The *Better Book* Approach for Education Research and Development

A final peer-reviewed version of this article will appear in:
Teachers College Record
Vol. 123, No. 2
www.tcrecord.org

James W. Stigler,[1] Ji Y. Son,[2] Karen B. Givvin, Adam Blake, Laura Fries, Stacy T. Shaw, Mary C. Tucker

Abstract

This article describes a new approach for education research and development - the *better book* approach - and reports on our initial development and application of the approach in the context of introductory college-level statistics. The *better book* approach leverages recent advances in technology that have made it possible to house learning materials such as textbooks online; to make them interactive (by including, for example, embedded formative assessments); to make them personalized so that different content can be given to different students; and to continuously update them based on data generated by students' interactions with the materials. Using these technologies - and applying the routines and methods of open-source software development and improvement science - we have developed an approach in which a living textbook is continuously improved by a community of researchers, designers/developers, and instructors; and a technology platform (CourseKata.org) to support the collaborative improvement process. This approach overcomes some of the obstacles that have slowed translation of research in the learning sciences into improved educational programs and materials.

[1] University of California, Los Angeles (all authors except as noted)
[2] California State University, Los Angeles

The authors gratefully acknowledge the support of the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation (DRL-1229004) and the California Governor's Office of Planning and Research (contract OPR18115).

Corresponding Author:
James W. Stigler, UCLA Psychology Department, Los Angeles, CA 90095-1563, USA. Email: stigler@ucla.edu.

Applying research to the improvement of education is a notoriously hard thing to do. Despite huge advances in our fundamental understanding of human learning, harvesting the potential of these advances to yield sustainable improvements in student outcomes has proved an elusive goal. More recent attempts to remedy this situation—one thinks of the *What Works Clearinghouse*, and the view of education science embodied in that effort—have had little effect on the overall education landscape. Of course, all of this was predicted by Dewey in 1929, who foresaw a more modest role for scientific research in the improvement of education:

> No conclusion of scientific research can be converted into an immediate rule of educational art. For there is no educational practice whatever which is not highly complex; that is to say, which does not contain many other conditions and factors than are included in the scientific finding ... The value of the science … resides in the enlightenment and guidance it supplies to observation and judgment of actual situations as they arise (pages 19 and 31).

Although Dewey's view has proven accurate up to now, there are some signs, at least to us, that we might be ready to take a major step forward in our ability to successfully apply science to the improvement of education and learning. In this article, we start by reviewing some of the reasons why, up to now, this application of science has proven so difficult. We start with a dose of reality, in other words, reviewing what makes this endeavor hard. Then, based on a realistic understanding of why it's hard, we propose a new approach for applying science to improving education. This new approach does not ignore the realities, but does leverage advances in several fields—technology especially—in a somewhat novel way. Having worked to develop and apply this approach in a preliminary way, we report what we have learned so far about how to carry out this process.

Why It Is Hard

It seems like each new generation of researchers and practitioners needs to discover anew what Dewey observed in 1929. We often hear people remark, in consternation, that given how much we've learned about the science of learning, it's a shame that educators don't apply what we know in the classroom or in the design of educational programs. The fact is, they often do try, yet fail to get the results they expect based on the laboratory science. Here are some of the reasons why.

Teaching is a complex system. Research findings in the learning sciences mostly come from carefully designed and controlled laboratory studies. Individual studies seek to isolate the effects of specific variables, assessing the effect of each on learning outcomes. But education, where we hope research findings will apply, is quite different from the laboratory. Education is a complex system with many interacting parts: teachers, students, content, curriculum, homework, assessments, culture, belief systems, motivations, and so on. Even if a variable makes a reliable difference in the lab, it may have little or no effect in the context of such a complex system. Ideally, we would study learning in real settings over long periods of time; learning in complex domains usually takes weeks, months, or years, and is not well-modelled by a one-hour lab experiment. Yet finding ways to study teaching and learning in more ecologically valid settings is fraught with challenges.

Teaching is highly contextual. Even well-designed curriculum materials will produce wide variation in student outcomes, both across classes and across individuals within classes. Figuring out how to make any set of materials work for all students—and to reduce variation among students to an acceptable range—is largely a problem of implementation: adjusting the program to fit the specific context. This is what skilled teachers do. Yet, though skilled teachers innovate every day, we have no means of capturing what they learn so that it can be shared with others.

Education R&D is highly siloed. Contributors with at least three kinds of expertise are required to produce effective educational programs: researchers, designers/developers, and practitioners on the ground. Currently, these kinds of expertise are represented by three distinct groups of people, each working largely independently of the other. Although there are many reasons for this divide, it is clear that isolation holds us back. Researchers are well-equipped to provide theories and methods for addressing the implementation problems identified by practitioners, and for revising their theories accordingly. Yet, by the time a program is turned over to the practitioners, researchers are no longer involved. Meanwhile, developers have the technical skills to build and deploy learning technologies at scale, but rarely have access to the type of "on the ground" insights needed to guide meaningful improvement on a continuous basis.

Many practitioners attempt to perform all three roles: to develop their own materials and carry out independent research projects within the confines of their classrooms, but by and large these results do not get shared with the broader profession—and sometimes not even with the teacher's own school site. It is an unrealistic expectation that a teacher be able to perform all three professional roles and do their job well.

Teaching is a cultural activity. Cross-cultural research brings us face to face with another incontrovertible fact: not only is teaching a complex system, but it is a cultural activity. It is governed by daily routines that evolve over long periods of time, are supported by widely shared beliefs, and are highly resistant to change (Gallimore, 1996). This makes innovative educational approaches or programs very hard to evaluate. Unless you can implement a program and sustain it at some scale, you cannot even study whether, and how, it is effective (Ostrow, Heffernan, & Williams, 2017). And the more innovative a program is, the less likely it is to fit with existing cultural routines, and thus the more difficult it is to get it up, running, and implemented at scale.

Summary. All of these factors together make it very hard for science as we know it to yield reliable improvements in student outcomes. The most innovative programs involve a redesign of the system, not just one component part; massive effort is often required just to get such programs off the ground.

What Has Changed

With this dire assessment before us, we nevertheless find ourselves with renewed optimism that we can exceed Dewey's expectations. Why? Because the world has changed. We have learned a lot, and technology has revolutionized our view of what is possible, both for research and for education.

Technology. The advent of online learning is a game changer in terms of the affordances it provides for research and development in education (Stigler & Givvin, 2017). Research on teaching and learning has long been hampered by the organization of schooling. When one teacher teaches a course to dozens or hundreds of students, and when that teacher is free to construct his or her own curriculum and instruction, it is hard to tie student outcomes to specific features of the instruction.

Putting learning resources online—whether they are implemented fully online or in blended learning environments—enables us to do things we simply could not do before. Let us consider our *better book* approach. First, students' interactions with learning objects generate data that tell us what students are doing as they work their way through the book - something we never could know with a hard-copy textbook. By embedding formative assessments *inline* with the instructional materials we can get new insights into how students interpret the materials—even specific text passages or questions—which can be used to guide improvements in the online materials.

Second, because at least part of the instruction is delivered to students at their computers or mobile devices, individual students—even those who are members of the same face-to-face class—can be randomly assigned to get different versions of the content. This makes it possible to do experimental research in the context of real educational settings, something that was nearly impossible to do before using research designs that required random assignment of schools and classrooms, not individual students, to different instructional conditions (Ostrow, Heffernan, & Williams, 2017).

Third, online materials, because they are easy to update, provide a means of storing what we learn through efforts to improve the program (cf., Morris & Hiebert, 2011). If practitioners, for example, find a paragraph that leaves students mystified, they can work to rewrite it, and perhaps even consult with learning scientists to help them craft a solution. If their rewrite is judged better based on testing, it can replace the original paragraph and be available immediately to the next group of students. In this way, online learning promotes dialogue between researchers, developers, and practitioners, and offers a novel setting for collaboration and innovation.

Advances in improvement science. At the same time as these technological advances, we have seen advances as well in methodologies for improving complex systems. The roots of improvement science grew out of the work of Deming, Juran, and Shewhart after World War II (Kenney, 2008), the first to apply statistical process control to the improvement of complex systems. These ideas became the basis of the lesson study movement in Japanese primary and middle schools (Lewis, 2015).

In recent years, we have seen these ideas popularized by authors such as Mike Rother in his book *Toyota Kata* (2009), developed further (Langley et al., 2009), and applied to health care (Kenney, 2008). Most recently, these ideas have taken hold in education (Bryk et al., 2010, Lewis 2015). Recognizing that teaching and learning are complex systems has naturally led to an interest in methodologies for improving the performance of such systems, and we now have success stories to validate the approach in education (Lewis, 2015).

New protocols for collaboration. If improvement science methodologies have the potential to lead to incremental improvements in online learning resources, wouldn't it be great to involve large communities of researchers, designers/developers, and practitioners in making such improvements? Yes. But if we do get more people involved, how do we keep them all working toward the same goals? And how do we integrate a myriad changes in the materials without creating an incoherent mess?

Pioneers of the open software movement have created solutions to this exact problem (Nielsen, 2011). Thanks to distributed version control systems (DVCS; e.g., Git, Mercurial), thousands of software engineers can contribute to the development of a common software project and track their changes. DVCS, and the concepts that underlie them, give us new tools and processes for implementing continuous improvements in educational materials. Bringing together these advances and applying them to the problem of education R&D is the focus of our current work and the goal of this article.

The *Better Book* Approach

Starting in 2017, we set out to leverage these recent advances to develop a new approach for education research and development, one that would lead to the continuous improvement of educational programs over time. Our strategy was to "learn by doing." Our vision: to build *a better book*. There are many layers to this phrase, which we will lay out in some detail before reporting on our progress and learning so far.

Learn by Doing: Building a Better Introductory Statistics Book

As part of our learn-by-doing strategy, we decided to work on creating, implementing, improving, and scaling a college-level course. We are, primarily, researchers. Our research focus is on how students come to understand complex domains—things that are hard to learn. We decided to focus on introductory statistics because it is a complex domain, difficult to learn, taught in every institution and across many departments, and critical for navigating the modern world. Statistics, especially at the introductory level, is generally taught as a series of isolated concepts and procedures: *p*, t-test, *F*, chi-square, ANOVA, regression, and so on. Although students are somewhat successful at learning these "bits" of knowledge, they often fail to grasp the coherent structure that underlies the field of statistics and data analysis, resulting in knowledge that doesn't easily transfer to new situations.

Our innovative statistics curriculum—which we set forth in detail in Son et al. (in preparation)—is grounded in research in the learning sciences. Briefly stated, we base our course on what we have called the *practicing connections hypothesis* (Fries et al., in preparation) with the explicit aim of fostering transferable learning.

The practicing connections hypothesis, which we as researchers seek to investigate, suggests that we might produce more flexible, transferable knowledge if we consistently give students opportunities to practice the connections that make statistics a coherent domain (Fries et al., in preparation). To begin this work, we decided to test our ideas by building an innovative introductory statistics course in which we teach students, from the beginning, to understand all of statistics in terms of the General Linear Model.

Business as Usual

It is worth noting, at the outset, how a research-based theory such as the practicing connections hypothesis might currently be expected to find its way into educational practice. As pointed out previously, researchers, designers/developers, and instructors currently work in silos. As researchers, our role is to develop knowledge and theories about how people learn. We conduct studies—often in the laboratory—and then publish them in academic journals. Our earnest hope is that someone who creates educational curriculum and materials might read our articles and put our ideas to good use. But in general, that's not our job as researchers.

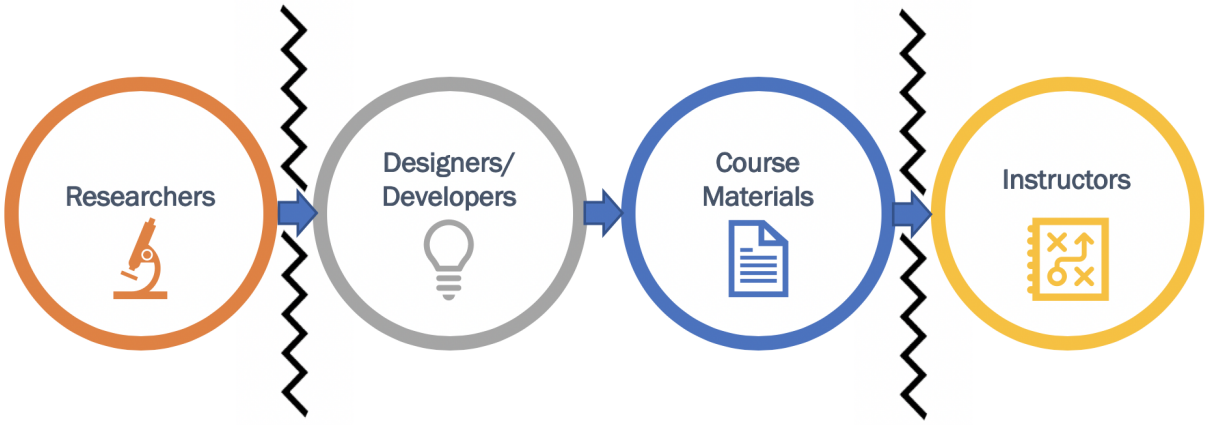

Figure 1: Business as usual: A sequential process.

Designers and developers of educational curricula and materials are often quite interested in research, but not in the details. They are looking for exactly what Dewey warned they would not find: a "rule of educational art" that they can adhere to as they build their products. But Dewey was right, and so there is slippage from research to development. And researchers who study teaching and learning are generally not in the same departments as the professors who develop educational materials, and so aren't able to assist in translating their research into products.

Once products get launched, they are essentially sold (or distributed free) in "as is" condition for practitioners to implement. With a few exceptions, there are no feedback loops whereby changes, or knowledge about how best to implement the products, can find their way back to the designers and developers. This is especially concerning given that most of what it takes for an educational product or program to succeed is the know-how supplied by practitioners. Needless to say, researchers, by this time, are long gone from the process.

The Better Book Vision

In our project we are trying to replace this business-as-usual approach with a new one that leverages the technological advances and methodological developments outlined above. Key elements of our vision are:

1. Create an innovative set of online materials (we think of this, currently, as an online textbook), with which students interact in order to learn. Fully instrument the online book so that it generates data relevant to students' interactions and learning.

2. Grow the number of students and instructors using the book. This is critical to drive improvement: more students means more data, and across universities, more diversity among students. We accomplish this by providing the materials free of charge, and by making them easy to adopt (in addition to being effective for learning).

3. Engage a community of researchers, designers/developers, and practitioners to work on continuous improvement of the materials and their implementation. Transform the relationship among the three legs of the R&D stool from one that is sequential and siloed to one in which all three work collaboratively over time (see Figure 2).

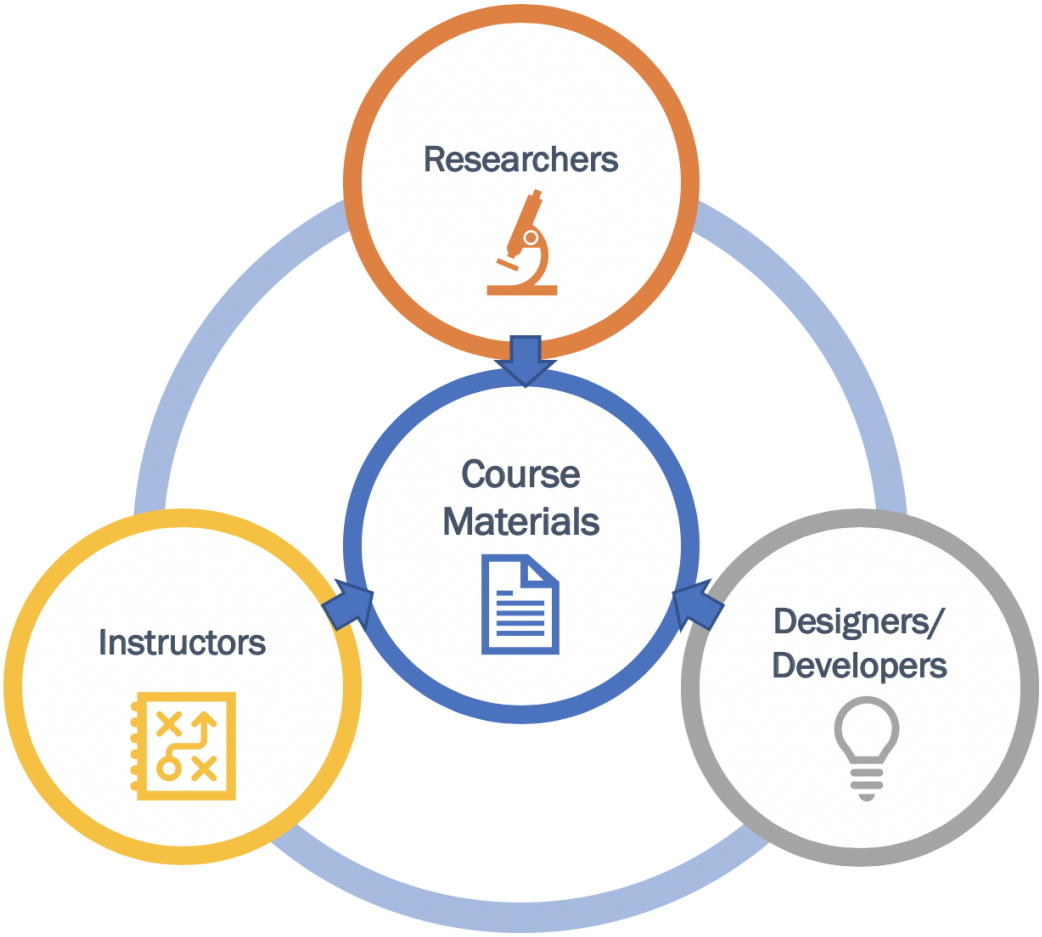

Figure 2: The Better Book vision.

In our initial learn-by-doing phase we are working to implement this new R&D approach. As we go, we are developing supports (technology, settings and routines) to make the process more effective and more shareable with others who want to apply the *better book* approach to a different set of online materials.

This approach opens up some new opportunities for both research and education. On the research front, we will have developed a research context for studying—in an ecologically valid setting—the processes through which students come to understand core concepts in complex domains—processes that play out over weeks and months. Researchers involved in this process will not be left to wonder if their discoveries will be able, someday, to help improve education. Instead, their discoveries will be immediately applied, and if useful, built into future versions of the materials, accessible to future students (c.f., Morris & Hiebert, 2011).

On the education side, we will have created a way to engage instructors in the improvement of educational materials such as our online textbook. Practitioners get a respected voice at the table, rather than being the passive recipient of materials. When they discover that something does not work, there are technologies, settings, and routines to help close the feedback loop between instructors and curriculum developers.

Building out this ambitious and fully-instrumented research infrastructure for the study of learning in complex domains provides a sort of "space shuttle" for educational research. Just as scientists with specific research interests worked to get their experiments included on the space shuttle, learning scientists interested in understanding how students learn in complex domains will be drawn to the opportunities of an improvement community focused on deep learning in a real course.

Learning to Improve

With this vision in mind, we set out to make it happen. We have read widely in the field of improvement science, as applied in a number of fields, and participated in improvement projects. But our specific vision - the continuous improvement of an online textbook - will require some adjustments. In this section we outline our strategy for innovation and improvement, and report on our progress thus far.

Much of our focus at this point will be on the technology platform (*CourseKata*) we are building to support the work. In later reports we will focus more on the settings and routines that are required to support collaborations among the stakeholders involved in the work.

Building Version 1.0 of the Statistics Book: The Innovation Phase

Our first step was to create Version 1.0 of our online statistics textbook. This task initially fell to two of us (Son and Stigler), learning scientists armed with a theory and actively engaged in teaching undergraduate introductory statistics courses at two different universities. In this "innovation phase" of the project, our goal was to implement our vision and get it working in our own classrooms with our own students.

The Minimum Viable Product. Our vision for this book was innovative in both content and pedagogy (Son et al., 2018; Fries et al., in preparation). We approached this vision with a shared understanding: the more innovative the approach, the more investment it will take to build Version 1.0. The reason for this traces back to the idea that teaching is a system. If an innovation is incremental, affecting only one component of a system, then it can be developed and dropped into the system to replace the current component. But if it is a major innovation,

affecting the system itself, thent an entirely new instructional system would have to be built just to test it.

Our statistics course differs in significant ways from most introductory statistics courses. To get it up and running, therefore, we not only had to write a textbook, but also had to develop other components of the system such as assessments that aligned with our new learning goals. Knowing that our eventual goal is to get other instructors using our online book, we had to build enough of a complete system so that others could implement and test. (This is similar to what Ries, 2011, and others have termed a "Minimum Viable Product," or MVP.)

Early design decisions. If our only goal was to improve our own courses, it would have been simpler. But we knew from the beginning that we wanted to build a networked improvement community focused on improving our online book. This knowledge led to some early design decisions that, no doubt, made life a little more complicated during the innovation phase. Here are the main decisions that impacted the development of our technology platform (described later).

- **Interleaved pedagogy.** We wanted our book to be a fully featured learning environment, aligned with our pedagogical theory (practicing connections). We thus looked for a platform on which we could build a book that would interleave text, graphics, videos, R coding exercises, and formative assessment questions all on a single page (see Figure 3). What we found is that none of the existing learning management systems (LMS) are designed to deliver this kind of pedagogy.

- **Delivered through any LMS.** At the same time, we thought that delivering the book through widely-used LMS platforms would make it easier for instructors to adopt the book and join our community. Thus, we identified two web applications that we could embed within the book's pages that could create the interleaved pedagogy we wanted: DataCamp Light, for R data analysis exercises; and Learnosity, for embedded assessments.

- **Learnosity for assessments.** Learnosity (learnosity.com) is the only proprietary platform we included in our design. We chose to use it for several reasons. First, it is a relatively mature and robust platform with a strong API, meaning we can easily embed it in our custom technology platform later. Most important, though, is that it is a cloud-based solution that, from the start, keeps student response data separate from students' identifying information. This makes it possible to deliver identified data back to instructors (through the LMS), and de-identified data to researchers. It also frees student data from the LMS. We don't need to figure out how to get the data out of different LMSs, but can simply store data externally from the beginning.

- **Using *Markdown* and *Git* for distributed version control.** In order to make our content easily transportable between different LMSs, we decided to write the book in markdown (https://en.wikipedia.org/wiki/Markdown). Because markdown files are plain text, they can be stored on a GitHub repository and tracked using the open and widely used *Git* distributed version control system (the strategy used for most open-source software development projects today).

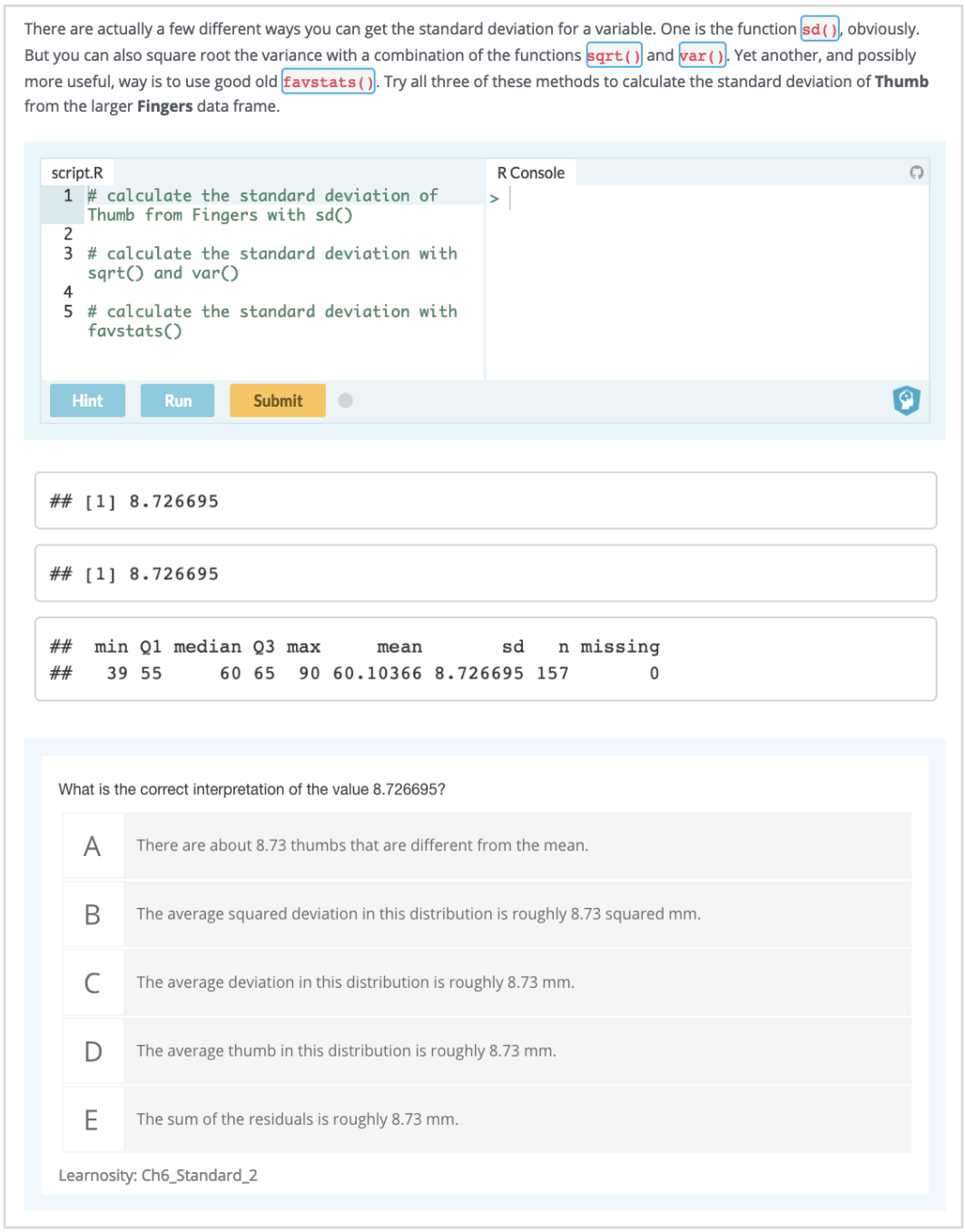

Figure 3: An example page from our interleaved online textbook, showing text, an R coding exercise, some R output, and a formative assessment question.

By the end of the innovation phase we had created a course that we were delivering on the Canvas LMS. We were using Git workflows to manage collaborative changes to the materials—processes that were overkill for our small team, but which we established with future expansion in mind. The end result was Version 1.0 of our book, which we managed to teach three times, to more than 500 students, during the first 1.5 years of our project — twice at UCLA and once at Cal State LA. The results of these initial implementations were encouraging (see Son et al., in preparation).

However, the process of getting the content into Canvas was extremely labor intensive. We also had no way to easily share the course with others, port the book to a different LMS, or get the student data out of the Learnosity data base — a process that required software development in order to accomplish. We needed a platform to manage all of this, one that did not exist at the time. But before we describe the platform we have built, let us further define what we need in order to support the next phase of our work, the implementation and improvement phase.

Time and space for innovation. Although our goal was to eventually develop a networked improvement community focused on improving our book, we recognized early on that it would be difficult to create an initial prototype of an innovative vision with a team that is too large or widely varying in perspectives. Our core team had to share a common vision of the MVP, and have the patience to bring it to life before getting into questions of its effectiveness or how to improve it.

The Implementation and Improvement Phase

Supporting adoption. To realize our vision we need to grow the number of instructors and students using the online textbook: the more users there are, the more data will be generated and the more, presumably, we can use the data to make improvements in the book. We made several decisions designed to make adoption easy.

We decided to give the book away for free to any instructor (and their students) who are willing to join our networked improvement community. But a free book is not enough: we also need to make the book easy for the instructor to use, or at least easier than using a normal textbook. We accomplished this by making it possible for instructors to get reports on their students' interactions with the content, and to use these data both for formative assessment and for automatic grading of students. We also have realized the importance of professional development for instructors. To feed growth, we will need to provide opportunities for new instructors, who may not have deep training in statistics, to learn about the ideas in our book, and the options for implementation. This is something we plan to attend to soon. Right now we are recruiting early adopters (the brave ones!).

The improvement framework. As we start to build a wider networked improvement community, we have begun to lay out a framework to guide our improvement efforts. The framework we started with is an adaptation of the *Toyota Kata* framework as described by Rother (2009, 2018). The word *kata* is from the Japanese and means *routine*. Rother summarizes the improvement routine as practiced at Toyota like this:

> Briefly put, the continuously repeating routine of Toyota's improvement kata goes like this: (1) in consideration of a vision, direction, or target, and (2) with a firsthand grasp of the current condition, (3) a next target condition on the way to the vision is defined. When we then (4) strive to move step by step toward that target condition, we encounter obstacles that define what we need to work on, and from which we learn.

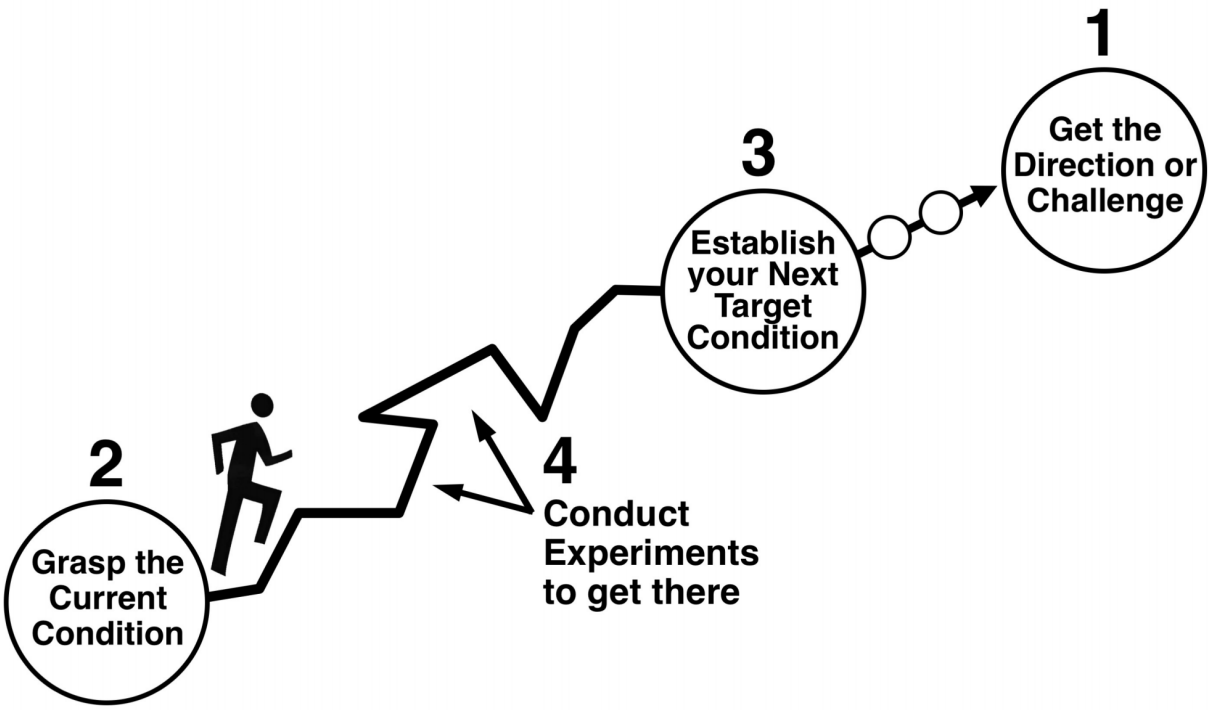

Figure 4: The *Toyota Kata* Model (from Rother).

Based on this framework, and on other information in the growing field of improvement science (e.g., Langley et al., 2009; Bryk et al., 2015), we set out to design and implement our *better book* R&D approach.

We realized that responsibility for the overall vision or challenge (Step 1) falls with our core innovation team. Although we started out thinking of ways to democratize the process of improvement — e.g., making it easy for anyone to submit changes to our master book — we were re-oriented in our thinking during a visit from Mike Rother in November of 2018. Mike reminded us of the critical role that an overarching vision plays in keeping the process on track. Indeed, a unified vision, advocated by the core innovation team, is even more important in the context of a large and diverse networked improvement community of the sort we are seeking to grow. Our vision is this: deep, flexible, transferable knowledge of statistics for all students. We want to teach an advanced course, but make it accessible to everyone, even those with poor preparation in mathematics.

Steps 2, 3, and 4 of the *Toyota Kata* framework are the core iterative parts of the improvement routine, and we envision small teams consisting of researchers, designers/developers, and instructors working together on these steps. As reported by Rother (2009), it is important to spend time on Step 2, trying to understand what in the current condition might be preventing us from achieving our vision. Often, just understanding the nature of the problem is enough to suggest a viable solution. Whereas the vision or challenge is longer term, the target condition (Step 3) is a short term goal, something that might be achievable within weeks or months. The target condition, from a research perspective, is like a detailed hypothesis of what you expect to see happen through the iterative experimentation cycles (otherwise known as PDSA, or Plan, Do, Study, Act cycles). Experimenting toward the target condition leads to new

discoveries, and new solutions, which in our case will be stored in improvements to the online textbook and its implementation.

The importance of measurement. The success of an improvement effort must be defined by the outcomes of the system you are trying to improve. Although there is often one primary outcome of interest—which in our case might be deep understanding of statistics—it also may be important to measure other outcomes. For example, we may want to improve understanding of statistics, but we also want to make sure that gains in understanding are not achieved by sacrificing the enjoyment students might get from doing data analysis, an outcome that may have huge impact on students' career trajectories. As Bryk et al. (2015) wrote, "You can't improve what you don't measure." So the very things we want to improve should be the things we measure.

In the context of learning, we also need to measure both the immediate effects of our textbook and the long-term effects. In essence, how long does learning last? We know that details will fall away over time, but hopefully students will be left with some kind of residue (Davis, 1992). Even if students forget specific details, hopefully they will be better prepared to re-learn the concepts we taught them, or more quickly learn more advanced concepts.

Using outcome measures to guide improvement is a commonsense idea. The more sensitive the outcome measures are, the easier it will be to see whether a change you are testing is making a difference in the right direction. But another, less intuitive, part of improvement science is the focus on reducing variability of outcomes to within "acceptable limits" (Langley et al., 2009). In our project, we want students' overall level of understanding to increase. But we also want to make sure that improvements bring everyone up to some desired level, not just students in one part of the distribution. In an education context, this means that much of our focus must be on students in the bottom end of the distribution, if only because getting the top students up to speed turns out to be relatively less challenging.

Finally, outcomes are not the only thing we need to measure. We also want to develop process measures (Langley et al., 2009), that indicate the underlying mechanisms through which students learn. For instance, when and at what rates do students complete the online assignments? Do they complete practice quizzes and when? How much do they write in response to open-ended questions, and how does that change over time? Process measures help us to get beyond trial-and-error testing of specific changes to test theories of teaching and learning. A good theory can be used again and again, in different contexts, to improve outcomes.

The need for a large item bank for assessments. Although the course we built included more than 1200 embedded assessments, we needed to develop summative assessments as outcome indicators. We assess students with 5 quizzes and a final exam over the course, which we implemented in either a quarter or a semester of instruction. Initially we created our own assessments, and we administered them independently of the online book. But as other instructors started to use our book, they too needed assessments.

Developing a very large and psychometrically sound item bank is currently one of our biggest needs. We need the number of items to be large enough that students will find it pointless to

try and guess which ones will be on the exam. We also want to make the quizzes and exams available through the online book platform (see more later on our CourseKata platform.)

Testing improvements. In the *Toyota Kata* framework, key stakeholders experiment their way towards a "target condition." This process involves identifying the most important obstacle preventing them from getting there, figuring out ways to remove the obstacle, and then testing the resulting solutions one by one. This is an iterative process, repeated multiple times over a short time period.

One challenge of implementing this framework in our context is that the process we are trying to improve spans an entire course; it is not something that can be easily repeated over and over again as we try to improve it. Some parts are repeated, such as the ordering of elements on a page (e.g., do you explain then ask questions, or ask questions prior to offering explanations?). But other things can only be tested once per class (e.g., should we start with the sampling distribution of the mean, or start, instead, with the sampling distribution of the difference between two means?).

We are beginning to see that different kinds of changes require different standards of evidence before they get adopted into the master version. If someone spots a spelling error in the book, we don't need to test it; just getting one other person to agree that the word should be changed is enough to warrant a change in the master. But other changes must be backed up by evidence. For some changes, correlational data might be enough to warrant the change. But for others, a true random-assignment experiment might be required before we institute the change for everyone. One of our goals in this project is to be able to easily generate the kind of data needed to support the improvement process.

Managing improvements through *Git*. We decided to put all of our content in markdown text files and to store them on GitHub. But learning how to use GitHub for our specific project turned out to be much more challenging than the decision itself. Using Git is as much a cultural practice as it is a piece of software, involving new vocabulary and new ways of working. We had to fundamentally change the way we think about and implement versioning of content.

Now a year and a half into the project, we have started to feel the benefits of this new way of working. Whereas we previously would have turned on track changes as we worked on a Word document or on a Google Doc, now we routinely "branch" the repository that holds our book, which means making a complete copy of the book on our own computer and, at the same time, on GitHub. Once you have your own branch, you can make any changes you want without affecting the work that others are doing. Periodically you can "rebase" your own branch onto the master branch to make sure your branch is up to date and not in conflict with recent changes to the master. You can share your branch with a collaborator, and work together on a change, before merging it into the master.

Maybe the most valuable concept in the Git workflow is that of a *pull request*. When you have marshalled enough evidence to warrant merging your changes into the master branch, you initiate a pull request. This, in essence, is a request that the team overseeing the improvement process review your work and decide whether or not to approve the merge (the adoption of

your change into the master). The pull request process serves the function of peer review and structures the interactions and thinking that go into evaluating a change. If a change is approved, a merge is performed. Sometimes merge conflicts will arise, which must be examined and resolved. But everything is documented, and you are left without that nagging feeling that something must somehow have gotten lost in the process.

The role of instructors. In our team, we often say, "The idea is 20%; implementation is 80%." This idea translates into an important role for instructors in the R&D process. A researcher may have a new theory; a designer may have a new idea. But ultimately, instructors will be needed to figure out how to make the theory and the idea work in practice. We have a few thoughts about the role of instructors in our R&D approach.

First off, not all instructors need to be on the R&D team engaging in the iterative improvement process. Some instructors—and we need as many of these as possible—will just want to use the online book with their students and nothing more. This is perfectly fine. Every instructor who uses the book with their students generates valuable data that can be used by the R&D team in the improvement process.

Another thought, which comes from Douthwaite (2002), is that the role of instructors will undoubtedly change as the online book matures. As more students from different backgrounds and with different interests take the course, the more know-how will be required by the instructor in order to get the most out of the course for every student. This know-how will primarily be developed by instructors, in the field, and not by researchers or designers/developers. As the book evolves, instructors will play a greater role in figuring out ways to improve student outcomes. And as more instructors teach the course, more can be gained by studying variations in implementation among them. What we learn from this variation can be captured as improvements in the online materials. This information will also inform our understanding of the science of implementation.

Finally, it is important to socialize instructors—the ones who do join the R&D community—into a new and more sophisticated way of seeing their own role in the community. In other approaches for research/practice collaboration, instructors see themselves as bit players in the process of figuring out *if* a new program works (perhaps as part of a randomized controlled trial). We want to change their mindset from one of helping *to judge if* a program works to one of helping *figure out what it will take* to make a program work for their students.

The OER movement. Although we give our book away for free, and although we are creating a community of instructors implementing the book in their classrooms, our project does not fit seamlessly into what has come to be known as OER, or the Open Educational Resources, movement. The OER community seems to have some values in alignment with ours, but others that are not in alignment.

For example, one assumption commonly voiced is that content for basic courses (such as introductory statistics) is all the same, that it can be commoditized, and that fundamentally there are no differences in quality, just in price, across different textbooks. Although we do agree that variation among textbooks is often minimal, we also think that it doesn't have to be

this way, and that quality differences are real and significant. Furthermore, we believe that innovation in textbooks is unlikely to happen in an environment in which anyone can mix and match chapters from different books into their own unique versions.

The emphasis on individuals creating derivative books based on their own taste, in our view, can work against innovation and the improvement of quality over time. Letting everyone do their own thing can, of course, yield some good ideas. Yet, unless the changes are incorporated into an explicit research design, it will not be easy to know which changes result in better student outcomes. The mix-and-match approach also ignores the fact that true innovations may require coherence in design across an entire course, something that can be provided by a visionary author on a mission to realize their idea, but less likely to emerge from a democratic process. Innovation needs its own space to grow, because that is where many of our largest leaps forward will come from.

In our R&D approach, we try hard to prevent instructors from making changes in the textbook unless those changes are part of a well-thought-out research design. In work similar to ours, Hiebert and colleagues at the University of Delaware (Hiebert & Morris, 2009) have adopted a group norm: If you are an instructor using the common materials, you don't have to use the materials exactly as designed. But if you decide to change them, you have to make sure there is some way to test that change so that everyone can learn from your experience.

The role of researchers. Researchers, too, have unique contributions to make to the networked improvement community, a point that may seem obvious given that we are developing a new R&D approach. In fact, we see an expansive and symbiotic relationship possible between researchers and the rest of the networked improvement community. It is clear from what we have written above that researchers bring critical skills to the table when it comes to implementing the *Toyota Kata* improvement framework. Researchers, in general, are trained specifically to do the kind of work envisioned in Step 2 (understand the current condition). They have theories, which provide lenses through which to see the current system; and they have methods for analyzing the current condition in light of their theories, and also in a more open discovery mode. The *Toyota Kata* framework is, fundamentally, a process based on scientific inquiry, and researchers are generally well-suited to undertake such inquiry.

Researchers also may have real contributions to make as they work with designers/developers and instructors to test their theories in the classroom. And in fact, being able to work in such settings is a major draw for researchers who, by tradition, have been relegated to publishing their findings in journals in hopes that someone will read and use them. Most researchers in the learning sciences want their work to make a difference. A networked improvement community of the sort we envision gives them access to a test bed, and an ecologically valid research site, in which they can work to get their ideas out into the world.

But researchers will also need to change the way they define their own research interests. Whereas traditionally, researchers see themselves on a mission to advance their theory, and thus their standing in the academy, they need to shift their focus from their own theory to a large problem shared by many stakeholders. As the improvement work progresses, researchers

may need to change their theories, or even their overall research focus, based on the problems that are of highest priority to the community.

The CourseKata Platform

In the previous section we reported on our progress in our "learn by doing" approach to building a continuously improving online textbook for introductory statistics. We started with little more than a vision, then set about finding out what it would take to implement the vision for a single online book. Along the way we confronted a number of challenges, to which we rigged up solutions (basically, with the education equivalent of duct tape). But as our goal is to create an approach that could be adopted by others, we used our experiences as a starting point to identify the kinds of support - especially technology - that would make the approach easier to implement moving forward.

The result of this process is the CourseKata platform, now in its second major version, that we are using to author and deliver our online book, and to support our continuous improvement work. We are continuing to develop this platform, and will make it available to others who want to replicate the work we are doing with other online materials. We will provide a brief description of the platform here.

Requirements in Brief

We realized early that the approach we were developing would be difficult to implement without technology supports. As we labored hard to get the approach up and running, we also kept track of what would make it possible to implement the approach at scale. This list, borne of the school of hard knocks, became the requirements we used to guide the development of the CourseKata platform. Here are the requirements we have developed so far:

- **Authoring and publishing:** We need to be able to author content, consistent with our pedagogical model; easily and automatically preview the pages as they will look to students; store the content in the cloud so that it can be shared and accessed by collaborators; keep track of different versions of the course; and manage the process of testing and improving the content.
- **Distribution:** We need to be able to easily distribute the current version of the book to instructors so that they can deliver it to their students through their preferred LMS; and to track the classes, so we can link student-generated data to instructors and institutions.
- **Data collection:** We need to be able to collect and store students' responses and interactions with the book while protecting students' privacy and confidentiality; and to store students' responses to summative assessments (quizzes and exams).
- **Data delivery:** We need to be able to provide detailed data back to instructors in real time so they can use it to guide instruction and grade students; and back to the R&D team to feed into the improvement process.
- **Experimentation:** We need to be able to conduct random assignment experiments, within classes, in which different students get different versions of the online book.

These requirements guided the development of the CourseKata platform, now in its second major iteration. We will briefly describe the platform in terms of the four functional systems suggested by these requirements.

System 1: Authoring and Publishing (Figure 5)

Our first priority in developing the platform was to automate the authoring and publishing process, which was highly labor intensive in our innovation phase. As illustrated in Figure 4, content developers can now work in markdown, using GitHub as a collaboration and version control platform. As the GitHub files are updated, CourseKata re-generates the HTML pages for the online textbook, and stores them on the Amazon cloud. The latest development version can be immediately previewed on the CourseKata.org website, as can release versions.

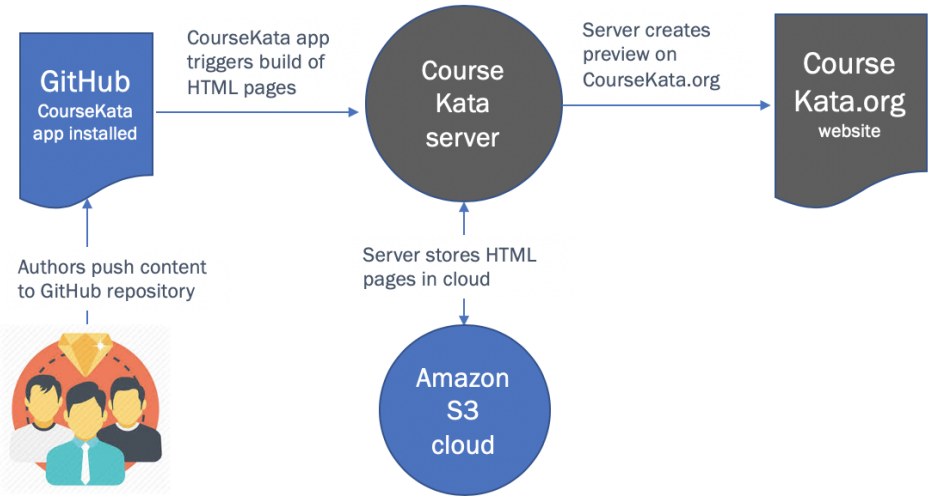

Figure 5: CourseKata content authoring and publishing system.

System 2: Content Distribution (Figure 6)

Once an instructor is granted an instructor membership on CourseKata.org, they can go into the website and create a class (defined as a group of students and a particular release version of the book). Once they have created a class, they can download the course in Common Cartridge format, and import it into their LMS (in the current version we support Canvas, but other LMSs will be added soon). Each instructor gets a unique consumer key and shared secret (i.e., passwords needed for installing the cartridge into the LMS), which allows us to track each instance of the course. Instructors can enroll students in their class just as they normally would. If they teach the course again, they simply need to go back to CourseKata.org and set up a new class.

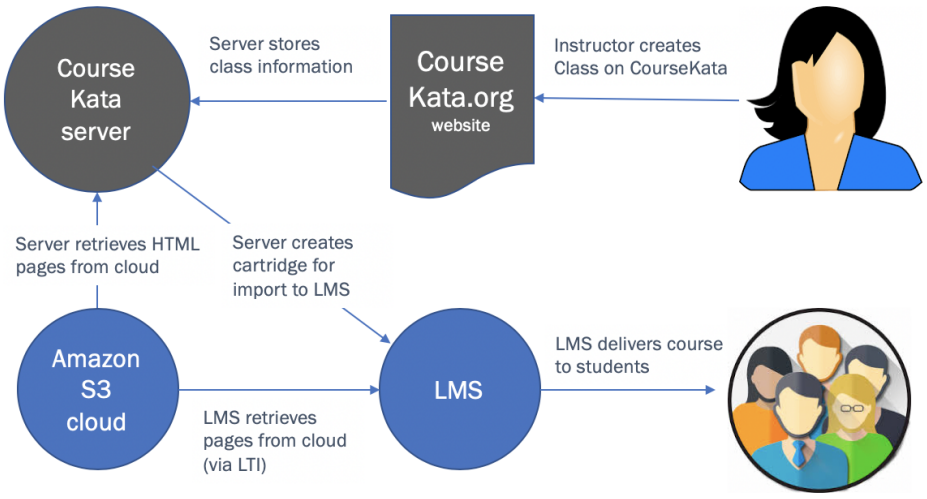

Figure 6: CourseKata content distribution system.

System 3: Teaching, Learning and Data Collection (Figure 7)

Once the instructor has enrolled students in the class, the course can begin. Students take the course through the LMS. All student responses are sent by the LMS to the CourseKata server. The server stores the data in the cloud, but also summarizes the data, sending reports back to both the students and the instructor through a special My Progress page in the LMS.

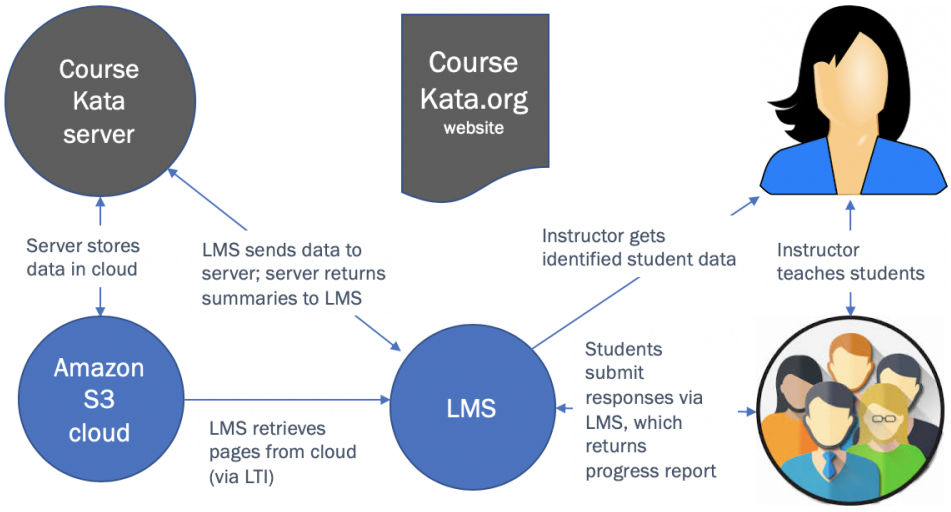

Figure 7: CourseKata teaching, learning and data collection system.

System 4: Research (Figure 8)

Finally, we have built a research system that currently serves two functions. First, it allows researchers to download and analyze de-identified student data in order to inform the improvement process. And second, it enables researchers to set up random assignment experiments on CourseKata.org. The ability to run experiments within classes of students is one of the most exciting features of the CourseKata platform. Researchers can make an experiment branch of the course, in which they alter the content in some way (this could be

something as small as changing a picture, inserting a video, or changing the way an explanation is written).

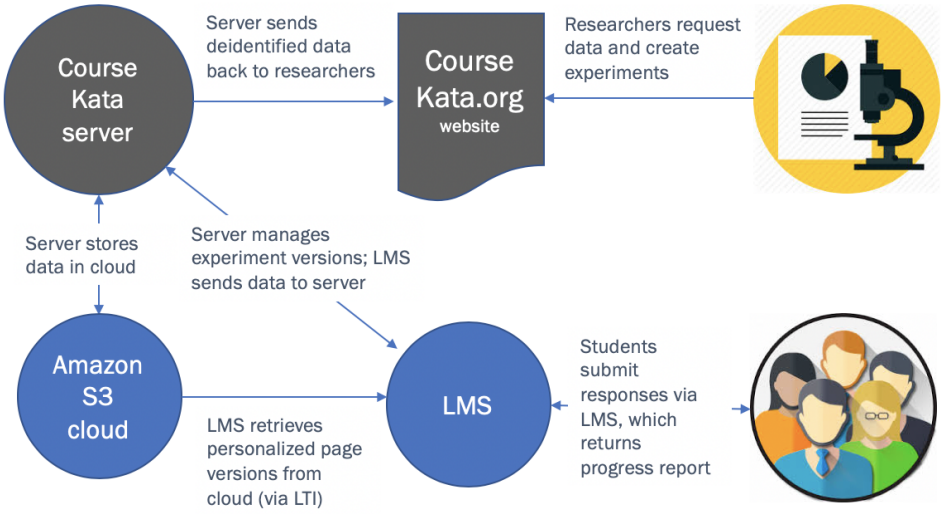

Figure 8: CourseKata research system.

After researchers add one or more classes into a study, which they can do on CourseKata.org, CourseKata will randomly assign each student in the class to get the latest release version or the experiment version of the pages. CourseKata delivers the assigned page to each student, tracks the data for students in the experimental versus the control condition, and gives researchers an easy way to download the experiment data for analysis.

Discussion and Next Steps

Our project is very much a work in progress. However, we wanted to outline our vision now—including some of the challenges and how we have met them—before they recede into distant memory. We also wanted to describe the web platform we have designed and built to support our vision. Although the current version of the platform is only being used for our introductory statistics project, it is a scalable platform that can handle many courses, each with its own networked improvement community. We look forward to helping the next communities get started. Meanwhile, here are some thoughts we have about our next steps in the project.

But First: What About RCTs?

One question researchers often ask when hearing about our project is: Does it work? At face value, post-class surveys show promising reactions from students in the course. Students found the textbook and embedded coding exercises were the most important tools for learning the content of the course (average rating ~4.5/5), as opposed to lectures and other face-to-face instruction (average rating ~3/5), a finding that suggests students are comfortable taking ownership of their learning. A vast majority of students indicated that they would recommend the course to others (~⅓ of students rate a 100% recommendation, and ~⅔ gave a recommendation of 70% or greater). But perhaps most importantly, a core concern of creating a *better book* is fostering learning that transfers to new content and contexts. Though this course only teaches modelling through simple linear regression, students were able to

accurately extend their knowledge to answer questions on multiple regression regarding notation, proportional reduction in error, and partitioning of sums of squares for testing individual terms (Son, Blake, & Stigler, in preparation).

Of course, this does not answer the "gold standard" question, "Have we compared what our students learn with what students in traditional courses learn?" As scientists, we've been socialized into this view ourselves, wanting to compare new programs with the "standard practice" that most students are engaged with through a randomized controlled trial (RCT). Based on what we are learning, and on some of the principles of improvement science, we believe that randomized controlled trials have an important role to play in improving education. But we also believe that they can be conducted too early, and kill off important innovations before they have been given a chance to realize their full potential.

Education is a cultural system. Building a new system is an incredibly hard thing to do, and it takes time. If we conduct an RCT too early, it isn't even clear what we are testing, as most early-stage products and programs produce high variation in outcomes. Once we bring variation under control, to within acceptable limits—evidence that we understand the system we have created well enough to control it—then it makes sense to compare the new program to others. Before that, we need to give innovation the time and space it needs to grow.

When to Improve and When to Fork

Building out an innovative program requires leadership. Someone needs to own the vision, and make sure that it is realized. If the process is democratized too early, there is great danger that the innovative vision will not be realized, but whittled away by the onslaught of too many voices.

Steve Jobs reportedly observed that the concept cars of the future you would see at the auto show almost never ended up being produced. Though exciting as concepts, by the time they went through the engineering, manufacturing, marketing, and sales departments they were slowly whittled back to something more resembling the cars of today than the cars of tomorrow. This is something to watch out for in a networked improvement community. Once it gets too large, it may be even more difficult to live up to the innovation that inspired it than it was in the initial stages of the project.

This is a tension that must be managed. Sometimes, the improvers are right; the system does need tweaking to work. But other times, the right approach is to encourage someone with a conflicting vision to "fork" the course (a bit of Git terminology) and launch out on their own. This is okay, and eventually, once both visions are realized, might create the conditions for really testing, with RCTs, the differences between the two visions.

Beyond Improvement: Personalization

We have described two phases of our project: the innovation phase, and the implementation and improvement phase. There is a third phase on the horizon, and that is personalization. Our focus in the improvement phase is on improving the online book for all students, which we

define as an increase in average performance on outcome measures, and a decrease of variability to within acceptable limits.

Sometimes, however, achieving these goals may require that different students get different input, i.e., that the instruction be personalized based on a model we can create of students who are taking the course. Fortunately, today's web technologies allow us to do that. We can, for example, learn about who our students are in terms of their interactions with and responses to the content, and then make adjustments in the book we deliver based on this knowledge.

We saved personalization until later because, no matter how you approach it, you need more data—and hence more students—in order to discover reliable differences in the exact content different students can benefit from most. In the last decade, terms like *machine learning*, *deep learning*, and *neural nets* have become prominent buzzwords in the field of personalization science. These types of (relatively) advanced artificial intelligence hinge on having a vast pool of training data and the ability to identify key variables that will perform well on data in the wild. That said, there have been some notably promising achievements in the field, like Amazon's eerily spot-on suggestions of what you might like to buy next.

Machine learning, however, is not necessarily the only model for how personalization should be done in education, though it is the model most commonly being pursued. In our view, some of the things we learn can be parsed into the "bits" that are needed in order to do machine learning. But some things—and especially complex domains like statistics—cannot be learned well if broken down into a long list of micro-objectives. Someone can learn all the pieces—*p*, chi-square, *F*, ANOVA, and so on—yet still not understand the deep structure of statistics. For this reason, we take a more deliberate approach to personalization. Instead of slicing up statistics into bits, and then letting algorithms generate personalized recommendations of brief learning resources to address each bit, we believe we must really come to understand the variation among our students, develop and test hypotheses about how different students learn, and then design targeted instruction that is aligned with our evolving theories. We aren't saying this is the only way to approach personalization, only pointing out that we are going against the tide in this sense.

Next Steps and an Invitation

We are continuing to strive toward our vision of the continuously improving textbook. As we learn, we take notes, and then reflect on what kinds of supports — in terms of technology, settings, and routines — will be needed in order to enable others to do this work. We are continuing to develop our technology, and we are writing a handbook of sorts that others can use to help us in this journey.

Our experience working to improve education and learning has been from the perspective of university-based researchers. Although many young people go to graduate school because they want to do research that makes a difference in education, they often emerge frustrated, having narrowed their audience from the world at large to the small group of researchers who do research in their own field. Through a process of socialization, they have somehow traded their goals for improving education for goals more narrowly focused on getting peer-reviewed publications in top journals. We believe that the better book approach provides a pathway for

researchers such as these to bring their talent, training, and ideas to bear on solving important educational problems in the world. It does not ask them to give up their identities as researchers, but instead shows how they can use the theories, methods, and findings of their field to contribute to the goal of improving learning in complex domains.

We have met many instructors and designers/developers of curriculum materials who would similarly be motivated to join the kind of community we envision. Curious instructors want to test ideas inspired by research, and they want a way to share what they learn with other instructors and with researchers, who, without detailed knowledge of implementation, only have part of the story. Similarly, many designers and developers want to contribute their imagination and ingenuity to a process of discovery that can lead to lasting improvements in instructional design and learning. And sometimes the three legs of our stool - research, practice, and design - end up in the same person.

If these ideas resonate with you, we invite you to join our community. If statistics is your thing, you can jump right in. Our introductory statistics book is now available to any instructor who wants to use it. As our user base grows, our CourseKata platform provides a testbed for researchers wishing to refine their theories, but also a place where findings can be immediately translated into improvements. And we can always use the talents of designers and developers who see ways of making incremental improvements to the book. Of course we are just learning how to organize this work, which we see as a long-term cultural change in the way we do education research and development. We welcome all who want to help us take the next step forward, using this better book approach to improve education.

References

Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). *Learning to improve: How America's schools can get better at getting better*. Cambridge: Harvard University Press.

Davis, R. B. (1992). Understanding "understanding." *Journal of Mathematical Behavior, 11,* 225-241.

Dewey, J. (1929). *The sources of a science of education*. New York: Liveright.

Douthwaite, B. (2002). *Enabling innovation: A practical guide to understanding and fostering technological change*. New York: Zed Books.

Fries, L., Son, J. Y., & Stigler, J. W. (2019). *Practicing connections: A framework to guide instructional design for developing understanding in complex domains.* Manuscript in preparation.

Gallimore, R. (1996). Classrooms are just another cultural activity. *Research on classroom ecologies: Implications for inclusion of children with learning disabilities*, 229-250.

Hiebert, J., & Morris, A. K. (2009). Building a knowledge base for teacher education: An experience in K–8 mathematics teacher preparation. *The Elementary School Journal*, *109*(5), 475-490.

Kenney, C. (2008). *The best practice: How the new quality movement is transforming medicine*. New York: Public Affairs.

Langley, G. J., Moen, R., Nolan, K. M., Nolan, T. W., Norman, C. L., & Provost, L. P. (2009). *The improvement guide: A practical approach to enhancing organizational performance*. San Francisco, CA: Jossey-Bass.

Lewis, C. (2015). What is improvement science? Do we need it in education?. *Educational Researcher, 44*(1), 54-61.

Morris, A. K., & Hiebert, J. (2011). Creating shared instructional products an alternative approach to improving teaching. *Educational Researcher, 40*(1), 5–14.

Ostrow, K.S., Heffernan, N.T., & Williams, J.J. (2017). Tomorrow's EdTech today: Establishing a learning platform as a collaborative research tool for sound science. *Teachers College Record, 119* (3).

Nielsen, M. (2011). *Reinventing discovery: The new era of networked science*. Princeton University Press.

Ries, E. (2011). *The Lean Startup: How today's entrepreneurs use continuous innovation to create radically successful businesses*. New York: Crown Business.

Rother, M. (2009). *Toyota Kata: Managing people for improvement, adaptiveness and superior results*. Blacklick: McGraw-Hill Professional.

Rother, M. (2018). *Toyota kata practice guide.* McGraw Hill Education.

Son, J. Y., Ramos, P., DeWolf, M., Loftus, W., & Stigler, J. W. (2018). Exploring the practicing-connections hypothesis: Using gesture to support coordination of ideas in understanding a complex statistical concept. *Cognitive Research: Principles and Implications*, *3*(1), 1.

Son, J. Y., Blake, A., & Stigler, J. W. (in preparation). A practicing-connections approach to introductory statistics.

Stigler, J. W., & Givvin, K. B. (2017). Online learning as a wind tunnel for improving teaching. *New Directions for Evaluation*, *2017*(153), 79-91.