

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

KORINN OSTROW

Worcester Polytechnic Institute

NEIL HEFFERNAN

Worcester Polytechnic Institute

JOSEPH JAY WILLIAMS

Harvard University

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 (S12-SSE&SS1: 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.*

Teachers College Record Volume 119, 030306, March 2017, 36 pages

Copyright © by Teachers College, Columbia University

0161-4681

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.*

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 (Institute of Education Sciences [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 non-replicability, contributing to a growing crisis of faith in RCEs spanning numerous scientific fields (Achenbach, 2015; Ioannidis, 2005; 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 (e.g., 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 (ITS), computer-aided testing platforms, and adaptive learning applications offering new and unique approaches to learning, heralding 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. Therefore, the National Education Technology Plan predicted that these platforms would play a key role in personalizing educational interventions (U.S. Department of Education, 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 and 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, Maldonado, et al., 2015; Williams, Ostrow, 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 (Heffernan & Williams, 2014) 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 such as 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 they 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 (Kelly, Heffernan, Heffernan, et al., 2013; Koedinger, McLaughlin, & Heffernan, 2010; Mendicino, Razzaq, & Heffernan, 2009; Miller, Zheng, Means, & Van Brunt, 2013; Singh et al., 2011; Soffer et al., 2014). Most recently, SRI International reported the results of an efficacy trial of ASSISTments, showing that the platform caused large, reliable learning gains on standardized assessments (Rochelle, Feng, Murphy, & Mason, 2016).

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 seventh-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 (e.g., 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 (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) can be manipulated to incorporate experimental modifications.

Figure 1. Example of a problem viewed within the builder: Notice that the interface allows creation of the problem itself, answers (both correct and incorrect), and tutoring strategies, and the navigation menu in the top right corner allows the user to navigate from editing a main problem to editing feedback.

PRA4XBW - Stem and Leaf - Mean Edit name

Details View Problem Test Drive New Copy

Problem Type: Standard problem

Parent: Problem 764022
[No tags currently assigned]
[Tag Skills to Problem](#)

+ New Main Problem

Main Problem 1

Hint1 Hint

* Click and drag main problems to re-order(may require refresh)

Font Sizes

B
I
U
S
☰
☷
A
A

🔗
🖼️
📄
↶
↷
📊
√
Ω
x₂
x⁺
👁️
<>

The following stem and leaf plot shows the number of shoes sold each week at a store. According to this plot, what is the mean of shoes sold each week?

Shoes Sold Each Week

Stem	Leaf
2	1, 1, 3, 6, 6, 7
3	1, 6, 8
4	8
5	2, 3

(Round to the nearest hundredths place)

Save Problem Body

Problem Type: Algebra

Answers What's this?

✓ 33.5 [Edit](#) [Delete](#) [Drag](#)

+ New Answer

Tutoring Strategies What's this?

Hint1 Hint [Disable](#) [Edit](#) [Delete](#)

+ New Strategy

Figure 2. 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 of 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.

Problem ID: PRA459W [Comment on this problem](#)

Find the mean (average) of the numbers below.

11, 4, 13, 8

To find the mean, add all the numbers and divide the sum by how many numbers there are.

First, find the sum of the numbers.

Sum = $11 + 4 + 13 + 8 = 36$

[Comment on this hint](#)

Problem ID: PRA459W [Comment on this problem](#)

Find the mean (average) of the numbers below.

11, 4, 13, 8

Type your answer below (mathematical expression):

[Submit Answer](#)

Problem ID: PRA459W [Comment on this problem](#)

Find the mean (average) of the numbers below.

11, 4, 13, 8

It seems you have found the SUM instead of finding the AVERAGE.

11, 4, 13, 8
 $11 + 4 + 13 + 8 = 36$

To find the average, first find the sum and then divide by the number of items you added.

$11 + 4 + 13 + 8 = 36$

A A A A
 1 2 3 4 = 4

Type your answer below (mathematical expression):

36

Sorry, try again: '36' is not correct

67% [Show hint 1 of 3](#)

[Submit Answer](#)

Problem ID: PRA4AUZ [Comment on this problem](#)

Find the mean (average) of the numbers below.

11, 4, 13, 8

Type your answer below (mathematical expression):

[Submit Answer](#)

Problem ID: PRA4AUZ - 1104185 [Comment on this problem](#)

Looks like you could use some help. Let's look at a similar Example Problem.

Example Problem:

Find the mean (average) of the numbers below.

20, 0, 9, 15, 10, 6

Step 1 of 3

To find the mean, add all the numbers and divide by how many numbers there are.

The first step is to sum up the numbers that were given.

$20 + 0 + 9 + 15 + 10 + 6 = 60$

Now it's your turn. Find the sum of all the numbers in **your** problem: 11, 4, 13, 8

Enter the sum in the box below.

Type your answer below (mathematical expression):

[Submit Answer](#) [Show answer](#)

Problem ID: PRA459W [Comment on this problem](#)

Find the mean (average) of the numbers below.

11, 4, 13, 8

To find the mean, add all the numbers and divide the sum by how many numbers there are.

First, find the sum of the numbers.

Sum = $11 + 4 + 13 + 8 = 36$

Next, we need to know how many numbers we have. For this problem, there are 4 numbers.

Finally, divide the sum of all the numbers, 36, by how many numbers there are, 4.

Mean = $\frac{36}{4} = 9$

Type in 9.

Type your answer below (mathematical expression):

[Submit Answer](#)

0% [Show hint 1 of 3](#)

ASSISTments also offers optional features such as the Automatic Reassessment and Relearning System (ARRS), which helps to reassess student retention following skill builder mastery (Xiong & Beck, 2014), 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 seven, 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 on 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 expansion of 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.

Through NSF funding (Heffernan & Williams, 2014), 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

Figure 3. 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 whether 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.

Student/Problem [Unanonymize]	Average Data driven	PRAHE5Y Data driven	PRAHE5Z Data driven	PRAHE52 Data driven
Problem Average	60%	27%	61%	84%
Common Wrong Answers		1/9^10, 56% +feedback	1/5^13, 58% +feedback	
Correct Answer(s)		1/3^10	1/5^3	1/16^2
XXXXX *	50%	✗ 1/9^10	✗ 1/5^13	✗ 1/16^2
XXXXX *	45%	✗ 1/9^10	✓ 1/5^3	✓ 1/16^2
XXXXX *	55%	✓ 1/3^10	✗ 1/5^13	✓ 1/16^2

reports detailing anonymized study participation (Ostrow et al., 2016). These reports, as shown in Figure 4, provide basic analyses, including bias assessment (examining attrition across experimental conditions) and simple hypothesis testing on post-test performance. Researchers are also provided a student covariate file, detailing student information collected prior to study participation (e.g., 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.

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., 2016). 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 in post-test performance.

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.

Condition	Started (n)	Completed (n)	Completed (%)
Group A – Experiment 1	109	80	73.39
Group B – Experiment 2	87	60	68.97
Group C – Control	99	89	89.90
Total	295	229	77.63

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.

	Completed (n)	Posttest Score*
Group A – Experiment 1	80	34.40 (4.34)
Group B – Experiment 2	60	32.95 (3.89)
Group C – Control	89	44.11 (3.72)
Total	229	37.15 (3.98)

* 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

- 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).
- 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).
- 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).
- 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).

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 such as 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 such as 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 lack 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 (see, e.g., Williams & Williams, 2013; Williams, Paunesku, Haley, & Sohl-Dickstein, 2013). Instead, work with less regard for improving specific interventions has evaluated the platform's efficacy in schools (Murphy, Gallagher, Krumm, Mislevy, & Hafter, 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.

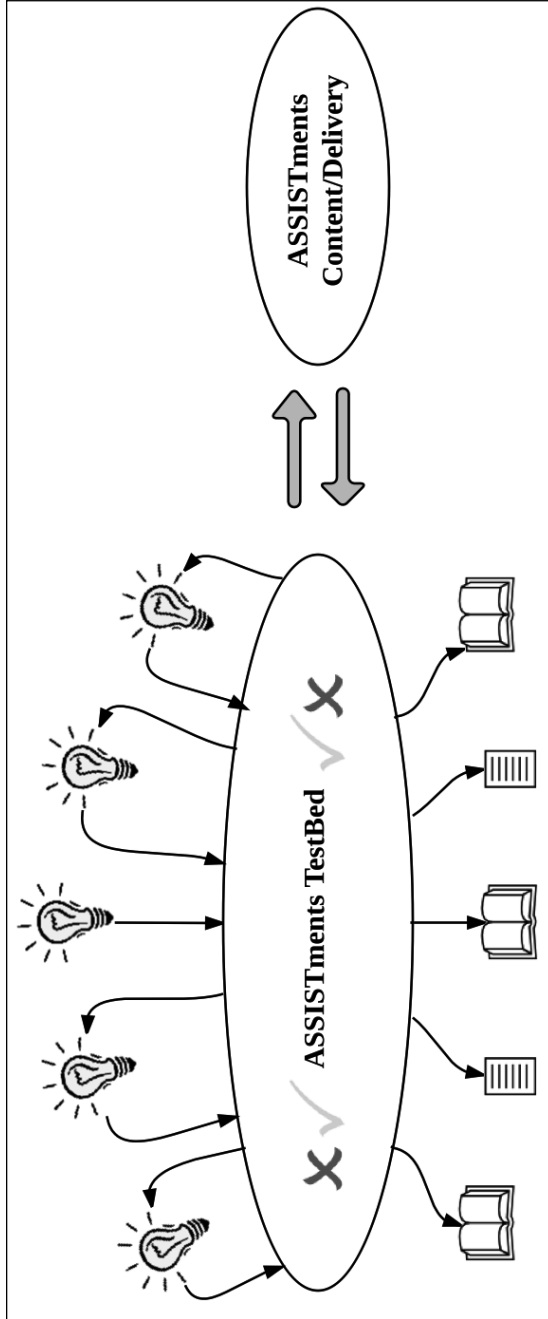
The application of stringent research methodologies to improve learning technologies and educational outcomes is severely lagging. This deficit is what makes the ASSISTments TestBed so unique. The TestBed guides researchers through the process of running practical RCEs by leveraging ASSISTments' content 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, whereas 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.

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 University. Since its

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.



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 (Heffernan & Williams, 2014).

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 versus delayed feedback (Fyfe, Rittle-Johnson, & DeCaro, 2012)
- Comparing the types of hints provided adaptively to learners (Stamper, Eagle, Barnes, & Croy, 2013)
- Comparing levels of feedback, from guided to open (Sweller, Kirschner, & Clark, 2007)
- Comparing “what you see is what you get” with interaction (Keehner, Hegarty, Cohen, Khooshabeh, & Montello, 2008)
- Prompting for comparison of analogous problems and worked examples (Jee et al., 2013)

Sequencing and spacing

- Changing schedules and procedures for practice sessions and quizzes (Roediger & Karpicke, 2006)
- Testing the effectiveness of pretesting prior to instruction (Richland, Kornell, & Kao, 2009)
- Spacing skill content (Pashler, Rohrer, Cepeda, & Carpenter, 2007)
- Examining testing effects (Butler & Roediger, 2007)

Self-regulated learning and 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, Holmes, Day, & Bonn, 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 and dynamics for peer assistance (Walker, Rummel, & Koedinger, 2011)
- Examining how confidence affects performance in early algebra (Mazzocco, Murphy, Brown, Rinne, & Herold, 2013)

Assessment

- Examining computational models used to diagnose learner state (Rafferty & Griffiths, 2014)
- Examining computational methods for assessing affective states (Ocumpaugh, Baker, Gowda, Heffernan, & Heffernan, 2014)
- Examining forgetting (Storm, Bjork, Bjork, & Nestojko, 2006)

Motivation

- Embedding motivational videos from teachers (Kelly, Heffernan, D'Mello, Namias, & Strain, 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, Khan, & Schacter, 2013)

Mathematics education

- Comparing representational formats in supporting mathematics learning (Rau, Aleven, Rummel, & Rohrbach, 2012)
- Investigating effective presentations of worked examples in mathematics (Booth, Lange, Koedinger, & Newton, 2013)
- Examining strategies for learning fractions (Cordes, Williams, & Meck, 2007)
- Testing images of manipulatives versus virtual manipulatives (Mendiburo, Sulcer, Biswas, & Hasselbring, 2012)

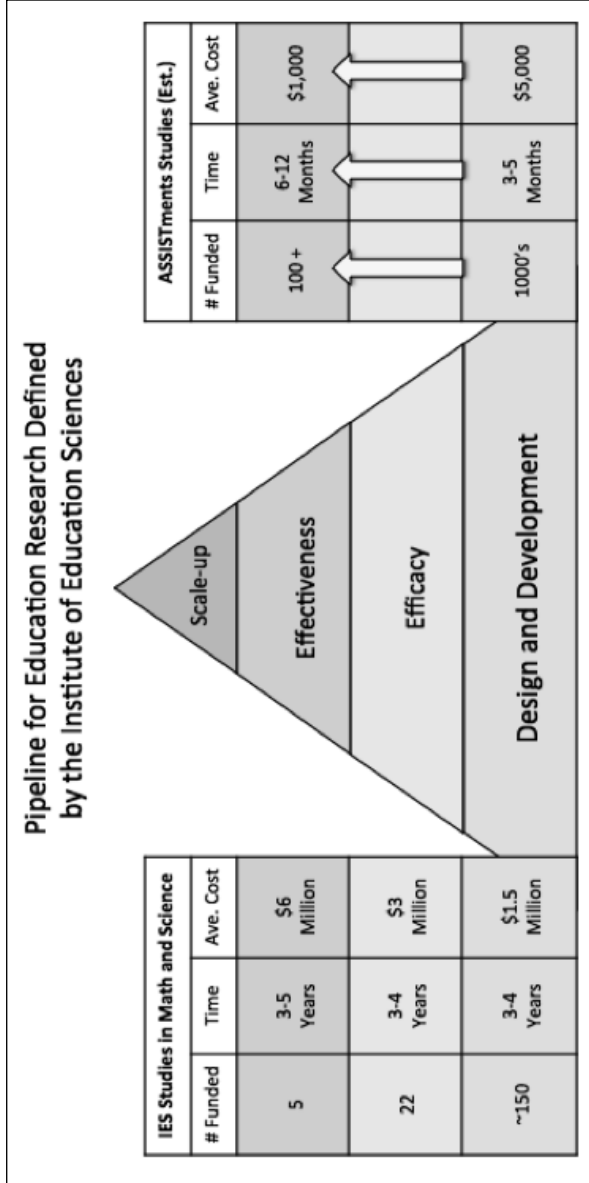
By building these types of collaborative scientific tools, the cost of funding educational research could be drastically reduced. For instance, the

Institute of Education Sciences (IES) currently funds efficacy trials for promising interventions that cost an average of \$3 million and can involve more than 50 schools. Larger and more stringent effectiveness trials carry a median cost of \$6 million. 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.

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 in manipulating prebuilt 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 (e.g., 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 APIs 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

Figure 6. Pipeline for education research as defined by the IES (2015) 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.



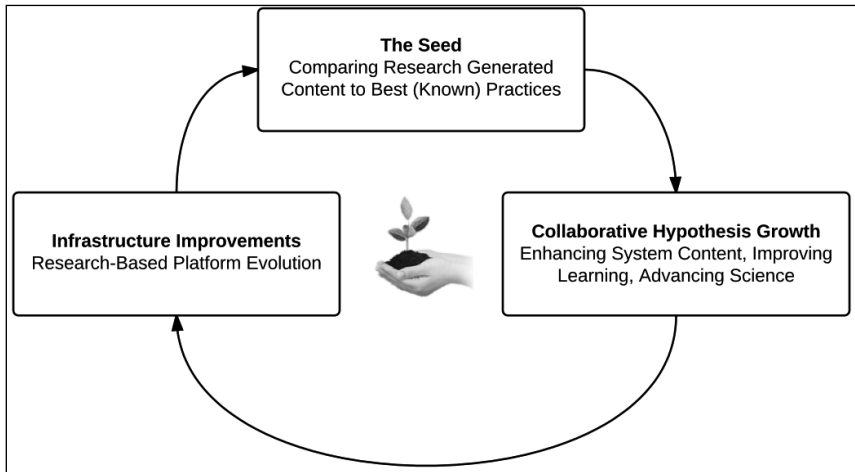
better measure the reliability and generalizability of results. Collaborative research goals that crosscut platforms may finally usher in the tipping point of educational technologies (Bush & Mott, 2009; Gladwell, 2002) as researchers grow to understand What works best? For whom? When?

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 of 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 expansion 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 publications. New hypotheses form and grow as results are observed, naturally evolving until they push 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.

Ever-expanding progress is a core concept for effectively marketing commercial products, but it 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

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.



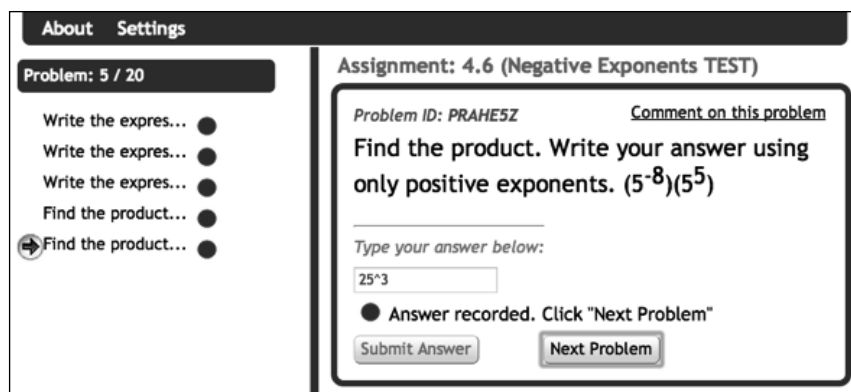
benefits made possible by the cycle of perpetual evolution. The following sections step through this cycle, defining exemplary research at each stage as conducted within ASSISTments and the ASSISTments TestBed.

THE SEED: COMPARING RESEARCH-GENERATED CONTENT TO BEST (KNOWN) PRACTICES

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

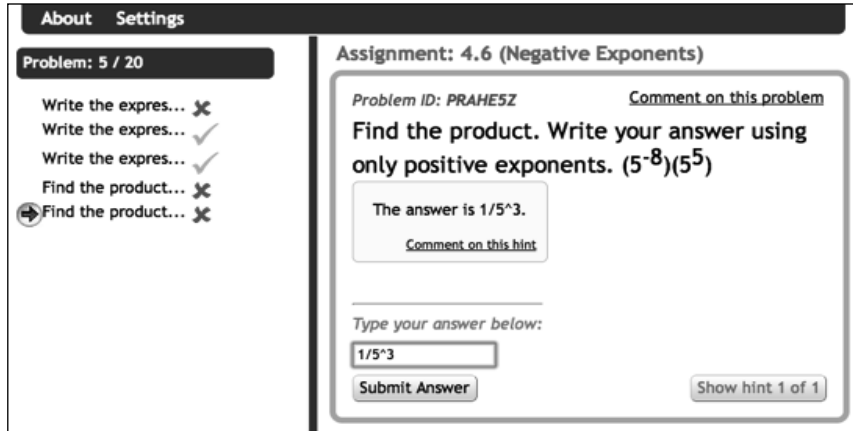
using ASSISTments reports and simply read answers aloud as students corrected their work. The teacher then worked through requested problems on the board. Students in the experimental condition received immediate correctness feedback while completing their homework, as shown in Figure 9. The next day 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. Control condition as experienced by the student (Kelly, Heffernan, Heffernan, et al., 2013): Students were not told whether their answers were correct or incorrect, an approach that mirrors traditional homework. This study implemented problem triplets, or sets of three questions per skill, providing multiple opportunities to display skill knowledge.



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 post-test (an effect size of 0.52), whereas 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, and Heffernan (2013)

Figure 9. 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.



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 et al. (2009) examined the effectiveness of mathematics homework with scaffolded tutoring in comparison to traditional paper-and-pencil homework. Students who received adaptive scaffolding showed significant learning gains over those following traditional homework procedures. Razzaq, Heffernan, and 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 (R. Kim, Weitz, Heffernan, & Krach, 2009; Shrestha 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 seeds (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


Ostrow and Heffernan (2014) expanded on the “feedback is good” hypothesis to examine the effectiveness of various feedback mediums. 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 consisted of a researcher working through each feedback step while referencing images on a whiteboard. Students received feedback through scaffolds, by either requesting assistance or answering a problem incorrectly. Learning gains on the second question were compared 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.

Results of an analysis of 89 students who 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 (Howe, 2008; Kittur et al., 2013). Crowdsourcing and learnersourcing (J. 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.

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.

[Comment on this problem](#)

The Pythagorean Theorem can be used to solve for side c . In this problem, the value of a is given as 6 feet and the value of b is given as 22 feet. Plug in the values of a and b into the Pythagorean Theorem and we can solve for c .



$$a^2 + b^2 = c^2$$

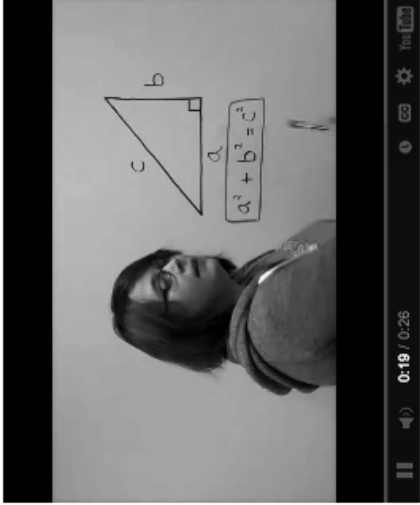
Try again to answer the main problem above. Remember to round your answer to the nearest tenth place.

Type your answer below (mathematical expression):

[Submit Answer](#)

[Break this problem into steps](#)

[Comment on this problem](#)



After you've watched the video, try again to answer the main problem above. Remember to round your answer to the nearest tenth place.

Type your answer below (mathematical expression):

[Submit Answer](#)

[Break this problem into steps](#)

Many of the studies that best define collaborative hypothesis growth are currently under way 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 2 x 2 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- to 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 reassigned to either video or text feedback.

For a sample of 78 middle school students who 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 who were able to choose their feedback medium showed significant improvements over students who 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 who were provided choice, regardless of whether or not the student actually ended up

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.

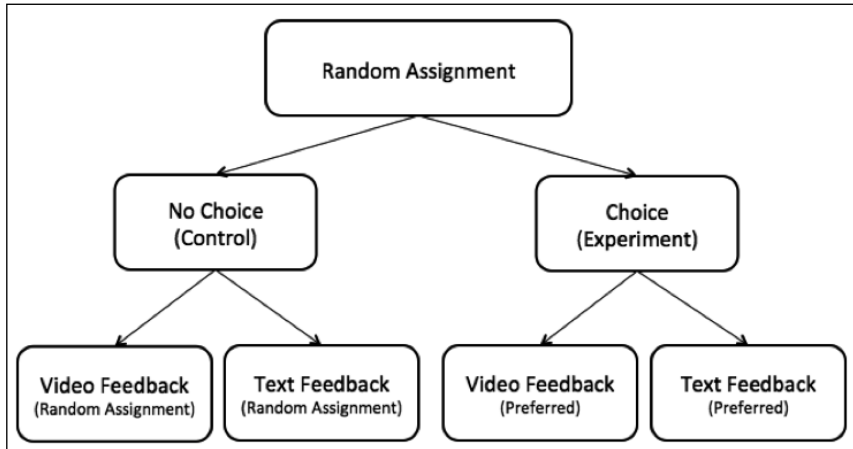


Figure 12. Student preference prompt guiding medium routing for those in the experimental condition (Ostrow & Heffernan, 2015).

Assignment: ReRoute

Problem ID: *PRAXCHQ* [Comment on this problem](#)

This problem set is a little bit different. We want to give you some say in how you learn!

Would you prefer:

Hints and feedback that use **text** to help you when you feel stuck.

OR

Hints and feedback that use short **videos** to help you when you feel stuck.

Select your answer below.

Select one:

I prefer text feedback!

I prefer video feedback!

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 (Heffernan & Williams, 2014) 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 and 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 include only students with the technological capacity to view video content. Although, 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.

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), thereby restricting the inner workings of the routing system from student view and removing the risk of undue penalties 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 13. Updated choice design replicating Ostrow and 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.

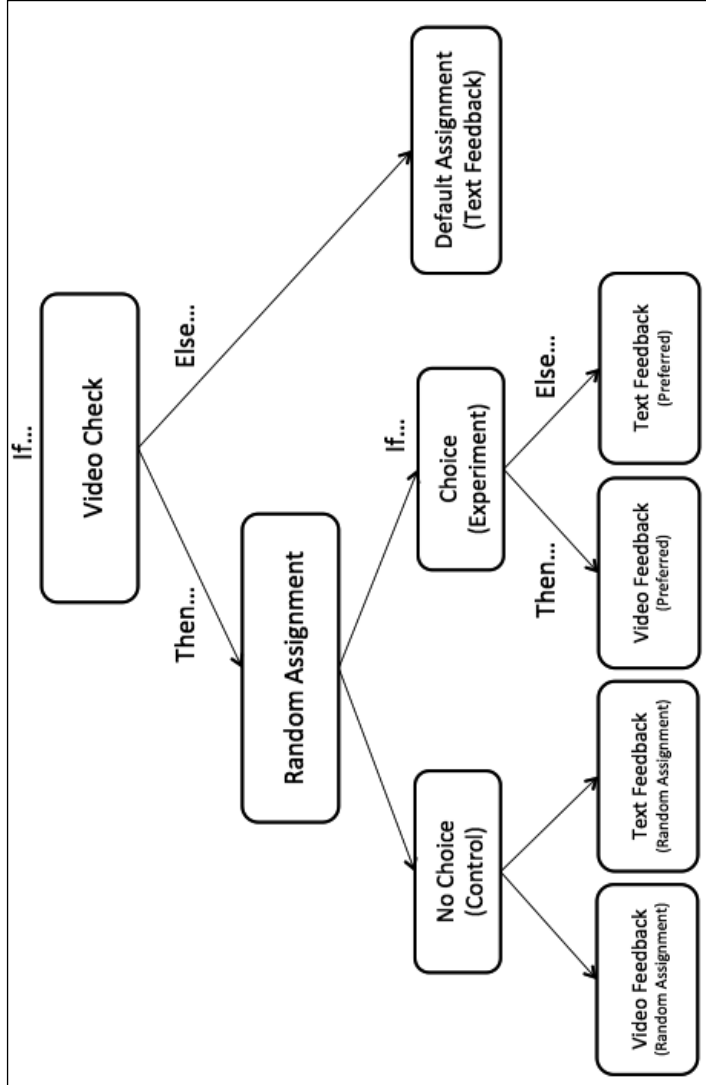
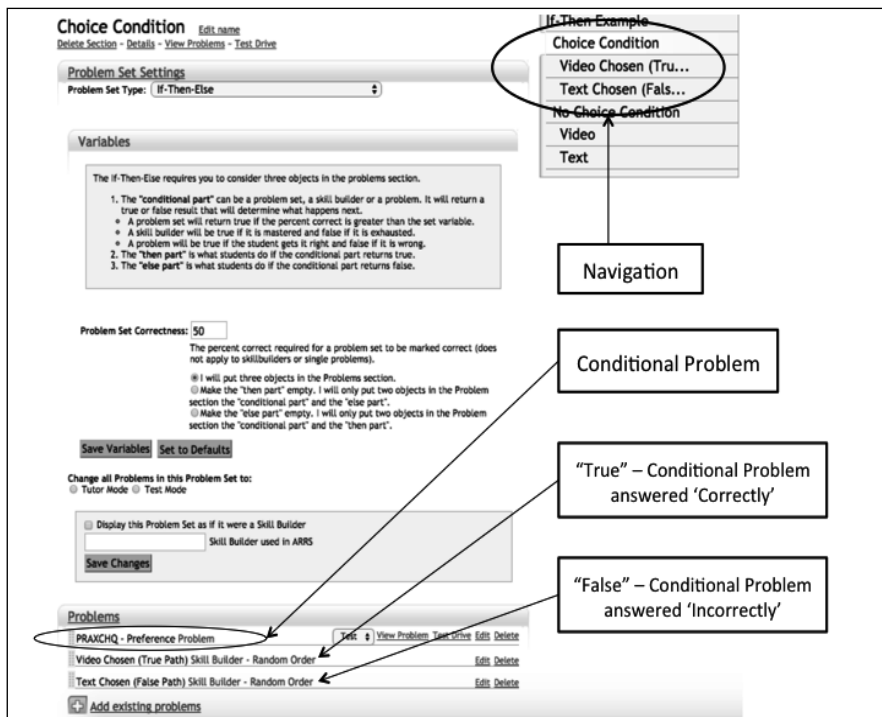


Figure 14. Researcher’s view while constructing a study using if-then routing within an ASSISTments problem set; study design shown here mirrors that in Figure 13.



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 teachersourced and learnersourced feedback. The platform has 25,000 vetted mathematics problems that were created by WPI and Carnegie Mellon University. In addition, teachers have added over 100,000 problems to the platform, many of which 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 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 is an algorithmic approach, rooted in the theory of sequential design (Robbins, 1952), to the exploration–exploitation tradeoff. 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 such as 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 preexisting 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 used 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 have a broad impact on 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.

ACKNOWLEDGMENTS

We would like to thank the 150+ students at WPI who have helped create ASSISTments and who 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. and L.P.B.O.

REFERENCES

- Achenbach, J. (2015, August 27). Many scientific studies can't be replicated: That's a problem. *The Washington Post*, Retrieved from <https://www.washingtonpost.com/news/speaking-of-science/wp/2015/08/27/trouble-in-science-massive-effort-to-reproduce-100-experimental-results-succeeds-only-36-times/>
- Adjei, S. A., & Heffernan, N. T. (2015). Improving learning maps using an adaptive testing system: PLACEments. In C. Conati, N. Heffernan, A. Mitrovic, & M. F. Verdejo (Eds.), *Proceedings of the 17th International Conference on Artificial Intelligence in Education* (pp. 517–520). Cham, Switzerland: Springer.
- Arnold, K. M., & McDermott, K. B. (2013). Free recall enhances subsequent learning. *Psychonomic Bulletin & Review*, *20*(3), 507–513.
- Belenky, D. M., & Nokes-Malach, T. J. (2013). Mastery-approach goals and knowledge transfer: An investigation into the effects of task structure and framing instructions. *Learning and Individual Differences*, *25*, 21–34.
- Bernacki, M. L., Byrnes, J. P., & Cromley, J. G. (2012). The effects of achievement goals and self-regulated learning behaviors on reading comprehension in technology-enhanced learning environments. *Contemporary Educational Psychology*, *37*(2), 148–161.
- Booth, J. L., Lange, K. E., Koedinger, K. R., & Newton, K. J. (2013). Using example problems to improve student learning in algebra: Differentiating between correct and incorrect examples. *Learning and Instruction*, *25*, 24–34.
- Bush, M. D., & Mott, J. D. (2009). The transformation of learning with technology: Learner-centricity, content and tool malleability, and network effects. *Educational Technology*, *49*(2), 3–20.
- Butler, A. C., & Roediger, H. L., III. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, *19*, 514–527.
- Chernyak, N., & Kushnir, T. (2013). Giving preschoolers choice increases sharing behavior. *Psychological Science*, *24*(10), 1971–1979.
- Cordes, S., Williams, C. L., & Meck, W. H. (2007). Common representations of abstract quantities. *Current Directions in Psychological Science*, *16*(3), 156–161.
- Ehrlinger, J., & Shain, E. A. (2014). How accuracy in students' self perceptions relates to success in learning. In V. A. Benassi, C. E. Overson, & C. M. Hakala (Eds.), *Applying science of learning in education: Infusing psychological science into the curriculum*. Retrieved from Society for the Teaching of Psychology website: <http://teachpsych.org/ebooks/asle2014/index.php>
- Fyfe, E. R., Rittle-Johnson, B., & DeCaro, M. S. (2012). The effects of feedback during exploratory mathematics problem solving: Prior knowledge matters. *Journal of Educational Psychology*, *104*, 1094–1108.
- Gladwell, M. (2002). *The tipping point: How little things can make a big difference*. Boston, MA: Back Bay Books.
- Heffernan, N., & Heffernan, C. (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*(4), 470–497.
- Heffernan, N. T., Ostrow, K. S., Kelly, K., Selent, D., Van Inwegen, E. G., Xiong, X., & Williams, J. J. (2016). The future of adaptive learning: Does the crowd hold the key? *International Journal of Artificial Intelligence in Education*, *26*(2), 615–644. doi:10.1007/s40593-016-0094-z
- Heffernan, N., & Williams, J. (2014). Adding research accounts to the ASSISTments' platform: Helping researchers do randomized controlled studies with thousands of students. National Science Foundation, SI2-SSE, Award #1440753. Abstract retrieved from http://www.nsf.gov/awardsearch/showAward?AWD_ID=1440753
- Howe, J. (2008). *Crowdsourcing: Why the power of the crowd is driving the future of business*. New York, NY: Crown Business.

- Institute of Education Sciences. (2013). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. Retrieved from U.S. Department of Education website: http://ies.ed.gov/ncee/pubs/evidence_based/randomized.asp
- Institute of Education Sciences. (2015). Request for applications: Education research grants (CFDA Number: 84.305A). Retrieved from U.S. Department of Education website: https://ies.ed.gov/funding/pdf/2016_84305A.pdf
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med*, 2(8): e124. doi:10.1371/journal.pmed.0020124
- Jee, B., Uttal, D., Gentner, D., Manduca, C., Shipley, T., & Sageman, B. (2013). Finding faults: Analogical comparison supports spatial concept learning in geoscience. *Cognitive Processing*, 14(2): 175–187.
- Keehner, M., Hegarty, M., Cohen, C., Khooshabeh, P., & Montello, D. R. (2008). Spatial reasoning with external visualizations: What matters is what you see, not whether you interact. *Cognitive Science*, 32, 1099–1132.
- Kehrer, P., Kelly, K., & Heffernan, N. (2013). Does immediate feedback while doing homework improve learning? In C. Boonthum-Denecke & G. M. Youngblood (Eds.), *Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference* (pp. 542–545). Palo Alto, CA: AAAI Press.
- Kelly, K., Heffernan, N., D’Mello, S., Namias, J., & Strain, A. (2013). Adding teacher-created motivational video to an ITS. In C. Boonthum-Denecke & G. M. Youngblood (Eds.), *Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference* (pp. 503–508). Palo Alto, CA: AAAI Press.
- Kelly, K., Heffernan, N., Heffernan, C., Goldman, S., Pellegrino, G., & Soffer, D. (2013). Estimating the effect of web-based homework. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Proceedings of the 16th International Conference on Artificial Intelligence in Education* (pp. 824–827). Berlin, Germany: Springer.
- Kim, J. (2015). *Learnersourcing: Improving learning with collective learner activity* (Doctoral dissertation, Massachusetts Institute of Technology). Retrieved from <http://juhokim.com/files/JuhoKim-Thesis.pdf>
- Kim, R., Weitz, R., Heffernan, N., & Krach, N. (2009). Tutored problem solving vs. “pure” worked examples. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 3121–3126). Austin, TX: Cognitive Science Society.
- Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., & Horton, J. (2013). The future of crowd work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (pp. 1301–1318). New York, NY: ACM.
- Koedinger, K. R., Baker, R. S., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. In C. Romero, S. Ventura, M. Pechenizkiy, & R. S. J. d. Baker (Eds.), *Handbook of Educational Data Mining* (p. 628). Boca Raton, FL: CRC Press.
- Koedinger, K. R., Booth, J. L., & Klahr, D. (2013). Instructional complexity and the science to constrain it. *Science*, 342: 935–937.
- Koedinger, K. R., McLaughlin, E., & Heffernan, N. (2010). A quasi-experimental evaluation of an on-line formative assessment and tutoring system. *Journal of Educational Computing Research*, 4, 489–510.
- Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., . . . Klemmer, S. R. (2013). Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 20(6), article 33. doi:10.1145/2505057
- Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., . . . Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms (NCSEER 2013-3000)*. Washington, DC: U.S. Department

- of Education, Institute of Education Sciences, National Center for Special Education Research.
- Mazzocco, M. M. M., Murphy, M. M., Brown, E. C., Rinne, L., & Herold, K. H. (2013). Persistent consequences of atypical early number concepts. *Frontiers in Psychology*, *4*, 486.
- Mendiburo, M., Sulcer, B., Biswas, G., & Hasselbring, T. S. (2012). Interactive virtual representations, fractions, and formative feedback. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Proceedings of the 11th International Conference on Intelligent Tutoring Systems* (pp. 716–717). Berlin, Germany: Springer.
- Mendicino, M., Razzaq, L., & Heffernan, N. T. (2009). Improving learning from homework using intelligent tutoring systems. *Journal of Research on Technology in Education*, *41*(3), 331–346.
- Miller, G. I., Zheng, Y., Means, B., & Van Brunt, J. (2013). *Next generation learning challenges Wave II: Final evaluation report* (Contract #20462, Work order #18). Menlo Park, CA: SRI. Retrieved from <https://docs.google.com/file/d/0B2X0QD6q79ZJU9Kd2JuVTN0VWhTYVRhX254QV85Njdqc1Vj/edit?pli=1>
- Murphy, R., Gallagher, L., Krumm, A. E., Mislevy, J., & Hafter, A. (2014). Research on the use of Khan Academy in schools: Research brief. Menlo Park, CA: SRI. Retrieved from http://www.sri.com/sites/default/files/publications/2014-03-07_implementation_briefing.pdf
- National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core state standards*. Washington, DC: Authors.
- National Research Council (2002). *Scientific research in education*. Washington, DC: The National Academies Press. Retrieved from <http://www.nap.edu/catalog/10236/scientific-research-in-education>
- Ocuppaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. (2014). Population validity for educational data mining models: A case study in affect detection. *British Journal of Educational Technology*, *45*(3), 487–501.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). doi:10.1126/science.aac4716
- Ostrow, K. S., & Heffernan, C. (2016) *ASSISTments TestBed resource guide: The if-then-else section type*. Retrieved from <http://tiny.cc/IfThenElse>
- Ostrow, K. S. & Heffernan, N. T. (2014). Testing the multimedia principle in the real world: A comparison of video vs. text feedback in authentic middle school math assignments. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 296–299). London, U.K.: International Educational Data Mining Society.
- Ostrow, K. S., & Heffernan, N. T. (2015). The role of student choice within adaptive tutoring. In C. Conati, N. Heffernan, A. Mitrovic, & M. F. Verdejo (Eds.), *Proceedings of the 17th International Conference on Artificial Intelligence in Education* (pp. 752–755). Cham, Switzerland: Springer.
- Ostrow, K.S., Schultz, S.E. & Arroyo, I. (2014). Promoting Growth Mindset Within Intelligent Tutoring Systems. In CEUR-WS (1183), Gutierrez-Santos, S., & Santos, O.C. (eds.) *EDM 2014 Extended Proceedings*. In Ritter & Fancsali (eds.) NCFPAL Workshop. pp. 88-93. Retrieved from http://ceur-ws.org/Vol-1183/ncfpal_paper03.pdf
- Ostrow, K. S., Selent, D., Wang, Y., Van Inwegen, E., Heffernan, N., & Williams, J. J. (2016). The Assessment of Learning Infrastructure (ALI): The theory, practice, and scalability of automated assessment. *Proceedings of the 6th International Conference on Learning Analytics and Knowledge*, 2016, 279–288. doi:10.1145/2883851.2883872
- Pashler, H., Rohrer, D., Cepeda, N., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review*, *14*(2), 187–193.

- Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Proceedings of the 28th Conference on Neural Information Processing Systems*. Retrieved from <https://web.stanford.edu/~cpiech/bio/papers/deepKnowledgeTracing.pdf>
- Prinz, F., Schlange, T. & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, 10, 712. doi:10.1038/nrd3439-c1
- Rafferty, A. N., & Griffiths, T. L. (2014). Diagnosing algebra understanding via Bayesian inverse planning. *Proceedings of the 7th International Conference on Educational Data Mining*, 351–352.
- Rau, M., Aleven, V., Rummel, N., & Rohrbach, S. (2012). Sense making alone doesn't do it: Fluency matters too! ITS support for robust learning with multiple representations. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Proceedings of the 11th International Conference on Intelligent Tutoring Systems* (pp. 174–184). Berlin, Germany: Springer.
- Razzaq, L., & Heffernan, N. (2010). Hints: Is it better to give or wait to be asked? In V. Aleven, J. Kay, & J. Mostow (Eds.), *Proceedings of the 10th International Conference on Intelligent Tutoring Systems* (Part I, pp. 349–358). Berlin, Germany: Springer.
- Razzaq, L., Heffernan, N. T., & Lindeman, R. W. (2007). What level of tutor interaction is best? In R. Luckin, K. R. Koedinger, & J. Greer (Eds.), *Proceedings of the 13th Conference on Artificial Intelligence in Education* (pp. 222–229). Amsterdam, The Netherlands: IOS Press.
- Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, 15(3), 243.
- Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5), 527–535.
- Rochelle, J., Feng, M., Murphy, R. & Mason, C. (2016). Online Mathematics Homework Increases Student Achievement. *AERA OPEN*. October-December 2016, Vol. 2, No. 4, pp. 1–12. doi: 10.1177/2332858416673968
- Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1(3): 181–210.
- Roll, I., Holmes, N. G., Day, J., & Bonn, D. (2012). Evaluating metacognitive scaffolding in guided invention activities. *Instructional Science*, 40(4): 691–710.
- Shrestha, P., Wei, X., Maharjan, A., Razzaq, L., Heffernan, N. T., & Heffernan, C. (2009). Are worked examples an effective feedback mechanism during problem solving? In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1294–1299). Austin, TX: Cognitive Science Society.
- Singh, R., Saleem, M., Pradhan, P., Heffernan, C., Heffernan, N., Razzaq, L., . . . Mulchay, C. (2011). Feedback during web-based homework: The role of hints. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Proceedings of the 15th International Conference on Artificial Intelligence in Education Conference* (pp. 328–336). Berlin, Germany: Springer.
- Soffer, D., Das, V., Pellegrino, G., Goldman, S., Heffernan, N., Heffernan, C., & Dietz, K. (2014). *Improving long-term retention of mathematical knowledge through automatic reassessment and relearning*. Poster presented at the Division C—Learning and Instruction / Section 1c: Mathematics. American Educational Research Association Conference. Retrieved from <https://drive.google.com/file/d/0B2X0QD6q79ZJX2hkdFNIZ2tWVXlsOFg1Rk1QcUhZS0MxYkh3/edit>
- Stamper, J., Eagle, M., Barnes, T., & Croy, M. (2013). Experimental evaluation of automatic hint generation for a logic tutor. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Proceedings of the 15th International Conference on Artificial Intelligence in Education Conference* (pp. 345–352). Berlin, Germany: Springer.

- Storm, B. C., Bjork, E. L., Bjork, R. A., & Nestojko, J. F. (2006). Is retrieval success a necessary condition for retrieval-induced forgetting? *Psychonomic Bulletin & Review*, 13, 1023–1027.
- Sweller, J., Kirschner, P. A., & Clark, R. E. (2007). Why minimally guided teaching techniques do not work: A reply to commentaries. *Educational Psychologist*, 42(2), 115–121.
- Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences*, 110(16), 6313–6317.
- U.S. Department of Education. (2010). *Transforming American education: Learning powered by technology: The national educational technology plan*. Washington, DC: Office of Educational Technology.
- Walker, E., Rummel, N., & Koedinger, K. R. (2011). Designing automated adaptive support to improve student helping behaviors in a peer tutoring activity. *International Journal of Computer-Supported Collaborative Learning*, 6(2), 279–306.
- Walkington, C. (2013). Using learning technologies to personalize instruction to student interests: The impact of relevant contexts on performance and learning outcomes. *Journal of Educational Psychology*, 105(4), 932–945.
- Wang, Y., & Heffernan, N. (2014). The effect of automatic reassessment and relearning on assessing student long-term knowledge in mathematics. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Proceedings of the 12th International Conference on Intelligent Tutoring Systems* (pp. 490–495). Cham, Switzerland: Springer.
- Whorton, S. (2013). *Can a computer adaptive assessment system determine, better than traditional methods, whether students know mathematics skills?* (Master's thesis). Retrieved from <http://www.wpi.edu/Pubs/ETD/Available/etd-041913-095912/>
- Williams, J. J. (2013). Improving learning in MOOCs with cognitive science. In Pardos & Schneider (Eds.) Part I: Workshop on Massive Open Online Courses (moocshop). In Walker & Looi (Eds.) *Proceedings of the Workshops at the 16th International Conference on Artificial Intelligence in Education*. pp. 49-54. Retrieved from <http://ceur-ws.org/Vol-1009/>
- Williams, J. J., Maldonado, S., Williams, B. A., Rutherford-Quach, S., & Heffernan, N. (2015, March). How can digital online educational resources be used to bridge experimental research and practical applications? Embedding in vivo experiments in “MOOClets.” Report presented at the Spring 2015 conference of the Society for Research on Educational Effectiveness, Washington, DC. Retrieved from <https://eric.ed.gov/?id=ED562268>
- Williams, J. J., Ostrow, K., Xiong, X., Glassman, E., Kim, J., Maldonado, S. G., . . . Heffernan, N. (2015). Using and designing platforms for in vivo educational experiments. In B. Woolf, D. M. Russell, & G. Kiczales (Eds.), *Proceedings of the Second ACM Conference on Learning @ Scale* (pp. 409–412). New York, NY: ACM.
- Williams, J. J., Paunesku, D., Haley, B., & Sohl-Dickstein, J. (2013). *Measurably increasing motivation in MOOCs*. In Proceedings of the 1st Workshop on Massive Open Online Courses at the 16th Annual Conference on Artificial Intelligence in Education. Memphis, TN., July 2013. Retrieved from <https://sites.google.com/site/moocshop/home/moocshop1>
- Williams, J. J., & Williams, B. (2013). *Using randomized experiments as a methodological and conceptual tool for improving the design of online learning environments* (Working Paper). Retrieved from the Social Science Research Network: <http://ssrn.com/abstract=2535556>
- Xiong, X., & Beck, J. E. (2014). A study of exploring different schedules of spacing and retrieval interval on mathematics skills in ITS environment. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Proceedings of the 12th International Conferences on Intelligent Tutoring Systems* (pp. 504–509). Cham, Switzerland: Springer.

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 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 on 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 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 UC 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,” in *ACM Learning at Scale*, and “Revising Learner Misconceptions without Feedback: Prompting for Reflection on Anomalous Facts,” in *Computer-Human Interaction*.