THE HAMILTON PROJECT

POLICY MEMO 2016-01 | MARCH 2016

Learning What Works in Educational Technology with a Case Study of EDUSTAR

Aaron K. Chatterji and Benjamin F. Jones

MISSION STATEMENT

The Hamilton Project seeks to advance America's promise of opportunity, prosperity, and growth.

We believe that today's increasingly competitive global economy demands public policy ideas commensurate with the challenges of the 21st Century. The Project's economic strategy reflects a judgment that long-term prosperity is best achieved by fostering economic growth and broad participation in that growth, by enhancing individual economic security, and by embracing a role for effective government in making needed public investments.

Our strategy calls for combining public investment, a secure social safety net, and fiscal discipline. In that framework, the Project puts forward innovative proposals from leading economic thinkers — based on credible evidence and experience, not ideology or doctrine — to introduce new and effective policy options into the national debate.

The Project is named after Alexander Hamilton, the nation's first Treasury Secretary, who laid the foundation for the modern American economy. Hamilton stood for sound fiscal policy, believed that broad-based opportunity for advancement would drive American economic growth, and recognized that "prudent aids and encouragements on the part of government" are necessary to enhance and guide market forces. The guiding principles of the Project remain consistent with these views.

THE HAMILTON PROJECT

THE HAMILTON PROJECT

Learning What Works in Educational Technology with a Case Study of EDUSTAR

Aaron K. Chatterji
Duke University

Benjamin F. Jones
Northwestern University

MARCH 2016

NOTE: This policy memo is a proposal from the author(s). As emphasized in The Hamilton Project's original strategy paper, the Project was designed in part to provide a forum for leading thinkers across the nation to put forward innovative and potentially important economic policy ideas that share the Project's broad goals of promoting economic growth, broad-based participation in growth, and economic security. The author(s) are invited to express their own ideas in policy memos, whether or not the Project's staff or advisory council agrees with the specific proposals. This policy memo is offered in that spirit.

BROOKINGS

Abstract

Despite much fanfare, new technologies have yet to fundamentally advance student outcomes in K–12 schools or other educational settings. We believe that the system that supports the development and dissemination of educational technology tools is falling short. The key missing ingredient is rigorous evaluation. No one knows what works and for whom. This policy memo articulates general principles that should guide the evaluation of educational technology; these evaluations have the promise to fill in critical information gaps and leverage the potential of new technologies to improve learning. We also present a case study of a new platform, EDUSTAR, conceived by the authors and implemented with a national nonprofit organization. The results from the platform pilot examples reveal several lessons for the future of educational technology.

Table of Contents

CHAPTER 1. Introduction

Technological advances are creating large opportunities in education. In the same way that educators have long employed printed textbooks, blackboards, exercise sheets, and other teaching tools, modern teachers have at hand an expanding toolkit with a new generation of digital learning activities (DLAs; e.g., multimedia exercises, instructional videos, educational games, etc.). The spread of computing devices and the Internet have widened access to these tools, both in schools and at home. Among the many potential transformative benefits of new educational technologies, observers point to distance learning applications, where effective training may become increasingly accessible despite local resource limitations. Personalized learning activities are also seen as particularly promising; in these activities, individuals with different skill levels, learning styles, or learning abilities might be closely matched to the specific products that are most effective for them.

But despite considerable promise, we argue that the current infrastructure to support the adoption and dissemination of educational technologies is inadequate. Teachers, parents, schools, and students have no way to know "what works" and "for whom," creating a virtual fog of mobile apps, videos, and other digital content. With no convincing and cost-effective way to ascertain effectiveness, school district administrators, teachers, parents, and others are right to be skeptical in adopting new technology. Moreover, even when adoption occurs, unproven educational technologies could hamper or even defeat efforts to raise educational outcomes. Innovators and entrepreneurs are also hindered in this environment. When impact is unmeasured, those organizations with the most effective products cannot demonstrate their advantage, limiting both their customer base and the incentive to create new products in the first place (Berger and Stevenson 2007). In the absence of solid evidence, students are far more likely to be presented with technology that creates little or no benefit and is not worth the time or money.

In a 2012 Hamilton Project paper, "Harnessing Technology to Improve K–12 Education," we called for the creation of an Internet-based, educational technology evaluation platform to address these issues. We argued that rigorous and wide-scale evaluation of specific learning activities could help uncover what works and encourage the spread of increasingly effective technologies. Moreover, we argued that such evidence could feasibly be generated in a rapid and low-cost manner. In essence, the platform could extend methods into the educational technology sector that many of the world's leading digital companies (e.g., Google and Amazon) employ every day: rigorous evaluations of their offerings across their user base.

Over the past three years we have launched such a platform, focusing on DLAs in primary and secondary education. Building on lessons learned and looking toward the future, this policy memo first discusses a set of five principles that we believe apply to diverse contexts where DLAs can be effectively utilized, including primary and secondary education, higher education, vocational education, and workplace training. This policy memo then considers a case study, describing the development of the EDUSTAR platform and its application to primary and secondary schooling. We believe the informed adoption of technology can ultimately help address core challenges facing American education: raising the skills of the American workforce, reinvigorating education as a foundation for individual opportunity, and increasing the return on investment.

CHAPTER 2. Learning What Works

The effective use of new technologies in education will be greatly assisted by systematic evaluation of what works. Parents, teachers, and students currently choose DLAs without any objective evidence of whether they will achieve their intent: to help users develop mastery of specific concepts. While user ratings, direct-to-consumer marketing, and measures of user engagement can all be informative, without a mechanism to determine the actual effectiveness of new technologies, new educational products will not necessarily raise outcomes—and may even worsen them.[1] In this chapter, we define five key design principles to solve the evaluation challenge in a systematic fashion, thereby leveraging the potential of educational technology.

PRINCIPLE 1: RCTs are essential means for the rigorous evaluation of learning tools.

Randomized-controlled trials (RCTs) have become the essential approach for evaluating innovations in many sectors. For example, in medical innovation RCTs are generally required for new drug approval, and the results of RCTs already hold a privileged position in assessing educational interventions.[2] RCTs are considered the gold standard of evidence because they move beyond potentially misleading correlations to determine causative effects (Manzi 2012). By randomly assigning individuals to "treatment" and "control" groups that provide different interventions, the results of the two interventions can be compared to determine relative effectiveness.

Such rigorous evidence is especially important in the educational context, where the effectiveness of a tool may not otherwise be obvious and where existing opinions will vary. For example, while RCTs have shown that certain mathematics training tools substantially improve student outcomes (Banerjee et al. 2007; Barrow, Markman, and Rouse 2009; Wang and Woodworth 2011), a popular and widely adopted reading program had no positive effect on students' reading (Rouse and Krueger 2004). We discuss additional examples of nonobvious results from our own work in section 3.C. These examples demonstrate a common lesson from RCTs: the effectiveness of specific interventions is often not what one expects. These types of discrepancies further underscore the importance of rigorous evidence for decision-making.

PRINCIPLE 2: Evaluations of learning technologies must be rapid and continuous.

Technologies progress quickly, especially information and computer technologies. New DLAs are being introduced at high rates, and existing content is regularly updated. To provide value in the constantly evolving educational technology sector, it is therefore important that RCTs be conducted rapidly and on an ongoing basis. Rapid and continuous evaluation will enable quick determinations about which among available products stack up best so the best learning technologies can be deployed more widely to students. Rapid evaluation will also provide timely feedback to software developers on how to improve their learning tools.

PRINCIPLE 3: Evaluation systems built on existing, user-friendly content platforms have substantial advantages.

When an evaluation like an RCT is undertaken from scratch, it tends to be expensive in both dollars and time. For any one-off evaluation, the researcher must search for willing partners, negotiate with schools or other educational organizations, set up and undertake the intervention with students, gather data, and invest in analysis. While the benefits of RCTs are large, implementation costs tend to limit their use in practice. Moreover, the long duration from conception to results makes it difficult to provide rapid feedback in the evolving technology landscape, thus undermining Principle 2.

We can do substantially better by building evaluation platforms, rather than one-off evaluation projects. Results will come far more quickly and cheaply through a platform that can be set up once and then used to run many evaluations. Additionally, and fundamentally, building the evaluation system on top of existing content platforms makes it far more straightforward to recruit participants for product evaluations. In particular, an existing, user-friendly platform—one that already attracts many users—can allow for the rapid and rigorous evaluation of learning tools.

Examples from technology companies such as Google and Amazon are illustrative. These companies are running

hundreds of RCTs daily through their standard user interfaces. Here the user engages a standard activity that is of value to the individual (e.g., an Internet search), and is randomly assigned to one of multiple versions of that activity to assess its relative success. By working with existing users and their needs, as opposed to recruiting people solely for the purpose of an evaluation, the costs (in dollars and time) for running RCTs drops dramatically. In the education context, numerous existing platforms exist that provide learning content (e.g., Khan Academy, PowerMyLearning, and many others). Running RCTs in tandem with these platforms creates large opportunities to learn what works.

PRINCIPLE 4: Scale unlocks transformative opportunities.

To provide information in a systematic manner that can substantially improve the use of educational technology, significant scale is required. Most directly, a large user base is necessary if one wishes to test large numbers of products. For example, in K–12 education the library of DLAs is already large, extending across numerous specific skills and teaching objectives, with standards that differ by grade (see, e.g., the Common Core State Standards); evaluating a substantial portion of available learning activities thus requires very large scale. Second, scale facilitates the capacity to undertake rapid evaluations and keep up with evolving technology, per Principle 2. Third, building and refining the evidence on learning technologies is best done across a large, diverse set of participants. With a large and diverse set of participants per product trial, a given RCT can precisely determine any differential impacts across student subgroups. The evaluation system can ask not only what works on average, but also what works for whom, thus advancing opportunities for personalized learning. In sum, the greater the scale of the platform, the more informative and potentially transformative it can be.

PRINCIPLE 5: The evaluator must be trusted and report the results transparently.

The results of the evaluations will be most impactful if they come from a trusted source. The public may view with diminished confidence product tests reported or performed by those with a private stake in the outcome. Third-party and nonprofit organizations may therefore play key roles in conducting the evaluations and in reporting the results. Organizations like Consumer Reports can provide useful models.

A related design principle is transparency. Transparent reporting of the methodologies and findings will facilitate trust. Furthermore, it befits evaluators to report findings using plain language, making straightforward summaries (as well as more-detailed information about methodologies and results) readily available. Transparent reporting will help the product evaluations reach a wide audience and expand their impact.

CHAPTER 3. The Case of EDUSTAR

To illustrate these principles in action, we describe the case of EDUSTAR, a new platform for evaluating educational technology. In a 2012 Hamilton Project paper (Chatterji and Jones 2012), we called for the creation of the EDUSTAR platform to conduct rapid RCTs to determine which technologies work best for which kinds of students, and then to disseminate the results through a public Web site. Over the past three years we have built the initial platform and begun to use evidence in a rapid feedback cycle to improve the impact of educational technology on student learning. The platform has since undertaken numerous trials of DLAs, demonstrating that EDUSTAR can produce important insights into "what works." The early results show that the platform holds substantial promise for enhancing student learning and advancing our understanding of educational technologies. We now describe the EDUSTAR platform and its use, lessons learned, and then propose an agenda to further improve the platform's impact on student learning.

A. What Is EDUSTAR?

EDUSTAR is a Web-based program that evaluates the results of DLAs. These DLAs include short instructional videos or tutorials, interactive exercises, and learning games. In line with Principles 1 and 2, EDUSTAR conducts rapid RCTs, the evidentiary gold standard of evaluation for assessing what works. EDUSTAR is currently moving to a scalable platform, designed to undertake many evaluations of individual DLAs.

EDUSTAR is embedded within the Web site of PowerMyLearning, a national nonprofit organization. They operate a free online platform called PowerMyLearning Connect, which provides a library of DLAs offered by a wide variety of providers and acts as a central access point for teachers, students, and others interested in utilizing the products. In line with Principle 3, this technology platform provides a large, existing user base on which to build and scale an evaluation system. The EDUSTAR pilots were conducted at PowerMyLearning's partner schools, establishing a direct link for feedback from teachers and students.

EDUSTAR's primary goal is to determine the relative effectiveness of distinct DLAs and to communicate those findings to teachers, parents, and students so that they can make informed decisions about which DLAs to use (Chatterji and Jones 2012). From its inception EDUSTAR aimed to provide rigorous, continuous, low-cost, and rapid evidence about the efficacy of DLAs. In line with the five principles of chapter 2, when brought to scale EDUSTAR will conduct large numbers of RCTs to evaluate many DLAs, and share these results publicly. The target audience for the results includes anyone who uses or may use DLAs, including students, teachers, parents, principals, and school district administrators.

In this role EDUSTAR provides two critical functions: First, it provides a means for the best learning activities to diffuse. Both in schools and at home, teachers and parents can direct students to the learning activities that have the most impact. Second, by providing rigorous and third-party evaluations, the platform provides a path to the marketplace for the developers of high-quality learning tools, who can now signal the quality of their products at low cost. This second role can help overcome large barriers to entry for innovators in the learning technology space and thereby accelerate both the diffusion of effective activities and the creation of new activities (Berger and Stevenson 2007; Chatterji and Jones 2012).

B. Pilot Testing EDUSTAR

We successfully piloted the first RCTs in classrooms during the 2013–14 school year and expanded to 21 product tests by the conclusion of the 2014–15 school year. In the current 2015–16 school year, the system is conducting an additional 56 product tests, marking the beginning of the scaling phase, which we will discuss in section 3.E. The EDUSTAR pilots have all focused on DLAs that target mathematics skills taught in grades 6 through 8.

Students access the EDUSTAR system through any Web browser on a computer with Internet connectivity. For the EDUSTAR pilot, students log in to the PowerMyLearning Connect platform and click on a learning exercise, which includes an RCT. In the initial pilots, the students proceeded through the following three steps:

1. The student answers a set of either six or ten multiple-choice questions. All of these pre-exercise questions target a specific skill (e.g., multiplying fractions or finding equal ratios) that the DLA is intended to address. The pre-exercise questions measure a baseline of knowledge for each student.

2. Each student is randomly assigned to a treatment condition. Depending on the trial, assignment could be to one of several DLAs designed to teach the same specific

skill, or to either a selected DLA or the control group with no DLA. Students assigned to a DLA can choose to opt out of the activity at any point.[3]

3. The student answers a second set of six or ten post-exercise multiple-choice questions that are similar (but not identical) to the pre-exercise questions.[4]

For each user, the platform captured the responses to each question, the learning activity used, and the time spent at each step.

C. Rapid RCTs: What Works?

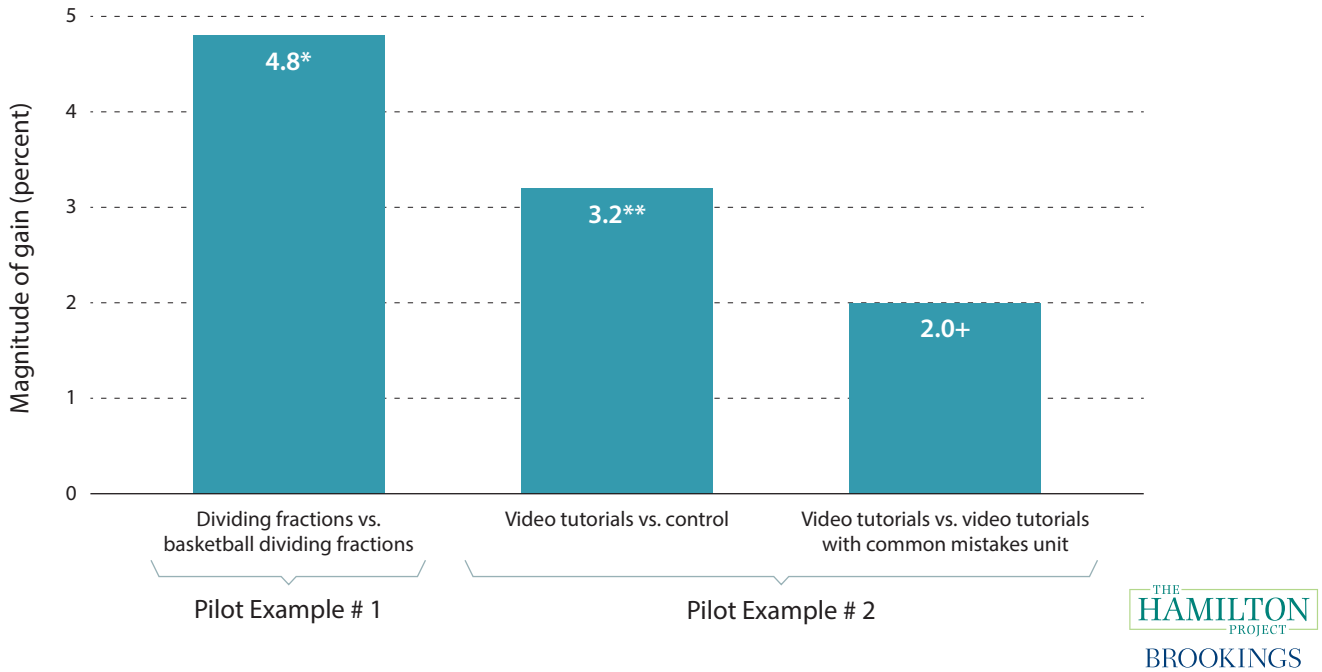
As an example of EDUSTAR's primary function of deploying randomized control trials (RCTs), we describe two pilot examples comparing digital learning activities (DLAs) targeting the same skill. The measure of success is the change in score for a given student, comparing the post-exercise result to the pre-exercise result. The change in score is transparently analyzed as the gain (in percentage points) in questions that are answered correctly.

The first RCT example compares two DLAs intended to help sixth-grade students learn to divide fractions. In this trial, 544 students participated and were randomly assigned to either the DLA "Dividing Fractions" or the DLA "Basketball Dividing Fractions." In the head-to-head trial, the gains were significantly greater among those assigned to "Dividing Fractions."[5] Interestingly, this advantage came despite the fact that students assigned to the other DLA, "Basketball Dividing Fractions," spent 4.5 minutes longer, on average, engaged with their assigned activity. In other words, the basketball program, which has game-like features, captured student interest much longer but taught them less.[6] This suggests that "Dividing Fractions" is a higher-quality activity on at least two dimensions: it teaches more in less time.[7]

The second RCT example investigated two variations of short instructional videos produced by digital learning developer LearnZillion. The instructional videos targeted mathematics skills by running three RCTS: one for each of the sixth, seventh, and eighth grades with a total of 1,457 participants.[8] These RCTs were intended to test something of great interest

FIGURE 1.

EDUSTAR Results for Two Pilot Examples



Note: ** and * indicate p-values of 0.01 and 0.05, respectively. + indicates a p-value of 0.101. We use the entire sample of students exposed to the treatment, with robust standard errors. There are 544 observations for pilot example #1 and 2,417 for pilot example #2 (1,457 in the video tutorial vs. control case, and 960 in the video tutorial with common mistakes vs. the video tutorial without common mistakes case). Restricting to students who fully complete the pre-exercise and post-exercise quizzes and/or those who spent relatively more time on the DLA continue to show gains in the 4–5 percent range with statistical significance at the 90, 95, or 99 percent levels, depending on specifications (whether or not school fixed effects and clustering by school are used). In pilot example #2 we pool the video tutorials across grades that do not include the common mistakes unit.

to teachers and product developers: What happens when students are forewarned about common mistakes? In each RCT, students were randomly assigned one of two video tutorials—the first treatment group watched a four-to-five-minute video tutorial designed to teach a math skill, while the second treatment group watched an otherwise identical video that included an additional 20 second section forewarning students about mistakes that students typically make. In order to test the effect of the instructional video compared to no intervention, a third group watched neither video but completed the pre and post multiple-choice exercises.[9]

The aggregated results from these RCTs are displayed in figure 1. Students who received the video tutorial *without* being forewarned of common mistakes saw the largest gains in learning, with gains in their test scores compared both to the control group who viewed no video and to those who viewed the video including the common mistakes unit. This pattern was the same across all three RCTs. While this set of RCTs demonstrates that the video without the common mistakes unit was more effective on average, it also provides suggestive evidence that forewarning students about mistakes may be counterproductive. In other words, while there may be a good time to provide students with information about common mistakes, these results suggest that doing so when introducing a new topic may disrupt learning.

D. Lessons from the Pilots

During and following implementation of the EDUSTAR pilots, comments from teachers were extremely valuable for improving the user experience. Data from the pilots themselves have also provided rich information about how students and teachers use the platform during an evaluation. We have incorporated feedback when possible to improve the tool while retaining the scientific rigor of the RCTs.

From the beginning of this effort, we have taken steps to ensure that EDUSTAR is an enhancement to existing classroom pedagogy and does not disrupt or hamper learning. When we originated our 2012 proposal, we discussed this concern with a variety of individual teachers and representatives from teachers' unions and concluded that we could design EDUSTAR to be an asset to teachers in the classroom as they planned and implemented their lessons. Having now run pilots with actual teachers and students, we feel even more confident that EDUSTAR will have a positive impact in the classroom. Fundamentally, and in line with Principle 3, by building the system on top of an existing platform that teachers were already using to access DLAs, EDUSTAR was able to align more naturally with classroom activities. Teachers varied in the ways that they integrated EDUSTAR into their lessons, sometimes using it as a diagnostic pre-assessment, sometimes using it as an evaluative post-assessment, and sometimes incorporating its use into regular "computer sessions" where students worked independently to complete various exercises and assessments. In all cases, students were engaged in content deemed by the teacher to be relevant to the skills they were learning in class.

Feedback from teachers has also been valuable in fine-tuning the platform design to better align with classroom goals and teacher interests. For example, teachers expressed substantial interest in automated feedback that provides snapshots of their students' mastery of a particular formative skill. The multiple-choice questions thus turn out to serve a dual purpose: not only do they serve to evaluate the DLA, but they also provide teachers with a view of mastery in their own classrooms. Without this information from EDUSTAR, the teacher would have to gather the information through an alternative assessment such as written exercises, through a purchased software product, or through other means. The platform is now designed to provide these automated snapshots to teachers.

As another example, providing a placebo (i.e., randomly assigning some children to receive no DLA) was less attractive to many teachers than designs where every student receives at least one type of treatment. Furthermore, some teachers prefer that each student ultimately participate in the same set of DLAs. Taking these preferences into account, we were able to redesign the experimental protocol so that all students ultimately experience both DLAs being tested, but in a random order.[10]

The current version of the platform implemented in the 2015–16 school year takes these design features into account. Students now engage in learning "missions" that are selected by the teacher at the appropriate time to meet the class's pedagogical needs. From the user's perspective, these missions consist of a sequence of practice exercises and DLAs aimed at mastering particular formative skills. Teachers can also view reports that provide feedback as their students progress through various activities.

E. The Next Phase: EDUSTAR 2.0

The EDUSTAR pilots have provided several insights about the design features that can make such an evaluation system function effectively, and these insights are incorporated in the current implementation of the platform. We remain convinced that evaluation of educational technology must be rigorous (preferably using RCTs), rapid, and continuous, and ideally that it be built on an existing platform to ease implementation. EDUSTAR, as currently designed, is aligned with these principles. The critical next step is to scale the platform to acquire more users and test more products. In doing so, we will also better enable EDUSTAR to become a trusted and transparent evaluator of educational technology tools. By scaling the

platform, we can also move closer to facilitating personalized learning, which represents a significant opportunity to improve student outcomes. We now describe the next phase of development for EDUSTAR in detail.

1. Expanding the User Base

At its current scale, EDUSTAR can already be used for prototyping and for some academic research on learning, as further described in section 3.F.1. By the end of the 2015–16 school year, the EDUSTAR pilots will have tested, since inception, 77 learning activities, engaging more than 10,000 students in more than 40 different schools.

While achieving these milestones is important to validate the platform's effectiveness in running rapid RCTs and aligning with user needs, meeting EDUSTAR's transformative vision will require much larger scale. For example, a platform that performed 100 RCTs per week could provide systematic, up-to-date information to consumers across a large library of DLAs, by subject area and grade level. At a scale of 100 RCTs per week, with 500 students per trial, the platform would require approximately 2,000 participating classrooms with students spending an hour of computer time per week, or 8,000 classrooms with students spending an hour per month. This scope can be easily achieved by partnering with a single large U.S. school district, such as New York City or Miami-Dade County, though the medium-term plan is to expand access to the platform to students in multiple medium-sized and large school districts. Extending EDUSTAR across schools with substantially different characteristics will encourage rigorous assessments of the generalizability of particular DLAs across school settings. The platform can also be scaled outside of classrooms, and future iterations of the platform may experiment with allowing the Web site's large number of home users to participate in RCTs using the same user interface currently employed in classrooms.

2. Expanding and Automating the Product Evaluations

A key step going forward is to expand the range of DLAs tested to cover a wider range of formative skills and grade levels served by EDUSTAR. At scale, primary and secondary school teachers will be able to engage the Web-based system and choose, at any time, the formative skill they are working on, and enroll the class in an appropriate learning "Mission," where the RCT is being conducted.

Moving forward, EDUSTAR's potential will be leveraged most fruitfully when "non-promising" trials (i.e., where a given activity or versions of activities show little effective difference in student learning after a certain number of classroom implementations) are discontinued relatively quickly while "promising" trials (i.e., where one DLA appears to be particularly promising in early phases of the trial) receive additional investigation. This type of optimization can be largely automated: the next implementations of the platform will mechanize the statistical analysis of RCTs so that the results will be available in tabular and graphical form immediately and without requiring the additional time or staffing costs for analysis.

3. Making Results Useful and Usable

As discussed in chapter 2, models like Consumer Reports suggest the value that independent, nonprofit entities can play as trusted evaluators. Trusted reporting can be further assisted by transparency, which in the EDUSTAR case includes (1) clear statements of the research methods in plain language, (2) standardized presentation of the data analysis underpinning any product trial, and (3) a simple rubric for the results that is easily understood by all users. By building out EDUSTAR's reporting functionality, the product testing results can be seamlessly communicated to teachers, students, parents, and school systems, encouraging more-informed choices about learning tools.

A "star" system (e.g., from one to five stars) is one natural candidate for an easily digestible rubric, characterizing the impact of the DLA, with more-detailed statistical analysis also made available to interested users. Moreover, the automation of these systems will greatly facilitate timely reporting. Reporting will indicate whether products have heterogeneous treatment effects (e.g., when a product works especially well for students with lower initial skill levels). This information should be presented in a transparent manner, again using the star system, with additional, detailed statistics available for interested users. Such reporting can better leverage the capacity for personalized learning, as we discuss further in section 3.E.4.

In addition to being trusted, transparent, and timely, it is important that the results of the public product tests be widely accessible. A public Web site would be the natural vehicle. An open question in the EDUSTAR case is whether to add reporting functionality directly to the PowerMyLearning platform, to use an existing third-party entity in the educational technology space to provide reporting, or to pursue both options.

4. Meeting the Promise of Personalized Learning

Greater scale will allow EDUSTAR to provide personalized learning recommendations to students. As we amass more results from the rapid RCTs, we may find that different activities are more effective for different students. In particular, the initial platform implementation allows comparisons across students with different initial skill levels at a given task, as

measured by the pre-exercise questions. Currently the RCTs can distinguish whether a given alternative is differentially effective for students with low, medium, or high initial skill levels. Starting in 2015–16, the platform also collects data about student metacognition—i.e., the student's self-perception of her skill level at a given task—allowing further analysis and increased matching of students to the DLA that is predicted to be most effective for them. Additionally, the platform could be extended to incorporate other student characteristics that could be helpful in matching to the most effective DLA option, such as students' interests or learning styles.[11]

F. Other Benefits of EDUSTAR

In addition to the students, teachers, parents, and school administrators who will use EDUSTAR to navigate through options to find the most effective DLA, there are two other constituencies that will find substantial value from EDUSTAR: DLA content developers and researchers in education and learning sciences. There are significant differences in how they will utilize the platform to meet currently unmet needs. We now consider the benefits of EDUSTAR for these groups, and the implications for scaling and funding models.

1. Content Developers

The EDUSTAR platform can provide insights to content developers as they prototype new activities or improve the effectiveness of existing tools. An example of prototyping can be seen in the second pilot example comparing two versions of LearnZillion's video tutorial, which found greater impacts in the version that omitted the common mistakes unit. This type of "Version 1 versus Version 2" testing can provide tremendous value to the content developer and encourages a "hypothesis-driven" approach to development. Rather than build a product with all the desired features from the very beginning, this method emphasizes building a prototype to test with users, learning from the RCTs about what features are most effective, and iterating toward the highest-quality eventual product. The lower the costs of testing new ideas and the more precise the feedback, the more effective this process becomes. This prototyping function is also expected to accelerate effective innovation, thus improving the quality of learning activities that reach the marketplace. At scale, EDUSTAR can additionally provide developers with a ready-built, rapid, rigorous, and low-cost means for improving the design of DLAs, leading to higher-quality products while providing a new entry point for aspiring educational technology entrepreneurs.

2. Education Researchers

Another important application for the platform is to assess different mechanisms and processes for learning. By varying the content shown to students, researchers can ask and answer questions about fundamental learning techniques and pedagogical strategies. By building a large platform to assess educational technologies, we also can create a valuable opportunity to learn what works in education more broadly. For example, the multimedia and interactive dimensions of DLAs often permit multifaceted assessment of cognition and skill development, with potentially generalizable lessons about the effectiveness of various teaching approaches across visual, auditory, and text-based communications and interactions. The EDUSTAR platform can help education researchers investigate fundamental questions with sample sizes and speeds that heretofore would not have been possible. For example, do audio or text-based explanations have a larger impact? To what extent do rewards or games influence student engagement and learning? What is the optimal length of video instruction, and how does that vary by student age? What is the right balance between listening and practicing? What presentation approaches hold students' attention most effectively? The LearnZillion pilot, for example, provides a promising initial demonstration, where the broader teaching mechanism of interest is whether showing students common mistakes when introducing a topic helps or hinders their learning. The capability to run rapid RCTs through the EDUSTAR platform will provide researchers a means to understand learning mechanisms that is considerably more rapid and cost-effective. Much as developments in gene sequencers have rapidly lowered the cost of genetic research, and platforms like Amazon's Mechanical Turk have allowed social scientists to run larger studies at lower costs, we believe that EDUSTAR at scale could provide new opportunities to education researchers.

Our vision is to continue providing EDUSTAR as a free resource to schools and individual home users. The initial EDUSTAR platform has been developed with philanthropic support, and will continue to require additional support in the near term. In the long term, though, revenue streams from software developers or researchers may make the platform self-sustaining. We predict that, as EDUSTAR expands in scope and automates more of its processing, the cost of RCTs can fall to less than $1 per participant.

CHAPTER 4. Conclusion

New technologies hold enormous promise to transform education and raise outcomes. With the United States facing ongoing performance challenges in educating the next generation, the smart application of new technologies provides a promising opportunity to improve our K–12 education system. Around the world and in underserved communities in the United States, there are even greater opportunities to enhance access to content and learning outcomes for students who often lack financial resources, quality instruction, and supporting infrastructure. However, without transparency around what works and for whom, schools, teachers, parents, and students themselves do not have the information necessary to choose the right technologies amidst many competing options. EDUSTAR provides a path forward. By using RCTs continuously, rapidly, and at scale in a user-friendly framework, EDUSTAR can reveal which educational content is most effective—and whether that varies by student characteristics and learning contexts. By developing a reputation as a trusted evaluator that provides transparent results, EDUSTAR can help in disseminating the most effective learning activities to the students who will benefit from them the most. The platform also has the potential to help content developers and education scholars understand why specific tools work, with applications to settings beyond K–12 such as workplace training and higher education. Armed with this information, we can provide teachers and their students with the very best educational technology available, target the right content to the right student at the right time, and encourage effective innovation.

Authors

Aaron K. Chatterji

Associate Professor, Fuqua School of Business, Duke University

Aaron Chatterji, PhD, is an Associate Professor (with tenure) of Business and Public Policy at Duke University's Fuqua School of Business and the Sanford School of Public Policy. He previously served as a Senior Economist at the White House Council of Economic Advisers where he worked on a wide range of policies relating to entrepreneurship, innovation, infrastructure and economic growth. For the 2014-2015 academic year, Aaron was on leave as a visiting Associate Professor at Harvard Business School.

Chatterji's research and teaching investigate some of the most important forces shaping our global economy and society: entrepreneurship, innovation, and the expanding social mission of business. He has received several awards, including an inaugural Junior Faculty Fellowship from the Kauffman Foundation to recognize his work as a leading scholar in entrepreneurship, the Rising Star award from the Aspen Institute for his contributions to understanding the intersection of business and public policy, and the Strategic Management Society Emerging Scholar award for his research in strategy.

His research has been published in leading academic journals and been cited by *The New York Times*, CNN, T*he Wall Street Journal*, and *The Economist*. He has authored several op-ed pieces in top newspapers, including *The New York Times* and *The Wall Street Journal*, appeared on national TV and radio, and was recently profiled in *The Financial Times* and *Fortune*.

Chatterji is a term member of the Council on Foreign Relations and previously worked as a financial analyst at Goldman Sachs. He received his PhD from the Haas School of Business at the University of California at Berkeley and his BA in Economics from Cornell University.

Benjamin F. Jones

Gordon and Llura Gund Family Professor of Entrepreneurship, a Professor of Strategy, and Faculty Director, Kellogg Innovation and Entrepreneurship Initiative

Benjamin Jones is the Gordon and Llura Gund Family Professor of Entrepreneurship, a Professor of Strategy, and the faculty director of the Kellogg Innovation and Entrepreneurship Initiative. An economist by training, his research focuses largely on innovation and creativity, with recent work investigating the role of teamwork in innovation and the relationship between age and invention. Professor Jones also studies global economic development, including the roles of education, climate, and national leadership in explaining the wealth and poverty of nations. His research has appeared in journals such as *Science* and *The Quarterly Journal of Economics* and has been profiled in media outlets such as *The Wall Street Journal, The Economist,* and *The New Yorker.*

Professor Jones served in 2010-2011 as the senior economist for macroeconomics for the White House Council of Economic Advisers and earlier served in the U.S. Department of the Treasury. In 2011, he was awarded the Stanley Reiter Best Paper Award for the best academic article written by a Kellogg faculty member in the prior four years.

Endnotes

1. Goolsbee and Guryan (2006) find that computers alone do not have positive effects, which points to the importance of content and understanding what specific content actually works.
2. See the U.S. Department of Education's "What Works Clearinghouse" at http://ies.ed.gov/ncee/wwc/.
3. DLAs are typically three to five minutes long, but in principle could be any length.
4. As discussed in Chatterji and Jones (2012), data privacy is a major concern and EDUSTAR is designed to maintain data privacy and confidentiality. All EDUSTAR data analysis is conducted using anonymized student data, and no personally identifiable data are released in any form.
5. The 4.8 percentage point gain in the score is equivalent to 18 percent of one standard deviation in the pre-exercise quiz scores across the students engaged in this RCT.
6. This result is important in its own right, as it shows that assessing programs based on the length of student engagement can be quite misleading for assessing effectiveness in learning.
7. This RCT could provide even more information in a larger trial. Even though "Dividing Fractions" on average outperformed "Basketball Dividing Fractions," there may be some subsets of students for whom the relative impact is reversed and greater gains would be received from the basketball DLA. Knowing this, and thus being able to predict which DLA is likely to best serve a given student, will enable construction of curricula that are more customized moving forward. We discuss these possibilities in chapter 3.E.4, with regard to opportunities to enhance personalized learning.
8. Each grade-specific video taught a different concept: the grade 6 videos taught dividing by fractions, the grade 7 video taught integer addition using a number line, and the grade 8 video taught multiplying exponential expressions. In each case the length of the video tutorial was between 4 and 5 minutes.
9. Once the RCT was completed, the students in the control group viewed the video that included the common mistakes portion.
10. An additional insight from the pilots concerns the appropriate number of pre- and post-exercise questions. More questions per student provide a more precise estimate of the individual student's knowledge, but answering more questions requires additional time—which might be better spent on other activities—and requires more-sustained attention from students, who sometimes opt to speed through longer question sets and resort to guessing. The results of testing different numbers of questions suggest that gains in precision are modest after the first few, partly due to increased guessing. Furthermore, the drop in precision per user can be readily compensated overall by including more users per trial. The 2015–16 iteration of the platform therefore asks three pre-exercise questions and three post-exercise questions; we will continue to experiment to find the optimal number.
11. As noted, all individual student data from using EDUSTAR are anonymized and will remain confidential. The product test results report about DLAs, summarizing performance gains only across large numbers of users. Furthermore, as students remain the primary users and beneficiaries of EDUSTAR, care will be taken to ensure that they are protected from poorly designed learning activities, particularly those still in the design phase. (This will be accomplished in part by screening all DLAs for basic quality before they are implemented in a learning Mission, and in part by discontinuing non-promising trials through the automated steps discussed in chapter 3.E.2.)

References

Banerjee, Abhijit, Shawn Cole, Esther Duflo, and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics* 122 (3): 1235–64.

Barrow, Lisa, Lisa Markman, and Cecilia E. Rouse. 2009. "Technology's Edge: The Educational Benefits of Computer-Aided Instruction." *American Economic Journal Economic Policy* 1 (1): 52–74.

Berger, Larry, and David Stevenson. 2007. "K–12 Entrepreneurship: Slow Entry, Distant Exit." Wireless Generation, New York.

Chatterji, Aaron, and Benjamin Jones. 2012. "Harnessing Technology to Improve K–12 Education." Hamilton Project Discussion Paper 2012-05, The Hamilton Project, The Brookings Institution, Washington, DC.

Goolsbee, Austan, and Jonathan Guryan. 2006. "World Wide Wonder? Measuring the (Non-) Impact of Internet Subsidies to Public Schools." *Education Next* 6 (1): 60–66.

Manzi, Jim. 2012. *Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society*. New York: Basic Books.

Rouse, Cecilia E., and Alan B. Krueger. 2004. "Putting Computerized Instruction to the Test: A Randomized Evaluation of a 'Scientifically Based' Reading Program." *Economics of Education Review* 23 (4): 323–338.

Wang, Haiwen, and Katrina Woodworth. 2011. "Evaluation of Rocketship Education's Use of DreamBox Learning's Online Mathematics Program." SRI International, Menlo Park, CA.

Acknowledgments

We would like to thank Elisabeth Stock, Mark Malaspina, Hussham Khan, and the entire staff of PowerMyLearning for their work developing the EDUSTAR platform, and we are very grateful to the Laura and John Arnold Foundation, the Bloomberg Philanthropies, The Broad Foundation, and the Bill & Melinda Gates Foundation for supporting this effort. We also wish to thank Diane Schanzenbach, Gregory Nantz, and the staff of The Hamilton Project for helpful guidance in developing this policy memo, and to thank Linyi Zhang of Northwestern University for excellent research assistance.

THE HAMILTON PROJECT

ADVISORY COUNCIL

Highlights

Aaron Chatterji of Duke University and Benjamin Jones of Northwestern University introduce a set of five key principles to guide the development of effective evaluation tools for educational technology. They also include an update on EDUSTAR, a Web-based platform for evaluating digital learning activities that they first proposed in their 2012 Hamilton Project paper.

The Principles

Principle 1: Randomized Control Trials (RCTs) are essential means for the rigorous evaluation of learning tools.

Principle 2: Evaluations of learning technologies must be rapid and continuous.

Principle 3: Evaluation systems built on existing, user-friendly content platforms have substantial advantages.

Principle 4: Scale unlocks transformative opportunities

Principle 5: The evaluator must be trusted and report the results transparently.

The Case of EDUSTAR

Since 2012, Chatterji and Jones have partnered with a nonprofit organization to build the initial platform. They have undertaken numerous trials of digital learning activities and have begun working with school districts to scale the platform. Early results from their pilot tests demonstrate that the platform has potential to serve as an important resource for students, parents, teachers, and school administrators looking for effective digital learning activities. Ultimately, they envision that the platform will provide significant value to content developers and education researchers as it becomes widely available.

THE HAMILTON PROJECT

1775 Massachusetts Ave., NW
Washington, DC 20036

(202) 797-6484

BROOKINGS

Printed on recycled paper.

WWW.HAMILTONPROJECT.ORG