biology, and statistics have also built material to harness the powers of the platform in their own classrooms. Content is built at the problem level, as shown in Figure 1. The problem builder allows teachers and researchers to design questions and tutorial strategies using a simple interface that allows for the inclusion of text, graphics, and hypermedia elements. The builder is unique in that it allows for efficient content design without extensive knowledge of computer programming. Questions can then be combined to form problem sets for assignment to students. Teachers commonly use ASSISTments to assign classwork and homework with immediate feedback and rich tutoring, but can also turn off feedback elements to assign content as a test or quiz. Use of ASSISTments has been shown to reliably improve students' learning in comparison to traditional paper and pencil approaches (Mendicino, Razzaq, Heffernan, 2009; Koedinger, McLaughlin & Heffernan, 2010; Singh et al., 2011; Kelly, Heffernan, Heffernan, et al., 2013; Miller, et al., 2013; Soffer, et al., 2014). Most recently, SRI International reported early results of an efficacy trial of ASSISTments, showing that the platform caused large, reliable, learning gains on standardized assessments (Rochelle, et al., 2015).
Figure 1. An example of a problem viewed within the builder. Notice the interface allows creation of the problem itself, answers (both correct and incorrect), and tutoring strategies. The navigation menu in the top right corner allows the user to navigate from editing a main problem to editing feedback.

In addition to building content, teachers and researchers are able to access an extensive library of prebuilt content and textbook material. Full problem content is available for more than 20 of the top 7th grade mathematics texts in the United States, delivered without infringing on copyright. Teachers can select from prebuilt problem sets, or use and alter copies of content to develop their own problem sets. There are two primary types of problem sets within ASSISTments. A linear problem set has a predetermined number of problems and the assignment is considered complete when the student has finished all problems, whether or not the answers are accurate. Alternatively, in a Skill Builder problem set, students must solve problems selected at random from a skill pool until reaching a predetermined threshold of mastery (i.e., answering three consecutive questions accurately on first attempts). Although the system default is three problems, mastery can be redefined to include any number of consecutive accurate problems. In both types of problem sets, assistance can vary to include correctness feedback, tutoring specific to particular problems, or worked examples depicting solutions to isomorphic problems. Tutoring strategies include hint messages, scaffolding problems (used to break a problem down into steps), and mistake messages (feedback tailored to common wrong answers). Hints, scaffolds, and mistake messages are compared in Figure 2. If researchers do not wish to design their own content, over 300 ‘Certified’ Skill Builders (tailored by the ASSISTments Team to the Common Core State Standards for Mathematics (NGACBP & CCSSO, 2010)), can be manipulated to incorporate experimental modifications.
Figure 2. A comparison of hints, a scaffold problem, and a mistake message in response to the same problem content. Three hints are shown on the left, as requested by the student. In the middle, the student provided an incorrect response and was automatically given a scaffold with a worked example on how to solve a similar problem. On the right, a mistake message is provided in response to a specific wrong answer, with detailed tutoring on strategy revision.

ASSISTments also offers optional features like the Automatic Reassessment and Relearning System (ARRS), which helps to reassess student retention following Skill Builder mastery (Xiong & Beck, 2015), and PLACEments, a prerequisite skill training system that allows teachers to create skill tests that pinpoint and help to alleviate knowledge gaps (Whorton, 2013). When a teacher elects to use ARRS, after completing a Skill Builder, students are given a series of single question reassessment tests, scheduled 7, 14, 28, and finally 56 days after the initial learning experience to estimate skill retention. If students fail to answer the reassessment question accurately, they are provided support to relearn the material through a secondary Skill Builder. Research has shown that ARRS significantly enhances longitudinal skill understanding and student assessment (Soffer, et al., 2014; Wang & Heffernan, 2014). Like ARRS, PLACEments is also connected to Skill Builder content. PLACEments acts as a computer adaptive test that taps into a hierarchy of prerequisite skills to personalize the remediations a student should receive based on performance in an initial skill test. Research has shown that PLACEments is a useful tool for isolating learning gaps that can also help to strengthen curriculum through a stronger understanding of prerequisite skill relationships (Adjei & Heffernan, 2015).

As an assessment tool, ASSISTments offers teachers a myriad of student and class reports that allow an expanse in classroom practices through actionable data. An example of an Item Report, the most commonly used report, is shown in Figure 3. This report has a column for each problem and a row for each student, as well as various summaries of student and problem performance. The report can be made anonymous (as shown in Figure 3) for teachers to use in the classroom to facilitate discussion. This report also allows teachers to pinpoint areas of struggle through common wrong answers (errors that were made by at least 10% of students in the class). In Figure 3, only 27% of students answered the first problem accurately, with 56% of students sharing the common wrong answer of 1/9^10. This offers an opportunity for discussion that may be lost on students grading their own homework using traditional classroom
methods. Teachers can also work with students to design mistake messages (like that shown in Figure 2) for future students who attempt the problem and share the same misconception.

Figure 3. An excerpt from an anonymized Item Report. Students are listed in the first column, followed by average performance and then specific performance on each question within the problem set. Teachers can see if the student answered correctly or incorrectly, the response given, whether a tutoring strategy was used, and common wrong answers as measured across the entire class. Common wrong answers are actionable; teachers and students can work together to provide a mistake message for future students.

Through NSF funding (SI2-SSE&SSI: 1440753), reports for researchers have grown far more complex than teacher reports, providing numerous formats of raw performance data with rich student, class, and school level covariates, as well as a number of automated analyses. Through the ASSISTments TestBed, and specifically through the Assessment of Learning Infrastructure (ALI), researchers are provided weekly automated reports detailing anonymized study participation (Ostrow, et al., In Press). These reports, as shown in Figure 4, provide basic analyses including bias assessment (examining attrition across experimental conditions) and simple hypothesis testing on posttest performance. Researchers are also provided a student covariate file, detailing student information collected prior to study participation (i.e., prior performance average), and four formats of raw data logged by the ASSISTments tutor as students work through the assignment. ALI’s reporting and researcher communications make the TestBed easier for researchers to use, streamlining research at scale.
The Assessment of Learning Infrastructure (ALI)

Completion Rates
Students that have started your study: 329
Students that have completed your study: 251

Bias Assessment
Before analyzing learning outcomes, we suggest first assessing potential bias introduced by your experimental conditions (i.e., examine differential attrition). The table below reports the number of students that have completed your study, split out by experimental condition.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Started (n)</th>
<th>Completed (n)</th>
<th>Completed (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group A – Experiment 1</td>
<td>109</td>
<td>80</td>
<td>73.39</td>
</tr>
<tr>
<td>Group B – Experiment 2</td>
<td>87</td>
<td>60</td>
<td>68.97</td>
</tr>
<tr>
<td>Group C – Control</td>
<td>99</td>
<td>89</td>
<td>89.90</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>295</strong></td>
<td><strong>229</strong></td>
<td><strong>77.63</strong></td>
</tr>
</tbody>
</table>

**NOTE:** A significant difference was found between observed and expected completion rates across conditions, $\chi^2 (2, N = 295) = 13.467, p < .01$. This means that a selection effect may have occurred. Hypothesis testing with regard to posttest scores has not been conducted out of an abundance of caution.

Mean and Standard Deviation of Posttest Score by Condition
To examine learning outcomes at posttest, an analysis of means was conducted across conditions. The table below reports mean posttest score and standard deviation for each condition. This information was sourced from our automated posttest sub-report.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Completed (n)</th>
<th>Posttest Score*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group A – Experiment 1</td>
<td>80</td>
<td>34.40 (4.34)</td>
</tr>
<tr>
<td>Group B – Experiment 2</td>
<td>60</td>
<td>32.95 (3.89)</td>
</tr>
<tr>
<td>Group C – Control</td>
<td>89</td>
<td>44.11 (3.72)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>229</strong></td>
<td><strong>37.15 (3.98)</strong></td>
</tr>
</tbody>
</table>

* Presented as Mean (SD).

Raw Data Files
Raw data files contain the logged information for each student that has participated in your study. We provide this data in a variety of formats, as explained below, to assist in your analytic efforts. We use Google Docs to share these files with you. If you would like to process these files manually, we recommend downloading the CSV file of your choice and saving the file as an Excel spreadsheet or workbook to retain formatting and formulas. If you will be passing the file directly to a statistical package, downloading the CSV to a convenient location should suffice.

For a field glossary and tutorials on how to read each type of file, visit our Data Glossary.

Historical Data
**Covariate File** - A collection of useful covariates for the students participating in your study. This file includes student level variables (i.e., gender), class level variables, (i.e., homework completion rates), and school level variables (i.e., urbanicity). Click here for a tutorial on how to link this file to your experimental data.

Experimental Data
1. **Action Level** - One row per action per student; the finest granularity. Students participating in your study have performed 13,655 actions (e.g., beginning problems, attempting to answer problems, asking for tutoring, and eventually completing problems).
2. **Problem Level** - One row per problem per student. Students participating in your study have completed 2,280 problems. The flow through a single problem incorporates many actions, resulting in a coarser data file (fewer rows).
3. **Student Level** - One row per student; the coarsest granularity. Columns are laid out in opportunity order to depict the student’s progression through the problem set. Problem level information is expanded to one column per problem per field (column heavy).
4. **Student Level + Problem Level** - One row per field per student. Columns are laid out in opportunity order to depict the student’s progression through the problem set. An alternative view of student level information (row heavy).

Figure 4. The Assessment of Learning Infrastructure (ALI) provides researchers with logged data from students participating in RCEs within the ASSISTments TestBed (Ostrow, et al., In Press). This automated report is generated weekly and/or at the request of the researcher, and presents analyses and raw data. Analyses include a Chi-squared test comparing the observed and expected sample distributions, simple hypothesis testing, and an analysis of means on posttest performance.
Technology Supported Randomized Controlled Experimentation

Through the ASSISTments TestBed, researchers are able to design minimally invasive RCEs within easily accessible and highly used educational content delivered by ASSISTments, while receiving organized reports detailing student performance to streamline the analysis of learning interventions. This type of open research environment is rare within learning technologies. The common use for RCEs or A/B testing within popular technologies is to optimize user experience or prolong user interaction. For instance, Google experiments with advertisement location to maximize ad traffic without diminishing the user experience. Similarly, gaming application creators like Zynga conduct A/B testing to optimize their games in a way that will retain users while promoting ad space. Although these approaches are consistent in marketing, few large-scale education platforms show an outward interest in examining learning interactions and optimizing learning gains. Massive Open Online Course (MOOC) platforms and large scale learning tools like Coursera, EdX, Udacity, openHPI, and Google’s “Course Builder” focus on delivering content, while spending little time or money thoroughly examining the effects of what they deliver. This argument is not intended to suggest a complete deficit of sound research, but instead to point out that few researchers have access to course data from these platforms to improve user interfaces or curriculum delivery. Even commercialized educational technologies lack open and easily accessible avenues for empirical research. For instance, the popular Khan Academy provides resources and support for select researchers to work through a process requiring substantial time and effort to understand the dynamics of the system. Creating and running an experiment within Khan Academy requires knowledge of the platform’s open-source code, the coding skills necessary to make modifications to implement experimentation, and progression through a standard code review process working alongside Khan Academy developers. Obtaining data files following an experiment is also heavily reliant on system programmers. To our knowledge, none of the A/B experiments that researchers have patiently conducted on Khan Academy have been formally published (e.g., see Williams, Paunesku, Haley & Sohl-Dickstein, 2013; Williams, 2014). Instead, work with less regard for improving specific interventions has evaluated the platform’s efficacy in schools (Murphy, et al., 2014) and prediction models for large-scale but secondary data (Piech, et al., 2015). Such major platforms should be reframed with a focus on open educational research at scale, or should at least support the open collection of anonymized data through APIs to inform EdTech policy.

The application of stringent research methodologies to improve learning technologies and educational outcomes is severely lagging. This deficit is what makes the ASSISTments TestBed so unique. The TestBed guides researchers through the process of running practical RCEs by leveraging ASSISTments’ content and user population. There are currently over 130 RCEs running within the ASSISTments TestBed. These studies are directed at solving practical problems within education and understanding best practices within technology driven learning. While these studies help researchers to identify evidence-based instructional improvements, findings also lead to the generation of new hypotheses that expand investigation or reroute postulated theories. Results from a single study may generate four new hypotheses, with the potential for exponential expansion as a line of research evolves. The results of these studies can also benefit ASSISTments: findings regarding best practices continuously improve the system’s content and delivery, while pinpointing areas for broad change through infrastructure improvements. Thus, a collaborative and open research infrastructure supports perpetual evolution on a small scale within the system and on a large scale across research communities.

Developing Collaboratives Around Shared Scientific Tools

To get the most out of educational technologies, learning platforms must be revolutionized into shared scientific instruments. Through the ASSISTments TestBed, ASSISTments is attempting to initiate this movement by stepping forward as the ‘Hubble Telescope’ of learning science. Unlike a static piece of equipment, the platform can be used to run multiple experiments simultaneously and researchers are able to improve the instrument for others through their experiences. Through this collaborative approach, as shown in Figure 5, researchers bring many ideas and hypotheses to the TestBed. Some of the studies designed around these hypotheses result in reliably positive effects while others are extended to form
stronger research questions. Through this process, researchers alter and enhance content and feedback within ASSISTments. Students and teachers benefit from stronger content while researchers expand their fields through refereed publications.

Figure 5. Research within the ASSISTments TestBed leads to knowledge of best practices, enhancements to student learning outcomes, and peer reviewed publications. Multiple iterations of hypotheses may arise, enhancing system content and strengthening content delivery as work progresses.

Realization of the platform’s value as a shared scientific tool has encouraged research at scale from universities including Boston College, Freiburg University, Harvard University, Indiana University, Northwestern University, Southern Methodist University, Texas A&M, University of Colorado - Colorado Springs, University of California - Berkeley, University of Maine, University of Wisconsin, and Vanderbilt. Since its inception, interest in the TestBed has continued to expand through a kickoff webinar, an AERA seminar, and well documented support for researchers made possible by NSF funding (SI2-SSE&SSI: 1440753).

By articulating specific challenges for improving K-12 mathematics education to a broad and multidisciplinary community of psychology, education, and computer science researchers, this funding allows researchers to collaboratively (and perhaps competitively) propose and conduct RCEs, at an unprecedentedly precise level and large scale. The following list highlights the broad spectrum of work that researchers have shown interest in examining further within the TestBed:

**Types of Feedback**
- Immediate vs. delayed feedback (Fyfe, Rittle-Johnson, & DeCaro, 2012).
- Comparing the type of hints provided adaptively to learners (Stamper, et al., 2013).
- Comparing levels of feedback from guided to open (Sweller, Kirschner, & Clark, 2007).
- Comparing ‘what you see is what you get’ vs. interaction (Keehner, et al., 2008).
- Prompting for comparison of analogous problems and worked examples (Jee, et al., 2013).

**Sequencing and Spacing**
- Changing schedules and procedures for practice sessions & quizzes (Roediger & Karpicke, 2006).
- Testing the effectiveness of pre-testing prior to instruction (Richland, Kornell, & Kao, 2009).
- Spacing skill content (Pashler, et al., 2007).
- Examining testing effects (Butler & Roediger, 2007).

**Self-Regulated Learning & Metacognition**
- Testing interventions to increase motivation and teach strategies (Ehrlinger & Shain, 2014).
- Examining how task framing changes what students learn (Belenky & Nokes-Malach, 2013).
- Examining metacognitive scaffolding provided in problem solving (Roll, et al., 2012).
• Testing the value of free recall (Arnold & McDermott, 2013).

Social Context and Interaction
• Adapting instructional materials to students’ personal & peer interests (Walkington, 2013).
• Embedding software & dynamics for peer assistance (Walker, Rummel, & Koedinger, 2011).
• Examining how confidence impacts performance in early algebra (Mazzocco, et al., 2013).

Assessment
• Examining computational models used to diagnose learner state (Rafferty & Griffiths, 2014).
• Examining computational methods for assessing affective states (Ocumpagh, et al., 2014).
• Examining forgetting (Storm, et al., 2006).

Motivation
• Embedding motivational videos from teachers (Kelly, Heffernan, D’Mello, et al, 2013).
• Incorporating messages to foster growth mindset (Williams, 2013).
• Examining the effects of goal setting (Bernacki, Byrnes, & Cromley, 2012).
• Examining the effects of student choice (Chernyak & Kushnir, 2013).
• Inserting quizzes and tests to maintain and guide student focus (Szpunar, et al., 2013).

Mathematics Education
• Comparing representational formats in supporting mathematics learning (Rau, et al., 2012).
• Investigating effective presentations of worked examples in mathematics (Booth, et al., 2013).
• Examining strategies for learning fractions (Cordes, et al., 2007).
• Testing images of manipulatives vs. virtual manipulatives (Mendiburo, et al., 2012).

By building these types of collaborative scientific tools, the cost of funding educational research could be drastically reduced. For instance, the Institute of Education Sciences (IES) currently funds “Efficacy Trials” for promising interventions that cost an average of $3M and can involve more than 50 schools. Larger and more stringent “Effectiveness Trials” carry a median cost of $6M. In the math and science domains, the IES has funded 22 Efficacy Trials and five Effectiveness Trials. Despite the high cost of funding this work, reliable positive implications for educational practice are rarely observed. Using adaptive technologies geared toward research, large-scale trials could be expedited at a fraction of the cost. The IES funding pipeline (IES, 2015) and the ASSISTments TestBed equivalent are depicted in Figure 6. Studies that were once restricted by the availability of funding could be considered through learning technologies.

![Pipeline for Education Research Defined by the Institute of Education Sciences](image)

**Figure 6.** The pipeline for education research as defined by the IES compared to a similar timeline for research within the ASSISTments TestBed. Educational technologies can be used as shared scientific tools to drastically reduce costs and enhance the efficiency with which educational research is conducted.
Much of the efficacy attained through use of the TestBed is due to student-level randomization (rather than traditional class- or school-level randomization), allowing experiments to be conducted within classrooms rather than across classrooms. This accrues drastically larger samples, increasing the power of analyses in order to better detect the reliable effects of interventions. The unique ability for student-level randomization, coupled with the scalability inherent to manipulating pre-built content of interest to a large user base, allows in vivo educational research to gain the minimally invasive A/B flavor often used in marketing. Studies within the TestBed also align with typical educational practice (i.e., students are never intentionally disadvantaged by a study design). This approach allows students to access and complete assignments, often without awareness that they are participating in research. Teachers are made aware of experimentation through a conventional assignment naming procedure that ‘tags’ experiments with ‘Ex.’ As data dissemination is carefully preprocessed to protect students’ identities, and students receive assignments that are within the definition of ‘normal instructional practice,’ this passive approach to research is IRB approved.

While low cost procedures may not hold for all educational investigations (i.e., the design of full learning programs or platforms that require significant funding), there are many benefits to cost-effective, efficient, and rigorous experimentation that can be conducted using educational technologies. Many unique features make ASSISTments capable of serving researchers as a shared scientific tool. However, ASSISTments is not the only platform with the power to drive a collaborative like the TestBed. The majority of learning applications have the capacity for data collection, and many could be restructured to offer the flexibility required for experimental content manipulation. Other platforms may also be capable of establishing an API to deliver preprocessed data, anonymized for student protection, to researchers conducting RCEs or even wishing to mine data. With similar research-based platforms in the field, it would also be possible for researchers to compare learning interventions across platforms to better measure the reliability and generalizability of results. Collaborative research goals that crosscut platforms may finally usher in the tipping point of educational technologies (Bush & Mott, 2009; Gladwell, 2002) as researchers grow to understand “What works best? For Whom? When?”

Collaborative Research at Scale Offers Perpetual Benefits

The power of the ASSISTments TestBed as a collaborative research tool did not come about overnight. As a learning platform, ASSISTments has pivoted numerous times in the past decade (Heffernan & Heffernan, 2014). The steady improvements from which the TestBed were largely driven by the results of pilot studies within the system. This growth and adaptation exemplifies perpetual evolution. Essentially, a simple hypothesis acts as the seed for an expanse of research that germinates through related ideas, eventually pushing the limits of the system until infrastructure improvements must be made to accommodate further questions - a cycle depicted in Figure 7. As the cycle begins, researchers form novel hypotheses that compare manipulations within the platform to best (known) practices (either comparable traditional classroom practices or previous versions of the platform’s material). Early results inspire collaborative idea expansion through replications and extensions of studies that serve to enhance system content and content delivery, while improving student learning and advancing the state of knowledge in the field through peer reviewed publication. New hypotheses form and grow as results are observed, naturally evolving until pushing the boundaries of the platform’s infrastructure. In response, scientifically validated infrastructure improvements can be tailored to research demand, forming the final stage of this cycle. New system features, a mark of evolution, allow researchers to start the cycle anew with novel hypotheses.
Figure 7. The cycle of perpetual evolution that stems from use of an educational platform as a collaborative research tool. An initial hypothesis comparing new methods to best known practices grows into a series of ideas that improve system content while benefiting students and advancing knowledge in the field. These ideas continue to grow until limited by the platform’s capabilities. Infrastructure improvements validated by previous findings and inspired by research demand can then be made to return the cycle to a fresh starting point, where new hypotheses can be formed.

Ever expanding progress is a core concept for effectively marketing commercial products, but is far less common in education. Education is a difficult rock to move, with teachers and administrators holding tight to traditional methods, and pushing back against the changes brought about by modern technologies (Bush & Mott, 2009). It is hardly surprising that most educational technologies lack collaborative research infrastructures. Administrators have not been focused on examining the effectiveness of new instructional strategies made possible by these platforms because most platforms have instead been tailored to simplify traditional teaching methods (Bush & Mott, 2009). As educators continue to grow more open to the possibilities of learning technologies, the value of collaborative research at scale will escalate. By establishing research environments like the TestBed, creators and users of educational technologies will learn of the unprecedented benefits made possible by the cycle of perpetual evolution. The following sections step through this cycle, defining exemplary research at each stage, as conducted within ASSISTments and the ASSISTments TestBed.

The Seed: Comparing Research Generated Content To Best (Known) Practices

In *Estimating the Effect of Web-Based Homework*, Kelly, Heffernan, Heffernan, et al. (2013) used ASSISTments to compare traditional mathematics homework (with delayed, next day feedback) to the same assignment featuring immediate correctness feedback. All students participating in this RCE used ASSISTments to complete their homework, with feedback settings differing between randomly assigned conditions. The research design included 20 questions delivered using skill triplets (i.e., three similar skill problems presented consecutively) to determine the effectiveness of correctness feedback. Students in the control condition did not receive feedback while completing their homework, as shown in Figure 8. Blue dots within the left menu show completed problems. The next day in class, the teacher reviewed the homework without using ASSISTments reports, and simply read answers aloud as students corrected their work. The teacher then worked through requested problems on the board. Students in the experimental condition received immediate correctness feedback while completing their homework, as shown in Figure 9. The next day in class, the teacher used data from the Item Report to determine which problems to focus on during the homework review, with an emphasis on common wrong answers shared by multiple students.
Figure 8. The control condition as experienced by the student (Kelly, Heffernan, Heffernan, et al., 2013). Students were not told if their answers were correct or incorrect. This approach mirrors traditional homework. This study implemented problem triplets, or sets of three questions per skill, providing multiple opportunities to display skill knowledge.

Figure 9. The experimental condition as experienced by the student (Kelly, Heffernan, Heffernan, et al., 2013). Students were provided immediate correctness feedback as they responded to each problem. The student in this example was able to self-correct and progress through the first skill triplet, but struggled with the second.

Analysis of 63 students suggested reliable improvements in student learning through the addition of correctness feedback. Students in the control group showed an average gain of 59% from pretest to posttest (an effect size of 0.52), while students in the experimental group showed an average gain of 74% (an effect size of 0.56). It should be noted that Cohen’s “rule of thumb” for interpreting effect sizes has been somewhat discredited as a measure for benchmarking the practical significance of effects, especially when working with researcher defined measures (Lipsey, et al., 2012). Instead, it is recommended that researchers compare growth attributed to an intervention to normative expectations. Comparing gains across conditions, this method suggests a reliable 15% increase in average learning gains. It is also possible to benchmark these findings against the results of similar studies, which have a mean effect size of 0.43 (Lipsey, et al., 2012), showing the clear strength of providing immediate correctness feedback as an intervention. Kehrer, Kelly & Heffernan (2013) replicated the positive effects of immediate correctness feedback observed in Kelly, Heffernan, Heffernan, et al.’s original work (2013).

Similar hypotheses examining the efficacy of feedback within ASSISTments have led to numerous publications over the past decade. Mendicino, Razzaq & Heffernan, (2009) examined the effectiveness of mathematics homework with scaffolded tutoring in comparison to traditional paper and pencil homework. Students that received adaptive scaffolding showed significant learning gains over those
following traditional homework procedures. Razzaq, Heffernan, & Lindeman (2007) suggested that adaptive scaffolding led to greater learning gains than on demand hints. Researchers observed an interaction between students’ proficiency levels and the effectiveness of feedback styles, with less proficient students benefiting from scaffolding and more proficient students benefiting from hints. Follow-up studies confirmed that on demand hints produced more reliable and robust learning in highly proficient students (Razzaq & Heffernan, 2010). Singh et al. (2011) then compared correctness feedback with on demand hints. Multiple trials consistently revealed that hint feedback led to significantly improved learning over correctness feedback alone. Research has also examined the content presented within feedback, through comparisons of worked examples and scaffolded problem solving (Shrestha, et al., 2009; Kim, et al., 2009) and investigations of motivational feedback (Kelly, Heffernan, D’Mello, et. al., 2013; Ostrow, Schultz, & Arroyo, 2014). Results suggesting the consistent benefits of feedback have allowed researchers working within ASSISTments to expand their questions from a seed - “Does immediate feedback help?” - to more detailed investigations - “What type of immediate feedback is most effective?”

**Collaborative Hypothesis Growth: Enhancing System Content, Improving Learning, Advancing Science**

In *Testing the Multimedia Principle in the Real World: A Comparison of Video vs. Text Feedback in Authentic Middle School Math*, Ostrow & Heffernan expanded on “feedback is good” hypotheses to examine the effectiveness of various feedback mediums (2014). Prior to this study, ASSISTments delivered feedback via text, altering color and typeface to draw students’ attention to significant variables and themes. This RCE pushed that boundary to compare learning outcomes when identical feedback was delivered using short video snippets. Outcomes of student performance and response time were measured across six problems pertaining to the Pythagorean theorem. All students received the same six questions in mixed orders, receiving three opportunities for text feedback and three opportunities for video feedback over the course of the assignment. As shown in Figure 10, feedback was matched across medium; videos were comprised of a researcher working through each feedback step while referencing images on a white board. Students received feedback through scaffolds, by requesting assistance or answering a problem incorrectly. Learning gains were compared on the second question, across students who received feedback on the first question. Following the problem set, students were asked a series of survey questions to judge how they viewed the addition of video to their assignment.

![Figure 10. A comparison of text and video feedback conditions as experienced by students (Ostrow & Heffernan, 2014). Isomorphic problems featured matched content feedback across mediums.](image)
Results of an analysis of 89 students that completed the assignment and were able to access video content revealed that video feedback increased the likelihood of accuracy on the next problem. Students spent significantly longer consuming video feedback but answered their next question more efficiently. Assessing self-report measures, 86% of students found the videos at least somewhat helpful and 83% of students wanted video in future assignments (Ostrow & Heffernan, 2014). Based on these findings, teachers and researchers have been recruited to create video feedback for Skill Builder problems to expand the amount of video content available within the system and allow for further examination into the effects of video. The ease with which teachers and researchers are able to record short video messages and upload them to the system suggests that this approach is a plausible avenue for crowdsourcing feedback (Kittur, et al., 2013; Howe 2008). Crowdsourcing and learnersourcing (Kim, 2015) feedback are future directions for the ASSISTments platform, as infrastructure improvements are required to optimally support, organize, and vet feedback collection at scale.

Many of the studies that best define ‘collaborative hypothesis growth’ are currently underway within the TestBed, examining the effectiveness of particular types of feedback. Numerous researchers are investigating what drives the apparent effects of video feedback by comparing various types of videos (e.g., recorded human tutoring, a “pencast” problem walkthrough with audio explanation, and peer videos with tutoring led by other students). Many of these studies are pushing ASSISTments’ technological boundaries, establishing a demand for specific infrastructure improvements that will help the system and its content to evolve.

**Infrastructure Improvements: Research-Based Platform Evolution**

Research on the efficacy of feedback mediums laid the groundwork for debates about the possible impacts of allowing students to choose between mediums. Without any real capacity to provide choice, ASSISTments was reaching a tipping point for infrastructure improvement. A pilot study was conducted by taking advantage of bugs in the system to ‘mock up’ student choice (Ostrow & Heffernan, 2015). This simple RCE examined interactions between student choice and feedback medium using a 2x2 factorial design, depicted in Figure 11. Two versions of a problem set on simple fraction multiplication were created, one incorporating text feedback and one incorporating video feedback. Short, 15-30 second video snippets were designed to be as comparable to text feedback as possible, in order to compare delivery medium. At the start of the assignment, students were randomly assigned to either the choice condition or the control condition. Those assigned to the choice condition were asked what type of feedback they wished to receive while working on their assignment, as shown in Figure 12, and were routed accordingly. Those assigned to the control were immediately re-assigned to either video or text feedback.

![Figure 11](image-url) **Figure 11.** Experimental design used to investigate student choice as a pilot study within ASSISTments (Ostrow & Heffernan, 2015). Prior to this study, students were not able to exert control over their assignments within the platform.
Considering a sample of 78 middle school students that completed this pilot, results suggested that feedback medium did not have a specific impact on learning gains within this context (contrary to results presented earlier on the efficacy of video feedback, suggesting that perhaps video is not effective for all age ranges or skill domains and beginning to answer “What works best? For whom? When?”). However, students that were able to choose their feedback medium showed significant improvements over students that were randomly assigned a medium. Students with choice earned higher scores on average, used fewer hints and attempts, and persisted longer than those not provided choice. Perhaps the most interesting observation: learning gains were higher in students that were provided choice, regardless of whether or not the student actually ended up requesting feedback during the assignment (Ostrow & Heffernan, 2015). These results became the driving force for a significant infrastructure improvement within the ASSISTments platform that would allow for conditional path routing. An If-Then routing structure was developed under the SI2 NSF grant (SI2-SSE&SSI: 1440753) to extend research capabilities within ASSISTments and the ASSISTments TestBed. Hypotheses regarding student choice, and other routing based research, can now be easily examined with greater validity and at scale.

A replication of the choice pilot by Ostrow & Heffernan (2015) was designed using the If-Then routing structure, as shown in Figure 13. The inclusion of conditional path routing helped to enhance the internal validity of video based research by allowing sample populations to be refined to only include students with the technological capacity to view video content. While in hindsight this feature seems like an obvious requirement for video based research, it was not possible within ASSISTments prior to If-Then routing. Thus, it is clear how this new feature has the potential to improve and expand research within the ASSISTments TestBed.
Figure 13. Updated choice design replicating Ostrow & Heffernan (2015) with an If-Then routing structure for greater internal validity. The initial If-Then statement assesses students’ technological capacity for viewing video content, while the second If-Then controls routing in the choice condition.

An example of how a researcher might go about building an ASSISTments problem set with simple If-Then routing is shown in Figure 14. The building process requires three elements: a conditional statement, a true path, and a false path (Ostrow & Heffernan, 2016). The conditional statement can include a problem or problem set, with an adjustable setting that guides path routing based on student performance as measured by completion or accuracy. If performance meets this preset threshold, the student is routed into the true path, or the second section in Figure 14 (“Video Chosen”). If performance does not meet this preset threshold, the student is routed into the false path, or the third section in Figure 14 (“Text Chosen”). In this example, the conditional statement is a single preference question, much like that shown in Figure 12. Video feedback is set as the ‘correct’ answer, routing students based on the “Then” clause, while text feedback is set as the ‘incorrect’ answer, routing students based on the “Else” clause. Students receive this problem in test mode (i.e., without correctness feedback, showing only a blue dot for completion) therefore restricting the inner workings of the routing system from student view and removing the risk of undue penalization for a ‘wrong opinion.’ Numerous studies now running within the ASSISTments TestBed implement If-Then routing in some capacity (e.g., as technical validation, as adaptive performance routing, to trigger interventions for struggling students, or to buffer sampling within intent-to-treat studies seeking to help only students with low skill proficiency). This simple infrastructure improvement completes an iteration of the cycle of perpetual evolution, opening new avenues for fresh ‘seed-level’ hypotheses to start the cycle anew.
Figure 14. The researcher’s view while constructing a study using If-Then routing within an ASSISTments problem set. The study design shown here mirrors that in Figure 13.

Future Directions of the ASSISTments Platform

It is difficult to advocate for a future consisting of research infused educational technologies without touching briefly on future goals for the ASSISTments platform. With a focus on disseminating the ASSISTments TestBed and enhancing its validity as a collaborative tool for sound science, the cycle of perpetual evolution will bring about a number of significant infrastructure improvements for ASSISTments in the near future. Perhaps the most immediate change, as suggested by the research presented herein, will be extending the platform to support teacher-sourced and learner-sourced feedback. The platform has 25,000 vetted mathematics problems that were created by Worcester Polytechnic Institute and Carnegie Mellon University. In addition, teachers have added over 100,000 problems to the platform (many that already include some form of feedback). The first step toward crowdsourcing feedback for these problems is to allow teachers to create tutoring strategies in support of content owned by others (rather than only in support of their own content). Differing teachers will offer differing solution approaches, which may help struggling students to see a problem from a different perspective. A select group of teachers and students have already recorded video feedback for use in a set of RCEs examining the potential benefits and obstacles of crowdsourcing feedback at scale. Eventually, this approach will be scaled to allow students to ‘show their work’ and provide explanations for their peers through a tool called PeerASSIST (Heffernan, Ostrow, et al., 2016). A task already appreciated by most mathematics teachers, showing work will help students to solidify their understanding of the content while creating feedback to benefit other users (Kulkarni, et al., 2013). The network effects inherent to teachersourcing and learnersourcing feedback will enhance system content at an impressive scale (Bush & Mott, 2009).
The implementation of crowdsourcing will naturally give way to another goal for the future of ASSISTments: establishing an automated process to select optimal feedback using contextual k-armed bandits. This approach, rooted in the theory of sequential design (Robbins, 1952), is an algorithmic approach to the exploration/exploitation trade-off. Essentially, with a pool of content available to students (i.e., many types of feedback), it is necessary to repeatedly sample the efficacy of assigned content in order to maximize the delivery of effective content while minimizing the delivery of ineffective content. The use of k-armed bandits will minimize detriment to students while allowing for the dynamic versioning of materials and setting the stage for personalized learning (i.e., algorithmically establishing “What works best? For whom? When?”). An important feature that will grow from the implementation of k-armed bandits will be the capacity to store user variables for lasting personalization. Variables like initial performance, particular student responses, or specific student characteristics could help to optimize content and feedback delivery for each student, both within and across assignments. The ASSISTments team expects that these goals will strengthen the platform and inspire new avenues for scientific inquiry.

In Conclusion: Infuse Educational Technologies with Collaborative Research to Promote Sound Science

Systemic change does not stem from a small number of large-scale RCEs funded by government grants, but instead from a revolution in thought surrounding the value of technology based learning applications. As shown herein, infusing pre-existing learning technologies with the capability to support RCEs is the first step in kick starting this revolution. From there, the platform can expand as a shared scientific tool utilized by a community of researchers collaborating to better understand the efficacy of educational interventions. ASSISTments bridges practice and research by enabling researchers to work collaboratively with teachers and students, and by providing unprecedented access to authentic learning environments and actionable classroom data. The collaborative nature of the ASSISTments TestBed gives way to a cycle of perpetual evolution that inspires continuous advancements to ASSISTments content while simultaneously advancing knowledge of best practices. Insights and innovations drawn from research findings can be incorporated into the system itself as well as future research, with each successive step building upon previous contributions.

Research infused platforms have the potential to drive inquiry for a diverse community of researchers through the low-cost, rapid iteration of valid, generalizable, and noninvasive investigations within authentic learning environments. Systems like ASSISTments can provide researchers with access to an extensive and diverse subject pool, an automated fine-grained logging of educational data, validated measures of student learning and affect, and automated data reporting and analysis to tackle the high stakes nature of typical education research. With similar research-focused platforms in the field, it would also be possible for researchers to compare learning interventions across platforms to better measure the reliability and generalizability of results. These platforms offer a unique opportunity for the synergistic growth of research and policy detailing best practices in education. If these platforms grow to welcome collaborative research, educational technology will reach its long awaited tipping point and begin to broadly impact the efficacy and validity of research across domains. Tomorrow’s educational technology demands a revolution in today’s approaches to research at scale: pave the way for sound collaborative science and the rest will follow.

Acknowledgements

We would like to thank the 150+ students at WPI that have helped create ASSISTments and that have used the platform to conduct research. We would also like to acknowledge funding from the National Science Foundation (0231773, 1109483, 0448319, 0742503, 1031398, and 1440753), GAANN, The Spencer Foundation, The U.S. Department of Education (R305K03140, R305A07440, R305C100024, and R305A120125), The Office of Naval Research, and the Bill and Melinda Gates Foundation. Our thanks to the Executive Editor for reviewing our submission prior to publication. The first author also extends her love and appreciation to S.O. & L.P.B.O.
References


