MEASUREMENT INCORPORATED
EXCELLENCE IN ASSESSMENT™

The Future of Testing

Michael B. Bunch, Measurement Incorporated

Prediction is very difficult, especially about the future – Niels Bohr

The Educational Testing Service Invitational Conference of 1985 had as its theme "The Redesign of Testing for the 21st Century." It has been a little over 25 years since the conference proceedings were published, and it is amazing to see that the future still looks a lot like it did 25 years ago: cognitive-based assessment, online assessment, widespread use of computer-adaptive testing, universal access to technology, and instantaneous reporting of test results. So many wonderful things, still within our view but just beyond our grasp!

Although the projections of the 1985 ETS conference were fairly modest in scope, Bob Linn sounded a note of caution. He predicted that the educational, psychometric, and technological breakthroughs of the 1980s and beyond would not be sufficient to cause fundamental change in educational assessment because they would have to overcome human reluctance to abandon a system that actually worked rather well, in terms of efficiency and cost effectiveness. We face the same issues today. Therefore, in this paper, I will (in the spirit of Charles Dickens' *A Christmas Carol*) examine a future that could be rather than a future that we believe must be.

The Role of Cognitive Psychology

The contrast between traditional psychometric approaches and contemporary cognitive approaches to assessment boils down to focus: Psychometrics poses tasks to permit mathematical modeling of current behavior, for either present or future purposes. Cognitive psychology poses tasks to find out what cognitive processes people use to solve them. To some degree, it may be fair to say that psychometrics does not care what goes inside someone's head, while cognitive psychology cares only about what goes on there. Each approach provides an important insight, but each also ignores something terribly important.

Cognitive psychology contributes most to assessment when it augments psychometric approaches. We want to know how much middle school mathematics children have mastered, partly to evaluate the quality of their middle school mathematics instruction and partly to understand (predict) who is adequately prepared for high school mathematics. Psychometrics does a perfectly good job of addressing both needs. But what exactly do those middle school students know? At what depth do they really know it? What else do they need to know, and how deeply do they need to know it? Who is going to need additional help to be successful in high school mathematics, and what kind of help are they going to need? Psychometrics isn't much help here. Cognitive psychology, however, exists to address just such questions.

Evidence-centered assessment combines the best qualities of psychometric and cognitive approaches. We still want to know what goes on inside people's minds, but we acknowledge that we have to deduce that by observing their behavior. Thus, we craft tasks and questions around external evidence that will allow us to make claims about what a person knows based on what that person does.

At the same time, we want to know what that person might be able to do in the future. When students leave high school, we want to know if they are likely to be successful in college and/or career. When they leave middle school, we want to know if they are likely to be successful in high school. In both instances, we know that the cognitive skills they have demonstrated in one setting will still be relevant in later life, either in their original form or as prerequisites to additional skills they will acquire along the way. Thus, while focusing on their thought processes (cognitive psychology), we are predicting future behavior (psychometrics). The quality of the tasks and questions we devise for this purpose will directly affect the quality of our understanding of present processes and our predictions of future behavior.

To a great extent, this approach to assessment design is not new at all. Industrial and organizational psychology has been using a very similar approach for decades, and with a much broader definition of cognition.

The moment we apply the term "college/career ready," we enter the world of the industrial/organizational psychologist, in that we are implicitly making a prediction of future performance. The I/O psychologist theorizes about a particular construct to create contrasting groups who should perform differentially on valid tests of that construct and then (using psychometric methods) eliminates tasks and questions that fail to differentiate until a factorially pure, valid test of the construct emerges. Tests of college and career readiness will be successful to the extent that we gather sufficient long-term evidence of success and use it to change or eliminate items (which may appear to be perfectly valid measures of the construct now) that do not support long-term claims.

> *"Evidence-centered assessment combines the best qualities of psychometric and cognitive approaches."*

By 2016, we should be seeing a rash of studies of first-year college performance relative to performance on high school tests. By 2020, we should start seeing a body of evidence that will permit test designers to make mid-course adjustments to those same high school tests. At the same time, we will see an emerging body of evidence that will permit developers of elementary and middle school tests to adjust their designs based on students' performances in high school. In all likelihood, both bodies of evidence will include non-cognitive factors and personal attributes we have not considered in the development of the current generation of tests, and we will have an opportunity to refine our definition of college/career readiness and even high school readiness and redesign the tests.

Beyond 2020, evidence-centered design will displace traditional educational assessment design paradigms, but only if certain conditions are met:

1. Evidence-centered design is seen not as a replacement for psychometric procedures by a purely cognitive-based approach but as a marriage of psychometric and cognitive approaches;
2. Graduate-level preparation in educational assessment recognizes this marriage and promotes it, along with a thorough grounding in both cognitive psychology and psychometrics;
3. Undergraduate teacher preparation stresses the relevance of evidence-centered assessment design to classroom instruction and provides at least a semester of bona fide instruction in test design, use, and interpretation.

The Role of Technology

Bob Linn warned us in 1985 that change would be met with resistance. Looking back to 1985, it is not so hard to see why someone might be less than excited about connecting an Apple II-e to a 300-baud modem and trying to take a test. A lot has changed since 1985, however, and the pace of technological change is increasing all the time.

Artificial Intelligence

A recent watershed moment illustrates not only the current state of technology but the resistance to it as well. Dr. Ellis Page pioneered a technology that allowed computers to read and score essays and demonstrated its viability nearly 50 years ago. For decades, Dr. Page's Project Essay Grade (PEG) remained an academic curiosity.

Fast forward to 2012. Automated scoring of essays by artificial intelligence (AI) has been commercialized. Nine vendors (eight companies and one university) participated in a demonstration of the viability of AI scoring of a wide variety of essay types over several grades and genres. Measurement Incorporated (MI) was one of those companies. The conclusion of the independent evaluation of all these AI scoring programs was that, overall, computers could score essays as well as humans could. Five of the companies (with MI in the lead) actually outperformed the human readers. In a follow-up demonstration in the summer of 2012, computers scored content-based responses to open-ended questions as well. Once again, MI took the lead.

The immediate response of many academics and journalists was that computers just can't appreciate a fine piece of writing or give counsel to a budding Steinbeck. Their objection to computer scoring of essays had nothing to do with the technical adequacy of the process or product; it was simply that the machine is not human.

The recording industry has faced similar criticisms for more than a hundred years. No matter how good recording became or how faithfully it reproduced the sound of the human voice (or instruments played by humans), people objected, simply because it was not human. An iconic set of television ads in the 1980s asked, "Is it live, or is it Memorex?" Today, having moved from cylinders to wax to vinyl to 8-track to mini-cassette to CDs to MP3 formats and beyond, few people worry whether the sound they are hearing is live or a recording. They don't care because they know they can listen to their MP3 players wherever they go. They can't do that with live music. Convenience, cost, and fidelity have won.

In the same way, we will eventually embrace automated scoring of essays, not because AI scoring is so superior to human scoring but because AI scoring can provide instantaneous and reliable scores. Humans can't do that. Or if they did, it would be terribly expensive. MI is already providing scores for one statewide writing assessment as well as instantaneous scores for commercially available tests. Other companies are providing writing assessments scored by computers for admissions purposes. Companies throughout the world are using AI to score the writing of their employees and prospective employees. Once again, convenience, cost, and fidelity will win.

In North Carolina and Connecticut, MI provides online writing exercises that are scored by PEG. Students can write as many essays as they like, on a variety of topics, and submit them to PEG for scoring. Scores on six dimensions of writing come back in two to three seconds. In addition, the scoring system directs students to tutorials designed to help them improve one or more of the six dimensions of their writing. The system also allows teachers to monitor the entire process, check the various drafts, and leave notes for the students. This system has become quite popular because it teaches as it tests and allows teachers and students to interact along the way.

"We will eventually embrace automated scoring of essays …because AI scoring can provide instantaneous and reliable scores."

Online Assessment

Other technologies, particularly online assessment, will also continue to advance – not because a particular state or group of states promotes online assessment but because we have become comfortable with the paradigm. Smartphones still drop calls and sometimes provide less audio fidelity than one might desire, but we have accepted their limitations because we so enjoy and appreciate their features, many of which are totally unavailable on conventional land-line phones or even older cell phones. Indeed, these added features are often the prime purpose of some people's smartphones – they text, they tweet, they surf the web, they take and upload pictures of themselves and their surroundings, they check their status, and they occasionally make a phone call. These other activities hardly represent pent-up demand; they are activities previously unheard of, and people engage in them now simply because they can.

In the same way, online assessment will prevail, and we will do things with these assessments we never thought of before, not because we wanted to but couldn't but because we can. Online assessment is positioned to replace paper testing in the same way that smartphones are poised to replace land-line phones. Many of our clients are now using our online test delivery system (MIST – Measurement Incorporated Secure Testing) interchangeably with paper-based testing, and some are using it exclusively. We are now able to offer item types (e.g., technology enhanced) and feedback that would be impossible with traditional paper-based tests. Other companies are having similar experiences.

CAT

The technology and psychometrics for computer adaptive testing (CAT) have existed for at least 50 years. Why hasn't it caught on? Actually CAT has caught on in a variety of settings but not yet in large-scale educational assessment. The Smarter Balanced Assessment Consortium (SBAC) is promoting it now and expects to see widespread use of CAT by 2015. This time, we think it will work, for two reasons:

1. Big money is promoting it – CAT has been another of those academic curiosities for the past 50 years because the expertise to create and maintain the software resides in the heads of very smart and therefore fairly expensive individuals, the technology to house the system is expensive, and the item banks required to make it work optimally are much larger and therefore more expensive than conventional item banks. All of these problems can be solved with money, and at least on the technology side of the equation, prices are dropping.
2. It is directly associated with online assessment – which is rapidly gaining acceptance. Students who begin their educational careers taking only online tests will not notice when one of those tests branches to new items based on their responses to the previous items. It will seem perfectly natural to them.

The remaining objection to CAT is the notion that somehow scores are not comparable unless the tests are identical. "You got a higher score than I did because you took an easier test," is easy to refute if the person making the claim understands psychometrics, but virtually impossible if not. However, we will eventually overcome even this objection as more and more people take online, computer-adaptive tests. The sheer weight and volume of CAT will ultimately overwhelm this objection long before logic will.

Digital Breadcrumbs

The U. S. Department of Education and the Federal Communications Commission have joined forces to publish the *Digital Textbook Playbook* to "advance the conversation toward building a rich digital learning experience" (p. 3). These two federal agencies, along with a host of schools, districts, nonprofit groups, and commercial vendors, have seen the future of education, and it is digital. The *Playbook* shows users how to make the transition to a digital learning environment, not at some distant time in the future but now.

Many schools and districts are already employing digital technologies to provide instruction, formative assessment, and grading for large numbers of students. Two North Carolina districts, for example, have replaced textbooks with laptops and smartphones. Students interact not only with the content on the districts' servers but with their teachers and one another, asking and answering questions, collaborating on projects, surfing the internet to find additional information, and uploading their homework to in-class smart boards and other digital devices.

As students interact with the system and one another, they leave a trail of digital breadcrumbs tracing their paths through the curriculum, just as social media users and retail customers leave digital breadcrumbs of their online behavior. Social media sites and retailers have already figured out how to turn their own users' and customers' digital breadcrumbs to economic advantage. Can educational testing be far behind?

> *"The power to track, predict, and alter is already in our hands."*

We are already using such information in the formative writing assessments in Connecticut and North Carolina described above. Can we, or should we use digital breadcrumbs for interim or summative assessment? How might knowing what students have downloaded, how long they spent with a given file, or how they phrased questions help us understand what's going on inside their minds? Is it even ethical to consider such an approach? This is not a hypothetical question. The power to track, predict, and alter is already in our hands. It's time for an open and serious discussion about how we will use it.

Classroom: The Final Frontier

Advances in psychometric theory over the past century have come largely in response to problems posed by large-scale testing programs. While there has been no lack of guidance for teachers to create better classroom tests, that guidance has focused primarily on the art of test construction, rather than the science. Recent advances in cognitive psychology, particularly as they relate to the multi-step approach to item and test construction, seem beyond the reach of typical classroom teachers, who have little undergraduate preparation in test construction, psychometrics, or cognitive psychology. If the promise of balanced assessment (with formative, interim, and summative components working together in harmony) is to be fulfilled, current and future teachers must receive adequate instruction in its basic tenets and appropriate supervision in their implementation.

In the future, classroom teachers can and should be full partners in the educational assessment enterprise. They will be if and only if we begin now to revamp undergraduate and graduate programs in colleges of education to include meaningful instruction in psychometrics, cognitive psychology, and the integration of evidence-centered assessment and instruction.

Mitigating Factors

As Bob Linn point out in 1985, in order to take advantage of breakthroughs in cognitive psychology and technology, we will have to address human resistance to change at multiple levels. Teachers, administrators, parents, testing companies, and the general public have a variety of reasons for clinging to the status quo. Understanding those reasons is a necessary first step toward change. With a few exceptions, the future still looks a lot like it did in 1985, at least as far as educational testing is concerned. The main difference is that now we have good reason to believe in it.

Suggestions for Further Reading

Educational Testing Service (1986). *The Redesign of Testing for the 21st Century*. Princeton, NJ: Author.

Lissitz, R. W. & Jiao, H. (2012). *Computers and Their Impact on State Assessments*. Charlotte, NC: Information Age Publishers.

Pellegrino, J. W., Chudowski, N., & Glaser, R. (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington, DC: National Research Council.