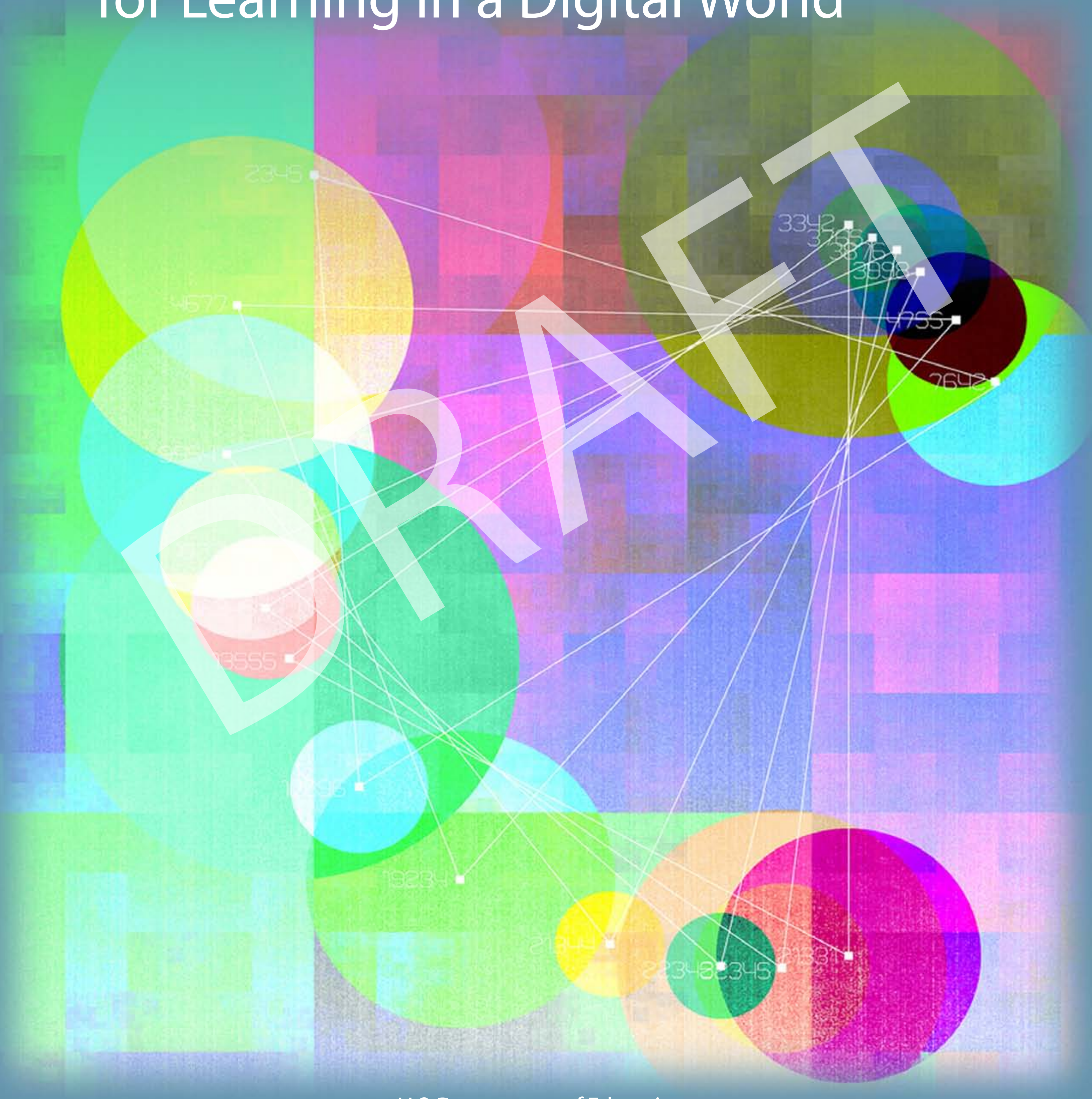
Expanding Evidence Approaches for Learning in a Digital World



U.S. Department of Education
Office of Educational Technology

Verso Page

Contents

Exhibits

Acknowledgments

This report was developed under the guidance of Karen Cator and Bernadette Adams of the U.S. Department of Education, Office of Educational Technology.

Barbara Means of SRI International led report development and drafting. She was assisted by Kea Anderson. Susan Thomas of S2 Enterprises LLC supported the development team with substantive editing, writing, and advice. Marianne Bakia, Marie Bienkowski, Geneva Haertel and Terry Vendlinski, also of SRI International, provided advice and insightful feedback on drafts. Eryn Heying provided research assistance and Brenda Waller provided administrative assistance. Patricia Schank and Ashley Lee supported the project website and provided graphics assistance. The report was edited by Mimi Campbell. Kate Borelli produced graphics and layout.

This report was developed collaboratively, in conversation with a Technical Working Group of learning technologies experts: Eva L. Baker (University of California, Los Angeles), Allan Collins (Northwestern University), Chris Dede (Harvard University), Adam Gamoran (University of Wisconsin), Kenji Hakuta (Stanford University), Anthony E. Kelly (George Mason University), Kenneth R. Koedinger (Carnegie Mellon University), David Niemi (Kaplan, Inc.), James Pellegrino (University of Illinois, Chicago), William R. Penuel (University of Colorado, Boulder), Zoran Popovi (University of Washington), Steve Ritter (Carnegie Learning), Russell W. Rumberger (University of California, Santa Barbara), Russell Shilling (Department of Defense, United States Navy), Marshall S. Smith (The Carnegie Foundation for the Advancement of Teaching), and Phoenix Wang (William Penn Foundation). The authors benefited tremendously from this group's expertise, engagement, and patience.

The authors thank those who provided formal input on drafts of the report. At the U.S. Department of Education, Elizabeth Albro (Institute for Education Sciences), Sue Betka (Institute for Education Sciences), Nadya Dabby (Office of Innovation and Improvement), John Easton (Institute of Education Sciences), Ruth Neild (Institute for Education Sciences), and Jim Shelton (Office of Innovation and Improvement) provided valuable input, as did Edith Gummer, Janice Earle, Barry Sloane, and Arlene de Strulle of the National Science Foundation.

The team also benefited from expert review and comments from Ryan S. J. d. Baker (Worcester Polytechnic Institute), Malcolm Brown (EDUCAUSE), Tony Bryk (The Carnegie Foundation for the Advancement of Teaching), Marc Chun (The William and Flora Hewlett Foundation), Ed Dieterle (The Bill & Melinda Gates Foundation), Stuart Elliott (National Research Council), Adam Ezring (The Council of Chief State School Officers), Andrew Gibbons (Brigham Young University), Margaret Hilton (National Research Council), Bradley Hull (National Association of State Boards of Education), James Kemple (New York University), Richard Laine (National Governors Association), Doug Levin (State Education Technology Directors Association), Karen Peterson (EdLab Group), Lisa Petrides (Institute of Knowledge Management in Education), Linda Roberts (American Association for the Advancement of Science), Thomas Smith (Vanderbilt University), Martin Storksdieck (National Research Council), William Tierney (University of Southern California), and Nick Warren (The Carnegie Foundation for the Advancement of Teaching).

During the development of the report, the authors solicited input at numerous conferences. Major themes from discussions with participants in sessions held at the following events were summarized by the report development team and help shaped the reports content: Digital Media and Learning (DML 2012); Consortium for School Networking (CoSN 2012); SxSW Edu 2012; American Association for Educational Research (AERA 2012); Learning Analytics and Knowledge (LAK 2012); SIIA Ed Tech Industry Summit; and the International Society for Technology in Education (ISTE 2012). In addition, Technical Working Group members sought input at Educon 2012 and at a Southern California meeting of Urban Education Dialogue.

Executive Summary

Relatively low-cost digital technology is ubiquitous in daily life and work. The Web is a vast source of information, communication, and connection opportunities available to anyone with an Internet connection. Most professionals and many students have a mobile device in their pockets with more computing power than early supercomputers.

Educators today can leverage these technologies to develop engaging and powerful learning experiences for their students and to provide professional tools, interactive content, and increasingly targeted feedback. The Internet is full of digital learning resources that students can interact with in authentic, meaningful ways that contribute to improved learning outcomes (U.S. Department of Education 2010a). From Khan Academy videos to the Massachusetts Institute of Technology's OpenCourseWare to a growing number of Massive Open Online Courses (MOOCs), millions of people now have access to learning resources developed by some of the foremost experts in their fields. Many of these resources are free and available for remixing or reuse for different learners and education contexts.

Some of the more sophisticated digital learning resources collect large amounts of fine-grained data as users interact with them, in real time and over time, as learning is taking place. Providers of these systems are not only analyzing these educational data to improve their systems, but also providing data for educators that can be used to help foster more effective practice at all levels of the education system.

Further, efforts are under way to create collaborations that result in combining data from digital learning systems with other data from multiple sources to gain broader insight into all the factors—both inside and outside school—that can affect educational outcomes for individual students. These combined datasets show promise for solving challenging problems such as improving high school graduation rates in specific communities.

Realizing the full potential of digital learning requires evolved thinking about education research and development (R&D) and evaluation. Specifically, realizing this potential requires

- digital learning innovations that can be developed and implemented quickly so every school has the chance to adopt them;
- continuous improvement processes to improve, adapt, and enhance these innovations as experience is gained in using them;
- use of the vast amounts of data that sophisticated digital learning systems collect in real time to ask and answer questions about how individual learners learn so the needs of every student can be met; and
- expanded approaches to evidence gathering for greater confidence that investments in cost-effective and cost-saving technology-based interventions are wise, producing the outcomes sought.

Purpose of This Report

Various stakeholders in the education community have different perspectives and needs, but all share an interest in understanding how to use digital learning systems, the data they generate, information, and evidence to address specific challenges in the U.S. education system. The opportunities digital learning resources create and the data they produce have important implications for each stakeholder group.

To address these implications, this report combines the views of education researchers, technology developers, educators, and researchers in emerging fields such as educational data mining and technology-supported evidence-centered design to present an expanded view of evidence approaches.

Specifically, this report focuses on evidence approaches that can be used individually or in combination to design education research and gather and analyze evidence produced by the vast amounts of data generated as teachers and students use digital learning systems. The evidence approaches are introduced and explained in the context of five specific education challenges. The approaches and the challenges described in this report are illustrative rather than exhaustive. The five challenges and ways evidence-gathering approaches based on digital learning system data can help address them are described below.

1. Making Sure Learning Resources Promote Deeper Learning

Digital learning can help meet new and more demanding expectations about what students need to learn, but technology-based resources and interventions must be up to the task.

Expectations about what all students should be able to understand and do are rising. In a global economy that demands innovation, people need the ability to transfer what they have learned to similar but different situations. Students today need to acquire critical thinking, problem solving, and communication competencies at levels that were expected of only the most highly educated students in past generations (Pellegrino and Hilton 2012).

Technology provides numerous new opportunities for educators, educational publishers, and developers to address these new more demanding standards with high-quality learning resources. Practices with a long history in commercial technology R&D can be brought to bear in developing these new digital learning resources. These design, development, and improvement practices, when applied to learning materials, can generate evidence of both usability and effectiveness.

Instead of a linear approach, these industry R&D processes are based on multiple cycles of rapid development and testing of effectiveness with constant feedback into redesign and further refinement. These efforts can be guided by the data generated when students interact with digital learning systems (U.S. Department of Education 2010a). These data can be used to inform rapid cycles of testing and refinement.

To meet the challenge of analyzing the vast amounts of data collected by digital learning systems, the emerging field of **educational data mining** is being combined with learning analytics to apply sophisticated statistical models and machine learning techniques from other fields such as finance and marketing to education. Educational data mining can address the question of how to refine a learning system or other type of learning resource and can provide the practitioner or researcher with information about learner behavior, achievement, and progression. It is less well suited to investigating the causal case for the effectiveness of a resource or intervention as a whole.

To get at effectiveness, digital learning systems provide an opportunity to conduct controlled random-assignment experiments (Shadish and Cook 2009) much more rapidly than was previously possible. In the software industry, a type of random-assignment experiment known as **A/B testing** is used to isolate variables by comparing two different versions of the same product or system (version A and version B) against each other by randomly assigning users to one or the other version. For example, one version of an online algebra course might have design feature A and the other design feature B, but the versions would be otherwise identical. Because researchers widely view random-assignment experiments as generating the highest quality evidence about what causes effects, technology-enabled rapid experiments may have especially broad appeal.

Educational data mining and rapid A/B testing can help developers refine and enhance digital learning systems, but they are less well suited for answering questions about the different ways digital learning systems are used in different contexts and how that affects outcomes. **Design-based implementation research** (DBIR) is an emerging research approach that is better suited for this kind of inquiry. DBIR seeks to bring research closer to practice by having researchers and practitioners jointly set a research agenda that incorporates insights from both the learning sciences and the wisdom of practice.

DBIR proponents work with their practitioner partners to lay out a theory of the implementation steps needed in the practitioners' context and study both implementation processes and outcomes simultaneously. In terms of evidence, they typically look for correlational patterns and use quasi-experimental designs rather than random-assignment experiments, although some DBIR studies do use experimental tests of different strategies for supporting implementation.

A hoped-for benefit of DBIR collaborations is that education practitioners will think about their activities as cycles of implementation, data collection, reflection, and refinement and constantly seek data and information to refine their practice.

2. Building Adaptive Learning Systems That Support Personalized Learning

Advances in technology-based learning systems enable customized strategies and content. Learning data collected by these systems can be used to improve the ability of the systems to adapt to different learners as they learn.

Digital learning systems are considered *adaptive* when they can dynamically change to better suit the learner in response to information collected during the course of learning rather than on the basis of preexisting information such as a learner's gender, age, or achievement test score. Adaptive learning systems use information gained as the learner works with them to vary such features as the way a concept is represented, its difficulty level, the sequencing of problems or tasks, and the nature of hints and feedback provided. Systems are now being developed to support personalized learning by incorporating options for varied learning objectives and content as well as for the method and pacing of instruction.

Advances in technology have heightened the possibility that digital learning systems can replicate dynamic adaptations used successfully by human tutors or even implement those and other methods more effectively than humans. Capabilities now available include fine-grained models of learner knowledge that are updated dynamically, micro-level tagging of both instructional content and learner actions with systems, and on-the-fly system adaptations to students' emotional states and levels of motivation.

The micro-level data digital learning systems collect about student interactions can be used to develop and validate learner categorizations based on those interactions rather than on membership in a demographic category with higher average risk. It now is possible also to combine insights from learning theory that suggest patterns to look for with large sets of detailed learning data.These new capabilities make the long-sought goal of differentiating instruction for every learner much more attainable.

Once important learner differences have been identified, digital learning systems can be revised to vary the experience provided for different kinds of students working in different contexts. Key to this is being able to determine what an individual learner knows and what he or she still needs to learn in a dynamic way throughout the learning process.

Perhaps the most clear and consistent difference between students is their incoming prior knowledge. Assessing and adapting to differences in prior knowledge require two types of models: one of concepts students must master (the expert model) and one of what individual students know about that domain (the learner model). Modeling learner knowledge is a dynamic process that resembles the kind of user knowledge modeling that has been used in adaptive hypermedia, recommendation systems, and intelligent tutoring systems. New machine-learning-based approaches to developing student knowledge models build on prior research in this area.

3. Combining Data to Create Support Systems More Responsive to Student Needs

Young people learn and develop in a wide range of settings. Better data use can help support the full range of student needs and interests—both inside and outside schools and classrooms—to improve learning outcomes.

Far too many students in the United States—especially those from low-income backgrounds—never finish high school. Academic and social disengagement from school are key factors associated with dropping out (Rumberger 2011). This disengagement is not typically associated with a single event; rather, it is a long-term, cumulative process (Newmann, Wehlage and Lamborn 1992).

Achieving progress in this area requires that schools be more responsive to students' needs and interests and take a more encompassing view of their lives. School administrators need to appreciate the fact that young people learn and develop in a wide range of settings, not just classrooms, and attend to the multiple aspects of their well-being. Technology provides opportunities for creating better support that can keep students engaged and progressing through school. These include the ability to collect different types of data and **combine and analyze data from multiple sources** in new ways to target intervention practices and programs and provide support for new practices and interventions.

State and district student data systems have improved greatly over the past decade in ways that permit examining an individual student's educational experiences and achievement over time, even if the student changes schools or school districts.

Some school districts now are also experimenting with combining data collected from students themselves with administrative and other datasets. For example, administrative data in student information systems are being cross-linked to records and events in learning management and digital learning systems. Those data, in turn, can be combined with data from social services agencies that students may engage with outside school, such as the juvenile justice system, the foster care system, or youth development programs.

Linking these various types of data can help schools explore relationships between students' conditions outside school and their in-school experiences and thereby develop early warning systems for predicting student risks. Increasingly sophisticated techniques for predictive analytics, which combine a variety of disciplines including statistics, data mining, and game theory, are also being used to investigate whether some student behaviors are predictors of school failure and dropping out.

This type of data can also be used to identify positive factors and student accomplishments. An emerging area of research is on environments that are "interest-driven," where young people choose to pursue activities (often outside school) that involve learning, deep engagement, and often the exercise of leadership (Heath and McLaughlin 1993). Interest, like deep content knowledge, develops over time and depends on the availability of guides and peers who can support its growth (Hidi & Renninger 2006). A related development is the use of "badging" systems that can capture and recognize the skills and abilities that students master when they pursue interest-driven routes to learning (Mozilla Foundation, Peer 2 Peer University and MacArthur Foundation 2012).

4. Improving the Content and Process of Assessment with Technology

Digital learning systems can collect data on important qualities not captured by achievement tests. These data can be used to measure more of what matters in a way that is useful for instruction.

The U.S. education system invests heavily in tests of student achievement that can be used to hold districts and schools accountable for whether students meet state proficiency standards. At the same time, supporting students' learning calls for additional types of assessment, including formative assessments administered in the courses of learning to provide information that teachers and students can use to guide future learning, assessments of 21st-century skills such as collaboration and innovation, and personal and affective qualities related to intellectual curiosity and persistence.

The fine-grained information about students' learning that newer digital learning systems collect can be used to construct measures of important learning outcomes and learning processes that have been difficult to capture with conventional state tests. What is more, these data can be mined to assess both cognitive and noncognitive skills—the latter being more oriented toward personal qualities such as conscientiousness and self-efficacy in college and the workplace. Specifically, digital learning systems provide the opportunity to measure these qualities on the basis of students' behavior in a learning system rather than through self-report.

Advances in assessment theory, notably **evidence-centered design** (ECD) and new statistical techniques and technology tools for supporting the use of ECD in assessment development, are making the assessment of complex cognitive components that are exercised in multiple subject matter contexts much more feasible than in the past. ECD is a systematic process in which the designer articulates (a) the competencies to be measured, (b) what would constitute evidence that a learner possesses those competencies, and (c) the situations or tasks that can be used to elicit that evidence. In the past, ECD had been labor intensive, but technology support systems for applying it to assessment development have recently emerged.

Combining ECD with assessments embedded in digital learning systems opens up possibilities for assessing features that are recognized as important but that could not be measured reliably and efficiently in the past (Pellegrino Chudowsky and Glaser 2001; Shute 2011). More of what educators really want to assess can be measured by mining the data produced when students interact with complex simulations and tasks presented in digital learning systems.

These measures require greater expertise to analyze, but that expertise can also be embedded in digital learning systems. Further, when assessments are embedded in digital learning systems, learners are assessed in the course of learning. Time no longer must be taken away from instruction to stop and measure how much has been learned.

Although the potential of embedded assessments is compelling, researchers are grappling with some open questions about them. For example, embedded assessments are tied to specific products, raising questions about whether performance on those assessments really predicts what students would do in other contexts. Also, embedded assessments often provide students with feedback, which means that the students are learning about a concept or how to execute a skill at the same time the system is attempting to gauge their competence in that knowledge or skill.

The application of ECD and data mining to learning systems for assessment purposes needs to be accompanied by the collection of evidence of validity and reliability. The successful development of psychometrically sound assessments that go on in the background as students use online learning systems would enable educators to begin asking whether these approaches obviate the need for some of the current end-of-year summative assessment. A number of research groups are working on how to make data gathered from online learning systems useful within accountability contexts as well as for individual learners and teachers (U.S. Department of Education 2010a).

5. Finding Appropriate Learning Resources and Making Informed Choices

Learning resources and materials play a critical role in achieving desired learning outcomes. Educators need better supports as they make decisions about which digital learning resources to adopt.

Now that digital learning resources are readily available on the Internet, many teachers and a growing number of schools are using them to expand the resources available for learning and to supplement or replace print-based materials such as textbook chapters and exercises. These digital resources give educators more choices, but they also raise the issue of how to ensure their quality and determine their effectiveness in achieving desired learning outcomes.

Besides the Internet, two other factors are driving the trend of teachers supplementing print-based textbooks and other materials with digital learning resources: easy-to-use creation and publishing tools that enable anyone to create, configure, aggregate, and modify learning materials and Internet-supported resources such as online repositories and communities that make it easier for educators to find and evaluate resources that might meet their needs.

User-generated learning resources and online repositories of learning materials generate more options for educators and also raise the question of the effectiveness of specific products and resources. The desire to continuously improve understanding of the usefulness of digital learning resources, including the degree to which evidence of effectiveness exists, is prompting technology developers, companies, government entities, and nonprofit organizations—separately and working together—to develop new ways of gathering and publishing information and evidence about these resources. Methods include

- aggregating user reviews, which is common in the consumer world and is now being used in many of the online education repositories and communities;
- aggregating user actions, such as rating, voting, and ranking; clicking, viewing, downloading, and sharing to social media; and actions connecting to the use of a learning resource in instruction, such as aligning, implementing in some context, and adapting resources to individual learners;
- user panels, which are sizable managed online communities (typically more than 5,000 members who are compensated in some way for their ongoing participation) that are used to provide prompt feedback; test a product's usability, utility, pricing, market fit, and other factors; and gather information about user needs and behavioral patterns for specific products or product categories for purposes of product improvement;
- expert ratings, reviews, and curation in which experts draw on both their specialized knowledge relevant to a product experience and their own experiences to evaluate and make recommendations about learning resources; and
- test beds, which in education can consist of a network of schools or classrooms and a community of researchers who have committed to working together and put the necessary infrastructure in place (for example, data sharing agreements and classroom technology).

The fragmented nature of many of the organizations providing access to these types of evidence approaches exposes the need for an objective, third-party organization that can serve as a trusted source of evidence about the use, implementation, and effectiveness of digital learning resources.

Summary, an Evidence Framework, and Recommendations

This report concludes with a summary, an evidence framework designed to support evidence decision making, and recommendations that can help accelerate progress in leveraging digital learning resources and data to expand evidence approaches.

The evidence framework provides actionable information about a wide array of resources and interventions, including the development and continuous improvement of digital learning resources and the incorporation of insights based on data from digital learning systems into education more broadly. The evidence framework is intended to help education stakeholders implement a process of planning, creating, choosing, and combining appropriate evidence-gathering approaches that could be useful under different circumstances. The framework has three components:

1. ***An Evidence Reference Guide*** *that summarizes the evidence approaches highlighted in this report as well as other evidence approaches widely used in education today. The Reference Guide includes the kinds of questions all the evidence approaches can help answer, the types of evidence that each can generate, and suggested uses.*
2. ***An Evidence Decision-Making Model*** *that can be used once a learning resource has been selected to make decisions about appropriate evidence approaches to use in conjunction with implementing the learning resource. The Decision-Making Model does not assume a linear staged model of R&D, which ties investment in collecting evidence of impact to product maturity and widespread use. Rather, the model suggests that gathering evidence is an ongoing process that extends beyond development and implementation of a learning resource depending on the factors of confidence in the improvement potential of the learning resource and its implementation risk.*
3. ***Scenarios*** *that illustrate how and when the various evidence approaches might be applied in situations familiar to education stakeholders.*

Recommendations

The following recommendations are designed to help education stakeholders turn the ideas presented in this report into action. Detailed explanations of each recommendation are in the Summary and Recommendations section of this report.

1. Developers of digital learning resources, education researchers, and educators should collaborate to define problems of practice that can be addressed through digital learning and the associated kinds of evidence that can be collected to measure and inform progress in addressing these problems.

2. Learning technology developers should use established basic research principles and learning sciences theory as the foundation for designing and improving digital learning resources.

3. Education research funders should promote education research design that establishes that digital learning resources teach aspects of deeper learning such as complex problem solving and promote the transfer of learning from one context to many contexts.

4. Education researchers and developers should identify the attributes of digital learning systems and resources that make a difference in terms of learning outcomes.

5. Users of digital learning resources should work with education researchers to implement these resources using continuous improvement processes.

6. Purchasers of digital learning resources and those who mandate their use should seek out and use evidence with respect to the claims made about each resource's capabilities, implementation, and effectiveness.

7. Interdisciplinary teams of experts in educational data mining, learning analytics, and visual analytics should collaborate to design and implement research and evidence projects. Higher education institutions should create new interdisciplinary graduate programs to develop data scientists who embody these same areas of expertise.

8. Funders should support creating test beds for digital learning research and development that foster rigorous, transparent, and replicable testing of new learning resources in low-risk environments.

9. The federal government should encourage innovative approaches to the design, development, evaluation, and implementation of digital learning systems and other resources.

10. Stakeholders who collect and maintain student data should participate in the implementation of technical processes and legal trust agreements that permit the sharing of data electronically and securely between institutions, complying with FERPA and other applicable data regulations and using common data standards and policies developed in coordination with the U.S. Department of Education.

11. Institutional Review Board (IRB) documentation and approval processes for research involving digital learning systems and resources that carry minimal risk should be streamlined to accelerate R & D without compromising needed rights and privacy protections.

12. R&D funding should be increased for studying the noncognitive aspects of 21st-century skills, namely, interpersonal skills (such as such as communication, collaboration, and leadership) and intrapersonal skills (such as persistence and self-regulation).

13. R&D funding should promote the development and sharing of open educational resources that include assessment items that address learning transfer.

14. The federal government and other interested agencies should fund an objective third-party organization as a source of evidence about the usability, effectiveness, and implementation of digital learning systems and resources.

Introduction

Education is the key to U.S. economic growth and prosperity and the best guarantee of the promises of the American dream. Fulfilling these aims requires raising expectations about what students should know and understand and embracing new strategies for improving learning outcomes so as to increase high school graduation rates and ensure college and career readiness for millions of Americans. One strategy for improving learning outcomes and educational persistence is applying digital technology to teaching and learning and other issues that can affect learning, such as lack of engagement or social and emotional connections to school.

Technology in education is not new: Experiments using computers in the classroom date back to the 1960s. What is new is the ubiquity of sophisticated, relatively low-cost digital technology in daily life and work. The Web is a vast source of information, communication, and connection opportunities available to anyone with an Internet connection. Most professionals and many students have a mobile device in their pockets with more computing power than the early supercomputers.

Technologies are being leveraged to develop engaging and powerful learning experiences and to provide professional tools, interactive content, and increasingly targeted feedback. The Internet is full of resources that students can interact with in authentic and meaningful ways that contribute to improved learning outcomes (U.S. Department of Education 2010a).

Digital learning has progressed greatly, and with it have come new opportunities and new challenges. Realizing the full potential of digital learning requires evolved thinking about education research and development (R&D) and evaluation. Specifically, realizing this potential requires

- digital learning innovations that can be developed and implemented quickly so every school has the chance to adopt them;
- continuous improvement processes to improve, adapt, and enhance these innovations as experience is gained in using them;
- use of the vast amounts of data that sophisticated digital learning systems collect in real time to ask and answer questions about how individual learners learn so the needs of every student can be met; and
- expanded approaches to evidence gathering for greater confidence that investments in cost-effective and cost-saving technology-based interventions are wise, capable of producing the outcomes sought.

Drivers of Change

Inspired by the explosion of innovation in consumer technology in the last decade, educational publishers and developers are creating a wide variety of digital learning resources for use inside and outside classrooms, at all grade levels and for learners of all ages. Consumers—students, teachers, parents, higher education institutions, and K–12 schools—are embracing learning technology in growing numbers.

The Khan Academy videos and Carl Wieman's simulations, for example, now have tens of millions of users. The Massachusetts Institute of Technology's OpenCourseWare initiative has spread to 60 nations and well over 150 institutions and has produced thousands of freely available college courses. More than half a million lectures from tens of thousands of courses are available on iTunes U. When an instructor at Stanford University offered his Introduction to Artificial Intelligence course free online—a Massive Open Online Course (MOOC)—160,000 people from around the world signed up (Thrun 2012). Widely publicized, this course gave rise to a new wave of MOOCs available through not-for-profit organizations and institutions of higher learning.

A ***Massive Open Online Course (MOOC)*** is designed to have large-scale participation—thousands or even hundreds of thousands of students—and be accessible free of charge to the public via the Internet. Higher education courses in a wide range of subject areas can be found on the websites of Coursera and Udacity.

Some of these new digital learning resources are sophisticated systems capable of collecting large amounts of fine-grained data as users interact with them, in real time and over time, as learning is taking place. Providers of these systems are beginning efforts to analyze these educational data. Such efforts hold promise for harnessing and sharing the information derived to improve the systems and the learning outcomes at all levels of education.

In addition, when educational data are combined with data from other sources, such as community and social services organizations that also serve children and youth, the opportunity arises to gain broader insight into students' lives, including factors outside school that can affect educational outcomes. These combined datasets can be used to solve problems that require community-wide supports, such as improving high school graduation rates.

Big Data in Education

The term that industry has coined to describe such large amounts of fine-grained data is big data. Big data denotes datasets that are large, complex, and difficult to store, search, share, analyze, and visualize using commonly available analytical tools and methods. But the term is more than just an indication of quantity and complexity. It also indicates the value of the information that can be derived by analyzing large datasets.

Big data is a term used to describe a dataset or collection of datasets so large and complex that standard data management tools have difficulty performing analyses and other tasks such as capturing, storing, searching, sharing, and visualizing information. Big data is often impossible to analyze on a single computer, requiring multiple servers running in parallel.

Depending on the goal, analyzing one large dataset can produce more accurate and actionable results than analyzing the same amount of data in smaller datasets. Examples include the ability to determine real-time traffic conditions to alert commuters to hazards or recommend faster routes, spot and act on global economic trends before a crisis occurs, and track the path of a disease and intervene to curb its spread before it becomes epidemic.

Viewed from this perspective, big data presents an opportunity for professionals in all fields to find new insights and answer questions that were previously beyond their reach.

Educators are just starting to appreciate the full potential of big data. For example, big data can be analyzed to create a picture of an individual learner's course of learning, not just the level of proficiency attained but the way the learner allocated his or her time and used system resources to attain that proficiency. It can also provide portraits of different learner types in a particular classroom or school or at a district, state, national, and even global level. Shared with individual learners, such findings can enhance their understanding of how they learn and where and how they could most profitably spend additional study time. The findings can give educators insight into the concepts students struggle with and individual student differences. Detailed information about variation across learners can be used to create alternative learning paths and supports that lead to more personalized learning, defined as instructional methods and pace tailored to the needs, preferences, and interests of different learners (U.S. Department of Education 2010a). Education researchers can use big data to test the applicability of principles of instruction derived from laboratory-based learning research in new, more authentic contexts and with more learners than ever before.

To help further understanding of how big data could be used to improve learning outcomes and the U.S. education system, new analytical disciplines and areas of expertise are evolving. For example, educational data mining combines conventional and new learning analytics in ways that make them useful for big data. A new type of professional is emerging as well, the educational data scientist. Lacking enough formally trained educational data scientists today, the education community is drawing on the expertise of interdisciplinary teams that often include analytics professionals from such fields as financial services or health care to fill the void.

New Players Bring New Perspectives

The growing availability and adoption of sophisticated digital learning systems are changing the nature of learning resources and who develops them, in addition to redrawing familiar development and distribution models.

For example, technology developers from disciplines other than education, such as search, gaming, mobile, and social technologies, are imagining and developing new digital learning resources that compete with print-based textbooks and other learning materials. They are working on tools for creating digital portfolios of students' work and gathering evidence of their competency attainments; establishing online repositories and communities for seeking and offering assistance with course content; creating games that engage players as they learn mathematics, science, and other subjects; and producing tools for building more authentic, engaging assessments. All these innovations are reaching wide audiences because of the power and low cost of the Internet as a distribution channel.

Educational data mining (EDM) professionals develop methods and apply techniques from statistics, machine learning, and computer science to analyze data collected during teaching and learning. EDM can be used to test learning theories and inform educational practice.

Many of these developers are new to the education market and bring a fresh enthusiasm, energy, and creativity to digital learning. They also bring R&D and evidence approaches and practices that are different from those of the established academic and government R&D communities. These should be considered in the effort to create innovative learning resources more rapidly and to expand the evidence approaches used to make decisions about which resources to adopt and how to improve them over time.

Opportunity for Expanded Approaches to Evidence

The most widely accepted model today for determining the impact of a learning resource or intervention consists of three stages of research: small investigations testing the principles behind a resource or intervention, somewhat larger studies testing its efficacy under ideal conditions, and effectiveness studies—large-scale multisite randomized controlled trials (RCTs) that test how the intervention works in the real world. Positive findings from each R&D stage are generally a prerequisite for the next.

Many academic and government education research communities consider the experimental design used in the latter two stages of the research model the only legitimate method of providing solid evidence of impact. This is largely because the experimental design involves randomly assigning study participants to test and control groups, enabling researchers to eliminate other possible explanations for observed effects. Researchers can thus conclude that the effects were caused by the intervention being tested.

Using this three-stage research model, the maturity of a learning resource, the scope at which it has been implemented, and the evidence of its impact grow together over time. Investigators commonly go through several rounds of small-scale studies over several years before concluding that a resource is ready for large-scale implementation and impact testing in an effectiveness study.

A **randomized controlled trial (RCT)** is a type of scientific experimental design. It is characterized by random assignment of study participants to a treatment group that receives the intervention being studied or a control group that experiences business as usual. If the only difference between participants in the two groups is whether or not they receive the treatment, any difference between the groups after treatment can be attributed to the treatment. RCTs are useful for ruling out competing explanations for observed effects of a given treatment.

The goal of an **efficacy study** is to test whether an intervention can produce a desired effect under ideal conditions.

The goal of an **effectiveness study** is to determine whether or not a desired effect can be produced in a range of real-world conditions.

Using an experimental model has the advantage of being able to establish a causal relationship between a practice or intervention and a learning outcome. The problem with applying it to digital learning resources, however, is that technology evolves at lightning speed. Developers cannot wait years to find out whether their products are effective: Most products would be obsolete long before the studies were completed. Similarly, education decision makers must make decisions today about whether and how to implement digital learning resources; they also cannot wait years for the results of a study.

In learning environments powered by technology, it is clear that there is both the need and the opportunity to create more and more timely guidance for developing, purchasing, and using digital learning resources. An important factor in leveraging this opportunity is accepting that the strongest level of causal evidence is not necessary for every learning resource decision.

Moreover, another important factor is accepting that a trade-off exists between having enough past use of a digital learning resource to have generated strong evidence of effectiveness and the extent to which the intervention is new and potentially transformational. If an idea has never been tried, justifying a high confidence that it will produce positive outcomes will be difficult. Yet if digital learning resources are implemented only when confidence levels are high, technology innovation will never occur in education.

Learning from Industry

To introduce innovations to users in a timely way in the commercial world, industry has evolved an R&D model in which an early-stage innovation—"a minimally viable product" (Ries 2011)—is launched and used on a massive scale, with data collection and analysis occurring simultaneously with widespread adoption rather than before.

The minimally viable product model involves specifying a product, building out its core idea and enough of its features to be useful, and deploying it to see how users react. As users engage with it, the product collects massive amounts of data about user interactions, which are then analyzed for insights into how to continuously refine and improve the product. This model transforms R&D into an iterative process with rapid design cycles and built-in feedback loops as opposed to a linear process with stages.

A **minimally viable product (MVP)** is not a minimal product but rather a model or strategy for accelerating the development of a product to shorten time to market. It is an iterative process of idea generation, prototyping, presentation, data collection, analysis, and learning. After launch at an early stage, an MVP is iterated to refine and improve it over time, based on user feedback.

When used to develop digital learning resources, this model severs the link between the maturity of an innovation and the scale at which it can be implemented and studied. The model frees early-stage digital products from having to be kept small scale. Because data collection can be embedded in the technology and data analysis can be partially automated, researchers can handle much larger datasets than was possible in the past. This enables them to ask and answer more and different types of questions about learning outcomes and how to improve the product.

This model has advantages when used to develop digital learning resources. When a resource is intended for use as part of formal education, however, educators and developers must be concerned with more than what learners do when using the product. They must also consider whether the learning demonstrated inside the product can be also observed in learners' actions *outside* the product—for example, in an independent performance assessment or in performing some new task requiring the same understanding or skill. This is necessary because while a student may demonstrate what appears to be understanding of fractions in a digital game, the student may not necessarily demonstrate that understanding in another situation. The ability to transfer what one has learned is a challenge in digital learning just as it is in face-to-face learning.

Educators and developers also need to be concerned about disentangling the multiple potential sources of observed learning differences. If the best math teachers gravitate toward a new technology-based resource for instruction, the strong performance of their students is not necessarily caused by that new resource but instead may be the result of the teachers' skills. To determine whether it enhances student outcomes, a digital learning system must be subjected to research designs in which outcome data are collected outside the system or in which other variables related to student learning, such as teacher skills, are carefully controlled.

The Role of Continuous Improvement

The underlying principle of continuous improvement is that a product or process is unlikely to be improved unless its intended outcomes can be defined and measured.

The continuous improvement process starts with identifying desired outcomes and entails collecting data on both the processes being put in place and their outcomes, interpreting those data to identify potential areas of improvement, and then trying out the revised process, collecting more data, and repeating the analysis, reflection, and refinement stages. It also involves getting a handle on costs so the ratio of outcomes to costs can be tracked over time.

Regardless of the specific tools used or the names applied to the stages of inquiry, continuous improvement processes all involve

- collaborative inquiry,
- collecting empirical data about processes and outcomes, and
- using insights gained from data to design improvements.

Technology developers with their roots in industry (which has embraced continuous improvement through such programs as Six Sigma, Kaizen, Lean, and other variations) believe that if they use processes and technologies that enable completing improvement cycles more rapidly, they can create dramatic improvements more quickly (Carlson and Wilmot 2006).

Tony Bryk, president of the Carnegie Foundation for the Advancement of Teaching, views rapid cycles of modification, analysis of results, and redesign as key to producing dramatic change while reducing risk (Bryk 2011). An alternative point of view is that continuous improvement processes can get in the way of innovation and must be put aside to create innovations that disrupt the status quo. Regardless of these different points of view and other influences, adoption of continuous improvement in education has been slow. This must change for education to benefit from new R&D approaches.

Building on Past Calls for Evidence

Efforts to produce evidence of the impact of learning resources are not new and not limited to digital learning resources. At the federal level, a focus on results dates to the passage of the *Government Performance and Results Act* and the establishment of the independent Coalition for Evidence-based Policy in the early 1990s.

The *No Child Left Behind Act,* passed in 2001, included the expectation that the increased accountability for the educational performance of all student subgroups that the law imposed would provide incentives to pay more attention to research on the effectiveness of educational practices. The need to report proficiency levels for student subgroups has led district, state, and federal education agencies to make substantial investments in student learning data systems with statewide student identifiers and information on students' demographic characteristics, achievement test scores, teachers, and grades (Data Quality Campaign 2012). For the last six years, the U.S. Department of Education has been funding states to develop such student learning data systems.

Education researchers are finding that having the ability to examine student achievement data longitudinally, they can investigate questions that had been very difficult to study, such as the long-term impact of having a poor teacher in a given grade. But there is a disconnect between what these systems can tell the researchers and what they most need to know. State and district data are collections of data on dependent variables—the outputs or effects—with little data on most of the independent variables that school systems can control—the inputs or causes.

Thus, most education data systems lack information on the nature of each student's learning experiences. Education researchers and education leaders want data that will help them go far beyond documenting whether significant gains in achievement test performance occurred to understanding how to better support learning for different kinds of learners and to identify the conditions under which particular curricula and programs are successful. Combining the data in these data systems with data from other sources will help fulfill this need.

Implications for Education Stakeholders and the Purpose of This Report

Various stakeholders in the education community have different perspectives and needs, but all share an interest in understanding how to use data, information, and evidence to address specific challenges in the U.S. education system. The opportunities created by digital learning resources and the data they produce have important implications for each stakeholder group:

- Education researchers must decide how to expand their approaches to R&D and evidence to reflect changing needs and opportunities created by technology and data.

- Developers of digital learning resources must decide how to integrate established basic research principles and learning sciences theory into their products.

- Education leaders, students, and their families must choose which of these resources to invest in.

- Teachers must decide how digital learning resources can support each and every student's learning progression.

- Funders and policymakers must determine appropriate criteria for their funding programs to leverage the opportunities offered by digital learning resources.

- Stakeholders at all levels must become both digitally and data literate in ways that are appropriate to their roles.

To address these implications, this report combines the views of education researchers, technology developers, educators, and researchers in emerging fields such as educational data mining and technology-supported evidence-centered design to present an expanded view of evidence approaches. These approaches can be used individually or in combination to design education research and gather and analyze evidence made possible by the vast amounts of data generated as teachers and students use digital learning systems.

The evidence approaches are introduced and explained in five chapters, each addressing a specific education challenge. Neither the approaches nor the challenges selected to illustrate them are meant to be exhaustive. The chapters and the challenges they address are:

Chapter 1: Making Sure Learning Resources Promote Deeper Learning

Digital learning can help meet new and more demanding expectations about what students need to learn. What can be done to ensure that technology-based resources and interventions are up to the task?

Chapter 2: Building Adaptive Learning Systems That Support Personalized Learning

Advances in technology-based learning systems enable customized strategies and content. How can the learning data that these systems collect be used to improve the systems' ability to adapt to different learners as they learn?

Chapter 3: Combining Data to Create Support Systems More Responsive to Student Needs

Young people learn and develop in a wide range of settings. How can data better be used to help support the full range of student needs and interests—both inside and outside schools and classrooms—to improve learning outcomes?

Chapter 4: Improving the Content and Process of Assessment with Technology

Digital learning systems can collect data on important qualities not captured by achievement tests. How can educators use the systems to measure more of what matters in a way that is useful for instruction?

Chapter 5: Finding Appropriate Learning Resources and Making Informed Choices

Learning resources and materials are critical in achieving desired learning outcomes. What better supports do educators need as they make decisions about which digital learning resources to adopt?

In addition, these chapters highlight six evidence approaches with great potential and on which headway is already being made:

1. ***Educational data mining and learning analytics applied to data gathered from digital learning systems implemented at scale***
2. ***Rapid A/B testing conducted with large numbers of users within digital learning systems***
3. ***Design-based implementation research supported by data gathered from digital learning systems***
4. ***Large datasets of different types from multiple sources, combined and shared across projects and organizations***
5. ***Technology-supported evidence-centered design of measures of student learning***
6. ***Data gathered from users of resources about a learning resource, how users have used it and their experiences using it.***

An Evidence Framework

This report concludes with a summary, an evidence framework designed to support evidence decision making, and recommendations that can help accelerate progress in leveraging digital learning resources and data to expand evidence approaches.

The evidence framework provides actionable information about a wide array of resources and interventions, including the development and continuous improvement of digital learning resources and the incorporation of insights based on data from digital learning systems into education more broadly. The evidence framework is intended to help education stakeholders implement a process of planning, creating, choosing, and combining appropriate evidence-gathering approaches that could be useful under different circumstances. The framework has of three components:

1. **An Evidence Reference Guide** that summarizes the six evidence approaches highlighted in this report as well as other evidence approaches widely used in education today. The Reference Guide includes the kinds of questions all the evidence approaches can help answer, the types of evidence that each can generate, and suggested uses.
2. **An Evidence Decision-Making Model** that can be used once a learning resource has been selected to make decisions about appropriate evidence approaches to use in conjunction with implementing the learning resource. The Decision-Making Model does not assume a linear staged

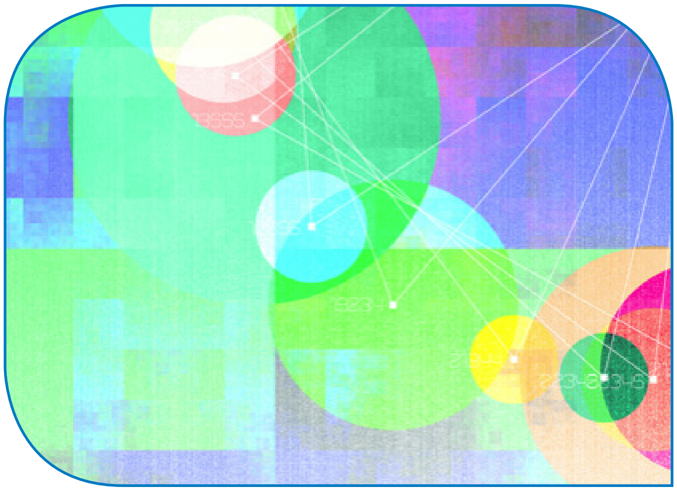
model of R&D, which ties investment in collecting evidence of impact to product maturity and widespread use. Rather, the model suggests that gathering evidence is an ongoing process that extends beyond development and implementation of a learning resource depending on the factors of confidence in the improvement potential of the learning resource and its implementation risk.

3. **Scenarios** that illustrate how and when the various evidence approaches might be applied in situations familiar to education stakeholders.

Now Is the Time

The need for expanded approaches to evidence that take advantage of and solve challenges created by digital learning resources is not new. What is new is the increased number and sophistication of digital learning resources and the vast amounts of data those systems generate while in use. Also new is the rapid rate of consumer adoption of these resources. These developments provide the opportunity to ask and answer these essential questions: What is appropriate evidence under which circumstances? How do we obtain it? How do we use it?

There is much work to do and much that education stakeholders can learn from each other to make the most of these new opportunities. Given the pace of innovation and adoption in digital learning, the time to act is now.



Chapter 1: Making Sure Learning Resources Promote Deeper Learning

Digital learning can help meet new and more demanding expectations about what students need to learn. What can be done to ensure that technology-based resources and interventions are up to the task?

Expectations about what all students should be able to understand and do are rising. In a global economy that demands innovation, people need the ability to transfer what they have learned to similar but different situations. Therefore, students today need to acquire critical thinking, problem solving, and communication competencies at levels that were expected of only the most highly educated students in past generations (Pellegrino and Hilton 2012).

Recent developments that recognize the importance of these competencies are the state-led development of Common Core State Standards (CCSS) and the Framework for K–12 Science Education from the National Research Council (NRC).

The CCSS initiative arose from a partnership between the Council of Chief State School Officers and the National Governors Association to provide a consistent, clear understanding of what students are expected to learn in K–12 English language arts and mathematics. Adopted by 45 states, the District of Columbia, and three territories, the standards encompass basic skills while raising the bar on expectations about what students need to learn to be prepared for life and work in a changing world.

In addition to the basic skills that have long been part of education standards, the CCSS for language arts require that students be able to "perform the critical reading necessary to pick carefully through the staggering amount of information available today in print and digitally." Similarly, the CCSS for mathematics stress the importance of such practices as "making sense of problems" and "constructing explanations" that students will need to be able to transfer learning to a range of content and situations.

The NRC Framework for K–12 Science Education identifies the key scientific ideas and practices all students should learn by the end of high school and calls for significant improvements in how science is taught. The overarching goal of the framework is to ensure that by the end of 12th grade, all students have an appreciation of science, the ability to discuss and think critically about science-related issues, and the skills to pursue further education and careers in science or engineering.

These new standards were crafted to reflect "deeper learning," defined by the Hewlett Foundation as the ability to acquire, apply, and expand academic content knowledge and also to think critically and solve complex problems, communicate effectively, work collaboratively, and learn how to learn. The latter aspects of deeper learning echo the business and research communities' call for "21st-century skills"—skills such as the ability to solve problems, innovate, and collaborate effectively as members of diverse, often geographically distributed teams. These skills include not only cognitive components but also noncognitive attributes such as grit, tenacity, and perseverance (U.S. Department of Education 2013).

Print-based learning materials (textbooks and worksheets), which dominated U.S. classrooms in the past, were not designed for this kind of learning. Now, a new generation of learning resources, many of them technology based, is being developed to address these more demanding standards. The quality of the resources associated with the new standards will be a major factor in determining what and how much is learned.

New Opportunities Provided by Technology

Technology provides opportunities for educators, educational publishers, and developers to address these new standards with high-quality learning resources. In addition, practices with a long history in commercial technology research and development can be brought to bear in developing these new digital learning resources. When applied to learning materials, these design, development, and improvement practices can generate evidence of both usability and effectiveness. Instead of having a linear approach, industry research and development processes are based on multiple cycles of rapid development and testing of effectiveness with constant feedback for redesign and further refinement. These processes promote both innovation and continuous improvement.[1] Given the challenge of designing resources for demanding new learning standards, developers of learning systems and resources, education researchers, and educators will need to work together with a commitment to producing better quality, more effective learning resources that support both basic skills and deeper learning.

These efforts can be aided by the data generated when students interact with digital learning systems. As students work, learning systems can capture micro-level data on their problem-solving sequences, knowledge, and strategy use, including each student's selections or inputs, the number of attempts a student makes, the number of hints and feedback given, and the time allocated across each part of the problem (U.S. Department of Education 2010a). These data can be used to inform rapid cycles of testing and refinement, provided that developers have the expertise to interpret them.

[1] Over the years, some curriculum and materials that are not technology based, such as Success for All and America's Choice, have used this iterative improvement principle, although at a far slower pace than is possible with digital learning systems.

Another advantage of digital learning systems is that they can be revised repeatedly, quickly, and economically. With the Internet as the hosting and delivery system, very little cost is associated with distributing updates and enhancements to users. Digital learning resources can thus be rolled out in more flexible ways, their effectiveness tested with existing users, and revisions made while the system is being used operationally.

The last five years have been a time of unprecedented interest in education by technology developers and venture capitalists. This interest is fueled by several factors: the availability of more powerful computers, advances in software and cloud computing, philanthropic and social business goals, and the belief that common standards could bring greater coherence to the education market. As a result, start-up companies and individuals are developing digital learning resources at a rapid pace. These new entrants in the education market bring an entrepreneurial vision driven by a desire to solve big problems quickly, venture funding, advanced programming skills, and cutting-edge data mining and analytics to the development of learning resources, opening the door to expanded approaches for gathering evidence.

Expanded Approaches for Gathering Evidence

For some years now, technology developers from industries other than education have been releasing products to users as soon as possible and then collecting and using data from the users to determine consumer preferences. Technology developers amass a large user base so they can collect and learn from data about how users respond to their product. In the commercial world, this approach can lead to faster development of better products at a lower cost.

This approach is now being extended to learning systems, with networks of teachers and/or curriculum experts providing ongoing reviews and analyses as learning system development progresses.

Educational Data Mining

As discussed, one advantage of digital learning systems is that they can collect very large amounts of data (big data) from many users quickly. As a result, they permit the use of multivariate analytic approaches (analyses of more than one statistical variable at a time) early in the life cycle of an innovation. But big data requires new forms of modeling for data that are highly interdependent (Dai 2011). Accordingly, the emerging field of educational data mining is being combined with learning analytics to apply sophisticated statistical models and machine learning techniques from such fields as finance and marketing (U.S. Department of Education 2012a).

The need for new techniques for mining data also is giving rise to a new type of professional: the learning data scientist. The field of data science emerged in the last few years, in parallel with the growth of big data. Data scientists, whose formal training may draw on computer science, modeling, statistics, analytics, and math, were first employed in marketing and finance but now have a place in education. Good learning data scientists are capable of both structuring data to answer questions and applying learning principles to select the right questions to study.

One of the key challenges of educational data mining is determining how best to parse learning interactions into right-sized components for analysis (Siemens and Baker 2012). Once the components are defined and identified, analysts can explore the records of learning interactions to find interesting patterns and relationships.

Educational data mining includes both bottom-up techniques, in which analysts look for interesting patterns in the data and then try to interpret them, and top-down approaches, with data collection and analysis shaped by a driving question or hypothesis.

Some practitioners advocate the former approach because of its ability to yield unexpected insights, but others stress the increased efficiency and interpretability of planned data collection and analyses. Most practitioners are coming to see the value of combining the two approaches.

Top-down approaches can be found in the work of both technology developers in industry and education researchers, but the two groups differ in that education researchers are more likely to be guided by concepts drawn from basic learning theory and research. In developing and studying learning technologies, education researchers often have the dual goals of creating an effective learning product and testing the applicability of a basic learning principle. Moreover, in the absence of existing empirical evidence about the effectiveness of different instructional design options, learning theory provides guidance that can increase the likelihood of making good design choices.

Learning theory is also important in the initial design of a learning technology. Without a basis in learning theory design principles, observing what students do as they move through an online curriculum is unlikely to reveal much about how to optimize learning for all students. The goal is not to find optimal pathways through bad content, but rather to design better content. The best way to achieve that initially is to draw on the extensive body of findings from learning science. Once content is improved, new technology-enabled data collection and analysis can be used both to improve the online curriculum and to test hypotheses about learning system design that extend existing research.

Uses of Evidence from Educational Data Mining

Educational data mining can address the question of how to refine a learning system or other type of learning resource and can provide the practitioner or researcher with information about learner behavior, achievement, and progression. It is less well suited to investigating the causal case for the effectiveness of a resource or intervention as a whole. However, even resources with causal evidence of effectiveness in particular settings often fail to have the same impact when applied elsewhere (Cronbach and Snow 1977). This is because education is a complex system, and any new intervention is likely to interact with different system components in a new setting in unforeseen and sometimes less effective ways. The ideal would be to have experimental tests of an intervention's impact in all the settings where it would be expected to be used. Such large-scale experiments are expensive and time consuming, however, so they are rarely done. (For an exception, see the sidebar *Scaling SimCalc and Testing the Generalizability of Measured Impacts*.)

There are two possible responses to this challenge. One is to try to create an intervention that works everywhere because all possible constraints of setting have been foreseen and accommodated. The other is to expect that an intervention will be used in somewhat different ways in different settings, possibly with different outcomes.

Rapid Random-Assignment Experiments

Another advantage of digital learning systems is that they provide an opportunity to conduct controlled random-assignment experiments (Shadish and Cook 2009) much more rapidly than was previously possible.

The purpose of randomly assigning study participants is to create two or more equivalent groups whose results can be compared. In randomized controlled trials (RCTs) in education, learners are randomly assigned to very different treatments or to an experimental treatment and a business-as-usual condition. For example, an RCT might involve one group of students taking an online algebra course and another group of students receiving face-to-face algebra instruction at school.

Scaling SimCalc and Testing the Generalizability of Measured Impacts

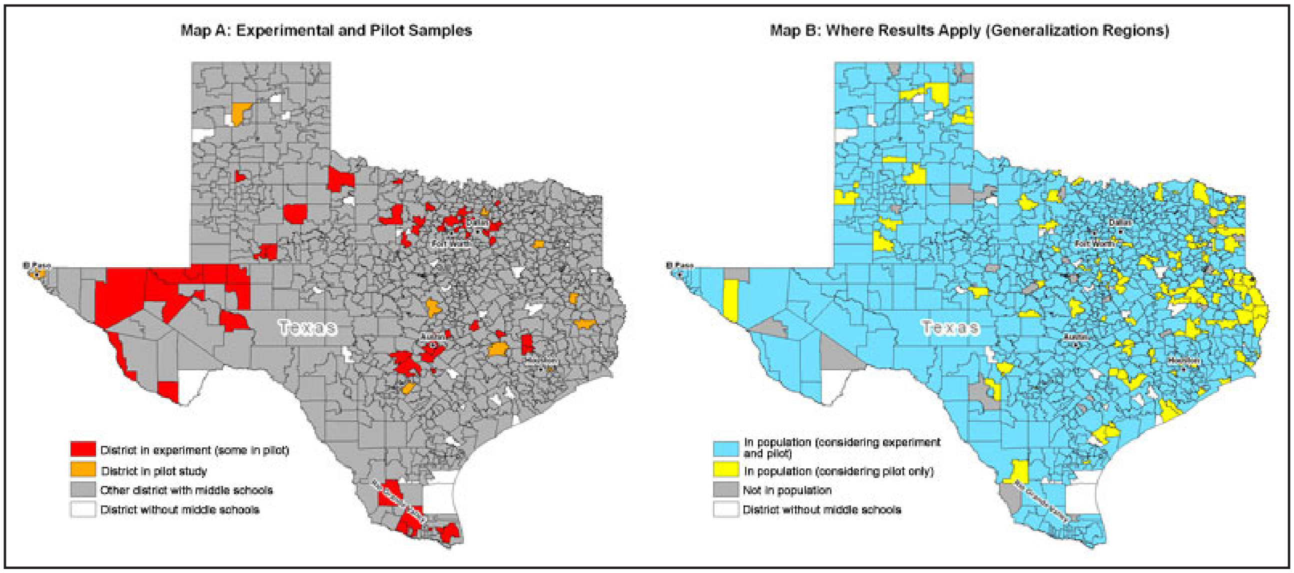
SimCalc MathWorlds® software has students create and analyze graphs that control animations of everyday experiences and things, such as a soccer game or a fish tank. Instruction is organized around having students make predictions, test those predictions using the software, and explain departures from their predicted outcomes, all supported by multiple representations (graphs, tables, equations, and animations). The mathematician Jim Kaput began SimCalc development at the University of Massachusetts, Dartmouth in the 1990s, and a number of small-scale studies had suggested that it could help low-income middle-schoolers acquire the rate-of-change concepts that form the basis for calculus. Jeremy Roschelle and his colleagues at SRI International posed the question of whether SimCalc could be effective at scale—that is, whether a large sample of typical teachers could implement SimCalc successfully—and whether positive effects would occur across variations in students, teachers, and settings.

Funding from an Interagency Educational Research Initiative grant supported a five-year test of this hypothesis in Texas. For this experiment, SimCalc was configured as a three-week software, curriculum, and teacher professional development package on the concepts of proportionality and rate of change. Random-assignment experiments were conducted with seventh- and eighth-grade teachers and their students. In all, the Scaling SimCalc Study involved over 150 teachers from 73 Texas schools. The research found that teachers assigned to the SimCalc condition spent more time teaching advanced math topics and that their students learned more, as measured by a carefully designed test of the proportionality, ratio, and rate-of-change concepts.

Randomly assigning classrooms to treatment and control conditions permits attributing observed differences to the treatment (that is, ensures internal validity), but it does not guarantee that the results are relevant to other classrooms (external validity). Educational researchers rarely discuss external validity, which is an important issue when the effect of an intervention varies and the sample participating in the experiment was not selected at random (as is nearly always the case).

Although the SimCalc research team had taken pains to recruit teachers for their study from all areas of Texas, they realized that they had not necessarily captured every kind of school context and student in the state. They worked with Larry Hedges and Elizabeth Tipton from Northwestern University to analyze the generalizability of the Texas results.

Hedges and Tipton had developed a method for quantifying a study's external validity that can be applied when good information on the characteristics of the population of interest is available. Their model estimates the proportion of a population to which research findings based on a sample can be generalized. Hedges and Tipton used information on the characteristics of the students and schools in the state of Texas as a whole and in the SimCalc study, along with propensity score matching (Rubin 1997). The images below show the results of their analysis: The SimCalc results from the research sites (left) are sufficient to generalize to the great majority of the Texas school population (right).


Map A: Experimental and Pilot Samples
El Paso
Dallas
Fort Worth
Texas
Austin
Houston
🟥 District in experiment (some in pilot)
🟧 District in pilot study
⬜ Other district with middle schools
⬜ District without middle schools
Map B: Where Results Apply (Generalization Regions)
El Paso
Dallas
Fort Worth
Texas
Austin
Houston
Rio Grande Valley
🟦 In population (considering experiment and pilot)
🟨 In population (considering pilot only)
⬜ Not in population
⬜ District without middle schools

In the software industry, a random-assignment experiment known as A/B testing is used to isolate variables by comparing two different versions of the same product or system (version A and version B) by randomly assigning users to one or the other version. One version of an online algebra course might have design feature A, for example, and the other would have design feature B, but the versions would otherwise be identical.

A/B testing allows for systematic comparison of particular features of an online system for design decisions. Two randomly assigned groups of users are given versions of the system that vary only in a defined way: One version has design feature A, and the other has design feature B, but the versions are otherwise identical. Researchers can then compare the experiences and outcomes of the two groups of users.

Historically, A/B testing has been used for market research, such as for comparing the sales or click-through results of two user interface designs or two versions of an advertisement. But increasingly it is being applied to digital learning research and development. The emergence of online learning resources that attract many users is making possible rapid collection of input on a scale that produces statistically significant results and comparison of relative outcomes from multiple versions during a short period. (For information about a project of this type, see the sidebar *Applying Multiple Forms of Evidence to Improve the Geometry Cognitive Tutor.*)

Sometimes A/B tests are conducted with a well-defined population of interest and sample participants who represent that population, as in the Geometry Cognitive Tutor example. For example, a study might assign all the eighth-grade algebra students in five school districts to take one of two forms of online eighth-grade algebra instruction with the goal of generalizing to eighth-graders in districts like the participating five.

In contrast, in an A/B test of two versions of a free online game for high-schoolers, researchers may make the game available to anyone who finds it online, with the result that they do not know anything about the characteristics of the players—their age, previous gaming experience, math concept knowledge, and so on. Developers of digital learning systems may not ask their users to provide any information about themselves because they do not want to discourage potential users with a sign-in process. In addition, they argue that the larger the pool of users, the less the importance of specific users' characteristics. The Khan Academy, for example, reports that it attracts enough users to run an adequately powered A/B test in a matter of hours, and typically it does so without collecting user information (See the sidebar on *A/B Testing and Rapid Improvement Cycles at the Khan Academy.*)

In A/B tests involving smaller groups of students, characteristics and prior achievement matter more, but rapid RCTs are still possible. At the Center for Advanced Technology in Schools at the University of California, Los Angeles, for instance, researchers ran 20 RCTs over an 18-month period to test various theory-driven hypotheses about learning game design. (See the sidebar *A/B Testing Using Samples with Known Characteristics.*)

Applying Multiple Forms of Evidence to Improve the Geometry Cognitive Tutor

This case illustrates the application of psychological principles to software design, design research, and A/B testing; subsequent commercial software development; and data mining to determine the effect of product refinement on the commercial software.

The Geometry Cognitive Tutor from Carnegie Learning has a number of exercises requiring students to calculate angle measures within a diagram using their knowledge of geometric theorems. In earlier versions of the program, the diagram was a static element. Students looked for angle relationships in it and entered angle values, along with the theorems leading to those values, in a table separate from the diagram (Aleven and Koedinger 2002).

Butcher and Aleven (2008) recognized that presenting the table and the diagram separately appeared to impose an additional, extraneous cognitive load on students, perhaps resulting in suboptimal learning. The underlying psychological principles have been described as the split-attention effect (Kalyuga, Chandler and Sweller 1999) and the contiguity principle (Mayer 1989). Butcher and Aleven conducted a series of design experiments applying these principles to the software. The result was a new version incorporating an interactive diagram in which the students entered calculated angles (and reasons for the calculations) directly in the diagram.

Butcher and Aleven used an A/B test to compare student performance between the "table interaction" and "diagram interaction" versions of the Geometry Cognitive Tutor. Tests immediately after use of the software favored the diagram interaction version but only for transfer items (which asked whether a particular angle could be calculated from the diagram, a kind of question that was not included in any of the tutored exercises). Delayed posttests indicated that students using the diagram interaction version better retained their knowledge of how to use geometric theorems to figure out angle values.

After the results of the A/B test were known, Carnegie Learning implemented the interactive diagram version of the Geometry Cognitive Tutor. Although the commercial version differed in some ways from the exact implementation that Butcher and Aleven had used in their research, Carnegie Learning attempted to preserve the educationally important aspects of the new design. Hausmann and Vuong (2012) compared data from students using the commercial table interaction version of the Geometry Cognitive Tutor and from those using the diagram interaction version. They found that students using the diagram interaction version were able to reach mastery in a shorter time than those using the table interaction version. The advantage was particularly strong for difficult steps in the problem.

A/B Testing and Rapid Improvement Cycles at the Khan Academy

The Khan Academy has grown from a collection of a few hundred YouTube videos on a range of math problem types created by Sal Khan himself to a digital learning system incorporating more than 3,000 videos and 300 problem sets geared to K–12 mathematics topics.

For each problem set, the Khan Academy system logs the number of attempts a user makes for each problem, the content of each answer, whether the answer was correct or not, and whether the system judged that the user had mastered the skill the problem set addressed. For each video, the system keeps track of the segment being used, the time when the user's viewing started and ended and any pauses or rewinding.

As an organization, Khan Academy combines technology research and development approaches with Wall Street-style financial analysis. Its Dean of Analytics, Jace Kohlmeier, was previously a trading systems developer at a hedge fund.

Khan Academy's open-source A/B testing framework enables the organization to randomly assign users to one of two or more versions of the software with one line of code. Developers can determine what percentage of their users they want to receive the experimental version, and a dashboard charts user statistics from the two treatment groups in real time.

Because Khan Academy has about 50,000 active exercise users doing several million problems each day, developers can accrue statistically significant data very quickly. For something with a large impact, Kohlmeier reported they can collect results in an hour (because large effects can be detected with small samples). But many of the Khan Academy's experiments involve changes with smaller effects and hence take longer. In addition, the organization likes to run experiments for a week or so because of user flow cycles; more adult and self-driven learners use Khan Academy in evenings and on weekends.

One of Kohlmeier's first projects with Khan Academy was to look at how the system determined that a learner had reached proficiency on a problem set topic. The system was using a simple but arbitrary heuristic: If the user got 10 problems in a row correct, the system decided the user had mastered the topic. Kohlmeier examined the proficiency data and found that the pattern of correct/incorrect answers was important. Learners who got the first 10 problems in an exercise set correct performed differently subsequently than did users who needed 30–40 problems to get a streak of 10.

Kohlmeier built a predictive model based on estimating the likelihood at any point during an exercise set that the next response would be correct. (Similar predictive models have been used in intelligent tutoring systems for some time.) The system was then changed to define mastery of a problem set as the point where a user has a 94 percent likelihood of getting the next problem correct.

This change in the system set a higher bar for mastery and meant that some users had to spend more time on an exercise set. By monitoring user data after making the change, Khan Academy analysts were able to see that users were willing to devote the extra effort. At the same time, the new criterion allowed fast learners to gain credit for mastering material after doing as few as five problems, enabling them to cover more material in a given time. The Khan Academy team used A/B testing to compare the old and the new models for determining mastery. They found that the new mastery model was superior in terms of number of proficiencies earned per user, number of problems required to earn those proficiencies, and number of exercise sets attempted.

Although a great proponent of A/B testing and data mining, Kohlmeier is also aware of the limitations of those approaches. It is difficult to use A/B testing to guide big changes, such as a major user interface redesign; too many interdependent changes are involved to test each possible combination in a separate experiment. In addition, system data mining is extremely helpful in system improvement, but to make sure the system is really effective, analysts need an external measure of learning.

A/B Testing Using Samples with Known Characteristics

With funding from the Institute of Education Sciences, the University of California, Los Angeles (UCLA) Center for Advanced Technology in Schools (CATS), under the leadership of Eva Baker, Greg Chung, and Keith Holyoak, has been conducting research and development on online games for middle school mathematics. The goal is to teach middle school math concepts (rational numbers, functions, and systems of equations) through online games that are both enjoyable and effective as tools for learning.

CATS game designers had a set of design principles for establishing a narrative, creating a playful environment, and providing different levels of challenge and reward. Yet they were not accustomed to thinking about how to design a game that would support academic learning that would carry over into what students do outside the game. The CATS education researchers asked their software developers to build the games in a way that would maximize flexibility, making it possible to manipulate and test various features as they went along.

The basic mechanics of the game were held constant, but game level setting and features of the user interaction were varied, with students assigned at random to play different variants of the game. CATS conducted A/B testing on 10 different variations of the game in a series of experiments over 18 months. Most of these studies involved 100–200 students. The different experiments tested variations in feedback and instruction, the incorporation of self-assessment, different scoring systems, the incorporation of collaboration, and different narrative structures.

Through their prior experience studying technology-based education interventions, the CATS researchers were very cognizant of the variations in hardware, teachers, network, and security across classrooms that could make testing the games difficult. To deal with these challenges, they bought a laptop cart they could move from school to school for data collection. Studies were run predominantly with students in math classes in grades 6–9 in urban schools with an ethnically diverse student body drawn from middle- to low-socioeconomic-status areas.

Having set up the game test bed, the CATS team collected experimental data very rapidly, with each experiment conducted over one week. Analyzing the data was more time consuming. In addition to analyses of variance and covariance, the researchers undertook some exploratory data mining. They looked for interesting clusters of behavior that might reflect concepts in the learning research literature.

Some of the A/B test findings were surprising. The addition of a more elaborate narrative, for example, increased students' enjoyment of the game but had no effect on learning. The research team used the insights gained through A/B testing to refine the games and the associated teacher professional development for use in an RCT with 80 classrooms conducted in 2012.

Uses of Evidence from Random-Assignment Experiments

Experiments with random assignment of students, classrooms, or schools to conditions using a new learning technology and some other approach, whether a variation on that technology or business as usual, provide the strongest demonstration that the innovation produces the observed outcome in that specific instance (Baron 2007). Because researchers widely view RCTs as generating the highest quality evidence, technology-enabled rapid RCTs may have especially broad appeal.

But random assignment is not always feasible. In those cases, quasi-experimental designs with statistical control for any preexisting group differences, regression discontinuity, and interrupted time series designs are useful tests of effectiveness.

Frank et al. (2011) noted that nearly all social science research designs are subject to potential bias, on the basis of either nonrandom sampling of participants, as is the case with most random-assignment experiments (albeit not with some A/B testing), or nonrandom assignment of participants to conditions, which is usually the case with quasi-experiments, regression discontinuity, and interrupted time series designs. In the latter case, some unknown, uncontrolled variable related to the study outcome could contribute to effects.

Frank et al. (2011) have developed a statistical technique for quantifying the amount of bias that would have to be present to invalidate the conclusion of either type of design. In several applications of their technique, Frank and his colleagues found that the amount of bias would have to be very substantial—in fact, larger than that which would have to be present in the random-assignment experiment they analyzed as a contrasting case. Consequently, they argue that if random-assignment experiments are the gold standard for establishing causal relationships, quasi-experimental designs with measurement and control for any preexisting group differences known to influence the outcome variable should be considered the silver standard.

When a learning technology produces huge effects (such as equivalent learning outcomes in half the time in the OLI statistics course documented in a 2008 study by Lovett, Meyer and Thille (2008), there are few credible competing explanations. When researchers do not need to rule out credible competing explanations, random assignment may not be necessary.

Design-Based Implementation Research

Educational data mining and rapid A/B testing can produce information for refining and enhancing digital learning systems, but they are less than ideal for answering questions about how digital learning systems are being used in different contexts and how implementation variations relate to differences in outcomes. An emerging research approach that is suited for this kind of inquiry is design-based implementation research (DBIR).

DBIR is an approach for investigating learning outcomes and implementation in tandem. It seeks to change the relationship between research and practice so that interventions are designed from the start with their ultimate uses in mind and are based on theories and methods from both the learning sciences and policy research. Penuel et al. (2011) articulated four core DBIR principles: focus on a persistent problem of practice; commitment to iterative, collaborative design; concern with developing theory and knowledge concerning both classroom learning and implementation processes; and concern with developing capacity for creating sustainable education system change.

Design-based implementation research (DBIR) is an emerging education research and development approach for contributing to the design or refinement of educational interventions that are usable, scalable, and sustainable (Penuel et al. 2011). DBIR was developed in response to concern that research-based educational interventions rarely are translated into widespread practice and that studies of interventions in practice put too much emphasis on implementation fidelity and not enough on understanding intervention adaptation.

With its roots in several decades of design research (Kelly, Lesh and Baek 2008), DBIR calls for sustained partnerships between developers, education researchers, and practitioners who jointly select a problem to work on and engage in multiple cycles of design and implementation decisions with data collection and analysis embedded in each cycle so that implementation can be refined based on evidence (Penuel et al. 2011).

DBIR is a complement to such techniques as educational data mining and A/B testing. One of its strengths—and a feature that the other two approaches lack—is the collection of information on what learners and their teachers, peers, and others in their environments are seeking to accomplish and what they are doing before, after, and during learning sessions. When the learning session includes digital interaction, the digital learning system can collect data automatically, and those data can be combined with the knowledge collected by practitioners or researchers in the offline world for a more complete picture.

The *Sara Solves It* video series for preschools is an example of how DBIR combines foundational research and observations of implementation to rapidly develop and improve a learning resource. (See the sidebar *Implementation Research and Rapid Prototyping of Digital Resources for Sara Solves It.)*

Another example of the complementarity of contextual and learning system data comes from the work of Carnegie Learning, a publisher of math curricula for middle school, high school, and postsecondary students. A school was using its tutoring system as part of a mandated school improvement effort. Examining data collected automatically by the tutoring system,

Implementation Research and Rapid Prototyping of Digital Resources for *Sara Solves It*

Preschool educators today are placing increasing emphasis on supporting the development of the foundational concepts for academic learning (e.g., National Research Council and Institute of Medicine 2000). Developers of public television content for children have found that having engaging characters that appear not just in a television show, but also in other media such as educational games and classroom digital activities can increase children's engagement, complement classroom activities, and facilitate learning (Linebarger, Taylor-Piotrowski and Vaala 2007; McManis and Gunnewig 2012).

With the goal of designing digital activities that enhance mathematics learning for preschoolers, educational content developers at WGBH teamed with researchers from EDC and SRI International in a design and development effort supported by the National Science Foundation. This effort, which uses creative assets from WGBH and Out of the Blue Enterprise's separate effort to develop the preschool mathematics television show *Sara Solves It*, illustrates the back-and-forth iteration that is characteristic of designed-based research and development.

EDC and SRI began with a thorough review of the learning research on the emergence of mathematical thinking in young children. From this review, they identified a set of key learning goals, such as subitizing—the ability to look at a set containing a small number of objects and automatically recognize the number as 1, 2, 3, 4, and so on. For each learning goal, the researchers identified associated knowledge and skills and a rationale for including it in the materials that WGBH would develop, including games designed for digital tablets.

With an initial set of requirements grounded in academic research, the specifications for common (nondigital) activities and complementary tablet-based games and digital activities were then developed in collaboration with the WGBH game designers. Research on subitizing suggested that young children would attend to sets of objects on a computer screen for only two seconds (Clements 1999). WGBH game designers insisted that to make a game that young children would find entertaining, the object sets would have to move and that if they did move, children would watch them for longer than two seconds. The team agreed to try this approach.

SRI researcher Phil Vahey reported that, "WGBH comes up with great game designs that are much more sophisticated and more on target as to what children will find fun than is typical of academic researchers. They incorporate engaging features and have graphics and game mechanics of much higher quality."

Once there is an agreed-on set of game requirements, WGBH developers produce a rough initial prototype based on characters from *Sara Solves It*. WGBH tries out the prototype with a few children in the Boston area while EDC and SRI each use it with five children in the New York City and San Francisco areas, respectively. EDC and SRI report their observations to WGBH, emphasizing their insights into whether students are actually learning math concepts in playing the game. A week later, WGBH provides the research teams with a revised version of the game, which the various organizations then try out with five more students.

On the basis of these first tryouts, the game developers undertake another round of revisions.

Carnegie Learning analysts could see that students in most classes were progressing as expected but that students in one class had stopped making gains midyear. When they brought this pattern of data to the attention of the school principal, they learned that the class that had stalled had lost its regular teacher and was being handled by a substitute. Seeing the data from the tutoring system, the principal realized that students in this class were suffering and decided that the plan to delay hiring a replacement teacher had to be changed as quickly as possible (Ritter 2012).

A relatively mature example of DBIR principles is the work of the Pathways project being led by the Carnegie Foundation for the Advancement of Teaching. This example illustrates the importance of implementation research in improving not just the design of a course with a strong technology component, but also the institutional practices in its implementation. The primary goal of this work is to improve outcomes for developmental mathematics students in community colleges in terms of entry into and success in college-level mathematics courses—a broader, more consequential objective than demonstrating that the online course per se produces mathematics learning. (For more information on the Pathways project, see the sidebar *Collaborative Research and Development on the Pathway to College Math*.)

Collaborative Research and Development on the Pathway to College Math

When researcher Tony Bryk became President of the Carnegie Foundation for the Advancement of Teaching, he wanted to increase the impact of education research (Bryk, Gomez and Grunow 2011). He and his colleagues argued that research should focus on a "persistent problem of practice" that, if solved, could have significant benefits for the education system. Bryk and colleagues often refer to these as high-leverage problems.

One such persistent problem is the developmental mathematics courses that students must take if they enter college not yet ready for college-level math. Many students believe developmental math courses just repeat their high school math experience. At some colleges, the lowest scoring students are required to pass as many as three semesters of developmental math courses (for which they do not earn credit) before being allowed to take credit-bearing college math courses. Not surprisingly, as many as 70 percent of these students become discouraged and fail to complete all the required developmental math courses. Without completing these requirements, they cannot earn a degree.

The Carnegie team defined its goal as doubling the number of students who earn college math credit within one year of continuous enrollment. To achieve this goal, the Carnegie team set out to collaborate with college administrators and instructors to redesign their approach to developmental mathematics by developing new courses and associated polices and then improving the new courses and practices by analyzing system data and feedback from implementation. They recognized that such an effort would need a collaborative community and an infrastructure to support its success.

Community and four-year colleges were invited to participate in a networked improvement community (NIC) for developmental math. A NIC is a group of people from multiple organizations committed to working together on a complex, high-leverage problem with a concrete target and a shared set of inquiry practices, including using what they build. The colleges that Carnegie convened agreed to collaborate with other colleges and with researchers and developers to implement the resulting new developmental math curriculum with their students, share data from their implementation, and participate in discussing implementation data and planning refinements. NIC participants recognized that as their work unfolded, new aspects of problems would become visible, and the NIC colleges found themselves working on emergent issues such as student engagement and persistence and the elimination of language that is a barrier to mathematics learning.

Collaborative Research and Development on the Pathway to College Math (Continued)

One of the important NIC activities was an analysis of the causes of the high failure rate for developmental math at their institutions. The collaborators found that many students were lost at the transition between multiple courses in a series, that the developmental math courses were not engaging, that many students had negative beliefs and attitudes about their ability to do math, and that many students' ties to peers, faculty, and programs of study were weak. Among the strategies that the group decided to apply to address these issues was consolidation of what had been multiple math courses into a single course emphasizing real-world problems from statistics. The Pathways project has worked on two courses: Statway, which deals with developmental math content in the context of statistics, and, more recently, Quantway, a course on quantitative reasoning and literacy.

The Statway development process illustrates how educators, developers, and researchers can collaborate to iteratively co-design a new intervention. A small group of academic researchers and curriculum developers produced the initial version of Statway. Community college faculty reviewed this initial version and informally tried out some of the lessons from it with their students in fall 2010. Ongoing conversations among researchers, course designers, and math faculty led to the conclusion that this first version needed a major reworking. A team of college math faculty members was brought to Carnegie to redesign the course, and the result was Statway Version 1.5, which was pilot-tested NIC-wide in school year 2011–12.

Statway uses the OLI course engine to support its homework platform. This course engine made it possible to obtain detailed learning data on students' engagement with individual problems and their persistence through the problem sets. Louis M. Gomez, a Learning Sciences professor at UCLA and Statway collaborator, expects that these data will enable the NIC to explore how various practices (implementation and context variables) make a difference in Statway outcomes and whether they vary by local setting.

When asked whether the Pathways project had conducted an efficacy study comparing Statway results with those for conventional developmental math sequences, Gomez explained,

> *We haven't done an experiment on Statway versus business as usual at a community college. Right now our goal is to improve Statway and have it be executed reliably in the variety of contexts that make up the NIC. We need to do more than convince ourselves that it works. All kinds of promising interventions are subjected to RCTs that show nothing; often because they're subjected to [experimental studies] too early. Equally important to work on is getting your intervention to work reliably across many different contexts. This is more important at this point than understanding whether Statway works better or worse than some other approach.*

Gomez pointed out that he does not view comparative experimental research as "wrong" but useful for answering a different kind of question. Having defined its task as improving rates of successful completion of developmental mathematics, the Carnegie team is more focused on understanding how to get Statway to produce this outcome in a range of college contexts (external validity) than on comparing it with alternative approaches in an experimental design (internal validity).

Gomez's colleague Paul LeMahieu noted that in the first year of Statway implementation, three times as many students earned a college math credit in one-third the time compared with historical averages at the participating colleges.

Tony Bryk, President of the Carnegie Foundation for the Advancement of Teaching, launched the Pathways project in part to create a concrete example of how education research can lead to educational improvement. Similar to technology developers from industry, who embrace continuous improvement, Bryk views rapid cycles of modification, analysis of results, and redesign as key to improvement. Bryk argues that improvement research should be structured as many rapid iterations of small changes, what he calls "rapid iterative small tests of change." His reasoning is that small changes can be implemented quickly, can be tested repeatedly in multiple contexts to make sure they are really improvements, and are unlikely to do harm (thus managing the risk associated with failure). By implementing many iterations in a short time, research collaborations can produce dramatic change through the accumulation of many small improvements.

Bryk further argues that traditional large-scale education research is most useful in few circumstances. He characterizes the research space in terms of three dimensions: confidence that a proposed change will lead to improvement (high or low); risk, or cost of failure (large or small); and the current situation with respect to stakeholders' receptivity to the change (resistant, indifferent, ready). Of the 12 possible combinations of these dimensions, in Bryk's view only two combinations (high confidence, indifferent audience, small cost; and high confidence, ready audience, and large cost) warrant a large-scale formal study (Bryk 2011).

Uses of Evidence from Implementation Research

DBIR proponents work with their practitioner partners to lay out a theory of the implementation steps needed in the practitioners' context and study the implementation processes and outcomes simultaneously. The evidence they typically seek is correlational patterns, and they use quasi-experimental designs rather than RCTs, though some DBIR studies include experimental tests of different strategies for supporting implementation.

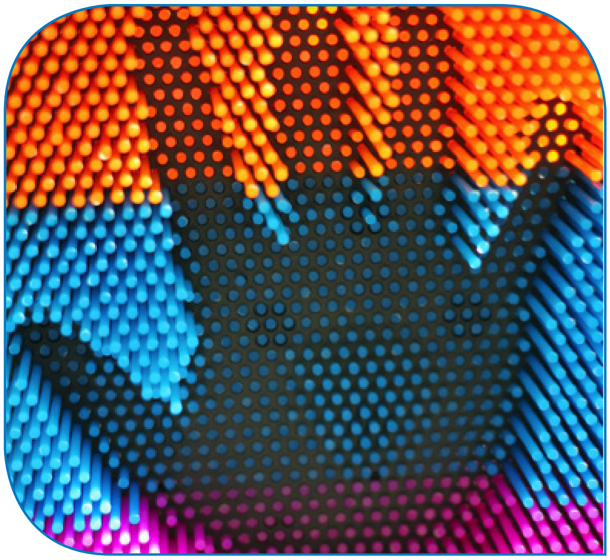
The Pathways project (described in the sidebar *Collaborative Research and Development on the Pathway to College Math*) has emphasized investigation of the relationships between specific changes in practices and changes in student completion rates for the developmental math sequence. The lack of alternative plausible explanations for dramatic changes in an outcome (in the Pathways project, dramatic differences from historical rates in the numbers of students qualifying for college-level mathematics by their second year of college) gives some credence to causal inferences, even in the absence of a random-assignment experiment. In some examples of DBIR, alternative plausible explanations exist for observed differences, and important decisions hang in the balance, making it appropriate to incorporate experimental studies into DBIR.

For example, a district implementing a new technology-based reading program might first ask the program developer for evidence that could assist in implementation. Then the district could determine whether it should invest in an expensive teacher professional development program offered by the technology developer to go along with a digital learning system. Interpreting a natural experiment, such as one that compared outcomes of students of teachers who chose to participate in the professional development with those of teachers who did not, would be difficult. This is because teachers who choose to participate in optional professional development activities may be more conscientious than other teachers or less adept with technology or more uncertain about their teaching skills. Any of these variables could influence student outcomes independently of the teacher professional development. In such a situation, an experimental design with teachers assigned randomly to mandatory professional development or to implement the digital learning system without the professional development would be the best way to determine the value of the teacher training experiences.

Whether or not they incorporate experimental designs, a hoped-for benefit of DBIR collaborations is that education practitioners will think about their activities as cycles of implementation, data collection, reflection, and refinement and constantly seek data and information to inform their practice. Classrooms, schools, and districts are not likely to launch a program of massive experimental research for its own sake, but they might seek university or other research partners when planning the implementation of major new initiatives. The network of colleges working with the Carnegie Foundation for the Advancement of Teaching in the Pathways project illustrates this approach. Key to this process is the collection of objective data on student learning.

Conclusion

This chapter describes some of the emerging approaches to collecting evidence of the effectiveness of a digital learning system and discusses the strengths and weaknesses of these approaches relative to those of other education research designs. Internet distribution of digital learning resources enables widespread use early in a product's life cycle, and data mining and A/B testing techniques generate massive amounts of data that can be used in rapid cycles of product improvement. There are limits to what can be learned solely on the basis of data captured within an online system, however. Experimental designs, including measures of the target learning outcomes external to the digital learning system, remain an important research tool as are studies examining the implementation of digital learning resources in different contexts.



Chapter 2: Building Adaptive Learning Systems That Support Personalized Learning

Advances in technology-based learning systems enable customized strategies and content. How can the learning data these systems collect be used to improve the systems' ability to adapt to different learners as they learn?

Adaptive instruction is not new. A form of it has existed since the days of Socrates. Since at least the 1980s, education researchers have viewed adapting instruction to students as a major factor in successful learning (Corno and Snow 1986). By that time, research had demonstrated the power of one-on-one tutoring, in which the tutor adapts learning experiences and the time provided for learning to the needs of the individual student (Bloom 1984).

Digital learning systems are considered adaptive when they can dynamically change to better suit the learner in response to information collected during the course of learning rather than on the basis of preexisting information such as a learner's gender, age, or achievement test score. Adaptive learning systems use information gained as the learner works with them to vary such features as the way a concept is represented, its difficulty, the sequencing of problems or tasks, and the nature of hints and feedback provided.

Adaptive instruction is related to individualized, differentiated, and personalized learning. Minimally adaptive learning systems offer individualized pacing, whereas more sophisticated systems differentiate the nature of learning activities based on student responses. Systems are now being developed to support personalized learning by incorporating options for varied learning objectives and content as well as method and pacing of instruction.

Although one-on-one sessions with a skilled human tutor who dynamically understands and responds to the person being tutored offer the most personalized experience, digital learning systems have advanced greatly in their ability to model the knowledge and competencies students should acquire and to diagnose and respond dynamically to learner needs. Good teachers are constantly assessing their students' understanding and level of engagement so that they can customize strategies and content for different students, although this is difficult to do for every individual student.

Individualized, Differentiated, and Personalized Instruction

Individualization, differentiation, and personalization have become buzzwords in education, but little agreement exists on what exactly they mean beyond the broad concept that each is an alternative to the one-size-fits-all model of teaching and learning. For example, some education professionals use personalization to mean that students are given the choice of what and how they learn according to their interests; others use it to suggest that instruction is paced differently for different students. In this report, we use the definitions from the National Education Technology Plan (U.S. Department of Education 2010a):

Individualization refers to instruction that is paced to the learning needs of different learners. Learning goals are the same for all students, but students can progress through the material at different speeds according to their learning needs. Students might take longer to progress through a given topic, skip topics that cover information they already know, or repeat topics they need more help on.

Differentiation refers to instruction that is tailored to the way different learners learn. Learning goals are the same for all students, but the method or approach of instruction varies according to the preferences of each student or what research has found works best for students like them.

Personalization refers to instruction that is paced to learning needs, tailored to learning preferences, and tailored to the specific interests of different learners. In an environment that is fully personalized, the learning objectives and content as well as the method and pace may all vary. Thus, personalization encompasses differentiation and individualization.

Teachers tend to vary learning approaches between classrooms serving students with different levels of prior achievement (Oakes 2005). Differentiating teaching within a classroom requires considerable effort and skill on the part of teachers and also a wide variety of resources spanning different levels of difficulty. When differentiation does occur *within* a classroom, it typically involves separating students into two or three groups based on skill fluency or degree of prior knowledge (Fuchs, Fuchs and Vaughn 2008).

In computer-based instruction, adapting the pace of introducing new material to individual learners began in the 1980s. Such mastery-based learning approaches were common in the learning systems many school districts used in those years with low-achieving students or students at risk. These systems provided instruction on sequences of skills, with the requirement that each student master a given skill before working on the next one. Although they adapted the amount of time a student spent learning material to the individual student's needs, these mastery learning programs still exposed all students to the same material presented in the same way.

New Opportunities Provided by Technology

Advances in technology have heightened the possibility that digital learning systems can replicate dynamic adaptations used successfully by human tutors or even implement those and other methods more effectively than humans. In fact, studies have shown that students taught by carefully designed systems used in combination with classroom teaching can learn faster and translate their learning into improved performance relative to students receiving conventional classroom instruction (Koedinger and Corbett 2006).

Capabilities now available in newer and more sophisticated digital learning systems include

- dynamically updated fine-grained modeling of learner knowledge that can be compared to a knowledge model of the concepts to be learned;
- micro-level tagging of instructional content, along with micro-level capture of learner actions within adaptive systems; and
- adaptations based on students' emotional states and levels of motivation.

For an example of a tutoring system that outperforms human tutors, see the sidebar *DARPA Develops a Digital Tutor to Train Navy IT Specialists*.

DARPA Develops a Digital Tutor to Train Navy IT Specialists

The Defense Advanced Research Projects Agency (DARPA) funded the development of a digital tutor to train information technology (IT) specialists in the U.S. Navy. When IT issues arise aboard a ship that cannot be resolved locally, the Navy incurs costs and loses time. Historically, training new IT specialists to the level of expertise to solve the Navy's more difficult IT challenges had required elite instructors, significant classroom time, and a few years' experience on the job. Using this model, the Navy was unable to train enough new IT specialists or train them quickly enough to the desired level of expertise. The Navy sought a digital tutor that would close the gap between the IT training goals and what the expert-led, classroom-based training could achieve.

First, the Navy designed and tested a new face-to-face instruction model on which the digital tutor would be based, aimed at realizing better training outcomes in less time compared with the then-current training model. This program was tested on a very small scale, with approximately 24 top experts training 15 new students mainly through one-on-one tutoring. The program was refined until its graduates could outperform fleet experts with an average of seven years' experience.

Once the team had achieved the sought-after training outcomes in the face-to-face tutoring, it developed the Digital Tutor (DT), which uses artificial intelligence to mimic the behaviors of the program's exceptional human tutors. The Digital Tutor was then used to train new students. In a series of tests conducted by the Institute for Defense Analyses (IDA), students who had completed one quarter of DT training (4 weeks of the 16-week program) outperformed not only students from the traditional training program, but also the instructors of those courses.

Students who had completed the 16-week DT program outperformed both graduates of the traditional 35-week IT training program and fleet IT experts with an average of 9.1 years' experience in a series of practical exercises, network-building tasks, and interviews conducted by a Review Board. They also performed better than graduates of the face-to-face tutoring program, but the difference was not statistically significant.

Compared with graduates of the traditional training program and fleet IT experts, the DT graduates successfully solved more problems and solved them more efficiently (were less likely to use unnecessary steps) and more securely (were less likely to cause harm or compromise the system). Of the three study groups, only DT graduates solved any of the problems with the highest difficulty rating (Fletcher 2011; Fletcher and Morrison 2012). Based on these assessment results, a 2012 report from IDA estimates that the "greater efficiency, absence of harmful errors, and ability to solve problems at the highest level of difficulty demonstrated by Digital Tutor students suggest both monetary and operational returns of substantial value to the Navy" (Fletcher and Morrison 2012, p. v).

Dynamically Updated Learner Models

Newer digital learning systems use artificial intelligence to go beyond a behavioral definition of mastery (e.g., whether a student responds correctly or incorrectly) to incorporate detailed cognitive models of the knowledge to be learned (Falmagne et al. 1990; Ritter et al. 2007). These systems base adaptations not just on whether a student responds correctly or incorrectly, but also on a model of the student's thinking compared with a target knowledge model (the domain model) with the goal of closing the gap. For example, instead of monitoring mastery of large topics such as "solving equations," new systems can monitor more fine-grained skills such as "solving an equation of the form –x = a." This makes learning more efficient.

These systems constantly update the model of a student's thinking as the student works with the system. On the basis of the learner model, the system adapts instruction, varying the pace of learning and the instructional content and methods. Such systems also can present explanations, hints, examples, demonstrations, and practice problems as needed by an individual learner and then reassess the student's understanding (Pellegrino, Chudowsky and Glaser 2001).

Micro-Level Data Capture Techniques

The increasingly sophisticated algorithms that power the adaptive capabilities of digital learning systems are capable of proposing ever finer adaptations. An example is the adaptive learning software from Knewton, a start-up that partners with publishers to offer course content on an adaptive platform. By tagging the content and tracking students' interactions with the content at a micro level, Knewton collects hundreds of thousands of data points per student per day. The Knewton software uses the micro data to improve its ability to adapt to different learners.

Knewton representatives explain that as the system learns how individual students learn—for example, what types of explanations they respond to best or what time of day they learn certain types of concepts more quickly—it becomes more efficient at presenting content in the way most likely to support a particular student's learning. Depending on how a student interacts with it, Knewton may provide text in shorter or longer versions and at greater or less complexity, offer more or fewer practice problems, and offer more textbook-like or more game-like modules (West et al. 2012)

As a student takes more courses in the Knewton platform, the system aggregates data about that student across those courses. Similarly, as more students take courses in Knewton, data mining will reveal patterns among students, with the promise of providing insights into students with a variety of characteristics. Knewton draws on a large population of students to do this; it expects 10 million enrollments in 2013 in courses offered through its largest partner, Pearson (West et al. 2012).

Adaptations Informed by Motivational and Affective Factors

Another example of groundbreaking work in building adaptive learning systems involves measuring and responding to motivational and affective factors as students work with digital learning systems. A team at the University of Massachusetts is combining data from sensors that detect learners' facial expressions and physical activity with data from the intelligent tutoring system Wayang Outpost to identify in real time whether a learner is feeling excited, confident, frustrated, or bored. The team has designed software characters or agents that behave differently depending on the learner's emotional state. This system adapts dynamically and can respond differentially to the same student at different times depending on his or her current emotional state. (See the sidebar on *Exploring the Role of Students' Emotions in Learning*.)

Exploring the Role of Students' Emotions in Learning

At the University of Massachusetts, Beverly Woolf and Ivon Arroyo have been using their intelligent tutoring system for geometry and statistics, Wayang Outpost, as a test bed for investigating the role of students' emotions or affect in learning. Extensive research has shown a relationship between students' conception of intelligence as fixed or expandable and how they view success and failure as having an influence on the learning challenges they will seek. It is also well established that a state of modest alertness (what psychologists refer to as arousal) enhances learning and that students tend to learn better when they feel an emotional closeness to their instructor.

The University of Massachusetts team wanted to see if they could make an intelligent tutoring system more effective by making it adapt to the student's emotional state. They assumed that a student's affect is dynamic, potentially changing over time as he or she works with the online learning system. One of the first challenges was conceptualizing the relevant aspects of student emotions and then determining ways they could be measured as students are learning on Wayang Outpost.

Studies with trained human observers watching students working on Wayang Outpost found that observing students having positive or negative feelings is possible, as is discerning students' arousal as revealed by physical activity, such as looking around the room instead of at the computer screen. Raters' judgments of students' emotions were correlated with how much mathematics students learned and with their responses to an attitude survey taken after the intervention was completed.

Further work involved developing sensors to detect students' facial expressions, movement in their chairs, the pressure they exerted on the computer mouse, and skin conductance (which varies with moisture level and is used in psychological studies as a measure of arousal). For each student, researchers combined data from these sensors with various types of data from the tutoring system, such as time spent on each problem, number of hints requested, and correct solutions. Machine learning techniques were used to discover how combinations of these online learning behaviors and sensor data related to student attitudes toward learning and toward math as indicated on post-intervention surveys (Arroyo et al. 2009). Once a predictive model was developed, it was tested on a new set of students; it predicted whether a student from the new sample would answer the next question correctly 75 percent of the time (Woolf et al. 2009).

Building on this work, the University of Massachusetts team set out to make the Wayang Outpost tutor sensitive to a student's affect. (The system was already adaptive in that it customizes problems and hints to an individual student's cognitive profile, gender, spatial ability, and speed of retrieving math facts.) The researchers implemented two animated agents, Jake and Jane, to work with students using Wayang Outpost. The revised system analyzes a student's emotional state as well as progress on the math content, and then the animated agent sends messages tailored to fit the student's combination of cognitive and emotional state.

For example, Wayang Outpost distinguishes between frustration and boredom. For a student who has become frustrated, Jake or Jane might say, "That was very frustrating. Let's move to something easier" or "Some students are frustrated by this problem. Let's look at some similar problems already worked out." For bored students who find the work difficult, the animated agent might move to an easier topic. For bored students who find the work too easy, the agent might say, "You seem to know this pretty well so let's move onto something more challenging that you might learn from" (Woolf et al. 2009).

The animated agents also adopt facial expressions that mirror the student's happiness or sadness. The University of Massachusetts research team is now evaluating whether affective agents perceived as caring can increase the likelihood that students will persist through frustrating portions of instruction and exhibit greater mastery of math content (Woolf et al. 2009).

As these examples show, learning can be adapted based on specific task performance, past work on similar tasks, dispositions, motivation, and preferences. What the system knows about a student increases as the student spends more time using it.

However, our ability to track factors that influence learning has outpaced careful research on which factors are worth tracking. An important challenge for researchers and learning system developers is to identify the factors of learning materials, supports, and pacing that make a difference in learning outcomes. Emerging systems will provide data to support these efforts.

Technology Supports for Teachers

Adaptive learning does not always require sophisticated digital learning and tutoring systems. Relatively simple technology supports can also be used to help classroom teachers dynamically adapt their instructional methods.

One example is student-response systems that facilitate rapid diagnostic assessment with respect to concepts. Early student-response systems used clickers, small devices with a few buttons for different response options; now systems may have students text from their mobile phones or choose answers using a Web-based system from their laptops or smartphones. Students' anonymous responses are displayed visually, often both to the instructors and the class.

This way, instead of getting an answer from a single student who raises his or her hand, a teacher can instantly see how every student in the class responds. If the teacher's questions are carefully crafted to elicit students' thinking, classroom communication systems can provide a window into each student's understanding of the concepts being discussed (Crouch and Mazur 2001). (For more information on the use of clickers to adapt instruction, see the sidebar *Using Clickers to Give Teachers Diagnostic Data for Adaptive Instruction*.)

For learning to be adaptive, teachers must not only gather this kind of formative assessment data, supported by either digital learning systems or classroom communication systems, but also have different instructional strategies to apply for those students who fail to demonstrate understanding.

Recent research by Penuel et al. (2012) demonstrated the positive effect of instrumenting a classroom with communication technology and training teachers in strategies for working with students who demonstrate different misconceptions as revealed by the formative assessment data. Dede and Richards (2012) have described additional examples of this kind of adaptive instruction and the infrastructure needed to support it.

Using Clickers to Give Teachers Diagnostics Data for Adaptive Instruction

The Contingent Pedagogies project team at SRI International has been working to help teachers assess their students' understanding of key science concepts and adapt their instruction accordingly. Students often bring problematic ideas to the classroom, and it is important to surface and address them in instruction to promote learning (National Research Council 1999).

Working with sixth-grade teachers from Denver Public Schools and the Investigating Earth Systems curriculum developed by the American Geological Institute and TERC, the SRI researchers designed a set of elicitation questions for teachers to ask their students after they completed one of the earth science investigations. The team had developed the questions using research they had done on problematic ideas students typically hold about the core ideas in the earth science curriculum.

The teachers' classrooms were equipped with clickers (a student response system) so that every student could respond to the question and the teacher could see and display a histogram of all the responses. For example, many students think that earthquakes happen during certain kinds of weather. If many students in a class answer elicitation questions in a way that suggests they hold this idea, the teacher can introduce a contingent activity in which students are asked to interpret tables and graphs of earthquakes around the world and then construct an explanation for the patterns they see in the data. Weather data are included, but so, too, are items like information on proximity to a plate boundary, so that students can construct a more scientific understanding of where earthquakes are likely to occur.

The classroom discussions that are incorporated into the Contingent Pedagogies approach give students the opportunity to engage in the scientific practices of argumentation and developing explanations. Contingent Pedagogies teacher training emphasizes two strategies for engaging students in productive discussions. The first is classroom norms, which make explicit the norms that scientists use when deliberating about ideas. One of these is "support claims with evidence." The second strategy is a set of talk moves, which teachers can use to elicit and probe student thinking and encourage students to weigh different perspectives in discussion. Prior research has shown that when teachers use these talk moves to promote student argumentation, students learn more effectively (Resnick, Michaels and O'Connor 2010).

To investigate whether the use of Contingent Pedagogies elicitation questions with clickers, along with the training in adaptive instruction and discussion facilitation, improves student learning, a field test was conducted with 19 teachers. Twelve received the Contingent Pedagogies professional development and tools; seven teachers served as a comparison group. Students in the classrooms of all 19 teachers took two sets of pre- and post-assessments on their understanding of the core earth science ideas targeted by the project. Controlling for students' pretest scores, students in the Contingent Pedagogies classrooms scored significantly higher than those in the comparison teachers' classrooms on the earth science posttest.

Expanded Approaches to Gathering Evidence

The data that digital learning systems collect provide opportunities for determining the effectiveness of the systems' adaptive capabilities. The micro-level data collected on student interactions can be used to validate learner categorizations based on those interactions rather than on membership in a demographic category with higher average risk. The data can also be mined to prescribe adaptations for different learner groups and individuals.

Implications of Big Data for Matching Learners with Instructional Approaches

The impetus to present different learning experiences to different individuals stems from the belief that certain characteristics predispose students to learn better from different modes of presentation. The concept has intuitive appeal, but solid evidence to validate it is sparse (Cronbach and Snow 1969; Massa and Mayer 2006; Koran and Koran 2006; Pashler et al. 2008). Historically, the success of experiments testing interactions between specific learner traits (aptitudes) and specific instructional approaches (treatments) has been very low. The fact that many more such experiments can now be conducted efficiently increases the likelihood of finding more of these interactions.

Earlier aptitude-treatment interaction research focused on adapting instruction to broadly conceived aptitudes or traits hypothesized to be stable in a given learner over time and across different tasks. More recent research suggests that stable learning traits are few and far between. The nature of a student's learning approach may vary from task to task and within a task as learning unfolds.

Rather than relying on prior student classifications, developers of today's adaptive learning systems identify student actions (or patterns of actions) at a micro level and in the context of specific tasks and then make adaptations and continue to collect data that may result in different adaptations as time goes on. The finer grained data available from these learning systems can lead to new insights into the variability and constancies in human learning.

Also possible is combining insights from learning theory that suggest patterns to look for with large sets of detailed learning data. These new capabilities make the long-sought goal of differentiating instruction for every learner much more attainable after empirical evidence has been obtained that validates both learner categorizations and instructional prescriptions.

Once important learner differences have been identified, digital learning systems can be revised to vary the experience for different kinds of students working in different contexts. Key to this is being able to determine what an individual learner knows and what he or she still needs to learn in a dynamic way throughout the learning process.

Automating the Development of Expert and Learner Models

Perhaps the most clear and consistent difference between students is their incoming prior knowledge. Assessing and adapting to differences in prior knowledge requires two types of models: one of concepts students must master—the expert model—and one of what individual students know about that domain—the learner model.

Developing expert models can be difficult because experts in all kinds of domains are surprisingly unable to articulate the knowledge and skills they use (Biederman and Shiffrar 1987). Furthermore, experts often have blind spots about student learning difficulties and trajectories (Nathan and Koedinger 2000a, 2000b). This can be likened to a soccer player's ability to use the right amount of curve on a corner kick without being able to explain how he does it or understand why another person might have trouble with the kick.

To address this challenge, researchers interview experts and novices as they work through complex problems in the domain. When used to design new instruction, these methods, known as cognitive task analysis (CTA), have led to large student learning gains over traditional instruction (Clark and Estes 1996). New learning technologies offer the possibility of reducing the effort associated with CTA through the use of data-driven automated approaches that can be more widely scaled and driven by more objective evidence.

Modeling learner knowledge is a dynamic process that resembles the user knowledge modeling that has been used in adaptive hypermedia, recommendation systems, and intelligent tutoring systems. New machine-learning-based approaches to developing student knowledge models build on prior research in this area.

One method for estimating students' knowledge development is Corbett and Anderson's knowledge tracing model (Corbett and Anderson 1995). Developed in the mid-1990s, it uses a Bayesian network approach for estimating the probability that a student knows a skill based on observations of him or her attempting to perform the skill.

More recently, Ryan Baker and colleagues proposed a new approach to modeling learner knowledge that uses machine learning to make contextual estimations of the probability that a student has guessed or slipped (that is, understood the correct procedure but made a careless error in executing it). Incorporating models of guessing and slipping into predictions of students' future performance has been shown to increase the accuracy of the predictions by up to 48 percent (Baker, Corbett and Aleven 2008).

Using Learning Data to Improve the System for Different Learning Profiles

The data a digital learning system collects can be used to improve the system itself. For example, Baker and colleagues have analyzed learner interaction data from adaptive learning systems for middle school math to distinguish between students who are attempting to game the system and those who are trying but still struggling, so that different strategies can be used with the two groups (Baker et al. 2004; Baker, Corbett and Koedinger 2006).

Baker and his fellow researchers were able to detect gaming behaviors (such as clicking until the system provides a correct answer and advancing within the curriculum by systematically taking advantage of regularities in the software's feedback and help) that were strongly associated with less learning for students with below-average academic achievement. They modified the system to detect this behavior and respond to these students by providing them with supplementary exercises, the use of which was associated with better learning.

A team at the University of Washington has similarly analyzed millions of players' behaviors in Refraction, an adaptive online math game it developed. When a student struggles to complete a level in Refraction, the system determines the likely source of that player's confusion based on other players' paths through the game and offers a different path. The ability to disaggregate Refraction learning data makes it possible to calculate an effect size for different subsets of students and gain more insight into learning and engagement processes.

To further understand how different players learned in the game, the University of Washington team also developed a tool called Playtracer that creates simplified visual maps of many players' moves through the system. The maps reveal points in the game where many players get stuck or make the same incorrect choice. The researchers can then develop a few possible fixes to the problems identified and apply A/B testing to find the best solution (Andersen et al. 2010; Liu et al. 2011). (A/B testing is defined and explained in Chapter 1.)

Results of such A/B manipulations can be examined for each type of learner to discover whether the same version of the feature is best for all learners or whether different variants produce better learning or more engagement for different learner types. Combining learning profiles and A/B testing creates the opportunity to find out whether there is a reason to adapt the nature of instruction for learners with different profiles or even for the same learners at different points in time (for example, when they feel anxious or bored as in the Wayang Outpost research). (For more information on the use of Playtracer to analyze Refraction, see the sidebar *Adapting Learning Games to Sustain Student Engagement*.)

Uses of Evidence from Adaptive Learning Systems

As the examples in this chapter illustrate, identifying situations in which adaptive instruction will be beneficial is well within our grasp. The more difficult challenge will be testing the generality of these learner categories and instructional principles. This will entail synthesizing findings across different learning systems and research groups, looking for patterns and combinations that have not been previously considered. Our understanding of human learning and our ability to adapt learning experiences for the needs of each individual can be expanded if developers extract and make available the system data they are using to diagnose learner types and validate adaptive instructional approaches.

Synthesizing data across learning systems and research groups is another area where technology can support advances. The DataShop at the Pittsburgh Science of Learning Center is an example of how data from multiple studies can be combined and made open to inspection by other researchers so that models can be reused and improved. DataShop makes 80 different datasets publicly available and hosts scores of others that researchers can request access to. DataShop also contains a set of analysis and reporting tools including standard reports of learning curves. Making learning system data and data modeling tools open and available for continuous improvement will help build a stronger knowledge base for designing adaptive learning experiences. (See the sidebar *Developing and Sharing Tools for Cognitive Modeling*.)

In addition to greater data sharing and transparency, the field also needs to develop a larger group of data mining experts with multidisciplinary training in statistics, computer science, machine learning, and cognitive science.

Conclusion

This chapter describes how the increasing sophistication of digital learning systems can support both the development and implementation of customized learning strategies and content for individual learners, including the ability to adapt to individual learners as they use a digital learning system. Capabilities now available in new learning systems are discussed, including fine-grained models of learner knowledge that are updated dynamically, micro-level tagging of both instructional content and of learner actions with systems, and the adaptations systems can make based on students' emotional states and levels of motivation. It also examines the implications of the "big data" that learning systems collect for matching learners with instructional approaches, including how this data might be used to assess the value of adapting instruction.

Adapting Learning Games to Sustain Student Engagement

Computer scientists at the University of Washington designed Playtracer, a tool that turns player-generated data into visual representations. They used it to analyze one of their own games, Space Rescue (an early version of Refraction). The goal of Space Rescue (like Refraction) is to select and place on the screen tools that redirect and split laser beams in a way that sends the designated fraction of a beam in the correct direction to reach all the targets visible on the screen (Andersen et al. 2010). Placing a tool on the board constitutes a move that changes the state of the game. In addition to the targets, players can direct lasers through bonus coins for optional extra points (Andersen et al. 2010).

Playtracer records the states and shows many players' paths through each level of a game as a map of nodes and vectors. The starting point and goals appear as nodes. The steps players took from the starting point appear as dots connected by vectors indicating the order of the steps. A large node means that many players arrived at a given point in the game. Playtracer's output can be tailored to display the data in different ways. It can show the path of only a single player or show comparisons of paths taken by those who completed a given level of game play and those who quit before reaching the goals.

The game designers made several changes to the game based on patterns they saw in the Playtracer output (Liu et al. 2011). In one level, for instance, they noted a cluster of activity associated with failure; most players who made that series of moves quit before completing the level. This led the designers to hypothesize they had increased the complexity too quickly from the previous level. They could then use A/B testing to compare a revision to that level against the previous version and analyze the Playtracer maps of players' success in each version to understand whether the revision was actually an improvement.

Analysis of Playtracer maps also led the developers to the surprising realization that players who collected the optional bonus coins along the way were more likely to quit than players who did not (Liu et al. 2011). In A/B testing, they found that players who sought the coins tended to try complicated approaches that probably increased their frustration, whereas those who played a version without coins tested simple approaches and found the solution.

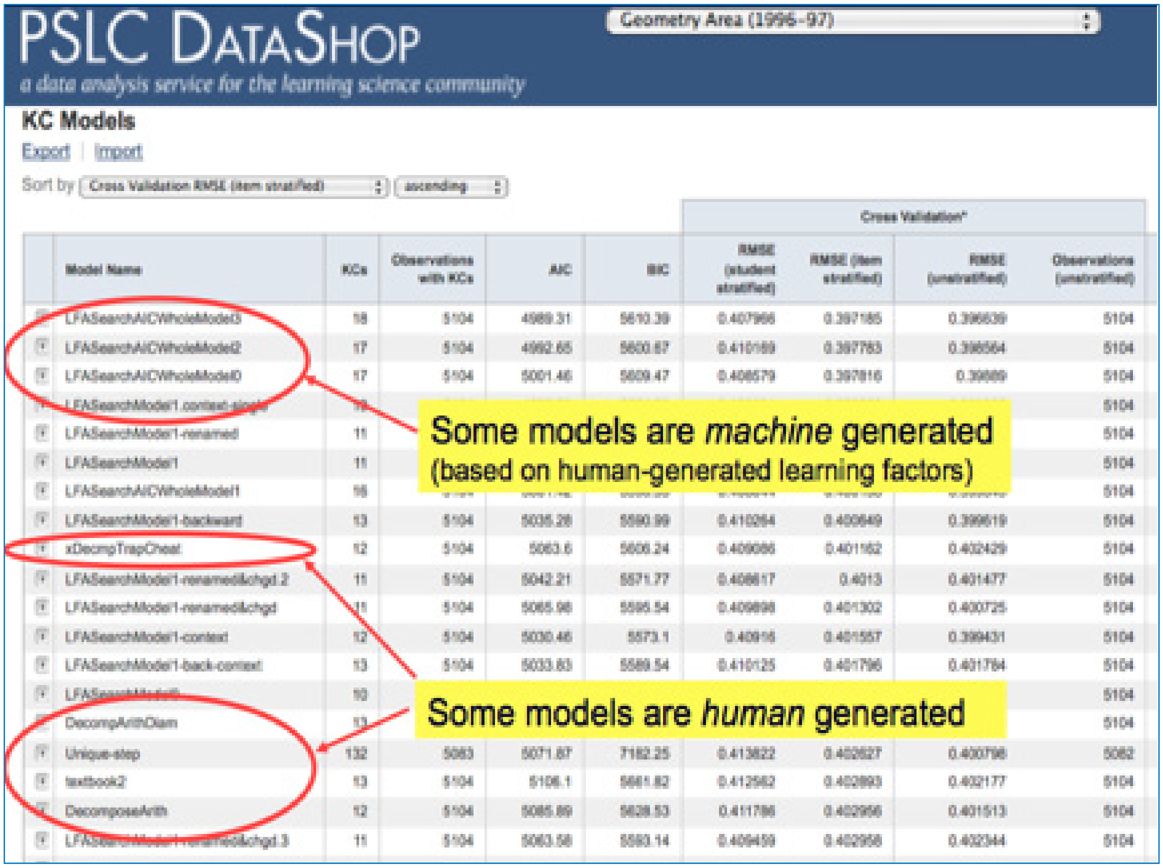
The design team's goal is not to eliminate player's confusion. Rather, the team wants the game to foster the kind of confusion that has been associated with ultimate mastery of a concept and deeper learning (Craig et al. 2004) rather than the confusion that leads to frustration and quitting. In Playtracer, activity that loops away from and back to the starting point can indicate that players have tested a logical hypothesis that was not a solution and then removed the pieces from the board to rearrange them in a different way.

The University of Washington team has also used Playtracer to analyze FoldIt, a protein-folding science discovery game it had also developed. In Playtracer maps of FoldIt play, the team saw that players who did not ultimately find good solutions often came very close without knowing it. Adding a message for users at that moment in the game could encourage them to persist and reach a successful solution (Liu et al. 2011).

Developing and Sharing Tools for Cognitive Modeling

At the Pittsburgh Science of Learning Center, the LearnLab's DataShop provides analysis tools to support the discovery of more accurate cognitive models of domain content, student skills, and learning trajectories (Koedinger, McLaughlin, and Stamper 2012).

DataShop provides analysis tools that can be applied to sets of learning system data. Some of the patterns that DataShop can detect were discovered bottom up through machine learning. Others were defined by human analysts. Hundreds of datasets from student use of educational technology in math, science, and language have been analyzed with DataShop to detect the presence of these cognitive models. DataShop's leaderboard, shown below, ranks discovered cognitive models for each of hundreds of datasets from student use of educational technology in math, science, and language.



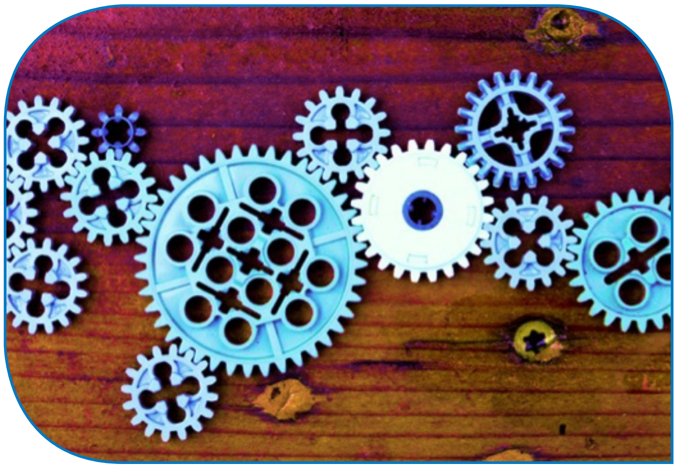
KC Models

Sort by: Cross Validation RMSE (item stratified) | ascending

					Cross Validation*			
Model Name	KCs	Observations with KCs	AIC	BIC	RMSE (student stratified)	RMSE (item stratified)	RMSE (unstratified)	Observations (unstratified)
LFASearchAICWholeModel3	18	5104	4989.31	5610.39	0.407966	0.397185	0.396639	5104
LFASearchAICWholeModel2	17	5104	4992.65	5600.67	0.410169	0.397763	0.398564	5104
LFASearchAICWholeModel0	17	5104	5001.46	5609.47	0.408579	0.397816	0.39889	5104
LFASearchModel1.context-single	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	5104
LFASearchModel1-renamed	11	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	5104
LFASearchModel1	11	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	5104
LFASearchAICWholeModel1	16	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	5104
LFASearchModel1-backward	13	5104	5035.28	5590.99	0.410264	0.400649	0.399619	5104
xDecmpTrapCheat	12	5104	5063.6	5606.24	0.409086	0.401162	0.402429	5104
LFASearchModel1-renamed&chgd.2	11	5104	5042.21	5571.77	0.408617	0.4013	0.401477	5104
LFASearchModel1-renamed&chgd	11	5104	5065.98	5595.54	0.409898	0.401302	0.400725	5104
LFASearchModel1-context	12	5104	5030.46	5573.1	0.40916	0.401557	0.399431	5104
LFASearchModel1-back-context	13	5104	5033.83	5589.54	0.410125	0.401796	0.401784	5104
LFASearchModel0	10	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	5104
DecompArithDiam	13	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	5104
Unique-step	132	5083	5071.87	7182.25	0.413622	0.402627	0.400798	5082
textbook2	13	5104	5106.1	5661.82	0.412562	0.402893	0.402177	5104
DecomposeArith	12	5104	5085.89	5628.53	0.411786	0.402956	0.401513	5104
LFASearchModel1-renamed&chgd.3	11	5104	5083.58	5583.14	0.409459	0.402958	0.402344	5104

Some models are *machine* generated (based on human-generated learning factors)

Some models are *human* generated



Chapter 3:
Combining Data to Create Support Systems More Responsive to Student Needs

Young people learn and develop in a wide range of settings. How can data better be used to help support the full range of student needs and interests—both inside and outside schools and classrooms—to improve learning outcomes?

Far too many U.S. students—especially those from low-income backgrounds—never finish high school. Without a high school degree, an individual's chances for employment are drastically reduced, as are lifetime wages, health, and prospects for staying out of the criminal justice system. Economists Hank Levin and Cecilia Rouse (2012) estimate that cutting the U.S. high school dropout rate by half would save taxpayers $90 billion a year, or $1 trillion over 11 years.

Academic and social disengagement from school are key factors associated with dropping out (Rumberger 2001, 2011). This disengagement is not typically associated with a single event; rather, it is a long-term, cumulative process (Newmann, Wehlage and Lamborn 1992; Wehlage et al. 1989). Moreover, disaffection with school is not limited to those who actually leave the system: A majority of high school students report being bored every day in class (Yazzie-Minz 2010). Many students fail to see the relevance of what they are asked to learn in their classes to the future lives they imagine for themselves.

Achieving progress in this area requires that schools be more responsive to students' needs and interests and take a more encompassing view of students' lives. School administrators need to appreciate the fact that young people learn and develop in a wide range of settings, not just classrooms, and attend to the multiple aspects of their well-being.

Young people learn and grow not just in school, but also at home and in interest-driven pursuits such as sports, music, and hobbies (Eccles and Barber 1999; Fredricks and Eccles 2006). Their successful development thus requires intellectual supports and a rich network of social and emotional supports so they can develop autonomy, competence, and a sense of belonging (National Research Council and Institute of Medicine 2002).

From a youth development standpoint, however, students' needs are often examined through a narrow lens. Education data systems track student attendance, incidents requiring discipline, grades, and achievement test scores. If a student is within an acceptable range on these measures, other indicators of difficulty are likely to go unnoticed, especially in large schools (McLaughlin, Irby and Langman 1994).

When districts and schools decide to proactively identify students for assistance, the criterion is usually membership in a demographic or status category such as poverty, ethnicity, or designation for special education. Other possible sources of difficulty are overlooked.

Looking back on individual negative student outcomes, such as incidents of school violence or dropping out, school administrators often realize that multiple warning signs had existed but that no one had the resources to put things together and respond to the warnings in time. Information that could have led to preventive action earlier was not captured in education data systems, not available in an aggregated form, or not examined and acted on.

At the other end of the spectrum, students who show great accomplishment, leadership, and collaboration skills in out-of-school settings may be overlooked for in-school leadership and learning opportunities because their schools do not recognize these accomplishments and capabilities (Hull and Schultz 2001).

New Opportunities Provided by Technology

Technology provides opportunities for creating better support that can keep students engaged and progressing through school. These include the ability to collect different types of data and combine data from different systems, analyze data in new ways to target intervention practices and programs, and provide support for new practices and interventions.

Researchers are finding that students themselves can be sources of data that the education system can use to predict achievement as well as the risk of dropping out. For example, student reports of how engaged they are in their classes and of the closeness of their relationships with school staff have proven to be connected to engagement and learning outcomes when aggregated at the classroom or school level (Bill & Melinda Gates Foundation 2012). Similarly, students' reports of the availability of a caring adult on the school staff are associated with more effective schools (Fredricks, Blumenfeld and Paris 2004; Wentzel 1997).

The perceived quality of students' relationships with their teachers is especially important as a foundation for engagement (Skinner and Belmont 1993; Skinner et al. 2008). Parents, siblings, mentors, and peers can also play critical roles in sparking and sustaining engagement in learning activities (Barron et al. 2009; Goldman, Booker and McDermott 2007).

The roles these different people play in supporting engagement are many. They include collaborating and providing resources or brokering connections to new learning opportunities, and they often are facilitated by access to technologies that support sharing and joint work (Barron et al. 2009).

The key is being able to combine these types of data with other data to further engagement and learning outcomes.

Expanded Approaches for Gathering Evidence

State and district student data systems have improved greatly over the past decade in ways that permit examining an individual student's educational experiences and achievement over time, even if the student changes schools or school districts.

For example, an increasing number of states now assign student identification numbers that stay with the student anywhere in the state, and state data systems typically contain more information on a student's background (that is, ethnicity, whether eligible for subsidized meals, English proficiency, disability status, date of birth, gender) as well as grade level, school attended, and state achievement test scores. Districts are also creating student data systems that include such variables as attendance, performance on district-mandated tests and benchmark exams, courses taken, grades, and teachers.

These improved data systems and the new data they house open up opportunities for schools and districts to partner with community and government agencies from other sectors to create linked datasets with more kinds of information about the circumstances of students' lives. Combining datasets from different agencies permits analyzing information on students' academic achievement, attendance, and other indicators of school success with information on their involvement in social services, the juvenile justice system, the foster care system, and youth development programming aimed at supporting students' social and emotional learning.

Linking these various types of data can help schools explore relationships between students' conditions outside school and their in-school experiences and thereby develop early warning systems for predicting students at risk. One example of linking data across agencies to better understand and address the issues young people face is the Youth Data Archive at Stanford University. For an illustration of the kinds of insights gained by combining data on individual youth across different institutional settings, see the sidebar *Linking Data from Different Service Agencies.*

The Promise Neighborhoods Research Consortium (PNRC)[2] also links different data systems to improve outcomes. Funded by the National Institute on Drug Abuse, this collaboration among university research centers, nonprofit organizations, and mental health service organizations has the mission of assisting policymakers in finding the most effective and efficient ways of helping high-poverty neighborhoods improve the well-being of their children and youth.

The PNRC notes that high-poverty neighborhoods often have high levels of "drug abuse, antisocial behavior, depression, academic failure, and intergenerational poverty" (PNRC 2012) and that research-based strategies exist for reducing all of them (National Research Council and Institute of Medicine 2009). The PNRC has developed a measurement framework to support communities in combining data from education data systems with information from surveys of households, teachers, students, and parents.

The PNRC website organizes these data into summaries for use in evaluating and refining community services. The website encourages community leaders to join with PNRC researchers to evaluate the well-being of their children and youth and to identify both unmet needs and supportive and protective factors within their communities. PNRC's review of research has identified practices that evidence shows have positive impacts on children and youth. The organization has also identified more than 55 policies that states and communities can adopt that have had positive effects on youth outcomes.

[2] The PNRC has no formal relationship with the U.S. Department of Education's Promise Neighborhoods grant program, but it views its own website and services as a potential resource for neighborhoods applying for or receiving these grants.

Linking Data from Different Services Agencies

Stanford University's John W. Gardner Center for Youth and Their Communities, in collaboration with the SPHERE Institute, houses The Youth Data Archive (YDA), an initiative linking data on individual youth across different institutional settings. Partners in the YDA are school districts, community colleges, local health departments, county offices of education, human services agencies, recreation and parks departments, and youth-serving nonprofit organizations. The YDA team develops agreements with nonprofit and government agencies in selected counties and communities in northern California and facilitates groups' investigation of youth data to improve services and youth outcomes.

The YDA has also been used in a collaboration between the Gardner Center and several agencies in San Mateo County, California, to analyze educational outcomes for court-dependent youth in foster care (Castrechini 2009). The frequent school and residence changes typical of this group of young people make tracking outcomes difficult without a tool like the YDA.

For this particular analysis, dependency records from Child Welfare Services were linked to educational data from several school districts. The analysis showed that outcomes for court-dependent youth were much worse than those for other children. In addition, the detailed records of the YDA revealed a relationship between the nature of a child's foster placement and educational outcomes. In general, outcomes were better for youth placed in involuntary family settings than for those placed in out-of-home settings (Castrechini 2009). As a consequence of the analysis and conversations about the findings facilitated by Gardner Center staff, the collaborating agencies recognized the need for greater academic support for foster youth, especially those placed in group homes and other nonfamily settings.

The work of the YDA illustrates how much more is needed in addition to creating a repository of data. The Gardner Center works in collaboration with local agencies to articulate a set of research questions of concern to the community and to identify data sources that could help address them.

Gardner Center staff members develop memoranda of understanding that detail the data to be included, analyses to be performed, and how and with whom analyses will be shared. All agreements comply with laws regarding the protection of privacy and human subjects. Participating organizations can withdraw from the YDA at any time and have their data removed.

Predictive Analytics and Early Warning Systems

Increasingly sophisticated techniques for predictive analytics, which combines a variety of disciplines including statistics, data mining, and game theory, are also being used to investigate whether some student behaviors are predictors of school failure and dropping out. Predictive analytics involves creating a quantitative model that can infer a predicted variable of interest (for example, the risk of dropping out) on the basis of some combination of other variables (predictor variables) drawn from available data systems (U.S. Department of Education 2012a).

Researchers have used predictive analytics with the Youth Data Archive, mentioned under "Expanded Approaches for Gathering Evidence," to examine chronic absence from school (Sanchez, Castrechini and London 2012). The California Department of Education tracks whether students are truant (defined as having three unexcused absences) but not the actual rate of absence for individual students. Understanding individual students' actual absence rates is important because those who work with young people in a variety of settings believe that chronic absence—whether excused or not—is a serious risk factor for disengaging and dropping out. The average daily attendance that a school reports to the state can be high and can mask the presence of a set of chronically absent students who may need a variety of different kinds of support, such as transportation, housing, or physical and mental health services, to be able to attend school consistently.

The YDA collaboration has defined chronic absenteeism as being missing from school 10 percent or more of the school year, with or without an excuse. The group examined three years of data in its linked data systems to find out how many chronically absent students were in their jurisdictions and investigate their characteristics and outcomes. The YDA analysts found that prior chronic absenteeism was a stronger predictor of future absenteeism than past suspensions or any demographic variables such as ethnicity or family income. They also found that students who were chronically absent during two or three years in middle school enter high school with significant gaps in mathematics achievement. This kind of information can be used to design targeted intervention programs—for example, providing additional mathematics support for students with high rates of absenteeism, even if their absences are excused.

Technology also provides new opportunities for collecting a broader set of student data at the classroom level, as exemplified by ClassDojo, a real-time behavior management system first made available in 2011. (For more information on ClassDojo, see the sidebar *Using Technology to Create Feedback Loops for Classroom Behavior.*)

Predictive analytics is described by Shmueli and Koppius (2010) as "statistical models and other empirical methods that are aimed at creating empirical predictions as well as methods for assessing the quality of those predictions in practice, i.e., predictive power" (p. 2). In education, predictive analytics is being used to identify struggling students and pinpoint student stressors early, with the goal of offering appropriate interventions and supports more quickly.

Using Technology to Create Feedback Loops for Classroom Behavior

From a teacher's standpoint, classroom management is a major portion of the job. Dealing with disruptive behavior can be time consuming and stressful. In addition, some students with behavioral issues have difficulty perceiving their own counterproductive behaviors, and feedback at the end of class may be too late. At the same time, positive classroom behaviors are not always recognized and reinforced.

ClassDojo, in beta version, is a real-time behavior management tool that teachers can use with a smartphone, tablet, or laptop computer. After their names have been entered into ClassDojo, each student is assigned an avatar. By clicking on a student's avatar and then clicking the appropriate behavior category, the teacher can enter data about a student's positive or negative behavior in real time. Built-in behavior categories include participation, helping others, insight, disruption, and tardiness. The teacher also has the option of adding behaviors.

Students receive the feedback on their positive and negative behaviors in real time. A positive behavior is acknowledged with a chime and a green badge that appears on the student's avatar, and a negative behavior is marked by a buzzer and a red badge. Students' avatars also receive or lose points based on their behaviors, which can motivate better behavior. Teachers have the option of allowing students to award or subtract points from each other's avatars based on behavior, a feature intended to stimulate class discussions about what is or is not appropriate behavior in a variety of situations (ClassDojo 2012).

ClassDojo also creates a summary report of all students' behavior during a class session and reports for each individual student. Individual student reports can be emailed to students and used as the basis for conversations with students and their parents to explore how behavior can be improved. These conversations may reveal behavior changes that may be by-products of other stressors in a student's life that can also be addressed.

A number of districts and school networks are starting to develop and use early warning systems based on applying predictive analytic models to student data systems. An example is the Achievement Reporting and Innovation System (ARIS) being used in New York City. ARIS is being extended to incorporate teacher-input grades, quiz scores, and other data.

Another example is the graduation and college readiness prediction work of the New Visions for Public Schools. (See the sidebar *Using Data to Help Keep Students on Track for Graduation*.)

In addition, a key application of predictive analytics is monitoring and predicting students' learning performance and spotting potential issues early so that interventions can be provided for students identified as at risk of failing a course or program of study (EDUCAUSE 2010; Johnson et al. 2010).

Because the use of digital learning systems was commonplace in higher education before it was in K–12 schools, colleges and universities are leading the way in combining real-time course-level data with information from student data systems to create more dynamic early warning systems. These systems can be used to identify a student's risk of failing a specific course in time for administrators to take corrective action. Examples from higher education include Purdue University's Course Signals system (Arnold 2010) and the Moodog system being used in courses at the University of California, Santa Barbara (EDUCAUSE 2010). (For more information about Purdue's Signals, see the sidebar *Using Current Course Data in the Signals Early Warning System.*)

Using Data to Help Keep Students on Track for Graduation

School attendance rates, credit accumulation, and grades are all good predictors of graduating from high school (Allensworth and Easton, 2007; Pinkus 2008). Defining good performance in these areas as being "on track" for graduation, the Consortium for Chicago School Research found that students who completed grade 9 on track were 3.5 times more likely than those not on track to earn a high school diploma in four years (Allensworth and Easton 2005). Conversely, failure to identify students not on track and to provide the necessary intervention early is associated with higher dropout rates (Allensworth and Easton 2007).

These findings have prompted a number of major school systems to develop benchmarks for being on track for graduation and qualification for college admission. In New York City, the nonprofit New Visions for Public Schools has developed and implemented an early warning system to provide timely information on students' progress not only toward graduation, but also toward qualifying for admission to college.

Since 1993 the nonprofit New Visions for Public Schools has opened 133 public high schools in New York City. Typically small and built around principles of student-centered education, New Visions schools have enjoyed higher graduation rates than New York City high schools as a whole (Foley, Klinge and Reisner 2007). In 2007 the New York City Department of Education made New Visions a Partnership Support Organization (PSO) with responsibility for supporting 76 district schools. A major New Visions focus as a PSO has been the development and implementation of the early warning system. The New Visions College Readiness Metric calls for earning 11 credits each year, acquiring specified numbers of credits in various core academic disciplines, and passing state Regents exams in four key academic areas with scores in English language arts and mathematics high enough to place out of remedial courses at the City University of New York (Fairchild et al. 2011).

Drawing on a single dataset combining elements from New York City Department of Education data and school records, separate tools have been developed for school staff, parents, and students to use in examining student progress. The School Snapshot identifies students who are off track for graduation, those who are on track but struggling, and those who are on track for a diploma but not for college entrance requirements. To provide more actionable, timely information, the School Snapshot pulls in new data after each grading period rather than waiting for end-of-course grades. A report of data aggregated at the school level summarizes school performance trends relative to district, state, and federal accountability requirements.

For parents and students, New Visions created the Ninth Grade Tracker and the College Readiness Tracker to provide an easy-to-understand visual display of an individual student's standing relative to graduation and college admission requirements. These tools are designed to help parents and students make sense of complicated high school graduation and college admissions requirements. They can use them to see subject areas where a student is doing well and those where he is weak.

In 2010 New Visions partnered with a commercial company (DataCation) to integrate its tools into a web-based environment that delivers real-time data to students, parents, and school staff and enables the latter to drill down from school-level reports of aggregate data to specific reports of detailed student-level data. Student profiles in this system include information on the services the student is receiving, family contact information, attendance, class schedule, grades, examination records, transcript data, and any anecdotal logs.

Using Current Course Data in the Signals Early Warning System

Like high schools worried about their graduation rates, colleges and universities try to maximize the proportion of their students who stay in school and graduate with a degree.

In addition to requiring students with weak academic preparation in language arts and mathematics to take developmental courses, colleges and universities generally offer a variety of supports through academic resource centers. Some also have special bridge programs to offer students a chance to ease in to the academic environment and its demands.

Students are identified for these courses and services generally by self-selection or on the basis of demographic variables (such as first-generation students, returning adult students, etc.) and the summary achievement data provided at the time of college application (test scores and grade point average).

Purdue University has worked on improving retention and graduation rates since launching the Purdue Academic Warning System (PAWS) in the 1980s. In fall 2007, it launched the Signals Project.

Purdue wanted to have a way of identifying student risk on the basis of much more near-real-time information so that it could respond to the dynamic nature of students' motivation and the challenges that they face (for example, illness, breakup of a relationship, family issues). This method would provide information in time to prevent course failure rather than dealing with challenges after the term ended.

The university had been having faculty report midterm grades through PAWS but found that the information came too late or was too incomplete to provide enough support in time for the student to succeed in the course. Commercially available digital early warning systems could automate the process, helping faculty identify students for assistance more quickly by combining course grade information with student demographics and summary achievement data, but these systems did not distinguish between students who were trying and failing and those who were not putting in the effort (Arnold 2010).

Purdue wanted a system that would provide more information about student behaviors and that could support inferences about how best to help different students. Most important, this information needed to be available to the faculty, the students, and their advisor.s

Mining data from the university's course management system (CMS), its student information system, and faculty grade books, Signals applies the Student Success Algorithm (SSA) developed by Purdue's John Campbell. The SSA generates a risk level for each student (high, medium, or low as represented by red, yellow, or green traffic lights) and provides specific information about what work that student has and has not completed. In this way, Signals provides warnings as early as the second week of the semester based on students' performance and effort in the course (Arnold 2010).

Instructors using Signals set up an intervention schedule with such elements as posting the signal on each student's CMS page, email messages or reminders, text messages, referral to an academic resource center, and face-to-face meetings. Signals provides faculty with sample email messages they can either edit or send to students as written.

Purdue has explored the effects of Signals on students' behavior and course completion. Arnold (2010) cited data from a biology course in which some sections used Signals and others did not. Students in sections using Signals were less likely than those in other sections to get a D or an F. Struggling students in the Signals sections were also more likely than struggling students in the other sections to seek help and to seek it sooner. The Signals team has found that even when faculty used the prewritten text messages provided by Signals, students receiving them felt that their professors were more caring and invested in their success.

By 2011, more than 17,000 students had experienced a course supported with Signals. Students in the courses using Signals have consistently had higher grades. In the 2009 cohort, for example, 22 percent of students in a course without Signals got a D or an F or withdrew compared with 16 percent for courses using Signals (Campbell, J.P. & Arnold, K., 2011). In more recent data, the fall 2007 cohort of Signals participants found an 18-point increase in the four-year retention rate and a nearly 13-point increase in graduation after five years (Campbell 2012).

Evidence Issues Associated with Predictive Analytics

An issue associated with the use of predictive analytics is demonstrating the validity of the predictive algorithm. From a resource allocation perspective, it is important to demonstrate that students identified as in need of services are in fact at higher risk than other students. If the algorithm does not identify students most in need, intervention efforts will not produce the maximum possible social benefit. Standard metrics exist for computing the predictive accuracy of an algorithm (that is, the number of correct predictions over total number of predictions) and quantifying this dimension so that the accuracy of different algorithms can be compared (Freitas 2002).

A broader evidence consideration is that predictive analytics is based on correlation. After discovering correlational relationships, researchers press for a deeper examination of the data and other available evidence to try to understand whether one of the variables actually causes the other or whether they just happen to occur together. For example, the fact that students who have many excused absences in middle school have poor math achievement in ninth grade does not prove that absenteeism causes poor math achievement. Additional research can build understanding of the multiple causes of absenteeism and inspire the design of interventions whose effectiveness can be studied. Nevertheless, from the standpoint of achieving better outcomes for students, identifying students at risk for certain kinds of problems early is an important start even if more research is needed to understand the causal mechanisms.

New student support interventions are likely to be inspired by correlations between certain factors and student outcomes and by what is known from research about how the outcomes typically emerge over time. Those organizations implementing the interventions should be tracking students' exposure to them and, at a minimum, comparing outcomes of those students receiving the intervention and of students with similar predictive profiles in earlier cohorts.

The Purdue University Course Signals system (see sidebar) is an example of a system that identifies students at varying levels of risk and institutes interventions in the form of feedback to the student and prompts and supports from the course instructor and the system itself.

No rigorous large-scale experiments have been done on the effects of Purdue's Signals. Yet the sheer size of the difference in the rate of completions with a grade of C or better for courses or course sections using Signals compared with historical data for numerous Purdue courses (Campbell and Arnold 2011) is highly persuasive for faculty and education administrators. Moreover, the many course outcome comparisons taken in aggregate make a strong case that Signals has positive effects for students in courses at Purdue. Other colleges and universities are now adapting and implementing Signals on their own campuses.

Emerging Options for Recognizing In- and Out-of-School Accomplishments

Thus far, we have discussed opportunities to use new data, combine data systems, and apply predictive analytics to identify student risk factors. Using data systems to identify protective factors and student accomplishments is also possible. An emerging area of research is on environments that are "interest-driven," where young people choose to pursue activities (often outside school) that involve learning, deep engagement, and the exercise of leadership (Heath and McLaughlin 1993). Interest, like deep content knowledge, develops over time and depends on the availability of guides and peers who can support its growth (Hidi and Renninger 2006).

Seminal research by Heath and McLaughlin (1993) described the interests and competencies developed by inner-city youth through extended participation in out-of-school activities such as drama, community service, and sports clubs. More recent work has described how afterschool engagement in creative use of technology (Barron 2006; Barron et al. in press) and technology design activities (Koch and Penuel 2007; Koch et al. 2009) can lead to the development not only of strong interests, but also the kinds of competencies that are valued in the 21st-century workplace. For many students from low-income and underrepresented communities, these activities and the competencies and recognition they gain through them appear to be important factors in building resilience to the challenges they face (McLaughlin 2000; McLaughlin, Irby and Langman 2001).

Schools are often unaware of the leadership and competencies that students who do not excel in the classroom have shown in these out-of-school settings (Koch et al. 2010; Lundh, Koch and Harris 2011). Lemke et al. (2012) are among those recommending that "the scope of valued learning outcomes be broadened to include" informal learning experiences including "learning by groups and whole projects as well as individuals" (p. 3).

The desire to see these accomplishments recognized by the school system, colleges, and employers is one of the factors fueling an increasing interest in using measures of competencies, rather than the amount of time someone sat in a classroom, as the metric of educational achievement. Already, a student can receive credit at some colleges if he or she earns a high score on an Advanced Placement test, even without taking an Advanced Placement course. Similarly, those who gain competencies through workplace experience or taking an online course can obtain certificates attesting to their capabilities that many employers, especially in the technology industry, value.

A related trend is the development of "badging" systems that can capture and recognize the skills and abilities that students master when they pursue interest-driven routes to learning (Mozilla Foundation, Peer 2 Peer University and MacArthur Foundation, 2012). In a badging system, some badges might be relatively easy to attain so that students remain motivated. Others might be earned only after students demonstrate mastery of fine-grained skills that are not formally recognized in a traditional classroom. In either case, badges could be collected and aggregated into online student portfolios that would document and certify their interest-driven achievements. Informal digital learning systems such as Khan Academy use badging systems, and traditional colleges and universities are now exploring their use in the context of the proficiency certificates awarded for completing Massive Online Open Courses (MOOCs). Badging proponents envision a time when employers might look to badge portfolios as a way of determining whether potential hires have acquired the tangible skills needed in their organizations.

In March 2012 the MacArthur Foundation and Mozilla announced the winners in that year's Badges for Lifelong Learning competition, which was designed to promote recognition of learning and proficiency gained outside formal schooling. Competition winners included a wilderness explorers badge system from Disney-Pixar; badges from the Manufacturing Institute recognizing the skills and competencies needed in modern manufacturing; badges from NASA for exploring robotics and science, technology, engineering, and mathematics topics; and a program to recognize the competencies that librarians must gain to meet the needs of today's adolescents from the Young Adult Library Services Association.

Certification of competency is a logical extension of the move to common state standards. Education researchers (Collins and Pea 2011) have suggested that certifications of competency be created for all the new Common Core State Standards, with national certification exams that students could take whenever they felt ready for them and regardless of how the competence was acquired. Such a system, if examinations were rigorous and their validity had been demonstrated, would certainly provide an alternative route for certifying students' college and career readiness (Collins and Pea 2011) to colleges and potential employers. (For more information about badges, see the sidebar *Creating Digital Badges to Recognize Student Learning and Accomplishments.*)

Creating Digital Badges to Recognize Student Learning and Accomplishments

Digital badges are a type of credential that can follow a student throughout life and be used in job and college applications. The origin of the concept is usually attributed to a Mozilla conference in Barcelona in 2010. Mozilla is an organization promoting open-source Web software. (It grew out of Netscape and continues to develop and release the Firefox Web browser and the Thunderbird email client.)

One argument for badges is that they provide much more information than standardized test scores or grade point averages. Viewing the metadata attached to a digital badge, a potential employer or collaborator should be able to see not only the competency that the badge represents, but also who awarded the badge and why and how the badge was earned (that is, evidence justifying the award).

Another argument for badges is that they represent a more well-rounded, lifewide view of a person's capabilities. Many students who are disaffected with school or have difficulty in conventional academic environments find a passion and a skill niche in out-of-school activities such as theater productions, sports teams, or volunteer work (Heath 1994). Doubtless, students learn and develop competencies in these environments, but the formal academic system seldom recognizes these accomplishments (Hull and Schultz 2001). Badges can be given for the kinds of competencies that are essential in real-world work and community that are seldom formally assessed or recognized within the school system.

Connie Yowell of the MacArthur Foundation, which has been a major supporter of the digital badges movement, envisions also using badges as a way to provide more alternative paths for students moving between multiple learning environments. She imagines a recommender system that looks at an individual's set of digital badges and then recommends some appropriate next learning experiences that would help build toward a career or academic success (Ash 2012).

Recently, the MacArthur Foundation, Mozilla, and the nonprofit organization HASTAC (Humanities, Arts, Sciences, and Technology Advanced Collaboratory) cosponsored a competition for proposals to develop digital badges. They received more than 90 proposals and selected 30 for grants to support developing the proposed ideas.

The sponsors hope that this effort will create a critical mass of digital badge opportunities. In the meantime, Mozilla is developing the Open Badge Infrastructure (OBI) to provide the technical scaffolding for badges (Ash 2012). The OBI will enable secure badge issuance and acceptance. By operating within the OBI, organizations will issue badges that cannot be counterfeited and that can be read and displayed by websites of other organizations also operating within the OBI. Learners will be able to earn badges across many organizations, websites, and out-of-school experiences.

Evidence Issues Associated with Digital Badges

The use of badges for recognizing competencies is in its infancy, and the key question is the weight this kind of recognition of learning and accomplishment will be given in school admission and hiring.

An advantage of badges over standardized test scores is that they typically provide much more detailed descriptions of what learners can do (for example, lay out a publication in InDesign or run a theater sound system) than standardized test scores (SAT Verbal score of 480). As with any assessment, those who would consider making decisions on the basis of badges would want to have evidence that the judgment of competence was made fairly and that the competence the badge was earned for would also be exhibited in new contexts. At present, we do not have empirical research bearing on these issues for badge systems, but it is likely that at least some potential users of badge information will be swayed first by the reputation of the organization or individual bestowing the badge and later by experience with learners who have received the badge certification.

Conclusion

This chapter focuses on how data from learning systems can be combined with data from other sources to support the full range of student needs and interests that affect learning outcomes. Much is known about what can be done to keep students engaged and progressing through school, but today students' needs are often viewed through a narrow lens. The chapter discusses how combining data from different agencies permits analyzing information on achievement, attendance, and other indicators of school success with information on students' involvement in social services such as the juvenile justice system, the foster care system, and youth development programs to create early warning systems for identifying at-risk students. It also addresses ways of using such data to recognize and reward positive learning and accomplishments both inside and outside school. Systems based on this array of data can better meet the needs and interests of each individual student by supporting students in the totality of their lives.

Chapter 4: Improving the Content and Process of Assessment with Technology

Digital learning systems can collect data on important qualities not captured by achievement tests. How can educators use the systems to measure more of what matters in ways that are useful for instruction?

The U.S. education system invests heavily in tests of student achievement that can be used to hold districts, schools, and, in some cases, individual teachers accountable for whether students meet state proficiency standards. All the states have implemented large-scale testing systems for this purpose, and technology will become part of most states' assessment systems within the next few years as the computer-based Next Generation Assessments connected to the Common Core State Standards (CCSS) go into effect. (See sidebar *State-Led Assessment Consortia for Common Core State Standards*.)

At the same time, supporting students' learning calls for additional types of assessment, including

- formative assessments administered in the course of learning to provide information that teachers and students can use to guide future learning;
- assessments of 21st-century skills such as collaboration, problem solving, and innovation; and
- personal and affective qualities related to intellectual curiosity, self regulation, and persistence.

Both educators and researchers have noted the importance of these kinds of assessment.

In the cognitive arena, formative assessments are needed that provide much more detailed information about how students think and approach problems, not just whether or not they arrive at a correct answer. Because state achievement tests are designed to measure a whole year's worth of academic progress and usually occur just once a year, they cannot serve this purpose. Moreover, large-scale assessments generally have not captured complex performances, such as science inquiry or the ability to design something under a complex set of constraints, although the PISA (Programme for International Student Assessment) tests have been an exception and the Next Generation Assessments currently being developed are striving to incorporate complex performances.

State-Led Assessment Consortia for Common Core State Standards

In an unprecedented step, 48 states and the District of Columbia signed a memorandum of agreement in 2009 with the National Governors Association and the Council of Chief State School Officers to participate in an initiative to identify a common set of standards for mathematics and English language arts. Working with the nonprofit Achieve and consulting experts, this partnership developed standards for each grade level, which were released in 2010. That same year, Race to the Top funds were awarded to two state-led assessment consortia to build assessments that could be used to measure students' attainment of the CCSS. The Partnership for Assessment for Readiness for College and Careers (PARCC) is a consortium of 23 states working with assessment development firms. The Smarter Balanced Consortium involves 25 states. Both consortia are scheduled to have their assessments ready for schools across the country to use in school year 2014–15.

Both assessment consortia face the challenge of developing assessment items that get at the deeper learning aspects of the CCSS and are planning to deliver their assessments via computer. Many of the assessment items contain multiple interrelated parts. Both consortia have made evidence-centered design central to their development process. Their assessment item formats include constructed response and "technology-enhanced" items that take advantage of the computer-based medium.

A publicly released grade 9 English language arts item from the Smarter Balanced consortium is shown below.

Stimulus Text: *Read these paragraphs from a student's report and then answer the question.*

Year-round Schools

Year-round schools are a better way to educate students than the traditional nine-month schedule. Students are more likely to remember information over short breaks than they are during a long summer vacation. One study conducted by a group that runs year-round schools showed that **their students had higher test scores than students who attended schools with a traditional schedule.** Many teachers say **they have to spend September and October reviewing material taught the previous year.**

Some people argue that students shouldn't have to go to school any longer than they already do, but with year-round schools students get the same amount of time off, it is just at different times during the year. Short vacations throughout the year give students and teachers much needed breaks and help keep them from burning out. This schedule actually gives families more freedom to plan trips since they aren't limited to traveling in the summer. In addition, ski resort owners say **that a longer break in winter is beneficial because people can spend more time skiing.** My friend says that **students won't mind attending school in the summer if they get to relax during their other breaks.**

Item Stem: Evaluate whether the evidence used in these paragraphs is relevant and comes from a credible source. Click on the highlighted statements and drag them to the appropriate boxes below.

Key and Distractor Analysis:

Not a credible source	Not relevant to the argument	Credible and relevant
their students had higher test scores than students who attended schools with a traditional schedule. **students won't mind attending school in the summer if they get to relax during their other breaks.**	**that a longer break in winter is beneficial because people can spend more time skiing.**	**they have to spend September and October reviewing material taught the previous year.**

Another concern is that academic assessments typically focus on subject matter content, whereas goals for student learning involve both content and cognitive processes, such as problem solving, reasoning, and explaining. Most educational standards documents are structured according to the subject matter to be covered, with the desired cognitive processes embedded within a statement about content (for example, "Students should be able to explain the mechanisms behind the water cycle"). As a result, assessment, instructional design, and claims about alignment between assessments and education standards tend to be driven by concerns about covering subject matter rather than concern with cognitive skills, including those that have been identified as 21st-century skills.

When subject matter content drives the design of assessments and learning materials and cognitive processing requirements are relegated to the background, the tendency is to neglect the higher order or complex cognitive components such as inquiry, problem solving, and explanation (Au 2007; Shepard 1991). The statistics and measurement models conventionally used in developing achievement tests and in interpreting test scores reinforce this tendency. Prevailing measurement models were developed to deal with assessments composed of independent items all sampling discrete skills or knowledge from the same domain. They were not designed to handle the interdependencies among a learner's actions in dealing with complex, multistep problems or inquiries; the presence of feedback after learner actions; or student learning during the course of assessment (Pellegrino, Chudowsky and Glaser 2001; Shute and Ventura in press).

Advances in assessment theory, notably evidence-centered design (ECD) and new statistical techniques and technology tools for supporting the use of ECD in assessment development, are making the assessment of complex cognitive components that are exercised in multiple subject matter contexts much more feasible. ECD and associated tools are being used in the development of the Next Generation Assessments of the CCSS and in learning system R&D. (See sidebar on *Evidence-Centered Design*.)

***Evidence-centered design (ECD)** is a view of assessment as an evidentiary argument— a process of reasoning from the necessarily limited evidence of what students say, do, and make in particular settings to claims about what they know and can do more broadly (Messick 1994).*

The ECD approach to developing assessments (Mislevy et al. 2003) entails the articulation of three models: the student competencies to be measured (the student or competency model), the evidence that will be used to make inferences about whether students exhibit those competencies (the evidence model), and the description of tasks that will produce that evidence (the task model). Together, these three models constitute the conceptual assessment framework, also referred to as the task template, which becomes the framework for developing the task or tasks that the student will see on the assessment.

ECD experts refer to assessment tasks rather than items. Assessment tasks may have multiple components or steps, and the student's response to each is used to model the assessment's estimate of that student's competence relative to one or more KSAs (knowledge, skills, and abilities) related to that step. The evidence model may require combining observed behavior on multiple steps within a task to generate evidence of a student's competence.

In addition, a number of learning researchers have noted that by intention standardized tests measure what students can do during a fixed time working in isolation, without information resources and tools at hand. These kinds of assessments cannot capture collaboration or the judicious use of digital information resources, two competencies on almost everyone's list of 21st-century skills. Performances that matter in work and civic life, on the other hand, involve working with others, using tools and multiple sources of information, and persisting over time with multiple opportunities for revision and refinement (National Research Council 1999).

Finally, we know that personal qualities related to intellectual curiosity, persistence, motivation, and interests can be just as important as subject matter knowledge in shaping students' lives (Almlund et al. 2011). More tools are needed also to assess students' passion for intellectual inquiry in various domains, the way they respond to setbacks and challenges, and the extent to which they have acquired strategies for supporting their own learning.

The increasing presence of digital learning systems and resources in classrooms creates opportunities for collecting these kinds of assessment data to supplement the data captured by conventional large-scale assessments. Learning systems can do this systematically, automatically, and on large numbers of students (U.S. Department of Education 2010a).

New Opportunities Provided by Technology

For many years, digital learning resources, such as computer-assisted instruction and now digital textbooks, have incorporated assessment modules that are very much like the questions at the end of the chapter in a typical printed textbook. Such online quizzes or practice exercises are used to assess student mastery or proficiency. More recently, online learning systems and resources have begun to collect and analyze more fine-grained information about learning processes, such as how quickly a student moves through a simulated environment or a sequence of problems, the amount of scaffolding and support the student needs, changes in a student's response time across problems, and the like.

Embedding assessments in digital learning systems opens up possibilities for assessing features that are important but that could not be measured reliably and efficiently in the past (Pellegrino, Chudowsky and Glaser 2001; Shute 2011). More of what educators really want to assess can be measured by mining the data produced when students interact with complex simulations and tasks presented in digital learning systems.

These measures require greater expertise to analyze, but that expertise can be embedded in digital learning systems. Moreover, the fact that indices such as response latency can be measured across hundreds or thousands of responses gives technology-based assessments a potential edge in generating measures that produce consistent results.

Further, when assessments are embedded in digital learning systems, learners are assessed in the course of learning. Time no longer must be taken away from instruction to stop and measure how much has been learned. If students are working with digital learning systems on an ongoing basis, the amount of course content that can be assessed and the amount of information about what and how they have learned will far surpass what is measured in a discrete test taken once a year.

To use an analogy from baseball, judging an individual student's academic prowess on the basis of a single test given at the end of the school year is like judging a baseball player's skill solely on the basis of performance in the home run derby. Baseball leagues use a much larger set of data; they compute batting, on base, and fielding error averages across an entire season and across an entire career. Digital learning systems and increasingly comprehensive data systems make it possible for education to adopt practices more similar to those used in baseball, to offer more data points for a fuller picture of a students' understanding.

When assessment is done continuously as part of the learning process, administrators can generate aggregate estimates of understanding and performance covering more concepts (that is, have greater content coverage), assess qualities that are difficult to capture in conventional multiple-choice tests (for example, problem solving and persistence), and do so with greater reliability than would ever be possible with once-a-year high-stakes assessments.

A concern with performance assessments has been the high cost of scoring complex performances that entail orchestrating multiple understandings and skills and difficulties in obtaining the needed reliability (Madaus and O'Dwyer 1999). Technology offers the promise of automating the scoring of complex performances, addressing issues of cost and reliability at the same time. A case in point is the automated scoring of student essays. (See sidebar *Intelligent Essay Scoring*.)

As more learning data are collected routinely for each student, opportunities will also arise to develop systems for aggregating learning data collected in different courses, settings, and time periods and to mine these data for new insights. Individual electronic medical records have become an area for rapid development and deployment in health care, and it is not far-fetched to imagine similar efforts over the next five years to create individual learning records that summarize a learner's experiences, learning processes, and accomplishments. DiCerbo and Behrens (2012) have described this concept of assessment information gleaned from an individual's interactions with a variety of digital learning systems and resources and synthesized into a cohesive view of his or her knowledge, skills, and other learning-relevant attributes. For example, the system of certification examinations for all the areas in the new CCSS proposed by Collins and Pea (2011) (discussed in Chapter 3) would produce a record of student proficiencies that is much more detailed and descriptive about what a student can do than achievement test scores or grade point averages.

Expanded Approaches to Gathering Evidence

Chapter 2 described how the fine-grained information about students' learning that newer digital learning systems collect is used to personalize learning. Here, we describe how such data also can be used to construct measures of important learning outcomes and learning processes that have been difficult to capture with conventional state tests.

Evidence-Centered Design

This chapter has noted the tendency in conventional assessment to neglect higher order or complex cognitive components such as inquiry, problem solving, and explanation (Au 2007; Shepard 1991). Traditional test item formats and measurement theory are more suited to capturing discrete bits of subject matter knowledge than to capturing the multistep, multifaceted complex performances that demonstrate deeper learning (see Chapter 1).

Intelligent Essay Scoring

Teachers have long believed that having students write about a topic is one of the best ways to obtain insight into what they do and do not understand about it. But grading essays for 30, 50, or 150 students is so laborious that most teachers make limited use of them as assessment tools.

Software that can score essays automatically (through use of "scoring engines" using artificial intelligence techniques) has been available for some time, and many essay-scoring products are available. To spur innovation and development in this field in time to support implementation of the Common Core State Standards, the William and Flora Hewlett Foundation has sponsored efforts to apply artificial intelligence to scoring essays and other constructed responses used for assessment purposes.

Hewlett sponsored research by Shermis and Hamner (2012) comparing the scores provided by nine commercially available automatic essay scoring engines and those produced by pairs of human scorers using state-developed scoring rubrics. The essays were student responses to writing prompts on six states' high-stakes writing assessments. Overall, each of the scoring engines produced essay scores very similar to those produced by human scorers.

At the same time, hoping to attract data scientists and machine learning experts to work on improving intelligent essay scoring, Hewlett offered $100,000 in prizes for the three scoring engines that could most accurately mimic human essay scorers. Kaggle, a platform that runs data prediction competitions in a variety of fields, ran the competition for Hewlett.

A set of varied student essays from 150 to 550 words long that had been scored by two humans were provided to contest entrants so that they could "train" their scoring engines, comparing their automatically generated scores with those from humans. Competitors were given three months to build and train their engines and then were given a new set of essays for the competition phase. The scores generated by scoring engines were compared with the consensus scores from the two human graders. The 11 people on the first-, second-, and third-place teams all came from non-education fields, including computer science, data analysis, and particle physics (Quillen 2012). Their engines used predictive analytics in addition to the computer's ability to process natural language.

Recently, assessment research and development has turned to evidence-centered design to guide the development of assessments capable of dealing with multiple competencies and complex performances (Mislevy and Haertel, 2006; Mislevy, Steinberg and Almond 2003). ECD is a systematic process in which the designer articulates (a) the competencies to be measured, (b) what would constitute evidence that a learner possesses those competencies, and (c) the situations or tasks that can be used to elicit that evidence. Assessment developers find that the ECD framework helps them connect what they want to assess to specific learner actions in complex task contexts (Shute and Ventura in press).

The use of ECD is not limited to assessments embedded in complex digital learning systems, but it is particularly useful for this purpose because conventional models come up short in these circumstances.

ECD can also contribute to the validity and quality of assessments. Its application requires considerable skill and careful documentation. Each task first must be carefully constructed to elicit a response related to a specific aspect of the student model. Then each step in the chain of reasoning must be linked to that specific aspect and to the specific actions the student takes in response to the task.

Mislevy and Haertel (2006) developed the computer-based system PADI to support this process. (For more information on PADI, see the sidebar *Design Patterns and Principled Assessment Designs for Inquiry.*)

Design Patterns and Principled Assessment Designs for Inquiry

Drawing on the framework of evidence-centered design (Mislevy et al. 2003; Mislevy and Haertel 2006), in the Principled Assessment Designs for Inquiry (PADI) project Geneva Haertel and Robert Mislevy developed a design pattern for the efficient development of multiple tasks that assess complex knowledge, skills, and abilities (KSAs).

Design patterns have several benefits for designers of complex assessment tasks. First, they facilitate the transition from knowledge about the domain to an operational assessment system. They keep the focus on the conceptual level and help the designer avoid moving too quickly to an item's technical elements. Complex psychometric concepts are not required to understand and use design patterns. The plain language in design patterns facilitates communication between content experts and assessment specialists.

Second, design patterns increase the validity of an assessment by explicating a structured assessment argument. Design patterns set forth the KSAs to be assessed, the performances or behaviors that reveal those constructs, and the tasks or situations that can elicit those performances. The discipline imposed in explicating the structured argument enhances the coherence of the assessment components and thereby the validity of evidentiary reasoning (Cronbach and Meehl 1955).

Third, design patterns facilitate decision making for task designers in assessment design. This design tool clarifies the explicit and implicit constraints and resources that will affect the design, development, and delivery of the actual assessment tasks.

Finally, design patterns can be generalized to a variety of content domains, grade levels, and student populations. For example, a design pattern on "observational investigation" provides a general design space that crosses different science domains and can be used to generate a family of assessment tasks.

Examples of some of the 100 design patterns that Haertel, Mislevy, and their colleagues have developed for science assessments are Experimental Investigation, Observational Investigation, and Model Revision in Model-based Reasoning. These design patterns and the others are in a library that is part of the Web-based PADI assessment design system. The online system not only supports the creation of design patterns, but also highlights associations among their attributes to support the task development process. In addition, the design system can generate a hierarchical picture of related design patterns. The efficiency of assessment task development can be improved by exploiting the relationships among similar design patterns.

The PADI design system also offers a linked glossary to help users with the language of ECD. Together, these technology-enhanced features of the online system make it easier for an assessment designer to exploit the connections among design patterns and task templates to design and develop assessment tasks.

SRI assessment researchers have used PADI and design patterns in a number of projects, demonstrating their utility in different assessment contexts. These projects include the development of a statewide science assessment, assessments for community college courses in economics and biology, and the design of alternative statewide assessments in reading and mathematics for students with significant cognitive disabilities.

Assessing Achievement During Learning

Educational accountability systems have directed the attention of schools and districts on students' performance on end-of-year state achievement tests. Whether or not a student's scores on these tests meet or exceed the proficiency requirement has consequences for superintendents, principals, and teachers. An entire industry has grown up around the provision of assessments that can be administered during the school year to identify students at risk of failing to score proficient on the end-of-year exam.

Critics point to the time this interim assessment is taking away from instruction, while advocates point to the usefulness of assessing during the school year when there is still time to give extra support to those students who need it.

With funding from the U.S. Department of Education, researchers at Worcester Polytechnic Institute and Carnegie Mellon University developed the Web-based ASSISTments system to address this issue. ASSISTments combines online learning assistance with assessment capabilities (Feng, Heffernan, and Koedinger 2009). While teaching middle school math concepts, ASSISTments uses information from learners' interactions with the system to provide educators with a detailed assessment of students' developing math skills.

When students respond to ASSISTments problems, they receive hints and tutoring to the extent they need them. The system helps students break hard problems down into subparts. Questions associated with the subparts are designed to elicit student responses that will reveal the reason why the student initially gave the wrong answer. Students can ask for stronger and stronger hints as needed to arrive at a correct problem solution. From the students' perspective, they are using ASSISTments to learn; there is not a time when learning stops for test taking.

ASSISTments treats data on how individual students respond to the problems and how much support they need from the system to generate correct solutions as assessment information. The ASSISTments system gives educators detailed reports of students' accuracy, speed, help-seeking behavior, and number of problem-solving attempts as well as their mastery of 100 middle school math skills.

ASSISTments research (Feng, Heffernan, and Koedinger 2009) has found that information on how students respond after an initial wrong answer predicts performance on the end-of-year state examination better than the number of problems a student got correct on his or her first try (the measure used by conventional interim assessments). By combining information on the number of items correct on the first try and the way the student worked with the system after a wrong answer, ASSISTments produced predicted MCAS (Massachussetts Comprehensive Assessment System) test scores with a .84 correlation to the scores the students actually obtained at the end of the year.

Using ECD and an ontology of critical cognitive processes such as problem solving, reasoning, and explaining (Koenig et al. 2010), assessment developers are creating design patterns and task templates for those cognitive processes that require students to demonstrate the process with different content. Reusable templates and objects, including validated scoring or judgment systems, are parts of this process (Vendlinski, Baker and Niemi 2008).

Moving cognitive requirements to the foreground in assessment task design in this way produces more coherence among assessment tasks and development is faster, reducing costs. Assessments designed using the same cognitive task models with different content within subject areas (and even between some subject areas, such as science, math, and social studies), provide greater coherence and facilitate transfer and generalization across topics, situations, and contexts. Teachers can then design or evaluate assessment tasks in terms of coverage of the cognitive demands of a state or national standard as well as its subject matter content. From the evidence side, a model can be tagged to indicate that it produces micro data for both content and cognitive learning progressions.

Mining Data from Learning Systems to Assess Cognitive Skills

The practical advantages of embedding assessments into digital learning systems are well illustrated by the ASSISTments project at Worcester Polytechnic Institute. Research with ASSISTments has demonstrated that information on how a student interacts with the learning software—in particular, how a student responds after answering a problem incorrectly—can improve predictions of future student mathematics performance (Feng, Heffernan, & Koedinger, 2009). (For more information on the ASSISTments project, see the sidebar *Assessing Achievement During Learning*.)

Using emerging educational data mining and learning analytics techniques to analyze learner log files from digital learning systems has potential for broadening the scope of educational assessment. Researchers are demonstrating that both these techniques can be used to analyze a series of actions not only within structured tasks like those in ASSISTments, but also within more open-ended exploratory environments in which learners take on avatars and interact with virtual characters and objects as they try to solve a realistic problem, such as identifying the cause of an epidemic in a 19th-century factory town or the cause for a sudden decline in the kelp population in an Alaskan bay. (For an example of such an exploratory environment, see the sidebar *Assessing Inquiry Skills in Virtual Environments.*)

Assessing Inquiry Skills in Virtual Environments

Chris Dede and his colleagues at the Harvard University Graduate School of Education have been studying the use of virtual worlds (immersive environments) for science learning and assessment.

This work began with River City, a re-creation of a city in the 19th century when scientists were just beginning to discover bacteria. Each student is represented as an avatar and communicates with other student avatars through chat and gestures. Students work in teams of three, moving through River City to collect data and run tests in response to the mayor's challenge to find out why River City residents are falling ill. The student teams form and test hypotheses, analyze data, and document their research in a report they deliver to the mayor.

Student inquiry activities in River City can be assessed by analyzing the research reports and also by looking at the kinds of information each student and each student team chose to examine and their moment-to-moment movements, actions, and utterances in the virtual environment. On the basis of students' actions in River City, researchers developed measures of their science inquiry skills, sense of efficacy as a scientist, and science concept knowledge (Dede, 2009).

Dede (2012) asserts that the open-ended nature of this kind of virtual environment more closely matches the kind of learning that happens in internships and the real world than either conventional classroom instruction or the more constrained interactions in online tutoring systems.

In the ongoing Virtual Performance Assessment project, the Harvard team is studying the feasibility of using simulation environments to assess hard-to-measure learning outcomes, such as science inquiry skills, in a way that would be suitable for use in accountability systems. The researchers' goal is to produce simulation-based assessments linked to national standards for science inquiry practices with demonstrated validity and reliability (Clarke-Midura, Dede, and Norton 2011).

The development team is using evidence-centered design and the PADI (Principled Assessment Designs for Inquiry) assessment design system to create multiple forms of the same assessment to demonstrate the feasibility of implementing this kind of assessment at scale (Dede, 2012).

As described by Zapata-Rivera (2012), researchers are applying educational data mining techniques to uncover interesting patterns in the digital log files generated by digital games. The virtual environments being developed at Harvard and the Newton's Playground game described below are examples of the fast-growing body of research on game-based assessment.

Mining Data from Learning Systems to Assess Non-Cognitive Skills

Research has demonstrated the importance of personal qualities such as conscientiousness and self-efficacy in college and workplace success (Almlund, Duckworth, Heckman, & Kautz, 2011; Pellegrino & Hilton, 2012). But education systems do not measure these noncognitive qualities explicitly. In research, these qualities are generally measured through self-report inventories. Yet such inventories are very susceptible to social desirability effects (the inventory takers tend to respond in ways that make themselves look good). An even greater concern is that inventory responses consistent with a trait are easy to fake if an inventory is used in a situation where consequences are attached to an individual's responses.

The availability of technology to create and support more sophisticated digital learning systems offers the opportunity to measure these qualities on the basis of students' behavior in a learning system rather than through self-report. For example, Shute and Ventura (in press) described how persistence (i.e., inclination to work hard even in the presence of challenging conditions) could be measured in a digital learning system. They pointed to the possibility of using learner actions, such as the average amount of time the learner chooses to spend on difficult problems, the number of retries after failure, and returning to a difficult problem after skipping it, as examples of the kinds of learning system data that could be used to construct a reliable measure of learner persistence. (For an example of one of the embedded assessments being developed, see the sidebar *Embedded Assessments in Newton's Playground.*)

Understanding how to support the development of these noncognitive skills and how to assess them are priorities for the U.S. Department of Education (Easton 2012). The Department has prepared a brief on grit, tenacity, and perseverance. Slated for release in January 2013, the brief summarizes current research on these skills and offers recommendations for R&D priorities in this area. The authors propose that grit, tenacity, and perseverance are teachable and made up of three components: academic mindsets (cognitive framings that support perseverance), effortful self-control, and strategies and tactics (such as adaptation). The brief recommends that students be given opportunities to develop these skills by pursuing optimally challenging longer term goals while having access to the supports needed to achieve the goals. It identifies further exploration of how perseverance functions in a wide range of settings and academic disciplines as research priorities, calls for design-based implementation research to connect theory and practice, and highlights the need for longitudinal studies.

Uses of Evidence from Embedded Assessments

Assessments embedded in learning systems, such as those featured in this chapter, have advantages for students, teachers, and education systems because they can measure important student outcomes that are not captured well by conventional assessments and do not require taking time away from learning to test for past learning. However, researchers are grappling with some open questions about embedded assessments.

First, embedded assessments are tied to specific products or environments, raising questions about the extent to which performance on them really predicts what students would do in other contexts. This problem applies to any assessment, but when assessments are embedded in particular digital learning systems they are particularly susceptible to being overly aligned with the content of those

Embedded Assessments in Newton's Playground

Valerie Shute, a researcher at Florida State University, is leading a project to design, develop, and validate unobtrusive assessments embedded in a digital game. The team is building the assessments inside Newton's Playground, a computer game designed as an assessment and learning environment for Newtonian physics.

Learners interact with a set of problems displayed as simple two-dimensional simulations. Each problem in the game challenges the learner to use Newtonian principles to get a green ball to move from its starting point to the location of one or more balloons. The learner can draw any of a set of objects (for example, a ramp, lever, or pendulum) on the screen, and those objects will "come to life" and move the ball according to Newton's three laws of motion. For example, a learner can draw a ramp on the two-dimensional space to change the direction of a ball in motion. Students work on problems and can retry them as often as they like. Some students retry problems they have solved to find a simpler, more elegant solution, which would earn them a gold trophy if successful.

Shute's research team is trying to assess both the extent to which students acquire the ability to apply Newtonian principles of motion correctly to novel problems and their persistence. The team conceptualizes persistence as the amount of time students will spend on problems they cannot readily solve. The challenge in designing this kind of assessment is the impossibility of predicting which problems will prove challenging for a given student. Using evidence-centered design, the R&D team created a difficulty rubric for the game's problems and is using it to systematically build problems of varying levels of difficulty so that they will be able to make sure that every student eventually faces problems that are difficult for him or her. (See Shute & Ventura, in press, for details of the rubric and illustrations of the games.)

Other conceptual models being developed by the team concern conscientiousness, physics concepts, and creativity. These outcomes are being modeled and will be measured automatically as students use Newton's Playground (Shute & Ventura, in press).

products. Theoretically, such embedded assessments might be specific to the particular learning system, and the learning system might introduce difficulties not relevant to the construct or constructs being assessed that affect estimates of students' competence. On a practical level, when assessments are integrated with a particular learning system, their use typically will be limited to classrooms using that system. Technical solutions to this limitation are feasible if assessments are carefully designed to avoid construct-irrelevant variance (for example, through the application of ECD) and developed as objects that can be embedded in any number of systems.

Second, unlike conventional assessments, embedded assessments often provide students with feedback. This is advantageous because students can learn from the feedback, but it means that the students are learning about a concept or how to execute a skill at the same time the system is attempting to gauge their competence in that knowledge or skill. Shute, Hansen, and Almond (2008) found that adding feedback within a system assessing high school students' ability to work with geometric sequences did not diminish the system's ability to assess student competence. More research of this nature is needed.

Whether assessment is conducted as a separate activity or occurs in the background during the course of learning online, the fundamental questions of reliability and validity apply. We must ask whether an assessment yields consistent results about a student's state of learning or competency (reliability) and whether the assessment provides adequate empirical evidence to support the inferences being made (validity). Modern thinking emphasizes that validity resides not in the assessment itself but in the strength of evidence supporting the inferences made on the basis of assessment results. Accordingly, establishing assessment validity, like establishing educational intervention quality, is not a one-time event but a continuous process of data collection and refinement (Cizek, Rosenberg and Coons 2008).

The application of ECD and data mining to learning systems for assessment purposes needs to be accompanied by the collection of evidence of validity and reliability. If efforts are successful to develop psychometrically sound assessments that go on in the background as students use online learning systems, educators can start to question the value of once-a-year achievement tests. A number of research groups are working on this issue of how to make data gathered from online learning systems useful within accountability contexts as well as for individual learners and teachers (U.S. Department of Education 2010a).

Conclusion

This chapter describes how the data collected by digital learning systems can be used to expand and improve both the content and process of assessment beyond student achievement tests that focus on subject matter content. For example, formative assessments administered in the course of learning can guide future learning and can provide insight into how students think and approach problems, not just the proportion of time they arrive at correct answers. Advances such as evidence-centered design and new statistical techniques and technology tools for supporting the use of ECD-based assessments embedded in digital learning systems are explored, as is mining data from learning systems to assess both cognitive and noncognitive skills.

Chapter 5:
Finding Appropriate Learning Resources and Making Informed Choices

Learning resources and materials are critical in achieving desired learning outcomes. What better supports do educators need as they make decisions about which digital learning resources to adopt?

With digital learning resources readily available on the Internet, many teachers and a growing number of schools are using them to expand learning and to supplement or replace print-based materials such as textbook chapters and exercises. Digital options include rich media, interactive textbooks, complete online courses, and supplemental materials.

While these digital resources give educators more choices, they also raise the issue of how to ensure their quality and determine their effectiveness in achieving the desired learning outcomes. The learning resources chosen are crucial in both what and how well students learn (Chingos and Whitehurst 2012; Schmidt et al. 2001).

Evaluating and choosing digital learning products can be daunting for many reasons. First, the Internet is a fast and far-reaching distribution channel, so the number of new products available grows every day. Second, many of the new products being offered are from sources new and unfamiliar to education decision makers rather than from tried-and-true suppliers. Third, some of the business models used by Web developers are new to education, for example, offering "fremium" versions of products at no cost but charging subscription fees for full-featured versions. Fourth, some of the attributes of digital resources—for example, that they can be continually refined and improved or even modified by users—make them a moving target when it comes to evaluating them.

As a result, one or both of two things can happen: Excellent and effective digital learning resources may be underused because educators cannot find them among all the choices available, and resources that are chosen may not be effective or may not fit within the constraints of a particular classroom or learning environment (for example, the length of the class period, curriculum context, or available bandwidth).

New Opportunities Provided by Technology

Besides the Internet, two other factors are driving the trend of teachers supplementing print-based textbooks and other materials with digital learning resources:

- easy-to-use creation and publishing tools that enable anyone to create, configure, aggregate, and modify learning materials (supported by Creative Commons licenses); and

- Internet-supported resources for educators such as online repositories and communities that make it easier for users to find and evaluate resources that might meet their needs.

Creative Commons (creativecommons.org) provides customizable copyright licenses free online for creators and authors of works ranging from writings and videos to songs and computer programs or images. Typically, works licensed under Creative Commons have copyrights that are less restrictive than the automatic "all rights reserved" copyright so that others may more freely share, use, and remix them.

User-Generated Learning Resources

Anyone can shoot an instructional video and upload it to YouTube. That is how the Khan Academy was started. Sal Khan was creating short videos to help a young cousin who wanted to improve her math scores (Thompson 2011), and he posted them on YouTube so she could view them easily. Today, millions of people around the world view and learn from Khan Academy videos.

Many user-generated resources are available for use and reuse at no cost. An entire movement, Open Educational Resources (OER), has facilitated the growth and distribution of these open user-generated materials. Modifications are already happening globally as people are taking open educational materials and programs from one nation and adapting them for the norms and needs of another. These changes can be as basic as language translation or as involved and sophisticated as improving the fit with specific learners' background knowledge, learning pace, or interests.

Open educational resources (OER) are teaching, learning, and research resources that reside in the public domain or have been released under an intellectual property license that permits sharing, accessing, repurposing – including for commercial purposes – and collaborating with others. These resources are an important element of an infrastructure for learning. Originating in higher education, OER forms range from podcasts to digital libraries to textbooks, games, and courses, and they are freely available to anyone over the Web. The OER movement was started by universities making their learning content available online free of charge (Smith 2009), and it is now well entrenched in K–12 education. In August 2012, the OER Commons contained more than 28,000 free openly licensed K–12 learning resources.

Online Repositories and Communities

Because users can modify OERs, they often tailor and combine them to create "best-of-breed" assemblages. For example, teachers, districts, and states are increasingly creating digital curricula by combining OER-based materials from multiple sources. A number of online repositories and communities are springing up to support these efforts, enabling users to search curated collections of materials, upload and share their own material, read and write reviews, create "playlists" of favorite resources, and interact with other users. (Short descriptions of some of these online repositories are in the sidebar *Examples of Pulling Together Learning Resources from Multiple Sources.*) In some cases, these interactions serve to improve and refine a learning resource over time. (For an example of a learning resource being improved in this way, see the sidebar *Collaborative Research and Development on the Pathway to College Math* in Chapter 1.)

Many of the online repositories and communities are using Internet-supported techniques to help users find resources that might meet their needs. Resources can be tagged according to established categories (for example, the educational standards they address or the grade level they were designed for). The sites themselves might categorize the resources, allow users to categorize them, or both.

Another new approach to capturing, sharing, and analyzing information about digital learning resources is the Learning Registry. Recently launched by the U.S. Department of Education and the U.S. Department of Defense, the Learning Registry stores data provided by numerous sources about the content of various learning resources hosted in different systems. Data published to the Learning Registry can serve as the basis for learning resource analytics to help recommend resources, detect trends in resource usage, and judge user experiences. (For more information on the Learning Registry, see the sidebar *Sharing of Information About Learning Resources Across Systems*.)

Examples of Pulling Together Learning Resources from Multiple Sources

BetterLesson is a curriculum-sharing platform containing more than 300,000 teacher-contributed preK–12 lessons that users can browse and search using key words and tools for creating collections. BetterLesson is free to individual teachers; school districts pay a subscription fee.

Gooru is a nonprofit organization with a free platform for students and teachers that offers access to a curated collection of 50,000 open educational resources for grade 5 through 12 mathematics and science. These resources range from digital textbooks to individual animations to games, all tagged to the Common Core State Standards and California science content standards they address.

LearnZillion is a learning platform that combines video explanations, assessments, and progress reporting. Each lesson highlights a Common Core Standard, starting with math in grades 3 through 9. The site offers more than 2,000 lessons created by teachers using a Web-based application. Lessons are free.

Open Tapestry is a website that allows users to find, organize, and share education resources. Users can adapt a variety of content retrieved on the website to suit their individual needs, as well as contribute new information. Users may also integrate Open Tapestry into their learning management systems.

PowerMyLearning is a platform developed by nonprofit CFY formerly (Computers for Youth) that has more than 1,000 digital learning activities. Free to teachers, PowerMyLearning lets them build a playlist of activities and add their own instructional text to introduce them.

Share My Lesson is an online portal created by the American Federation of Teachers that now contains more than 250,000 digital learning resources reviewed and prepared by 200 teachers. The lessons include OERs that can be remixed, reused and reposted. The portal also includes a community where teachers can pose questions or reactions to the resources.

Sharing of Information About Learning Resources Across Systems

The Learning Registry, an open-source software project, provides the technical infrastructure and community practices for sharing information about learning resources across systems. It does not impose standards for how to represent data but instead provides opportunities for communities to discuss and agree on real-world practices. In this way, the Learning Registry helps alleviate the problem of disparate metadata standards and missing metadata by changing the business model for digital content suppliers from hand-curation of metadata (the "library model") to tapping data streams from social networks and learning management systems (among others) to locate and identify resources (the "recommender model").

The Learning Registry began as a project funded by the U.S. Departments of Education and Defense to share information about learning resources from federal repositories such as the Smithsonian, the National Archives, and the Library of Congress. It has evolved into a mechanism for taking advantage of metadata and social metadata generated as educators and learners interact with online learning resources and systems, including learning object repositories, teacher portals, search tools, learning management systems, and instructional improvement systems. (Social metadata have been locked in to these separate systems.)

The Learning Registry enables the learning resource information created by one site to be shared with others. Learning resource data collected from these sources and published to the Learning Registry network can serve as the basis for learning resource analytics to help recommend resources, detect trends in resource usage, and judge user experience.

At present, the Learning Registry community is exploring new and interesting ways to use Learning Registry data, such as recommending resources, visualizing trending resources, and analyzing connections among resources. California's Brokers of Expertise and CTE (Career and Technical Education) Online sites and Florida's CPALMS site are now part of the Learning Registry network, sharing resources, ratings, and alignment data. North Carolina, Massachusetts, and Ohio have announced projects that will connect their instructional improvement systems to the Learning Registry. The National Science Digital Library and PBS (the Public Broadcasting System) have both connected to the Learning Registry network. The Learning Registry can also link to educator-generated or commercial resources. A list of early collaborators is on www.learningregistry.org.

As more data are published to the Learning Registry, the possibilities expand for using it to provide different kinds of evidence for learning resources. The Learning Registry affords a unique opportunity to help collect, amplify, and aggregate evidence for recommending learning resources.

The Learning Registry community is currently focusing on supporting data that reflect standards alignment because of the sharing across states possible with the Common Core State Standards. When a resource stored in a digital repository or created by a teacher (e.g., posted on BetterLesson) is aligned by a state or local education authority, that alignment provides evidence about the content it purports to teach. These alignments are being captured in the Learning Registry. When teachers searching for standards-aligned content at a state portal locate a resource, they can view how other state or local entities have aligned the resource as well as the other social metadata on actions such as the number or downloads and ratings for that resource.

The trends of creating large online collections of disparate resources and of teachers mixing and matching learning resources raise the issue of how well resources drawn from different places fit into a coherent whole. Research relating the quality of a curriculum to learning outcomes stresses the importance of curricular coherence (Schmidt and Houang 2012).

Because open resources come from many different places and were developed for many different purposes and kinds of users, they do not necessarily use consistent terminology or representations and gaps can be left in students' understanding. Creating coherent learning activities and curricula from diverse sets of learning resources requires considerable skill and effort. Some R&D groups are taking up this challenge. (For an example of such an effort, see the sidebar *Supporting the Creation of Coherent Curriculum Units Incorporating Digital Resources.*)

Supporting the Creation of Coherent Curriculum Units Incorporating Digital Resources

Under a grant from the National Science Foundation, the Institute of Cognitive Science and the University Corporation for Atmospheric Research at the University of Colorado developed the Curriculum Customization Service (CCS) to support instructional planning of middle and high school earth science teachers in the Denver Public Schools.

CCS is a Web-based system that contains a curriculum planning interface and resources from the grade 6 and grade 9 earth science curricula of the Denver Public Schools plus interactive digital resources from the Digital Library for Earth System Education. Resources within CCS include Top Picks, images, animations, and activities that call on students to use scientific data.

When using CCS for planning, teachers start with either a unit from the district curriculum or a specific learning goal. In response to their selection of a unit-learning goal, the system displays a set of key concepts and the instructional resources that support learning each of those concepts. Teachers can also identify instructional resources from outside the system and bring them in with a tag to the concepts they address.

In this way, teachers are supported in planning that works backward from the intended learning outcome (Wiggins and McTighe 1998) rather than the common practice of selecting an activity and then trying to find a place for it in the curriculum (Sumner et al. 2010). Teachers can create personal collections of resources they like and annotate resources with their thoughts on how to use them or the types of students they are most appropriate for. When teachers upload resources to CSS, they have the option of choosing to share them with other teachers in their district.

A field test with 124 teachers conducted in fall 2009 found that more than half the teachers reported using CCS digital resources as much as or more than their textbook materials. Participating teachers made heavy use of the resources that other teachers had uploaded and tagged for sharing and reported that CCS made it easier to find instructional resources relevant to their teaching (Sumner et al. 2010). The system was implemented in four school districts in 2010–11.

Expanded Approaches for Gathering Evidence

User-generated learning resources and online repositories of learning materials offer more options for educators but raise questions about the effectiveness of specific products and resources. As for any learning product, when evaluating digital learning resources, educators should consider multiple criteria, such as

- design variables, including alignment with standards, whether the resource addresses the desired learning outcome, and its accessibility for all students;
- product fit, including whether students find the resource engaging and whether the theory of learning underlying it matches a specific learning approach;
- implementation issues such as how easy the resource is to use, whether a teacher or student will have to undertake some preparatory training, and the kind of technical requirements it has;
- cost and time needed; and
- evidence of effectiveness and specifically effectiveness for students like theirs in settings like theirs.

The desire to continuously improve our understanding of the usefulness of digital learning resources, including the degree to which evidence exists of effectiveness, is prompting technology developers, companies, government entities, and nonprofit organizations—separately and working together—to develop new ways of gathering and publishing information and evidence about these resources. Methods include

- aggregating user actions;
- aggregating user reviews;
- user panels;
- expert ratings, reviews, and curation; and
- test beds.

Developed by consumer-oriented websites, these approaches have become a familiar part of consumer decision making of all kinds and are now being explored and applied to education.

Aggregated User Actions

Three types of user actions can be aggregated to form evidence of popularity: (1) rating, voting, and ranking; (2) clicking, viewing, downloading, and sharing to social media; and (3) actions connected to the use of the learning resource in instruction, such as aligning, implementing in some context, and adapting learning to individual learners. Information of this kind about digital learning resources is typically found in online repositories or communities.

When a teacher visits a repository or community and selects a learning resource, that action indicates interest in the resource. Students' use of that resource is captured as another data point about usage. When the teacher reflects on how well the resource worked with students and adds a rating or shares it with other teachers, more data are accumulated.

Further, the electronic record created when teachers or students rate, comment on, and download instructional resources can be analyzed through educational data mining. These records enable analysts to apply statistical models to identify groups of similar users for the purpose of recommending resources based on those selected or rated highly by other members of the same user category (Amershi & Conati 2009). These data are also mined to explore the relationships among the various resources selected by a particular user or cluster of users (Romero and Ventura 2010). In these ways, user activity can contribute to improving the ability to predict which learning resources a user new to the site will be interested in.

Aggregated User Reviews

Through online reviews, consumers can learn from the experiences of others—many, many others in some cases—before making a decision when shopping online, selecting a restaurant, planning a trip, or finding a doctor. Reviews can be found on the websites of large online retailers as well as on commercial online review sites such as Yelp, TripAdvisor, Angie's List, and Service Master.

In education, user reviews of digital learning products are now becoming available, typically on sites whose primary content is reviews or collections of resources with associated review and rating features. Many of the resource-sharing platforms also offer reviews and ratings features. Decision makers looking for information about which resources to adopt can benefit from user reviews, but, equally important, the reviews can be used by developers to improve their products. (Short descriptions of some of these review sites are in the sidebar *Digital Learning User Review Sites*.)

The quality of user reviews of digital learning resources as evidence depends on the number of reviews, the relevance of the reviews to a user's needs, and how transparent the website that hosts the reviews is about how it presents the reviews to users. The value of reviews increases exponentially with the number of people providing them.

Further, reviews are more useful when the categories or factors that reviewers use in providing comments are standardized from one review to the next. For example, when the categories are explicit, users can tell what underlying factors combine to create summary ratings. Understanding the various factors makes it easier for educators to determine the relevance of reviews to their needs. It is also helpful for those providing the ratings to have access to information on the time frame and sources of ratings, characteristics of raters, and any standardization of rating scales.

Digital Learning User Review Sites

EdSurge has short descriptions of products and how they are used, what content areas and grade levels they cover, a sense of who uses the them, costs and technical factors, and results. EdSurge serves both users and developers, presenting statements from vendors alongside users' comments that may corroborate or contradict what the vendors say.

Curriki lets users share and access educational information with the goal of lowering economic, political, and geographical barriers to learning. The resources available on Curriki are also reviewed by an in-house review team, and users may also rate and make comments about them.

Classroom Window enables educators to search for and review digital learning products. A "report card" review template asks reviewers to respond to a set of questions by selecting numerical rankings or terms from drop-down menus and to rank effectiveness for different kinds of learners rather than selecting just one overall ranking. The site aggregates information from reviews users post and sells it to developers to inform product improvements.

Edshelf lets users create and share collections of their favorite learning resources. It is also an app store that enables users to search for and purchase tools directly from the site. Users are invited to review tools and rank them according to such criteria as ease of use, pedagogical effectiveness, and student engagement.

User Panels

User panels are a relatively new practice in market research that enable market research firms, corporate brands, and public agencies to efficiently and cost-effectively conduct, gather, and share quantitative and qualitative research on a continuous basis. User panels are sizable managed online communities (typically more than 5,000 members who are compensated in some way for their ongoing participation) that are used to

- provide a prompt feedback loop to connect developers and targeted users to test and review product design from inception to launch;
- test a product's usability, utility, pricing, market fit, and other factors; and
- gather information about user needs and behavioral patterns for specific products or product categories for purposes of product improvement.

Many user panels use social media as key elements of their development and marketing strategies. BzzAgent is one example in the consumer world. BzzAgent runs an online community user panel of over four million members. Members try out products that are not yet on the market, such as new ice cream flavors or dish soaps. BzzAgent monitors community activity to provide feedback for the companies developing the products. The community members also help generate interest in new products being launched by sharing the products or discounts on the products with friends. Feedback from BzzAgent to its client companies can affect how products are designed and how they are marketed.

An important difference between a collection of user reviews and user panels is that panels are ongoing designed and managed market research studies. Panel members are invited to participate in studies or campaigns for which their input is expected to be relevant and valuable. Input takes place through conversation within and outside the panel, but panel members also complete surveys designed by researchers who have tested them for validity and reliability. Large collections of user reviews are also analyzed for patterns in user preferences but generally without controlling for user characteristics.

As yet, no user panels focus on digital learning, although some of the new and existing online communities such as EdSurge, Classroom Window, and Edmodo could be logical places to recruit interested and knowledgeable participants. Teachers in existing networks of schools or in districts already cooperating through procurement consortia could also be recruited for user panels.

As in the case of user reviews, discussed above, numbers add strength to user panels. Input from a several thousand people is more likely to lead to findings that can be generalized to various populations of interest.

Expert Ratings, Reviews, and Curation

In contrast to users, who typically report on their own personal experiences with a product in their reviews, expert reviewers draw on both their specialized knowledge relevant to a product experience and their own experiences. Some also present research findings. Moreover, experts may be more likely than regular users to provide complete, objective reviews and opinions about specific features of learning resources such as whether they are aligned with learning theory. Examples of sources of expert reviews in the consumer world are full-blown testing organizations such as the Consumers Union and specialized publications such as PCWorld or CNET.

Consumer Reports, published by the Consumers Union, has long been a household name for high-quality, objective expert reviews on everything from household products to cars. The organization develops its ratings by conducting its own product tests and large-scale consumer surveys.

Expert reviews of digital learning resources are becoming more widely available. One organization, Common Sense Media, has reviews and ratings for parents on all types of media aimed at children. Reviews are created by experts, although parents and teachers can also submit ratings. Common Sense Media launched a beta version of its "education ratings initiative" in March 2012, adding the educational value of the children's media it rates as part of its reviews.

The experts' reputations—including their partnerships or affiliations with any of the companies whose products they review or other potential conflicts—and the relevance of the review to a user's particular context help determine how useful reviews are as evidence. Ideally, complete information on any research results included in the reviews should be presented, including the characteristics and number of students involved, how the product was used, and the like.

Although it is not targeted for digital learning per se, another example of an expert resource devoted to providing evidence of the effectiveness of learning resources is the What Works Clearinghouse (WWC) of the Institute of Education Sciences. The WWC was established in 2002 specifically for identifying educational interventions for which rigorous evidence of effectiveness exists. Once an intervention has been chosen for WWC review, a systematic, well-documented process for locating studies and judging their quality is implemented (U.S. Department of Education, 2012a). One category includes digital resources. (More information can be found in the sidebar *What Works Clearinghouse Reviews of Digital Resources*.)

What Works Clearinghouse Reviews of Digital Resources

The WWC has published intervention reports on 45 digital learning interventions, of which 26 were found to have positive or potentially positive effects on at least one outcome.* The recently redesigned WWC website makes these intervention reports available to educators who can search for interventions addressing different outcomes (such as academic achievement, language development), grades, student populations (for example, English language learners, general), and intervention types (curriculum, supplement, practice).

The WWC is a useful source of information on the effectiveness of well-established digital learning interventions, but it is not feasible for it to address practitioners' every need with respect to evidence or to do so as quickly as educators require evidence. The WWC process for determining whether an intervention is effective depends on having publicly available research on the intervention to review. Given how long it typically takes to conduct rigorous studies and publish the results in peer-reviewed journals, the WWC's work cannot keep up with the supply of new digital learning resources.

Even when a WWC review is available, it may address a much earlier, less powerful version of a technology than that currently being disseminated. For example, seven of the 26 learning technology interventions with positive or potentially positive reviews on the WWC in August 2012 were no longer available at that time.

* For these interventions, 1,406 studies were reviewed and just 78 of them met WWC standards of evidence with or without reservations. Although 6 percent of studies sounds very small, this is the same percentage of reviewed studies meeting WWC evidence standards for educational interventions overall, suggesting that the studies of the effectiveness of digital learning interventions are neither more nor less rigorous than educational intervention studies overall.

Test Beds

The use of test beds for learning technology not only makes effectiveness evidence available, but also generates more evidence. In science research, a test bed can consist of a specialized environment along with the equipment and staff needed to run tests. For example, the National Oceanic and Atmospheric Administration (NOAA) supports test beds for various types of climate research. This type of test bed is an ideal environment. By contrast, test beds in education are real environments—regular classrooms, for example—in which certain conditions, such as data-sharing agreements among participating schools or districts, are in place.

Test beds in education can consist of a network of schools or classrooms and a community of researchers who have committed to working together and have access to the resources necessary to a given study (such as classroom technology). The Innovation Zone, or iZone, of New York City Schools, has a test bed within it. (See the sidebar *Researchers and Schools Collaborate on an Education Test Bed*.)

Researchers and Schools Collaborate on an Education Test Bed

iZone includes approximately 250 schools interested in personalizing student learning by using new practices and technology. In this test bed project, iZone has partnered with Research Alliance, EdSurge, ChallengePost, and IDEO to support developers in rapidly developing and testing selected technology-based instructional supports and featuring test results on EdSurge. With support from the U.S. Department of Education and the Bill & Melinda Gates Foundation, the New York City Department of Education plans to launch a prize competition for developers, calling for new digital learning resources that address key unmet needs identified by a diverse group of school stakeholders. Winners of the competition will be invited to beta-test products in iZone classrooms.

The Research Alliance, based at New York University, will work with participating schools and the developers to evaluate each product and to formulate recommendations for product improvement. The collaboration will produce an online *Consumer Reports*-style guide for learning technologies, with results of the research studies published online along with product reviews and other user feedback.

In its first years, the partnership will focus on challenges in middle school STEM and work just with iZone schools, with the goal of including other content areas and a broader network of schools as the program scales.

The network of PowerMyLearning schools that will participate in rapid online experiments comparing the effectiveness of alternative digital learning resources, currently being set up by the nonprofit CFY, is another test bed.

Existing networks of schools such as the League of Innovative Schools could also be used to create a large-scale test bed for learning technologies more quickly and efficiently than recruiting suitable schools or districts from scratch. Potentially, districts, developers, and researchers could find test bed participation attractive because, among other benefits, they can have access to some of the latest learning technology resources, and they can learn more about implementation strategies. For districts, direct input to the development process ensures that products are designed to meet their needs. Smaller districts or entities can gain more leverage as members in a larger network.

Prize competitions can provide incentives for developers to participate. The Hamilton Project, a Brookings Institute initiative that has also called for the establishment of a digital learning testing body, proposes that many developers would find this approach beneficial for the following reasons:

- guaranteed access to a large user base through the test bed network, which would be especially appealing to small or start-up developers that district procurement processes often rule out;
- the ability to beta-test innovations at a large scale;
- working with education researchers to incorporate what is known in the learning sciences into the design phase, possibly saving costly mistakes; and
- the possibility of positive results that could increase sales. (Chatterji and Jones 2012)

The most successful digital learning test beds will draw on the research expertise of both the education and technology fields. The former contribute a valuable learning sciences background, and the latter contribute expertise in rapid-testing methods familiar in technology but generally less well known in education research. Further, making the test bed data available for secondary analysis could enable others to complete additional testing, answering questions relevant to other kinds of schools or other groups of decision-makers such as funders or policymakers.

Additional Evidence Considerations

The online repositories and communities of digital learning resources described in this chapter— which today are practitioners' primary sources of access and information about digital learning resources—do not yet incorporate empirical evidence about the learning outcomes obtained when the resources they include are used. Therefore, three areas in particular merit further consideration in focused and collaborative research efforts:

- evidence of alignment with standards,
- evidence considerations when adapting learning resources, and
- prior evidence of effectiveness.

Alignment with Standards

Today's educators are acutely attuned to matching the teaching and the learning resources they use with the specific standards their students are tested on. Decisions about alignment are almost always made on the basis of expert judgment rather than on the basis of an empirical demonstration that the resource has an impact on a measure of the competency described in the standard. Too often, the criterion for alignment is that the resource is related to the standard in terms of content topic, not that the resource covers the full competency called for in the standard.

For example, one of the grade 8 Common Core State Standards (CCSS) for reading informational text is "delineate and evaluate the argument and specific claims in a text, assess whether the reasoning is sound and the evidence is relevant and sufficient; recognize when irrelevant evidence is introduced" (RI.8.8). A digital learning activity in which students read a series of articles about current events and answer questions about each article's main idea might be tagged as aligned with this standard. Although such an activity would require some of the competencies in standard RI.8.8, it would not capture them all unless it called for students to identify not only the author's overall position or argument, but also each claim made and the evidence provided for that claim and to make judgments about the relevance and sufficiency of the evidence for each claim. Few eighth-grade instructional activities developed before publication of the CCSS are likely to encompass all aspects of standard RI.8.8.

In addition, alignment is complicated by the fact that most intellectual tasks with any complexity or verisimilitude to the real world will call on the competencies expressed in multiple standards, not just a single standard. This means that one digital learning resource may address multiple standards, and keeping track of the opportunities provided to address all relevant standards with a teacher-assembled set of learning resources is complex enough to call for the use of technology tools.

Moreover, just as the validity of an assessment cannot be judged independently of how it is used (see Chapter 4), whether or not a digital learning resource really aligns with a standard will depend on how it is implemented. The same digital learning resource may be used in one classroom in a way that fully addresses a learning standard while in another classroom the teacher inadvertently provides students with so much structure and so many hints that the standard is not addressed at all.

Because learning materials are often selected on the basis of their alignment with standards, those who use alignment evidence should be able to find out who made the judgments and how those judgments were made. Digital learning resource users need more transparency about these processes than they currently have. An example of a resource that is moving toward this kind of transparency is CPALMS, Florida's platform for sharing lessons aligned with Common Core State Standards in English language arts and mathematics and with the Next Generation Science Standards. The platform includes information about the individuals who make alignment and other decisions about the lessons submitted to CPALMS and the rubrics used in making judgments.

Adapting Learning Resources

When educators decide to use digital learning resources (or indeed any educational intervention), one of the decisions they need to make is whether to use the resource as originally designed or to adapt it in ways they believe will better fit their needs.

Perspectives on this choice are varied within the education research community. In some cases, the need for adaptation is obvious, such as translating learning resources into a language known to the users. But many adaptations are less clearly warranted and may have unintended side effects. For example, the decision to use reading software once a week for 60 minutes rather than three times a week for 20 minutes could influence young learners' degree of engagement during software use and their learning outcomes. Moreover, if adoption of the software was justified on the basis of studies showing a positive effect under the three-times-a-week model, it might be unwise for a district or school to allow this kind of adaptation.

The debate over whether a proven practice should be adapted to a new context also occurs in the field of public health (U.S. Department of Health and Human

Services 2002). After some years of arguing that adapting evidence-based programs was not acceptable, public health researchers are now trying to articulate guidelines for ascertaining which components of a program should and which should not be open to modification to suit local conditions and cultural differences (O'Connor, Small and Cooney 2007).

In a similar vein, education researcher Ann Brown warned of the "lethal mutations" possible when practitioners began making changes to research-based innovations. Datnow, Hubbard, and Mehan (1998) presented a different perspective, arguing that the reason it is difficult to "scale up" externally developed education reforms is that the idea that a complex reform can be replicated in many sites is misguided. Implementation, they argued, should be thought of as "co-construction" between the developers of an education intervention (or set of learning resources) and those who will implement the intervention (including teachers and students). Educational implementation, they reasoned, "involves a dynamic relationship among structural constraints, the culture of a school, and people's actions in many interlocking sites or settings" (p. 2). Design-based implementation research (DBIR), described in Chapter 1, has been heavily influenced by this latter view of implementation as involving adaptation. As implementation of an innovation is studied across a range of contexts, DBIR seeks to identify the core elements of the innovation design and implementation that must be in place to obtain desired outcomes and distinguish them from elements that may be modified to fit local constraints or preferences without undermining effectiveness.

Key to building this kind of knowledge is a central repository of information about adaptations that have been made, the reasons they were made, and the context of their use (for example, broader curriculum context, accountability framework, teacher qualifications and support, student characteristics), as well as the outcomes experienced with the adaptation. Obtaining empirical evidence on useful and deleterious adaptations is an ongoing process that could be accelerated if multiple groups working with the same intervention shared their implementation and impact data, as is being done with the Pathway to College Math project described in Chapter 1.

Prior Evidence of Effectiveness

A decision that must be made when identifying appropriate learning resources is whether to rely on prior research for evidence of effectiveness. This decision is related to the issues of the weight to place on expert opinion and whether to make modifications to the resource or the way it is used.

Chapter 1 noted the value of experimental data testing the effectiveness of a digital learning resource but acknowledged the long time and sizable expenditure of resources required to collect experimental data across the full range of contexts over which one wants to generalize. In practice, this rarely happens. Moreover, differences in the way a learning resource is implemented can make a great difference in what students learn (Means 2010). Add in modifications to the learning resource itself and changes in the mix of learning resources of which it is a part, and there are grounds for avoiding uncritical acceptance of evidence of effectiveness from past studies.

The final chapter of this report suggests the need for continuing to monitor and improve the effectiveness of a learning technology. As noted in that chapter, some groups are exploring ways to start accumulating evidence of learning resource effectiveness more rapidly.

Collaborations among technology developers, education researchers, and practitioners offer promise for more rapidly accumulating evidence of learning resource effectiveness in different contexts and with different implementation strategies. Such collaborations are growing in number and offer promise for addressing the evidence issues described above.

Conclusion

This chapter discusses how the explosion in the availability of digital learning resources on the Internet gives teachers, schools, and students more choices while raising issues about how purchasers and users can ensure the quality of these resources. Included are descriptions of some of the new online repositories and communities that can help users find digital learning resources—which is difficult enough on its own given the number and the diversity of resources that exist on the Web—and expanded approaches for gathering evidence of their potential to improve learning and other outcomes. The roles, strengths, and weaknesses of such methods as aggregating user interactions with digital learning resources, aggregating user reviews, user panels, expert ratings, review and curation, and test beds are discussed. The fragmented nature of information sources and the potential for self-interested individuals and organizations to provide biased reviews expose the need for an objective, third-party organization that can serve as a trusted source of evidence about the use, implementation, and effectiveness of digital learning resources. A recommendation that the federal government fund such an organization is included in the recommendations section of this report.

Summary and Recommendations

To take full advantage of the opportunities digital learning resources present and the learning data they capture, education stakeholders should collaborate to adopt R&D and evaluation approaches that enable

- digital learning innovations that can be developed and implemented quickly so every school has the chance to adopt them;
- continuous improvement processes to adapt and enhance these innovations as experience is gained using them;
- use of the vast amounts of data that are collected in real time to ask and answer questions about how individual learners learn so the needs of every student can be met; and
- expanded approaches to evidence gathering for greater confidence that investments in cost-effective and cost-saving technology-based interventions are wise, producing the outcomes sought.

Before turning to the recommended actions for education stakeholders, this report offers an **Evidence Strategy Framework** with three components:

1. ***Evidence Reference Guide*** summarizing the six evidence approaches highlighted in this report as well as other approaches widely used in education today,
2. ***Evidence Decision-Making Model*** for deciding about appropriate evidence approaches to use in implementing a selected learning resource,
3. ***Scenarios*** illustrating how and when the various evidence approaches might be applied in situations familiar to education stakeholders.

Evidence Reference Guide

Exhibit 1 provides a summary of the evidence-gathering approaches discussed in this report, the kinds of questions they can help answer, and the types of evidence that each can generate as well as their suggested uses.

Exhibit 1. Evidence Reference Guide: An Expanded View of Selected Evidence Approaches

	Sample Questions	Evidence Approach	Resulting Evidence	Uses
Research Foundations	What features should digital learning resources have to promote better learning? Are the assumptions this resource makes about the nature of human learning consistent with the learning science theory and research literature?	Review the learning sciences research literature and syntheses derived from it. Analyze the resource or resource design in terms of these principles.	Indication of the degree to which the resource reflects what is known about how people learn	Useful in early stages of design Useful on an ongoing basis to help improve product by interpreting data patterns extracted through data mining Useful in making and/or evaluating claims made about a resource design
Collaborative Design	How can this digital learning resource and the classroom activities it will be embedded in be designed to promote the targeted learning outcomes? What constraints of the school and classroom (for example, length of class periods, mandated pacing schedules) will limit how and how much this learning resource will be used?	Co-design new digital learning resources and implementations through collaborations of teams of developers, education researchers, and individuals from the intended group of users (often teachers).	Documentation of the learning resource's theory of action, including learning design principles and implementation constraints	Combines high level of design skill with research insights and insights from the field Capitalizes on teachers' experience with large numbers of students and their understanding of the contexts in which school teaching and learning occur Useful in early stages of design and on an ongoing basis because it can help developers interpret data patterns Applies at each stage of design and adaptation because this is the refinement portion of the iterative cycle of analysis, refinement, and implementation
Rapid Prorotyping	How many people choose to use a freely available digital learning resource, and how much time do they spend with it? What features of the learning resource do people use? Do users appear to learn something? Where do users appear to get stuck or lose interest?	**1. Make early version of a product freely available online, collect data while in use, and mine data for insights.**	Data from many users of an early version of a digital learning resource showing its appeal, features people use, and where people appear to get stuck or lose interest. May include responses from pop-up quizzes as well as log data	Creates a user base providing data that can be used to inform rapid cycles of product improvement and retesting Data mining useful at product launch and throughout product lifecycle It helps to ask users for information about themselves so their characteristics and goals for system use are known

Exhibit 1. Evidence Reference Guide: An Expanded View of Selected Evidence Approaches, Continued

Sample Questions	Evidence Approach	Resulting Evidence	Uses	
Will modifying the digital learning resource in this way lead users to spend more time with it? Will this change in the learning resource result in faster learning or deeper mastery of the target competencies?	**2. Conduct rapid A/B experiments within the learning system.**	Data indicating which version of the learning resource, A or B, results in better learning, more time spent with it, or more positive user reviews	Provides for rapid collection of data justifying a causal interpretation that the manipulated feature produced the observed impacts Sufficient for making claims about what improves outcomes internal to the resource Ideally, part of ongoing system improvement process conducted throughout the learning resource's life cycle	**A/B Testing**
What aspects of the way this digital learning resource is implemented in these settings influence learning outcomes? How could the technology or implementation practices be refined to improve outcomes?	**3. Practitioners team with research partners to conduct collaborative design-based implementation research (DBIR) on the learning resource as it is used in different settings.**	Data on the contexts and implementation practices associated with obtaining improved outcomes through using the resource	Desirable as part of large-scale implementation of complex digital learning systems to maximize the likelihood that the innovation will be well implemented and to learn from each iteration cycle as part of a continuous improvement process Brings data-informed decision making to the level of local practice Can inform subsequent effectiveness studies but is important also for innovations on which effectiveness studies have been done to maximize local benefits from the innovation and to build knowledge of how to scale up the innovation without degradations in its impacts	**Design-based Implementation Research**
How can I tell whether students who perform well in the online system will do so also?	**4. Use technology-supported evidence-centered design (ECD) to develop valid assessments of learning outcomes targeted by the digital resource and applicable in other settings as well.**	Analysis showing the knowledge, skills, and other attributes needed in executing a type of task that includes but is not limited to the tasks presented by the digital learning resource	Necessary when an existing validated assessment well aligned with the learning resource focus cannot be identified	**Evidence-Centered Design**

Exhibit 1. Evidence Reference Guide: An Expanded View of Selected Evidence Approaches, Continued

	Sample Questions	Evidence Approach	Resulting Evidence	Uses
Assessing Transfer	Does using this digital learning resource result in better learning outcomes ?	Conduct small-scale experiments (efficacy studies) using learning outcome measures external to the resource.	Data showing whether using the learning resource results in improved performance on a measure of the targeted learning that is external to the resource Evidence that performance on the learning measures internal to the resource is correlated with performance on measures external to the resource that are theoretically tapping the same competencies	Tests whether the digital learning resource can produce a desirable outcome external to the resource in some generally conducive context Important to establish the relationship between measures of learning internal to the resource and some external valued outcome Can be important in (a) attracting funding to support scaling the intervention to more users/sites and (b) convincing new users/sites that the intervention is worth adopting
Linking Systems for Data Mining	What other aspects of this student's life are likely to be affecting performance in this course or program of study?	**5. Participate in collaborations linking data from learning systems, education records, and social services agencies.**	Fuller picture of individual students' support needs and their accomplishments and challenges outside the classroom Resource for exploring relationships between social services and education outcomes	Especially useful in working with vulnerable students who may be facing multiple barriers to completion of a course or education program Requires negotiation of data-sharing agreements and privacy protections
Ratings and Reivews	Does this digital learning resource possess the attributes users view as important? What digital learning resources do most people think are the best? Which resources are rated the best by people like me or people who care about the same features that I do?	**6. Consult user ratings from a variety of sources on the properties, contextualized use, and perceived effectiveness of digital learning resources.**	Data showing how widely used or well known a learning technology is Indication of whether the learning resource possesses the features that other users regard as important	Provides at-a-glance overview that directs potential users to products that have been used and highly rated by others Information on the time frame and sources from which ratings are taken, characteristics of raters, and any standardization of rating scales should be available to those reviewing the ratings

Exhibit 1. Evidence Reference Guide: An Expanded View of Selected Evidence Approaches, Continued

Sample Questions	Evidence Approach	Resulting Evidence	Uses	
Do students from a wide range of settings who use the digital learning resource perform better than those who do not on learning measures external to it (for example, performance assessments)?	Conduct larger scale, multisite experiments (effectiveness studies) using outcome measures external to the resource.	Evidence that on average the intervention involving the digital learning resource causes improved outcomes in the types of settings included in the study May use assessment tasks developed using evidence-centered design	Useful after the resource has demonstrated it can be implemented across a large range of contexts with positive results in terms of system-internal learning measures and that some evidence exists that these outcomes are related to an external outcome of interest Variations in the way in which the resource is implemented across different contexts may swamp any impacts of the technology per se	**Effectiveness Studies**
What insights about how people learn can be derived from a digital learning system's big data? How similar is the way people learn with different learning systems stressing different approaches or design principles?	Participate in research consortia that combine and analyze anonymized data from multiple studies.	Analyses examining the generality of learning design principles across different learning content and learning system designs	Amplifies the value obtained from the extensive data set generated through use of learning systems Enables other researchers to explore hypotheses that had not occurred to developers and supports the generality of instructional design principles by testing them with multiple data sets taken from multiple learning systems Developers may be reluctant to make their datasets public either through concerns about protecting learner privacy or for fear that some analyses will reflect poorly on their products	**Research Consorita**

Note: The six evidence approaches highlighted in this report appear in boldface.

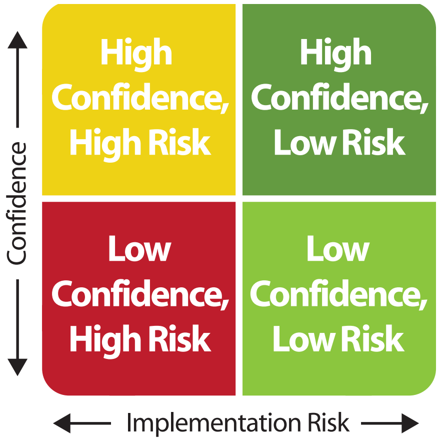
Evidence Decision-Making Model

The expanded approaches to evidence presented in the Reference Guide are intended to be useful for developing and improving digital learning resources and evaluating them for purchase and implementation. It is also important to think about using them to gather evidence as an ongoing practice throughout the life cycle of a resource. In fact, ongoing data collection and reflection are necessary as long as the potential for serious consequences exists if an intervention should fail. Even for a mature intervention for which extensive prior research has demonstrated effectiveness, education stakeholders should consider ongoing data collection and reflection as long as the stakes are high.

Therefore, the Evidence Decision-Making Model presented below does not assume a linear sequence of research types going from small-scale to large-scale, from weak to strong evidence, or from developmental to confirmatory data collections. Nor does it assume an endpoint to data collection and research on an intervention that entails significant risk (cost, time, or harm if the intervention should fail in a new implementation).

Exhibit 2 presents an overview of the Evidence Decision-Making Model, which draws on a conceptualization of the education research space offered by Tony Bryk, President of the Carnegie Foundation for the Advancement of Teaching (Bryk 2011), based on the factors of confidence and risk.[3] As used here, confidence is the belief that a digital learning system or resource will produce better outcomes than the status quo or will produce similar outcomes but less expensively, faster, in a more engaging way, or when used with different learners. Confidence can be based on empirical research, but it also can be based on the reputation of the organization that produced the system or resource, knowledge of the process used to develop it, or knowledge that many people are using it and believe it helps them. High confidence is not necessary in all instances, and the need for more evidence depends on the amount of implementation risk that trying a new product or approach would raise.

Exhibit 2.
Evidence Decision-making Model


High Confidence, High Risk
High Confidence, Low Risk
Low Confidence, High Risk
Low Confidence, Low Risk
Confidence
Implementation Risk

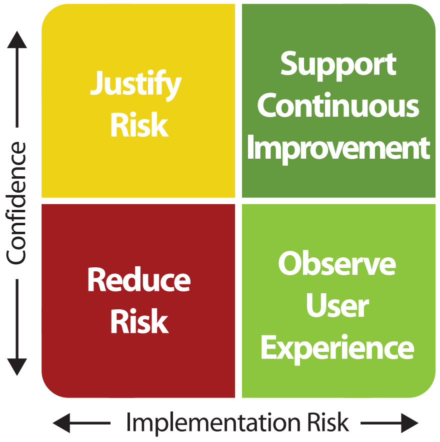
Implementation risk concerns the magnitude of the adverse consequences if a learning resource proves to be ineffective. Important aspects of risk are the size and scale of implementation, both in terms of numbers of students and time; the cost of purchasing and implementing the intervention; the extent to which the intervention will disrupt existing processes and practices; the amount of time it will take away from other valuable activities; and an overall consideration of how much there is to lose.

Once a decision has been made to purchase and implement a digital learning resource, the resource can be positioned in the appropriate quadrant of the model based on confidence in its improvement potential and estimated implementation risk. Depending on the placement of a resource in a particular quadrant, Exhibit 3 presents a simplified

[3] Bryk's conceptualization also includes receptivity as a third factor. Although influenced by Bryk's work, this decision-making model was developed with a tighter focus on digital learning. Bryk, of course, bears no responsibility for any limitations of the model presented in this report.

view of evidence goals that are appropriate to apply to evidence gathering as a resource is implemented. For example, when confidence in a learning resource is high and implementation risk is low, ongoing evidence gathering can be directed toward refining, improving, and enhancing the product over time as experience is gained with its use. However, when a resource engenders a high degree of both confidence and risk, empirical research of its effectiveness could be considered necessary to justify the initial and ongoing investment.

Exhibit 3. Simplified View of Evidence Goals



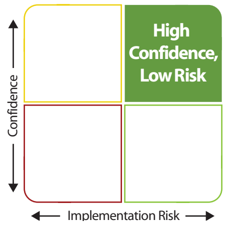
In contrast, when both confidence and implementation risk are low, the model suggests that costly and time-consuming empirical research is not necessary to proceed, depending on where and how the resource will be used. For example, little ongoing evidence gathering is called for when a product is to be used as a self-selected, optional activity a student chooses to help with study or practice. The user's own experience and satisfaction with the resource may be sufficient.

Finally, some learning resources generate low confidence among educators and have high implementation risk. For these types of resources, examining ways of staging implementation to reduce risk or rethinking its use altogether is best.

Scenarios

The following scenarios are designed to bring the uses of the Evidence Reference Guide and Decision-Making Model into sharper focus by describing their application in situations familiar to education stakeholders.

Scenario 1: Improving Homework Completion Rates in an Inner City School District



Situation

A nonprofit organization serving children and youth in an inner-city school district in a large Midwest city has been awarded a Promise Neighborhood planning grant. These one-year grants are given to nonprofit organizations and higher education institutions in distressed urban and rural communities to formulate comprehensive community-wide plans for preparing students for success in college and careers. In this neighborhood, about half the residents live below the poverty line, about two and a half times the citywide average.

Under its Promise Neighborhood initiative, this school district's residents, school principals, health care and social services providers, and other stakeholders have been working together to develop a multilayer approach to academic and life success for children and youth that will extend beyond the classroom and school hours. When the plan is completed, it will be comprehensive, focusing on building services for students from cradle to career.

In the meantime, the district faces a growing problem: Homework completion rates among high school students have fallen every year since 2008. They have dropped to such low levels that some teachers no longer assign homework without giving time in class to finish it. Although this practice improves homework completion rates, it reduces the class time teachers have for covering required curriculum.

Determined to find other solutions to the homework completion problem, the Promise Neighborhood directs its program manager to find out why homework completion rates are decreasing and to develop an intervention to ensure that the needs of every child who is having trouble getting homework done are met.

Understanding the Problem

Because schools and community services organizations have already begun working together under the Promise Neighborhood program, considerable progress has been made on solving the technical and legal challenges associated with combining and sharing individual student data across these various entities while still maintaining security and privacy. Data about each student in area schools are being collected continuously, including demographic information, health status, school attendance, grade point averages, test scores, classroom behaviors, and student self-reporting on well-being.

In addition, the College of Education at a nearby university has been training its graduate students on new data mining tools and learning analytics techniques that have equipped them to mine these often large combined datasets to seek answers to various educational research questions posed by area schools and community-based organizations.

The program manager starts his search for answers by asking the graduate students to mine the database to create profiles of students who routinely failed to complete their homework for each high school grade going back to 2008. Two relevant findings emerge: First, the most striking attribute distinguishing students who habitually fail to complete their homework from other students is that they are more likely to be from homeless families. Second, the absolute number of students from homeless families in the district today is much higher than it was in 2008.

Looking for the Underlying Causes

These initial findings drive the program manager to seek more information about the situations of students from homeless families. The literature on homelessness reveals that most homeless families are highly mobile—they stay in homeless shelters, cars, inexpensive hotels, or temporarily with families or friends. To the casual observer or even to schools and community service and youth organizations, it might not be obvious that these children are homeless. In addition, homeless families experience multiple upheavals in their living situations each year. This high degree of mobility makes it difficult for the students to establish daily routines or find a quiet place to study and do homework.

Interventions

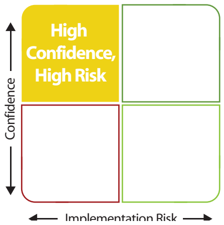
In his initial report to the executive director and the board of the nonprofit organization, the program manager recommends several immediate interventions:

- Inform high school principals and educators in area schools of the increase in students from homeless families and the corresponding decrease in homework completion rates to ensure that they are aware of the challenges these students face and encourage them to identify and intervene with such students in their schools. Help might include holding afterschool homework sessions on campus for students and/or engaging volunteer tutors to work with students one on one after school.

- Reach out to area homeless shelters in the area and parents of homeless families to reinforce the importance of providing quiet areas for children to study and do homework. Engaging volunteer tutors to participate in homework sessions in the shelters could complement this outreach effort, if appropriate.

- Partner with community service organizations providing services for children of homeless families to help them connect directly with students to determine what each one needs to stay focused on their studies, including homework. Depending on the nature of these organizations—for example, if they had appropriate facilities—they could be encouraged to hold their own afterschool homework programs with volunteer tutors attending.

Because it is unclear which of these interventions would result in the best outcomes, the program manager proposes to collaborate with the university to conduct formative research on how best to implement each strategy and longitudinal research to determine which intervention produces the greatest positive impact per dollar invested.



Scenario 2: Improving Middle School Math Scores While Implementing Common Core State Standards (CCSS)

Situation

A district superintendent whose seven elementary schools, three middle schools, and two high schools are slated to use the Next Generation Assessments of the Common Core State Standards in the 2014–15 school year is concerned. She knows from published reports that when Kentucky adopted new state tests in school year 2011–12 the share of students scoring proficient or better in reading and math dropped by roughly a third or more in both elementary and middle schools. She also knows that education administrators around the country are warning that because the CCSS encompass deeper learning competencies than most schools have been teaching and are more rigorous than most state standards, results the first year that the Next Generation Assessments are used are likely to shock the public.

The superintendent would like to get out ahead of what may be a problem for her district, and she is particularly concerned about the potential of a drop-off in middle school math scores, an area where her district struggles already. She asks her district curriculum coordinator to start now to research and implement a middle school math curriculum that is aligned with the CCSS and has evidence of effectiveness. She further specifies that because they will be adopting an entirely new math curriculum aligned with the new standards, it should be technology based so they can make the transition to digital learning at the same time.

To leverage the work in this effort, she also reaches out to other district superintendents in her region, inviting them to participate. Two other districts sign on to the effort and also assign their curriculum coordinators to work on the project.

The Process

The three curriculum coordinators agree that the involvement and buy-in of their district math teachers will be essential to the success of any curriculum they might choose. Thus, as a first step, they reach out to all the math teachers in the three districts and invite them to serve on anevaluation committee for the effort. Seven math teachers from various middle schools volunteer.

Given that the CCSS were adopted only recently, the participants agree that it will be a challenge to find a technology-based curriculum that is both aligned with the standards and has evidence of effectiveness. They start by making a list of the technology-based math curricula they know of that might meet these criteria. Each participant is also asked to solicit suggestions from colleagues. The result is a list of four products that might meet their needs.

To supplement this list, one of the curriculum coordinators searches various online educational repositories and websites, some of which have expert and/or user reviews and rating systems. After an unfiltered search produces several hundred possibilities, she applies three filters: (1) Is the product a full eighth-grade curriculum? (2) Is the product aligned with the math CCSS? and (3) Is it currently in use with a minimum of 100,000 students? These filters narrow the list to 17 products, including the four that were recommended by the evaluation committee.

Still, with a total of 17 products that might meet their criteria, the group decides to create a more detailed list of evaluation criteria to continue the weeding out process. Additional evaluation criteria include having a placement test that enables the system to map a learning pathway for each student, a variety of different kinds of learning activities, incorporation of universal design for learning (UDL) to support all students including those with special needs, a comprehensive reporting system with both formative and summative assessments that can be accessed by both teachers and students, and an unlimited school licensing option.

Applying these additional criteria to all 17 products, the evaluation committee settles on three products that have most of the desired features. The providers of the three products are sent a request for a proposal and invited to present to the group, stressing that evidence of effectiveness is a key to winning the bid.

After the three providers present to the group, it is clear that although all the products are aligned with many portions of the CCSS and are promising in many ways, none has evidence strong enough to justify the adoption of the curriculum across all the middle schools in all three districts.

Plan B

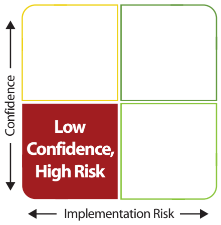
At this point, the district superintendents decide that even though the improvement potential of the math program is high, so are its implementation risks. Therefore, a different approach is needed.

Given that all school districts in the states that have adopted the CCSS will most likely be concerned about a drop in test scores, a technology-based middle school math intervention that helps students gain competence in areas of the new standards not stressed in middle school math in the past would probably appeal to a large market. The superintendents decide to see whether they can interest one of the providers of the three finalist products to enter into a partnership with the three school districts and education researchers from a nearby university on a multiyear R&D project with the goals of rapidly and simultaneously improving middle school math learning outcomes, producing appropriate evidence of the efficacy of the program, and refining the product.

To gather evidence of learning outcomes, the R&D project would include a pilot test in volunteer middle schools in the three districts in the first year, with efficacy studies producing data at the end of the first semester and again at the end of the first year. To drive continuous improvement, the project would include design-based implementation research initiated simultaneously.

Attracted by the prospect of a large and growing market for an improved product backed by evidence, one of the providers signs on to this forward-looking collaboration and the project launches.

Scenario 3: Meeting State University Requirements for Hands-on Chemistry Lab Experience in a Rural High School District



Situation

When a state university system changes its entrance requirements to include two science courses with hands-on laboratory experience, the superintendents of rural school districts know they will have difficulty ensuring that their students can meet those requirements. Many of them have only a single science lab or none at all. Given their high schools' small enrollments and budgets, building and maintaining two science labs in each school is not feasible.

Furthermore, given the long distances between communities and high schools in the district, requiring students to commute to a high school that does have lab facilities is not an option.

Like others in their state, these districts have been experimenting with online simulated lab experiences for their high school science courses. But now, with the change in entrance requirements for state universities, the superintendents in the state's rural districts realize they have more to do.

The superintendents collectively reach out to the state department of education, which authorizes a working group to tackle the problem and agrees to fund the effort. The decision is made to focus first on solving the problem for chemistry, which is expected to be the most difficult case.

Possible Solutions

The working group consists of state and district curriculum coordinators, volunteer high school chemistry teachers, and a group of education researchers from one of the state universities.

Next Steps

The working group consults the research literature and finds that the College Board has addressed the issue of whether lab simulations can be substituted for real-world lab experiences as part of its redesign of Advanced Placement science courses. Assessment experts hired by the College Board applied evidence-centered design to identify the essential skills, knowledge, and abilities (KSAs) that students need to gain in an introductory college-level chemistry course. Starting with this model of what students are expected to learn in chemistry labs, the College Board's assessment experts then specified the features of tasks that would elicit the KSAs.

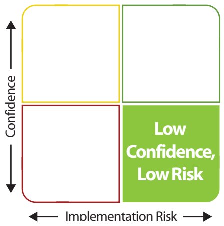
Using the information provided by the College Board AP science redesign team, the state's working group examines the subset of AP Chemistry KSAs that are elicited in laboratory-based activities. Comparing the task contexts for eliciting these skills with available virtual chemistry labs reveals that by combining several products and online simulations, all the target KSAs can be learned online. The working group makes a list of the laboratory KSAs identified by the College Board assessment experts and documents the coverage and assessment of each concept or skill in an online chemistry lab. The working group argues that if online labs are capable of addressing all the KSAs in AP Chemistry, they surely would meet the needs for a high school-level chemistry course. This document is submitted to the State Higher Education Board as part of a request for a waiver for the new requirement of completing a chemistry course with a physical lab component.

The Higher Education Board is impressed with the thoroughness of the working group's analysis but points out that some aspects of laboratory work are not included in the AP Chemistry test or in online simulations. These include characteristic odors of some key compounds and practices related to safety. The working group acknowledges this remaining concern and notes that the equipment necessary for acquiring this smaller set of skills is much less expensive than a fully stocked lab. The group develops a plan for installing these scaled-down chemistry labs in selected rural high schools to supplement the previously identified suite of online chemistry lab activities.

Several chemistry teachers from the district schools that do have full chemistry labs are given release time to participate on a task force to develop a curriculum for the scaled-down labs. Using the list of knowledge, skills, and competencies that had been identified as difficult to learn online, they pull together the lab activities they have been using to develop students' proficiency in those areas and turn to online collections of chemistry activities to identify additional options. They annotate each possible activity with the KSAs it covers and the equipment and amount of time it requires. Working with the annotated set of chemistry lab tasks, they design a curriculum covering all the essential KSAs that cannot be acquired in online lab simulations in a minimum amount of time and with the least expensive equipment.

To ensure the curriculum works as expected, the rural districts contact a faculty member at a nearby university who puts them in touch with a graduate student eager to conduct an evaluation of the labs for her master's thesis.

Scenario 4: Helping a Child Overcome Problems with Vocabulary



Situation

The mother of a seventh-grade boy is worried because her son is having trouble with vocabulary. When she reviews his vocabulary quizzes, his essays, and other writing homework, she notes that his teacher is consistently deducting points for vocabulary. Her son's grades in English are suffering as a result, and he is losing confidence in his ability to write.

The Search

Because her son loves computers, the mother wonders whether there is a game, app, or website that could help make learning new vocabulary words more fun and rewarding for him. She knows he will resist using something, especially outside school hours, unless he is having fun. So while being entertaining is an important factor, she also wants to make sure that whatever she chooses actually improves his vocabulary.

She starts her search on the Internet and through a series of searches of blogs written primarily by parents, she compiles a list of half a dozen games and apps on vocabulary that might engage her son. Next, she finds a blog post written by a middle school teacher on how parents can find and choose technology-based digital learning resources for their children to use at home. The blog describes several online communities that publish ratings and reviews of educational games, apps, and programs similar to the way consumer-oriented review sites collect and publish ratings of consumer products and services.

The mother visits these sites and after an hour or so of research, she has found ratings or reviews of all the resources on her short list. For several, there are as many as six or eight detailed reviews written by individual teachers or parents. In one or two cases, the product ratings are based on the aggregated comments of hundreds and even thousands of parents. The fact that so many people have indicated that they like certain products and why gives her confidence that she can trust the ratings.

On the basis of these ratings and reviews, she chooses two products that are highly rated and might appeal to her son.

A Solution

After school that day, she introduces both products to her son, who spends the next hour exploring them and choosing the one he likes best. An aspect her son likes in the app he selects is a digital badging system that rewards his progress in hitting certain milestones, similar to the way the video games he loves to play recognize him for advancing to the next level.

The mother and son decide together that he will use the app 30 minutes a day three times a week and that she will review his progress once a week. Her son tells her he's looking forward to earning as many badges as he can over the next month. She tells him she's looking forward to seeing improvement in his vocabulary and his English grade on his next report card.

Importance of Privacy Policy and Legal Issues

Several privacy policy and legal issues arise when educators, administrators, and researchers collect, store, analyze, and possibly release student data to third parties for data mining and analysis. *The Family Educational Rights and Privacy Act (FERPA)* permits use of identified student data for educational purposes. A full discussion of privacy and confidentiality is beyond the scope of this report, but new resources are available that address data management for education and research. These include the technical brief series from the Department's National Center for Educational Statistics (U.S. Department of Education 2010b). In addition, recent guidance on FERPA has helped clarify how institutions may use detailed and longitudinal student data for research, accountability, and school improvement under certain conditions. According to the Data Quality Campaign, new amendments to the existing FERPA regulations increase access to data for research and evaluation (including sharing across levels, such as from high school to college) while maintaining student privacy and parents' rights (Data Quality Campaign 2011).

One approach is to work at an agency level to examine aggregate patterns in data rather than individual-level data, as happens within projects of the YDA (Youth Data Archive, featured in a sidebar in Chapter 3). Projects in the YDA have made use of linked datasets to explore such questions as whether youth in foster care face especially difficult challenges in school compared with similar youth not in care or what pathways from high school to local colleges and community colleges yield greater success for various kinds of students. Another option is to use anonymized versions of datasets with individual-level data. Such datasets can be analyzed to help researchers and educational systems understand better how student difficulties in school are linked to problems outside school. Analyses of these data may also help identify protective factors that help keep students engaged and in school.

Institutional Review Boards

As the only formal mechanism for overseeing social research in the United States, Institutional Review Boards (IRBs) have an important role in protecting the rights and privacy of students and teachers when they participate in education research. But as organizations receiving federal research funding have sought to comply with federal IRB regulations, review processes have become more elaborate, time consuming, and costly. Moreover, organizations that do not receive federal funding, such as commercial developers of learning resources, are not subject to IRB requirements and therefore, do not have their data collection plans reviewed by a third party or experience the often lengthy delays associated with gaining IRB approvals.

In July 2011, the federal government issued an Advanced Notice of Proposed Rulemaking (ANPRM) as a first step to identifying ways the process could be streamlined without compromising needed protections, especially when the research being proposed involves minimal risk. The American Association of University Professors (AAUP) responded by proposing that researchers with relevant expertise could more easily distinguish proposals involving minimal risk than current IRB staff, who often have no research expertise in the fields relevant to the studies they approve (AAUP 2012). Rule changes are pending.

Commitment to Transparency

In the use of big data in education, ensuring privacy is of tantamount importance—and so is a commitment to transparency. The methods and sources used in collecting evidence about the effectiveness and implementation of digital learning resources should be shared. The value of claims of alignment with standards cannot be judged without knowing the qualifications of the people who performed the alignment and the process they used, for example. Similarly, the credibility of a claim that experimental evidence demonstrates the effectiveness of a technology hinges on details of the experiment's design and implementation (for example, the way students or classrooms were sampled and assigned to treatment or control groups, the rate of attrition from the two conditions, the way student learning was measured).

Although not all education stakeholders are research methods experts, making this kind of information public increases the probability that it will be reviewed and commented on by experts in much the same way that reviewers tag Wikipedia entries when they see the need for greater documentation or objectivity.

Recommendations

The following recommendations are designed to help education stakeholders turn the ideas presented in this report into action.

1. Developers of digital learning resources, education researchers, and educators should collaborate to define problems of practice that can be addressed through digital learning and the associated kinds of evidence that can be collected to measure and inform progress in addressing these problems.

In doing this work, collaborative teams should seek opportunities to structure the data collected by digital learning resources in ways useful as evidence. Learning technology developers should carefully define their systems' desired learning outcomes in the early stages of design and collaborate with education researchers to design data collection that will provide strong evidence that these goals have been achieved. Educators who make decisions about which learning systems to adopt should use evidence about learning outcomes and implementation as key criteria. An example of this type of collaboration that the U.S. Department of Education endorses is the move to identify and support regional innovation clusters' purposeful partnerships to break down domain silos and create connections between researchers, the commercial sector, and educators.

2. Learning technology developers should use established basic research principles and learning sciences theory as the foundation for designing and improving digital learning resources.

To assist in this endeavor, education researchers should make compendiums of research-based principles for designing learning systems widely available, more understandable, and more actionable for learning technology developers.

3. Education research funders should promote education research design that establishes that digital learning resources teach aspects of deeper learning such as complex problem solving and promote the transfer of learning from one context to many contexts.

Some of the most difficult skills to learn are those expected by high-performance workplaces—the ability to work with others to solve difficult problems and to be able to go beyond what has been taught to learn and master new things. The new Common Core State Standards address related competencies and also raise challenges in trying to judge the effectiveness of a given learning resource in helping students achieve competencies that will generalize across different materials and settings. This area is ripe for the articulation of goals, processes, and methods by which learning resources can help achieve these outcomes.

4. Education researchers and developers should identify the attributes of digital learning systems and resources that make a difference in terms of learning outcomes.

Learning technology developers have incentives to improve their own specific product but not necessarily to investigate and share general design principles for effective online learning. Collaborations between system developers and researchers with experience working with multiple digital learning systems can focus on generalizable principles and make sure that the world at large benefits from insights gained through data mining and A/B testing.

5. Users of digital learning resources should work with education researchers to implement these resources using continuous improvement processes.

In today's world of myriad digital learning resources and user choices about technology configuration and use, labeling a learning resource as one that does or does not "work" is an oversimplification. Users and adopters should expect to take an active role in planning technology implementations and collecting data that can be used in multiple ongoing cycles of implementation, analysis, and refinement. Technology developers should seek the resulting data and use this feedback to improve their products.

6. Purchasers of digital learning resources and those who mandate their use should seek out and use evidence with respect to the claims made about each resource's capabilities, implementation, and effectiveness.

Decision makers need to have or develop the expertise to locate and evaluate these kinds of evidence about the learning technologies being considered. This report provides some guidance on the kinds of questions purchasers should ask about the learning resource design and development process, the extent of usage in contexts like the purchaser's own, and the evidence of impacts on learning outcomes outside the system as well as on embedded formative assessments.

7. Interdisciplinary teams of experts in educational data mining, learning analytics, and visual analytics should collaborate to design and implement research and evidence projects. Higher education institutions should create new interdisciplinary graduate programs to develop data scientists who embody these same areas of expertise.

Educational data mining that incorporates learning analytics is a new field experiencing rapid growth (U.S. Department of Education 2012a). It draws on multiple disciplines including statistics, machine learning, and cognitive science. Experts in these areas report that one cannot learn the necessary combination of skills without access to large datasets and guidance from mentors.

8. Funders should support creating test beds for digital learning research and development that foster rigorous, transparent, and replicable testing of new learning resources in low-risk environments.

These test beds should be established and managed by intermediary organizations and through partnerships between government and industry that can bring together the required expertise, skills, and personnel. Funders should kick-start test beds by structuring funding programs that encourage them and cover some of the costs of setting them up. Digital Promise should continue to expand the League of Innovative Schools, and other programs like iZone should be designed and funded.

9. The federal government should encourage innovative new approaches to the design, development, evaluation, and implementation of digital learning systems and other resources.

The federal government through the U.S. Department of Education has proposed to create an Advanced Research Project Agency for Education (ARPA-ED). ARPA-ED would fund projects run by industry, universities, and other innovative organizations based on their potential to transform teaching and learning. ARPA-ED should fund directed development projects so progress can be accelerated and the essential activities of data aggregation and sharing across different research and evaluation efforts facilitated.

10. Stakeholders who collect and maintain student data should participate in the implementation of technical processes and legal trust agreements that permit the sharing of data electronically and securely between institutions, complying with FERPA and other applicable data regulations, using common data standards and policies developed in coordination with the U.S. Department of Education.

Digital learning systems create new opportunities for collecting large amounts of data that when aggregated and analyzed can contribute to our understanding of how people learn, how we can better support individual students' needs, and how we can improve our education system at all levels. These possibilities can be realized only if data are available to educational researchers and developers across systems and institutions with appropriate data security and privacy protections in place.

11. Institutional Review Board (IRB) documentation and approval processes for research involving digital learning systems and resources that carry minimal risk should be streamlined to accelerate their R&D without compromising needed rights and privacy protections.

In July 2011, the federal government issued an Advanced Notice of Proposed Rulemaking (ANPRM) as a first step to identifying ways the process can be streamlined without compromising needed protections. Changes to the rules regarding approval processes are pending.

12. R&D funding should be increased for studying the noncognitive aspects of 21st-century skills, namely, interpersonal skills (such as such as communication, collaboration, and leadership) and intrapersonal skills (such as persistence and self-regulation).

New research suggests that the development of 21st-century skills —a combination of cognitive skills, interpersonal skills, and intrapersonal skills — may relate to positive adult outcomes, such as increased earnings, better health, and greater civic engagement. Emerging evidence also suggests that 21st-century skills support transfer — the ability to apply something learned in one situation to a similar but different situation (Pellegrino and Hilton 2012). More research is needed on what factors contribute to students' development of 21st-century skills, as well as on how to assess students' acquisition of them. Multiple measures of learning outcomes can give a far richer picture of student learning than standardized tests alone.

13. R&D funding should promote the development and sharing of open educational resources (OER) that include assessment items that address learning transfer.

Open educational resources are increasing rapidly, but most have focused on learning materials rather than on assessments that could be used with any number of curricula. What we need now is to also develop performance assessment OERs that could be implemented in a variety of contexts as long as they target the same outcomes.

14. The federal government and other interested agencies should fund an objective third-party organization as a source of evidence about the usability, effectiveness, and implementation of digital learning systems and resources.

With so many sources of digital learning resources and the competing claims of different distributors, educators should have reliable, objective information not just about effectiveness but also about implementation issues and usability. To be useful, the information must be produced rapidly and at a low enough cost that a large number of digital learning products in each area can be continuously evaluated and information about their effectiveness reported on a regular basis.

References

AAUP (American Association of University Professors). 2012. “Current Research Review System Threatens Academic Freedom, New Report Says”. Association website. http://www.aaup.org/AAUP/newsroom/2012PRs/irb.htm.

Aleven, Vincent, and Kenneth Koedinger. 2002. “An Effective Metacognitive Strategy: Learning by Doing and Explaining with a Computer-Based Cognitive Tutor.” *Cognitive Science* 26:147–179.

Allensworth, Elaine M., and John Q. Easton. 2005. “The On-Track Indicator as a Predictor of High School Graduation.” Consortium on Chicago School Research at the University of Chicago. http://ccsr.uchicago.edu/publications/track-indicator-predictor-high-school-graduation.

Allensworth, Elaine M., and John Q. Easton. 2007. “What Matters for Staying On-Track and Graduating in Chicago Public Schools.” Research Report. Consortium on Chicago School Research at the University of Chicago. http://ccsr.uchicago.edu/sites/default/files/publications/07%20What%20Matters%20Final.pdf.

Almlund, Mathilde, Angela Lee Duckworth, James J. Heckman, and Tim Kautz. 2011. “Personality Psychology and Economics.” IZA Discussion Paper. Institute for the Study of Labor (IZA).

Amershi, Saleema, and Cristina Conati. 2009. “Combining Unsupervised and Supervised Classification to Build User Models for Exploratory Learning Environments.” *Journal of Educational Data Mining* 1(1):1–54.

Andersen, Erik, Yun-En Liu, Ethan Apter, François Boucher-Genesse, and Zoran Popovi .2010. “Gameplay Analysis Through State Projection.” In Proceedings of the *Fifth International Conference on the Foundations of Digital Games, 1–8*. FDG ’10. New York, NY: ACM.

Arnold, Kimberly E. 2010. “Signals: Applying Academic Analytics.” EDUCAUSE Quarterly 33(1). http://www.educause.edu/ero/article/signals-applying-academic-analytics.

Arroyo, Ivon, David G. Cooper, Winslow Burleson, Beverly Park Woolf, Kasia Muldner, and Robert Christopherson. 2009. “Emotion Sensors Go To School.” In *Proceedings of the 2009 Conference on Artificial Intelligence in Education: Building Learning Systems That Care: From Knowledge Representation to Affective Modelling,* 17–24. IOS Press.

Ash, Katie. 2012. “K-12 Marketplace Sees Major Flow of Venture Capital.” *Education Week*, February 1. http://www.edweek.org/ew/articles/2012/02/01/19venture_ep.h31.html?r=435323443.

Au, Wayne. 2007. "High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis." *Educational Researcher* 36(5) (June 1):258–267.

Baker, Ryan, Albert Corbett, and Kenneth Koedinger. 2006. "Responding to Problem Behaviors in Cognitive Tutors: Towards Educational Systems Which Support All Students." *National Association for the Dually Diagnosed (NADD) Bulletin* 9 (4): 70–75.

Baker, Ryan, Albert Corbett, and Vincent Aleven. 2008. "More Accurate Student Modeling Through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing." In *Intelligent Tutoring Systems: 9th International Conference, ITS 2008, Montreal, Canada, 2008. Proceedings.* ed. Beverley Woolf, Esma Aïmeur, Roger Nkambou, and Susanne Lajoie, Lecture Notes in Computer Science (LNCS) 5091:406–415.

Baker, Ryan, Albert Corbett, Kenneth Koedinger, and A.Z. Wagner. 2004. "Off-Task Behavior in the Cognitive Tutor Classroom: When Students 'Game The System'." *In Proceedings of ACM CHI 2004: Computer-Human Interaction*, 383–390.

Baron, Jon. 2007. "Making Policy Work: The Lesson From Medicine." *Education Week,* May 23. http://www.edweek.org/ew/articles/2007/05/23/38baron.h26.html.

Barron, Brigid, Caitlin Kennedy Martin, Lori Takeuchi, and Rachel Fithian. 2009. "Parents as Learning Partners in the Development of Technological Fluency." *International Journal of Learning and Media* 1(2) (May):55–77.

Barron, Brigid, K. Gomez, N. Pinkard, and Caitlin Kennedy Martin. In Press. *Cultivating Creative Production and New Media Citizenship in Urban Communities: The Digital Youth Network*. Cambridge, MA: MIT Press.

Biederman, Irving, and Margaret M. Shiffrar. 1987. "Sexing Day-old Chicks: A Case Study and Expert Systems Analysis of a Difficult Perceptual-learning Task." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 13(4):640–645.

Bill & Melinda Gates Foundation. 2012. *Gathering Feedback for Teachers: Combining High-Quality Observations with Student Surveys and Achievement Gains*. Princeton, NJ.

Bloom, Benjamin S. 1984. "The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring." *Educational Researcher* 13(6) (June 1):4–16.

Bransford, John D., and Daniel L. Schwartz. "Rethinking transfer: A simple proposal with multiple implications." *Review of Research in Education* 24 (1999):61-100.

Bryk, Anthony S., Louis M. Gomez, and Alicia Grunow. 2011. "Getting Ideas into Action: Building Networked Improvement Communities in Education." *In Frontiers in Sociology of Education,* ed. Maureen T. Hallinan, 127–162.

Bryk, Anthony. 2011. "Accelerating Our Capacity to Learn to Improve: Networked Communities Engaged in Improvement Research". Presentation for the Singapore National Institute of Education. May 30.

Butcher, Kirsten R., and Vincent Aleven. 2008. "Diagram Interaction During Intelligent Tutoring in Geometry: Support for Knowledge Retention and Deep Understanding." In *Proceedings of the 30th Annual Conference of the Cognitive Science Society,* ed. B.C. Love, K McRae, and V.M. Sloutsky, 1736–1741.

Campbell, John P. 2012. Personal communication to Eryn Heying. September 19.

Campbell, John P., and Kimberly E. Arnold. 2011. "Course Signals: The Past, Present, and Future Application of Analytics". *Presented at the EDUCAUSE 2011 Annual Conference*, Philadelphia, Pennsylvania.

Carlson, Curtis R., and William W. Wilmot. 2006. *Innovation: The Five Disciplines for Creating What Customers Want.* New York, NY: Crown Business.

Castrechini, Sebastian. 2009. *Educational Outcomes for Court-Dependent Youth in San Mateo County*. Issue Brief. John W. Gardner Center for Youth and Their Communities. http://cs210net.stanford.edu:81/resources/publications/JGC_IB_CourtDependentYouth2009.pdf.

Chatterji, Aaron and Benjamin Jones. 2012. "Harnessing Technology to Improve K-12 Education." Policy Brief 2012-5. Washington, DC: The Hamilton Project. http://www.brookings.edu/~/media/research/files/papers/2012/9/27-education-technology/thp_chatterjijones_edtech_brief.pdf.

Chingos, Matthew M., and Grover J. Whitehurst. 2012. *Choosing Blindly: Instructional Materials, Teacher Effectiveness, and the Common Core.* Washington, DC.

Cizek, Gregory J., Daniel Bowen, and Keri Church. 2010. "Sources of Validity Evidence for Educational and Psychological Tests: A Follow-Up Study." *Educational and Psychological Measurement* 70(5) (October 1):732–743.

Clark, Richard E., and Fred Estes. 1996. "Cognitive Task Analysis for Training." *International Journal of Educational Research* 5(25):403–417.

Clarke-Midura, Jody, Chris Dede, and Jill Norton. 2011. *Next Generation Assessments for Measuring Complex Learning in Science: The Road Ahead for State Assessments*. Cambridge, MA: Rennie Center for Education Research & Policy.

ClassDojo. 2012. "About." http://www.classdojo.com/about.

Clements, Douglas H. 1999. "Subitizing: What Is It? Why Teach It?" **In Teaching Children Mathematics.** Reston, VA: NCTM.

Collins, Allan and Roy Pea. 2011. "The Advantages of Alternative Certifications for Students." *Education Week*, October 19. http://www.edweek.org/ew/articles/2011/10/19/08collins.h31.html?print=1.

Corno, Lyn, and Richard E. Snow. 1986. "Adapting Teaching to Individual Differences Among Learners." *In Handbook of Research on Teaching*, ed. M. C. Wittrock. Washington, DC: American Educational Research Association.

Corbett, Albert T., and John R. Anderson. 1995. "Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge." *User Modeling and User-Adapted Interaction* 4(4):253–278.

Craig, Scotty D., Arthur C. Graesser, Jeremiah Sullins, and Barry Gholson. 2004. "Affect and Learning: An Exploratory Look into the Role of Affect in Learning with Autotutor." *Journal of Educational Media* 29:241–250.

Cronbach, Lee J., and Paul E. Meehl. 1955. "Construct Validity in Psychological Tests." *Psychological Bulletin* 52(4):281–302.

Cronbach, Lee J., and Richard E. Snow. 1969. *Individual Differences in Learning Ability as a Function of Instruction Variables: Final Report*. Palo Alto, CA: Stanford University School of Education.

Cronbach, Lee J., and Richard E. Snow. 1977. *Aptitudes and Instructional Methods: A Handbook for Research on Interactions.* Oxford, England: Irvington.

Crouch, Catherine H., and Eric Mazur. 2001. "Peer Instruction: Ten Years of Experience and Results." *American Journal of Physics* 69(9):970

Dai, David Yun. 2011. *Design Research on Learning and Thinking in Educational Settings: Enhancing Intellectual Growth and Functioning.* New York: Routledge.

Data Quality Campaign. 2011. "U.S. Department of Education Final FERPA Regulations: Advisory and Overview." Prepared by EducationCounsel, LLC. http://www.dataqualitycampaign.org

Data Quality Campaign. 2012. "*Leveraging Federal Funding for Longitudinal Data Systems - A Roadmap for States.*" Resource Library. http://www.

dataqualitycampaign.org/resources/fedfunding/.

Datnow, Amanda, Lea Hubbard, and Hugh Mehan. 1998. "*Educational Reform Implementation: A Co-Constructed Process*." Research Report 5. Washington, DC.

Dede, Chris, and John Richards, Eds. 2012. D*igital Teaching Platforms: Customizing Classroom Learning for Each Student. Technology, Education-Connections*. New York, NY: Teachers College Press.

Dede, Chris. 2009. "Immersive Interfaces for Engagement and Learning." *Science* 323 (5910) (January 2):66–69. http://www.sciencemag.org/content/323/5910/66.

Dede, Chris. 2012. "Interweaving Assessments Into Immersive Authentic Simulations: Design Strategies for Diagnostic and Instructional Insights." *Presented at the Invitational Research Symposium on Technology Enhanced Assessments*, May 7, National Harbor, MD.

DiCerbo, Kristen E., and John T. Behrens, 2012. "From Technology-Enhanced Assessment to Assessment-Enhanced Technology." *Presented at the National Council on Measurement in Education 2012 Annual Meeting,* Vancouver, BC.

Easton, John Q. 2012. "The Power of Measurement". Presented at the National Council on Measurement in Education. Inaugural Opening Plenary Session, April 14.

Eccles, Jacquelynne S., and Bonnie L. Barber. 1999. "Student Council, Volunteering, Basketball, or Marching Band: What Kind of Extracurricular Involvement Matters?" *Journal of Adolescent Research* 14(1):10–43.

EDUCAUSE. 2010. "*Next Generation Learning Challenges: Learner Analytics Premises*." http://www.educause.edu/library/resources/next-generation-learning-challenges-learner-analytics-premises.

Fairchild, Susan, Brad Gunton, Beverly Donohue, Carolyn Berry, Ruth Genn, and Jessica Knevals. 2011. "*Student Progress to Graduation in New York City High Schools: A Metric Designed by New Visions for Public Schools. Part I: Core Components*." New York, NY: New Visions for Public Schools.

Falmagne, Jean-Claude, Mathieu Koppen, Michael Villano, Jean-Paul Doignon, and Leila Johannesen. 1990. "Introduction to Knowledge Spaces: How to Build, Test, and Search Them." *Psychological Review* 97(2):201–24.

Feng, Mingyu, Neil Heffernan, and Kenneth Koedinger. 2009. "Addressing the Assessment Challenge with an Online System That Tutors as It Assesses." *User Modeling and User-Adapted Interaction* 19(3): 243–266.

Fletcher, J. Dexter, and John E. Morrison. 2012. "DARPA Digital Tutor: Assessment Data." Institute for Defense Analyses. www.acuitus.com/web/pdf/D4686-DF.pdf.

Fletcher, J. Dexter. 2011. "*DARPA Education Dominance Program: April 2010 and November 2010 Digital Tutor Assessments*." Institute for Defense Analyses. www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA542215.

Foley, Eileen M., Allan Klinge, and Elizabeth R. Reisner. 2007. "*Evaluation of New Century High Schools: Profile of an Initiative to Create and Sustain Small, Successful High Schools*." Final Report. Washington, DC.

Frank, Kenneth A., Minh Q. Duong, Spiro Maroulis, and Ben Kelcey. 2011. "*Quantifying Discourse About Causal Inferences from Randomized Experiments and Observational Studies in Social Science Research*." Austin, TX: University of Texas.

Fredricks, Jennifer A., and Jacquelynne S. Eccles. 2006. "Is Extracurricular Participation Associated with Beneficial Outcomes? Concurrent and Longitudinal Relations." *Developmental Psychology* 42(4) (July):698–713.

Fredricks, Jennifer A., Phyllis C. Blumenfeld, and Alison H. Paris. 2004. "School Engagement: Potential of the Concept, State of the Evidence." *Review of Educational Research* 74(1):59–109.

Freitas, Alex A. 2002. *Data Mining and Knowledge Discovery With Evolutionary Algorithms. Natural Computing Series Title II*. New York, NY: Springer.

Fuchs, Douglas, Lynn Fuchs, and Sharon Vaughn, eds. 2008. *Response to Intervention: A Framework for Reading Educators*. Newark, DE: International Reading Association.

Goldman, Shelley, Angela Booker, and Meghan McDermott. 2007. "Mixing the Digital, Social, and Cultural: Learning, Identity, and Agency in Youth Participation." *The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning* (December 1): 185–206.

Hausmann, Robert, and Annalies Vuong. 2012."Testing the Split Attention Effect on Learning in a Natural Educational Setting Using an Intelligent Tutoring System for Geometry." In P*roceedings of the 34th Annual Conference of the Cognitive Science Society*, ed. N. Miyake, D. Peebles, and R. P. Cooper. Austin, TX: Cognitive Science Society.

Heath, Shirley Brice, and Milbrey W. McLaughlin. 1993. *Identity and Inner-City Youth: Beyond Ethnicity and Gender*. New York, NY: Teachers College Press, Columbia University.

Heath, Shirley Brice. 1994. "Promoting Community-based Programs for Socialization and Learning." In *New Directions for Child Development,* ed. Francisco A. Villarruel and Richard M. Lerner, 63:25–34. San Francisco: Jossey-Bass.

Hidi, Suzanne, and K. Ann Renninger. 2006."The Four-Phase Model of Interest Development." *Educational Psychologist* 41(2):111–127.

Hull, Glynda, and Katherine Schultz. 2001. "Literacy and Learning Out of School: A Review of Theory and Research." *Review of Educational Research* 71(4) (December 1):575–611.

Johnson, Larry, Alan Levine, Rachel S. Smith, and S. Stone. 2010. *The 2010 Horizon Report*. Austin, Texas: The New Media Consortium. http://wp.nmc.org/horizon2010/.

Kalyuga, Slava, Paul Chandler, and John Sweller. 1999. "Managing Split-attention and Redundancy in Multimedia Instruction." *Applied Cognitive Psychology* 13(4):351–371.

Kelly, Anthony E., Richard A. Lesh, and John Y. Baek, ed. 2008. "Handbook of Design Research Methods." In Education: Innovations in Science, Technology, Engineering, and Mathematics Learning and Teaching. 1st ed. New York: Routledge.

Koch, Melissa, and William R. Penuel. 2007."Designing for Learning." *In CHI 2007 Conference Proceedings. Workshop on Converging on a Science of Design Through the Synthesis of Design Methodologies*. Reading, MA: ACM Press.

Koch, Melissa, Annie Georges, Torie Gorges, and Reina Fujii. 2010."Engaging Youth with STEM Professionals in Afterschool Programs." *Meridian: A Middle School Computer Technologies Journal* 13(1).

Koch, Melissa, William R. Penuel, Torie Gorges, and Geneva Haertel. 2009. "Embedded Assessment of Youth Learning in Informal Learning Environments." *Presented at the Annual Meeting of the American Educational Research Association.* San Diego, CA.

Koedinger, Kenneth R., and Albert Corbett. 2006. "Cognitive Tutors: Technology Bringing Learning Sciences to the Classroom." In *The Cambridge Handbook of the Learning Sciences,* 61–77. New York, NY: Cambridge University Press.

Koedinger, Kenneth R., Eileen A. McLaughlin, and John C. Stamper. 2012. "Automated Student Model Improvement." In *Proceedings of the 5th International Conference on Educational Data Mining*, ed. K. Yacef, O. Zaïane, H. Hershkovitz, M. Yudelson, and J. Stamper.

Koenig, Alan D., John J. Lee, Markus Iseli, and Richard Wainess. 2010. *A Conceptual Framework for Assessing Performance in Games and Simulations. CRESST Report* 771. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Koran, Mary Lou, and John J. Koran. 2006. "Aptitude-treatment Interaction Research in Science Education." *Journal of Research in Science Teaching* 21 (8).

Lemke, Jay, Robert Lecusay, Mike Cole, and Vera Michalchik. Forthcoming. "*Documenting and Assessing Learning in Informal and Media-Rich Environments: A Report to the MacArthur Foundation*." MacArthur Foundation.

Levin, Henry M., and Cecilia E. Rouse. 2012. "The True Cost of High School Dropouts." *The New York Times,* January 25, sec. Opinion. http://www.nytimes.com/2012/01/26/opinion/the-true-cost-of-high-school-dropouts.html.

Linebarger, Deborah L., Jessica Taylor-Piotrowski, and Sarah Vaala. 2007. *"Literature Review Part 3: Supplementing Television: What Enhances or Detracts from the Power of Television to Teach*." Childcare and Early Education Research Connections. http://www.researchconnections.org/childcare/resources/16194.

Liu, Yun-En, Erik Andersen, Richard Snider, Seth Cooper, and Zoran Popovi . 2011. "Feature-Based Projections for Effective Playtrace Analysis." *Presented at FDG '11*, Bordeaux, France.

Lovett, Marsha, Oded Meyer, and Candice Thille. 2008. "The Open Learning Initiative: Measuring the Effectiveness of the OLI Statistics Course in Accelerating Student Learning." Article submitted to the *Journal of Interactive Media in Education.*

Lundh, Patrik, Melissa Koch, and Christopher Harris. 2011. "Seeing Science as Part of Who You Are: Initial Impact of a STEM-focused Out-of-school Curriculum." *Presented at the Annual Meeting of the National Association for Research in Science Teaching.*

Madaus, George F., and Laura M. O'Dwyer. 1999. "Short History of Performance Assessment: Lessons Learned." *Phi Delta Kappan* 80(9):688–695.

Massa, Laura J., and Richard E. Mayer. 2006. "Testing the ATI Hypothesis: Should Multimedia Instruction Accommodate Verbalizer-visualizer Cognitive Style?" *Learning and Individual Differences* 16(4):321–335.

Mayer, Richard E. 1989. "Systematic Thinking Fostered by Illustrations in Scientific Text." *Journal of Educational Psychology* 81(2):240–246.

McLaughlin, Milbrey W. 2000. "*Community counts: How youth organizations matter for youth development*." Washington, DC: Public Education Network.

McLaughlin, Milbrey W., Merita A. Irby, and Juliet Langman. 1994. *Urban Sanctuaries: Neighborhood Organizations in the Lives and Futures of Inner-City Youth*. San Francisco: Jossey-Bass.

McManis, Lilla Dale, and Susan B. Gunnewig. 2012. "Finding the Education in Educational Technology with Early Learners." *Young Children* 67(3):14–25.

Means, Barbara. 2010. "Learning Technology in Context: Time for New Perspectives and Approaches." Prepared for the 2010 Aspen Institute Congressional *Program, Transforming America's Education Through Innovation and Technology.* Whistler, BC (Canada), August 18.

Messick, Samuel. 1994. "The Interplay of Evidence and Consequences in the Validation of Performance Assessments." *Educational Researcher* 23 (2) (March 1): 13–23.

Mislevy, Robert J., and Geneva D. Haertel. 2006. "Implications of Evidence-Centered Design for Educational Testing." *Educational Measurement: Issues and Practice* 25 (4): 6–20.

Mislevy, Robert J., Linda S. Steinberg, and Russell G. Almond, 2003. *On the Structure of Educational Assessment. Measurement Interdisciplinary Research and Perspective.*

Mozilla Foundation, Peer 2 Peer University, and MacArthur Foundation. 2012. "*Open Badges for Lifelong Learning.*" Mountain View, CA and New York, NY.

Nathan, Mitchell J., and Kenneth R. Koedinger. 2000a. "Teachers' and Researchers' Beliefs About the Development of Algebraic Reasoning." *Journal for Research in Mathematics Education* 31(2):168–90.

Nathan, Mitchell J., and Kenneth R. Koedinger. 2000b. "An Investigation of Teachers' Beliefs of Students' Algebra Development." *Cognition and Instruction* 18(2):209–237.

National Research Council and Institute of Medicine. 2000. *From Neurons to Neighborhoods: The Science of Early Childhood Development.* Ed. J.P. Shonkoff and D.A. Phillips. Committee on Integrating the Science of Early Childhood Development. Washington, DC: The National Academies Press.

National Research Council and Institute of Medicine. 2002. *Community Programs to Promote Youth Development.* Washington, DC: The National Academies Press.

National Research Council and Institute of Medicine. 2009. *Preventing Mental, Emotional, and Behavioral Disorders Among Young People: Progress and Possibilities.* Washington, DC: National Academies Press.

National Research Council. 1999. H*ow People Learn: Brain, Mind, Experience, and School. Expanded Edition.* Ed. Bransford, John D., Ann L. Brown, and Rodney R. Cocking. National Academies Press, Washington, DC.

Newmann, Fred M., Gary G. Wehlage, and Susie D. Lamborn. 1992. "The Significance and Sources of Student Engagement." In F. M. Newmann (Ed.), *Student Engagement and Achievement in American Secondary Schools*. New York: Teachers College Press, Columbia University.

O'Connor, Cailin, Stephen A. Small, and Siobhan M. Cooney. 2007. "Program Fidelity and Adaptation: Meeting Local Needs Without Compromising Program Effectiveness" (4). What Works, Wisconsin - Research to Practice Series (April).

Oakes, Jeannie. 2005. *Keeping Track: How Schools Structure Inequality*. New Haven, CT: Yale University Press.

Pashler, Harold, Mark McDaniel, Doug Rohrer, and Robert Bjork. 2008. "Learning Styles Concepts and Evidence." *Psychological Science in the Public Interest* 9(3) (December 1):105–119.

Pellegrino, James W., and Margaret L. Hilton, Eds. 2012. *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century. Committee on Defining Deeper Learning and 21st Century Skills. Division on Behavioral and Social Sciences and Education.* Center for EducationWashington, DC: The National Academies Press.

Pellegrino, James W., Naomi Chudowsky, and Robert Glaser, Eds. 2001. *Knowing What Students Know: The Science and Design of Educational Assessment.* Committee on the Foundations of Assessment. Board on Testing and Assessment: Washington, DC: The National Academies Press.

Penuel, William R., Barry J. Fishman, Britte Haugan Cheng, and Nora Sabelli. 2011. "Organizing Research and Development at the Intersection of Learning, Implementation, and Design." *Educational Researcher* 40(7) (October 11):331–337.

Penuel, William R., Savitha Moorthy, Angela DeBarger, Yves Beauvineau, and Katherine Allison. 2012. "Tools for Orchestrating Productive Talk in Science Classrooms." *The Future of Learning: Proceedings of the 10th International Conference of the Learning Sciences* (ICLC 2012). Sydney, Australia.

Pinkus, Lyndsay. 2008. "*Using Early-Warning Data to Improve Graduation Rates: Closing Cracks in the Education System.*" Policy Brief. Washington, DC.

PNRC (Promise Neighborhoods Research Consortium).

2012. “About the Promise Neighborhoods Research Consortium.” http://promiseneighborhoods.org/about/.

Quillen, Ian. 2012. “Hewlett Automated-Essay-Grader Winners Announced.” *Education Week,* May 9, sec. Digital Education.

Resnick, Lauren B., Sarah Michaels, and M. C. O’Connor, 2010. “How (well-structured) Talk Builds the Mind.” In *From Genes to Context: New Discoveries About Learning from Educational Research and Their Applications*, edited by D. Preiss and J. Sternburg, 163–194. New York: Springer.

Ries, Eric. 2011. *The Lean Startup: How Today’s Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses.* New York: Crown Business.

Ritter, Steven, John R. Anderson, Kenneth R. Koedinger, and Albert Corbett. 2007. “Cognitive Tutor: Applied Research in Mathematics Education.” *Psychonomic Bulletin & Review* 14 (2) (April): 249–255.

Ritter, Steven. 2012. Personal communication to Barbara Means. January 31.

Romero, Cristobal, and Sebastian Ventura. 2010. “Educational Data Mining: A Review of the State-of-the-Art.” *IEEE Tansactions on Systems, Man and Cybernetics,* Part C: Applications and Reviews 40 (2): 601–618.

Rubin, Donald B. 1997. “Estimating Causal Effects from Large Data Sets Using Propensity Scores.” *Annals of Internal Medicine* 127(8):757–763.

Rumberger, Russell. W. 2011. *Dropping Out: Why Students Drop Out of High School and What Can Be Done About It.* Cambridge, MA: Harvard University Press.

Rumberger, Russell.W. 2001. “Who Drops Out of School and Why.” *In School Completion in Standards-Based Reform: Facts and Strategies.* Washington, DC: The National Academies Press.

Sanchez, Monika, Sebastian Castrechini, and Rebecca A. London. 2012. “Exploring the Causes and Consequences of Chronic Absenteeism in a San Francisco Bay Area Community.” *Presented to the Annual Conference of the American Educational Research Association,* April 13-16, Vancouver, BC.

Schmidt, William H., and Richard T. Houang. 2012. “Curricular Coherence and the Common Core State Standards for Mathematics.” *Educational Researcher* 41(8) (November 1):294–308.

Schmidt, William H., Curtis C. McKnight, Richard T. Houang, Hsing Chi Wang, David E. Wiley, Leland S. Cogan, and Richard G. Wolfe. 2001. *Why Schools Matter: A Cross-National Comparison of Curriculum and Learning.* San Francisco, CA: Jossey-Bass.

Shadish, William R., and Thomas D. Cook. 2009. “The Renaissance of Field Experimentation in Evaluating Interventions.” *Annual Review of Psychology* 60(1):607–629.

Shepard, Lorrie A. 1991. “Will National Tests Improve Student Learning?” *Phi Delta Kappan* 73(3):232–38.

Shermis, Mark, and Ben Hamner. 2012. “Contrasting State-of-the-Art Automated Scoring of Essays: Analysis.” *Presented at the Annual National Council on Measurement in Education Meeting*, Vancouver, BC, April 14-16.

Shmueli, Galit, and O. Koppius. 2010. “*Predictive Analytics in Information Systems Research.*” Robert H. Smith School Research Paper No. RHS, 06-138. SSRN eLibrary. College Park, MD: University of Maryland.

Shute, Valerie J. 2011. “Stealth Assessment in Computer-based Games to Support Learning.” In *Computer Games and Instruction,* 503–524. Charlotte, NC: Information Age Publishing.

Shute, Valerie J., and Matthew Ventura. In press. "Measuring and Supporting Learning in Games: Stealth Assessment." White paper. MIT Series. The MacArthur Foundation. http://myweb.fsu.edu/vshute/pdf/white.pdf.

Shute, Valerie J., Eric G. Hansen, and Russell G. Almond. 2008. "You Can't Fatten a Hog by Weighing It -Or Can You? Evaluating an Assessment for Learning System Called ACED." *International Journal of Artificial Intelligence in Education* 18(4):289–316.

Siemens, George, and Ryan S. J. d. Baker. 2012. "Learning Analytics and Educational Data Mining: Towards Communication and Collaboration." In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 252–254. New York: ACM.

Skinner, Ellen A., and Michael J. Belmont. 1993. "Motivation in the Classroom: Reciprocal Effects of Teacher Behavior and Student Engagement Across the School Year." *Journal of Educational Psychology* 85 (4): 571–81.

Skinner, Ellen, Carrie Furrer, Gwen Marchand, and Thomas Kindermann. 2008. "Engagement and Disaffection in the Classroom: Part of a Larger Motivational Dynamic?" *Journal of Educational Psychology* 100 (4) (November): 765–781.

Smith, Marshall S. 2009. "Opening Education." Science 323 (5910) (January 2): 89–93. http://www.sciencemag.org/content/323/5910/89.

Sumner, Tamara, and C. C. S. Team. 2010. "Customizing Science Instruction with Educational Digital Libraries." In *Proceedings of the 10th Annual Joint Conference on Digital Libraries,* 353–356. Gold Coast, Queensland (Australia): ACM.

Thompson, Clive. 2011. "How Khan Academy Is Changing the Rule of Education." *Wired*, August. www.wired.com/magazine/2011/07/ff_khan/all/1.

Thrun, Sebastian. 2012. "University 2.0" presented at the Digital Life Design 2012 Conference, Munich Germany.

U.S. Department of Education. 2010a. *Transforming American Education: Learning Powered by Technology* (National Education Technology Plan 2010). Washington, DC. http://www.ed.gov/technology/netp-2010

U.S. Department of Education. 2010b. "SLDS Technical Brief 1: Basic Concepts and Definitions for Privacy and Confidentiality in Student Education Records". http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2011601.

U.S. Department of Education. 2012a. *Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics*. Washington, DC. www.ed.gov/edblogs/technology/files/2012/03/edm-la-brief.pdf.

U.S. Department of Education. 2012b. "Inside the WWC." What Works Clearinghouse Website. http://ies.ed.gov/ncee/wwc/InsidetheWWC.aspx#process.

U.S. Department of Education. 2013. G*rit, Tenacity, and Perseverance in 21st-Century Education: State of the Art and Future Directions*. Washington, DC.

U.S. Department of Health and Human Services. 2002. Finding the Balance: Program Fidelity and Adaptation in Substance Abuse Prevention. Conference Edition. Washington, DC: Substance Abuse and Mental Health Services Administration, Center for Substance Abuse Prevention.

Vendlinski, Terry P., Eva L. Baker, and David Niemi. 2008. "Templates and Objects in Authoring Problem-Solving Assessments." *CRESST Report 735*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Wehlage, Gary G., Robert A. Rutter, Gregory A. Smith, Nancy Lesko, and Ricardo R. Fernandez. 1989. *Reducing the Risk: Schools as Communities of Support*. Philadelphia, PA: Taylor & Francis.

Wentzel, Kathryn R. 1997. "Student Motivation in Middle School: The Role of Perceived Pedagogical Caring." *Journal of Educational Psychology* 89(3):411–419.

West, Darrell, Patte Barth, Karen Cator, and Jose Ferreira. 2012. "Data Analytics and Web Dashboards in the Classroom." Brookings Institution. http://www.brookings.edu/events/2012/09/04-classroom-analytics.

Wiggins, Grant, and Jay McTighe. 1998. "*Understanding by Design*." Association for Supervision and Curriculum Development.

Woolf, Beverly, Winslow Burleson, Ivon Arroyo, Toby Dragon, David Cooper, and Rosalind Picard. 2009. "Affect-aware Tutors: Recognising and Responding to Student Affect." *International Journal of Learning Technology* 4(3):129–164.

Yazzie-Minz, Ethan. 2010. "Charting the Path from Engagement to Achievement: A Report on the 2009 High School Survey of Student Engagement." Bloomington, IN: Center for Evaluation and Education Policy, Indiana University. http://ceep.indiana.edu/hssse/images/HSSSE_2010_Report.pdf.

Zapata-Rivera, Diego. 2012. "Embedded Assessment of Informal and Afterschool Science Learning." Paper commissioned by the National Research Council, Board on Science Education, Division of Behavior and Social Sciences and Education, Washington, DC. SOW 06-01-12 (NRC short thought paper). http://sites.nationalacademies.org/DBASSE/BOSE/DBASSE_071087#.UM-g5ba8FRw.

Photo Credits

Cover and Chapter 1: "datajockey" by medul.la. Available at http://www.flickr.com/photos/medul_la/5697365897/in/photostream/ under a Creative Commons Attribution 2.0. Full terms at http://creativecommons.org/licenses/by/2.0/deed.en. Modified by Ashley Lee and Kate Borelli.

Introduction: "bits" by sciascia. Available at http://www.flickr.com/photos/washingtonydc/4391696247/ under a Creative Commons Attribution 2.0. Full terms at http://creativecommons.org/licenses/by-nc-sa/2.0/deed.en. Modified by Ashley Lee and Kate Borelli.

Chapter 2: "Hand" by John-Morgan. Available at http://www.flickr.com/photos/aidanmorgan/2241368324 under a Creative Commons Attribution 2.0. Full terms at http://creativecommons.org/licenses/by/2.0/deed.en. Modified by Ashley Lee and Kate Borelli.

Chapter 3: "Gear" by avrene. Available at http://www.flickr.com/photos/enerva/4302079406/ under a Creative Commons Attribution 2.0. Full terms at http://creativecommons.org/licenses/by/2.0/deed.en. Modified by Ashley Lee and Kate Borelli.

Chapter 4: "LEGO Mindstorms" by jProgr. Available at http://flickr.com/photos/philossofia/7321746490/ under a Creative Commons Attribution 2.0. Full terms at http://creativecommons.org/licenses/by/2.0/deed.en.

Chapter 5: "Compass Inlay" by Steve Snodgrass. Available at http://www.flickr.com/photos/stevensnodgrass/4828932145/ under a Creative Commons Attribution 2.0. Full terms at http://creativecommons.org/licenses/by/2.0/deed.en. Modified by Ashley Lee and Kate Borelli.

Summary and Recommendations: "Light bulb vase" by entso. Available at http://www.flickr.com/photos/entso/6431413137/ under under a Creative Commons Attribution 2.0. Full terms at http://creativecommons.org/licenses/by-nc-nd/2.0/deed.en.

The Mission of the Department of Education is to promote student achievement and preparation for global competitiveness by fostering educational excellence and ensuring equal access.

www/ed.gov