Educational Researcher

http://er.aera.net

2007 Presidential Address—The End(s) of Testing

Eva L. Baker

EDUCATIONAL RESEARCHER 2007 36: 309
DOI: 10.3102/0013189X07307970

2007 Presidential Address

The End(s) of Testing

by Eva L. Baker

This expanded version of the 2007 Presidential Address for the American Educational Research Association approaches changes in assessment and accountability from the perspective of achieving balance. The author identifies unresolved conceptual and practical issues in the use of assessment in the schools and describes approaches intended to mitigate problems or improve practice. An example is given from formative assessment research that integrates some of these approaches (Center for Research on Evaluation, Standards, and Student Testing, at the University of California, Los Angeles). International examples of performance and assessment are also reviewed. The author suggests that secondary school assessment be modified to focus students on acquiring concrete Qualifications that certify important accomplishments, with a wide choice of Qualification areas. She argues that using concrete accomplishments and restructuring tests will help schools and their students to regain needed balance.

Keywords: accountability improvement; formative assessment; high school performance; international qualifications; testing policy

Take a moment and consider the following question. What needs to change in your life? I have a confession to make. You may be surprised to learn that I want a life with more balance. Over the years, I've asked many colleagues to name a professional who lives a balanced life, and always with the same result. Long stares and no names. For most of us, however, balance is a victim of competing values, looming obligations, unfulfilled ambitions, and in my case, wearable media. I think it no accident that balance is a metaphor for justice. The lack of balance takes different forms but is found across the entire economic spectrum—for all groups and all kinds of people—for the old, for the young, and—relevant here—for those in schools.

In particular, U.S. adults consistently report in surveys that their lives are out of whack (New American Dream, n.d.). I think we have inadvertently projected this imbalance onto the lives of schools.

So this talk—"The End(s) of Testing"—is about balance in testing. It is about what we have now; ways to move to a new equilibrium; and where we reconnect achievement to learning, equity to more than equal test scores, and students to their own paths. I'll use a few examples from other countries and liberally incorporate ideas from research here and around the world.

Consider this statement by Jerome Bruner (1996): "Life in culture is . . . an interplay between the versions of the world that people form under its institutional sway and the versions of it that are products of their individual histories" (p. 14).

Educational policies and practices are motivated in at least two major ways: by institutional requirements that convey uniform requirements and by opportunities that reflect local, group, and individual preferences, talents, and values. This is a modification of a construction by Tyler (1949) identifying societal, disciplinary, and personal needs and preferences as sources of goals. During different eras, different balances between institutional and personal values have been struck. Institutional or bureaucratic requirements are important, as Bruner (1996) noted, in serving as a common basis of competence. Accountability tests have swung education strongly toward institutional goals and away from those of the individual.

Schools around the world and especially in the United States are heavily tilted to institutionally instigated tests, exemplified in their accountability structures. At best, accountability requires the accounting of outcomes and the redress of inadequate or inequitable outcomes. For example, in many countries, untenable gaps exist on common measures of differences between the highest- and poorest-performing students. That we have institutional pressures in numerous countries pushing toward common goals is understandable during a period of growing population diversity. Such goals are one way—not without critics—to nurture national identity.

A fundamental policy question, especially urgent in secondary school, is, What should be the balance between promoting common performance and supporting different, either individual or locally valued, talents? Can common and divergent performance be encouraged simultaneously? What are the trade-offs for adding more options and flexibility to accountability? Can we add options without adding ways to fail? Expanding accountability options is a hard sell when significant group differences in performance remain. It is far easier to understand South Korea's policy initiative to incorporate applied problem solving and innovation in its schools, for South Korean students are already high on the ladder of comparative performance. To presage my conclusion, I think we must augment our usual accountability tests with other qualifications that link accountability to learning, that tap deeper individual

Educational Researcher, Vol. 36, No. 6, pp. 309–317
DOI: 10.3102/0013189X07307970

knowledge, and that measure ability to transfer to new contexts and applications. In addition to improving normative tests, we should consider encouraging the students to pursue their interests and talents at some depth. I would hope that many of these secondary school qualifications could be adapted from those used in other countries.

Despite the desire for options, we *need* common goals and measures to identify and fix inadequacies and inequities. Yet when limited, these accountability indicators have dramatic effects on teaching and learning. Cramped by requirements, we harden instruction, drop electives, and shorten time for in-depth engagement. Students' voices and choices are fainter. Since the world is flat in testing, too, test results everywhere strongly affect classroom realities and public perceptions of learning and schooling.

Balance is a goal, then, as well as a fundamental design principle essential to quality itself. Balance invariably involves tradeoffs, so what do we give up to gain educational balance? Let's start with what we have.

On Current Standards and Accountability Measures

Daily, or more often, researchers and education writers bemoan the technical shortfalls of tests now used in accountability, contending that the content of these tests is off, their sampling is odd, and deep cognition is missing. And the institutional tests don't seem to matter much to students who take them. I add to that litany (National Research Council, 2001) that these tests often ignore findings from learning research and transfer of knowledge, that is, the application of learning (Donovan, Bransford, & Pellegrino, 1999) to something other than another test. Tests only dimly reflect in their design the results of research on learning, whether of skills, subject matter, or problem solving. These test-design properties matter to researchers but rarely are observable in the tests because the naked eye is drawn to test format. Many professionals, and, I would guess, most of the public, don't make fine distinctions. They see tests on the same topic as interchangeable and a high score on any test as sufficient evidence of learning.

To my mind, the evidential disconnect between test design and learning research is *no small thing*. Think about it. It means, at worst, that tests may not actually be measuring the learning for which schools are responsible, thus gutting the basic tenet of the accountability compact. What if we set aside learning-based design and ask, "How well do any of our external tests work?" The answer is that we often don't know enough to know. We have little evidence that tests are in sync with their stated or de facto purposes or that their results lead to appropriate decisions. Nevertheless, we *act* as if tests were valid, in the face of weak or limited evidence. We make heavy and far-reaching decisions about schools and students, talk about gaps, and applaud progress. This excitement takes place with only fragments of the kinds of evidence called for in the testing standards published jointly by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1999). Education researchers conduct validity studies, when possible, and continue their plaintive call for high-quality, evidence-based tests. We know the importance of right inferences by teachers, administrators, and the government. We should be at least queasy about the quality of the interpretations, the meaning of gaps and improvement, and the resultant classifications of schools. Yet test validity languishes as a largely unexamined, prior question because of inexorable schedules and budget constraints. With tests of uncertain validity, adequate yearly progress (AYP), value-added, or other growth modeling analyses will have limited meaning in accountability interpretations.

So how did we get here? The wave of U.S. reform was stimulated, in part, by lackluster performance on international comparisons more than two decades ago (Beaton, Martin, et al., 1996; Beaton, Mullis, et al., 1996). Not surprisingly, the reform plan was to follow the international lead and design a quasinational system of standards and assessments. Despite awareness of huge differences in context and traditions (our 50-state autonomy in education, distributed curricula, independent teacher education institutions, and waning respect for those working in education), state and federal legislation enabled state standards and related tests (National Council on Education Standards and Testing [NCEST], 1992).

We see once more how tests exert power. Research in both schools and laboratories shows that testing improves learning of what is tested (Roediger & Karpicke, 2006), even in the absence of feedback, and that with feedback, learning improves more (Nyquist, 2003). When results are linked to sanctions, teaching moves to conform to the content boundaries of the tests. Because topics on tests are represented inequitably, to avoid sanctions, it is efficient to practice *test-like* items. Practice them a lot. Dan Koretz (2002) calls this "score inflation." You may call it smart. In either case, it is likely that test-specific content and formats are learned through test practice at the expense of the intended content and skill domains promised by standards. Some learning takes place, but institutional test anxiety uses up all the instructional oxygen in the classroom.

Coping Strategies and "Accountabalism"

While accepting the importance of accountability, educators in the United States and around the world have anticipated negative consequences of hyper-test consciousness. David Weinberger (2007, p. 54), in the *Harvard Business Review*, describes a process called "accountabalism" (think of Hannibal Lector with a checklist). Weinberger illustrates the consequences of a repair mentality that deals serially with each successive difficulty in accountability systems as if we could perfect the systems. Although not focused on schools, his analysis of accountability suggests that a set of chronological, piecemeal fixes does not really improve systems. Over time, repeated patching and spackling consume all remaining flexibility. Sound familiar?

Mitigations

In attempting to head off some of the negative side effects of accountability while still accepting its practical and political value, educators have advocated six tactics. The first tactic is to add more or different measures or indictors to accountability—in shorthand, to use *multiple measures* (Baker, 2003; Improving America's Schools Act, 1994)—although the original intent of the term was to allow students different ways to show their competence (NCEST, 1992). One approach adds a periodic test, say, of history, with rotating content to the system, to be incorporated into the accountability calculation.

A second tactic calls for *measures of opportunity to learn* (OTL) as part of accountability models (NCEST, 1992). OTL was early thought of both as a check on fair access to test-relevant instruction and as a shield to protect nontested content and activities, such as different courses. Scalable, valid methods for routinely assessing inside-the-classroom OTL remain an ongoing research quest (Aguirre-Muñoz & Baker, 1997; Clare & Aschbacher, 2001; Goldschmidt & Martinez-Fernandez, 2004; Niemi, 1996; Niemi, Wang, Steinberg, Baker, & Wang, in press). When there are efficient, replicable, and valid procedures for its measurement, the role of OTL in accountability is clear.

"Having tests worth teaching to," despite its grammatical failings, is the mantra of the third approach, *performance assessment*. Change the test and legitimate test preparation. Serve good rather than evil. Performance assessment calls for multistepped activities, usually imbued with some realism. More significantly, performance assessment design originally reflected research from teaching, from cognitive psychology, and from subject matter learning (Baker & O'Neil, 2006; Baxter & Glaser, 1998; Baxter & Mislevy, 2005; Darling-Hammond, 2000; Glaser, 1996; Linn, Baker, & Dunbar, 1991; Shavelson, Gao, & Baxter, 1993). In the heyday of performance assessment, models from abroad were emulated, and tasks were sometimes directly copied. After a period of handsome financial support, popularity waned and only vestiges of performance assessment, such as written composition, can be found on mandated tests—and in higher education (Klein, Kuh, Chun, Hamilton, & Shavelson, 2005). Why the demise? The list includes lack of political acceptance and structural integration, feasibility, technical quality, and credibility (because of overselling). Costs were also an important constraint.

A fourth and currently "hot" mitigating tactic is *formative assessment*. Assessment to support learning has been continuously in play since Lindquist (1951) and Skinner (1953) wrote about it, followed by Vygotsky (1962), Lumsdaine (1965), Markle (1967), Popham and Baker (1970), Baker and Saloutos (1974), and Herman, Osmundson, Ayala, Schneider, and Timms (2006). Black and Wiliam (1998), James (1998), and Heritage and Niemi (2007)—all originally from abroad, by the way—now invoke a process whereby teachers ask penetrating questions, pinpoint errors, give insightful and timely feedback, and use innovative teaching to help students. Not surprising has been the entry of commercially supplied benchmark or interim tests widely in use as formative assessment, in name if not in spirit. For use at intervals during instruction, interim or benchmark tests give previews of coming results based on test segments similar to the main assessment. The validity of these measures for instruction is rarely cited, other than predictive validity. Although the interim results are intended to influence teaching, because of strict instructional pacing, some teachers have little or no time to use them and must move on to the next topic, without addressing students' needs. Even with adequate time, formative assessment depends on teachers with expertise in their subject matter, understanding of their students, *and* sufficient fluency in alternate ways of teaching when students require a second or third opportunity. Teachers with these skills are often in short supply where students are most at risk.

A fifth tactic is *limiting the number of standards* to be tested by setting clear priorities: a less-is-more stance. The Commission on Instructionally Supportive Assessment (2001) prepared a report arguing that the use of fewer but more powerful standards will lead to valid measurement, more coherent teaching, and deeper student learning. Their proposal, authored principally by W. J. Popham, directly confronts the political bargains that underlie expansive lists of standards.

A sixth approach is *technology-based assessment*, that is, the use of procedures and hardware intended to provide smart assistance to teachers and learners, and it is my favorite. I'm the one, after all, with a Blackberry *and* a Treo, playing with my new iPhone, which complements my four iPods and five computers. Intelligent tutoring systems show us how to link assessment and learning models at a microlevel. Yet technology-enhanced tests have not yet stepped up to early expectations. In large-scale testing, they serve efficiency without capitalizing on their potential to leverage better design and higher fidelity experience (e.g., simulations), provide instant scoring of open-ended responses, or exploit students' ease with technology. They have been stuck in the useful but limited domain of computer-adaptive tests.

As a way forward to revitalize performance assessment, technology offers some hope for the discovery of more effective ways to assess validity with regard to learning and to reduce the cost of design and use. There has been excellent progress in computer scoring of open-ended written responses (Shermis & Burstein, 2003); with optical character reading, students' handwritten work can be scanned and directly marked by computer. Speech recognition technologies process oral language so that discourse of learners of all ages and language backgrounds can be partly analyzed (Hasegawa-Johnson & Alwan, 2002). Mapping of knowledge provides a rapid, easily designed, expertly scored approach (Herl, 1995). Expect to see more computer games and virtual worlds serving as assessment contexts (Mislevy, Steinberg, Breyer, Almond, & Johnson, 1999) as well as assessments embedded in common devices such as cell phones and game platforms—first, to reduce gaps in computer access and, second, to leverage motivation with fun and the familiar. In addition, the use of common devices will permit "anytime, anywhere" or "just-in-time" assessment (Fletcher & Tobias, 2007). To speed and improve assessment design, look to computer-assisted assessment authoring systems that can improve assessment design with built-in knowledge to help users make, mark, and manage tests (Mislevy et al., 2003; Vendlinski, Niemi, & Wang, 2005; Vendlinski, Niemi, Wang, & Monempour, 2005).

International Comparisons

If the United States' showing on international comparisons started us down this path, what do recent international tests tell us? The Organisation for Economic Co-operation and Development (OECD) has fielded the Programme for International Student Assessment (PISA), examinations that include cross-curricula content and the application of cognitive skills, such as problem solving. Their results are worth reflection (McGaw, in press, 2007; OECD, 2007; Schleicher, 2007). The vertical bars in Figure 1 show great performance variation between the best- and poorest-performing students in most countries on some 21st-century skills. This variation applies almost independently of average score but varies in degree in homogeneous or small countries. Why? Are current curricula not sufficiently attuned to future requirements—are students not taught to transfer learning or to apply cognitive skills to new situations? Don't all students deserve to learn these skills?

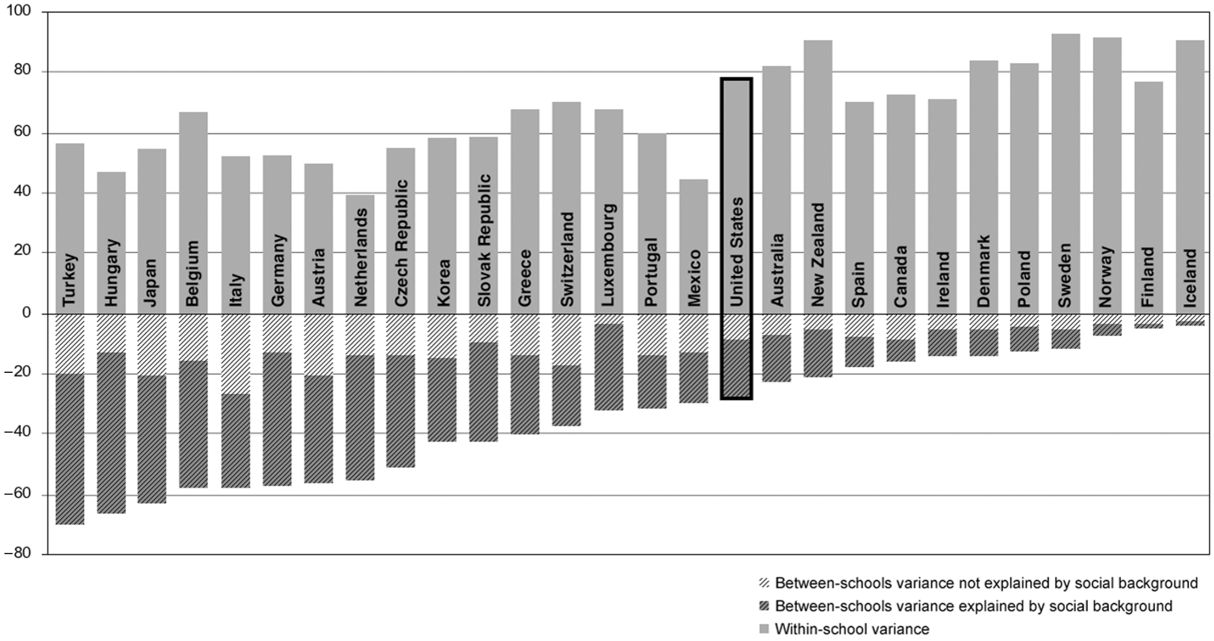

FIGURE 1. *Variance in student PISA 2003 mathematics achievement, partitioned into within- and between-schools components. Adapted from* Learning for Tomorrow's World—First Results From PISA 2003 *(p. 162, Fig. 4.1) by Organisation for Economic Co-operation and Development, 2004, Paris: OECD Publications. Copyright 2004 by OECD Publications. Adapted with permission.*

CRESST POWERSOURCE©: A U.S.–International Collaboration

I will now briefly touch on the system developed by the Center for Research on Evaluation, Standards, and Student Testing (CRESST), our own integration of a balanced formative assessment system. Its design is based on learning research, including schema development, explanation, narrative, and transfer. This system, CRESST POWERSOURCE©, is now in play, created by a team including David Niemi, Julia Phelan, Noelle Griffin, Joan Herman, K. C. Choi, Terry Vendlinski, Keith Howard, and me. Based on 40 earlier studies, POWERSOURCE© is funded by an Institute of Education Sciences (IES) award for our center. It is one of CRESST's major experimental projects. Set in middle school pre-algebra, it consists of multiple interim assessments of problem solving and explanation. POWERSOURCE© development begins with analyzing cognitive demands, then representing the relationship of these intellectual skills in a content ontology. The content includes "big mathematics ideas": the power principles, which we construe as schema, to be applied flexibly across topic and problem type. The ontology controls sampling and sequence and also serves as a learning performance aid to help teachers and students map where they are during instruction. Our assessments are based on research by Sweller (1988; Sweller, van Merrienboer, & Paas, 1988; van Merrienboer & Sweller, 2005), Mayer (Quilici & Mayer, 1996), and their colleagues (Atkinson, Derry, Renkl, & Wortham, 2000; Paas & van Merrienboer, 1994) on schema acquisition and worked examples; on explanation studies by Chi (Aleven & Koedinger, 2002; Atkinson, Renkl, & Merrill, 2003; Chi, 2000; Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Conati & Van Lehn, 1999; Siegler, 2002) and the CRESST team (e.g., Baker, 1994); and on transfer research by Bjork (Schmidt & Bjork, 1992) and Holyoak and others (Bassok & Holyoak, 1989; Catrambone, 1998; Catrambone & Holyoak, 1989; Gentner, Loewenstein, & Thompson, 2003; Gick & Holyoak, 1983; Novick, 1988; Novick & Holyoak, 1991).

Assessments are embedded in a kid-friendly narrative theme, and we are experimenting with formats derived from graphic novels. We predict that this longitudinal study will improve teacher content and pedagogical knowledge. We expect equal or superior student performance on state assessments but, because of their schema acquisition, higher performance on transfer tasks drawn from both the Key Stage 3 math tests in England and the PISA examination. This experiment is being replicated in South Korea, at that country's own expense, to generate comparative information about the use of learning research in designing assessments. With POWERSOURCE©, we illustrate assessment balance with a method that transfers to different subjects and student ages.

The International Stage of Examinations

An obvious question about U.S. tests, given their initial impetus, is how closely they resemble tests from abroad, such as those in New Zealand (see New Zealand Ministry of Education, 2007; New Zealand Qualifications Authority, 2007), Hong Kong (see Hong Kong Examinations and Assessment Authority, 2007), or Finland (see Finland Ministry of Education, 2007). The quick answer is, they don't, despite the use of similar test formats in some countries. But response format aside, most national testing differs substantially from the U.S. versions. Many countries make far greater investment in integrating curriculum, professional development, and assessment, thus removing the need for post hoc rationalization of alignment. Extensive care is given to the content of the test questions and to the model answers used to guide marking. (Of course, national testing systems are themselves subject to continuing and sometimes scathing debate.)

If you visit websites of education departments and ministries around the world (e.g., the New Zealand Qualifications Authority, http://www.nzqa.govt.nz/ncea/assessment/resources/visualart/index.html), you will see great variety. Notice the range existing *within* many examination systems. Secondary school students in other countries, whether on workforce or university paths, have many options. They select from a wide range of course-based examinations, in some countries close to 50 choices. Some options look like the familiar Advanced Placement or International Baccalaureate courses; others encourage high school students to pursue interests unconnected to traditional university or work requirements, for instance, in the visual and performing arts or environmental studies.

Before the Future

Before I sketch a way to attain balance for U.S. students and schools, let's return to our national practices in accountability and how we might deal with them. First, the existing tests: I say, let's leave them alone. They are resilient and embedded in our traditions, and changes to them are always temporary (they snap back like rubber bands), with trivial residues at best. To attain better balance, I'd have fewer tests, locating them during skill building in elementary schools and—designed to reflect particular domains—at the ends of courses. Even if their numbers are reduced, we still need to demand more complete studies of validity inferences and, in particular, of the tests' purported multiple purposes. With current accountability systems in the United States, I say, fix the elements of AYP to minimize harm, by considering the raw probabilities of failure related to numbers of groups, by partitioning tests into passable components, and by aggregating results in units larger than a year or two. Growth modeling is an obvious option if it can be established that the tests progress vertically in their complexity. It is appealing to consider new indicators or multiple measures to balance testing, but their integration into accountability numbers must avoid artifacts that put schools inappropriately at risk.

Change

So let us consider how the process of change seems to impinge on balance. What guidance have we been given by a key journalist?

> The scientific revolution that began 300 years ago has accelerated exponentially. It is moving so fast that the spread of knowledge defines our times. (Fareed Zakaria, *Newsweek,* November 28, 2005)
>
> More things will change in more places in the next 10 years than in the previous 100. (Fareed Zakaria, *Newsweek International,* May 29, 2006)

If our accountability system stays locked on producing only common (though needed) achievement, we will continue to widen the chasm between what is sanctioned and what student futures demand. So the matter of balance returns. Can we have balance and deal with the students we have now without wishing they were different? We can't wait until all high school students have developed proficient language skills, attended innovative preschools, or learned number facts from amazing third-grade teachers. So here and now, can we offer students options better suited to a new and changing work environment, to a life in a real-time society with amped-up connectivity? And can we teach them in a way that models flexibility and problem solving? Researchers Baker and O'Neil (2003), Glaser (1977), Perez, Cherniavsky, and Hamilton (2006), and Simon (1997), among others, have listed precise skill sets for a future brimming with choices and expectations, which I have compiled below:

- Adaptive problem solving
- Assessing and responding to risk
- Managing distraction and giving mindful, rotating attention to tasks
- Working alone, with self-management
- Playing changeable roles in real or virtual teams and groups

Posit these as a beginning for a new, 21st-century definition of educational quality. Start in secondary school, where we all have unresolved and growing problems. There, we should rapidly create a system of Qualifications to reflect 21st-century needs, to be available to all students, whatever their status on standards-based tests. My image of a Qualification is validated accomplishment, obtained inside or outside school. A Qualification means simply that, at various levels of challenge, a student has attained a certified, trusted accomplishment. Warmed-over performance assessment? I think not. Each Qualification is not a new test but an integrated experience with performance requirements. It might look like a course, or a collection, or a musical or sports performance. Some Qualifications, such as securing a certification in cardiopulmonary resuscitation or network management, may demand brief, intense involvement. Qualifications would be aligned, with integrated goals, tasks, learning experiences, criteria, and tests. Merit badge metaphors, like one proposed many years ago by Al Shanker (1988), may help convey the idea. It is a truth-in-advertising approach; instead of a set of "scores," a student possesses demonstrated accomplishments.

Qualifications seem to have many benefits. They provide clear venues for high school students to improve skills, apply and adapt knowledge, and acquire new learning. They support and credit emerging maturity to develop desired personal expertise.

To presage construction, an ontology (Chung, Delacruz, & Bewley, 2006; Chung, Delacruz, Dionne, & Bewley, 2006) or technical network of Qualifications will give structure to their features. Let me illustrate an ontological dimension of Qualifications related to learning. For example, as design criteria, each Qualification could represent the following:

1. Complex problem solving and reasoning
2. Flexibility and adaptive performance
3. Rich knowledge base
4. Schema or principle learning
5. Metacognition and self-monitoring
6. Communication, either explanatory or interactive

The Qualifications chosen by students should span a range of personal goals and, with any luck, help them to develop a passion in at least one area.

What could Qualification topics include? Certainly disciplinary inquiry, exemplified in various ways, perhaps in science fairs, competitions, or projects; then there are the performing or visual arts, community service, health care, interning in business, environmental studies, teaching the young, mentoring peers, and helping the old.

Who will offer these options? Some may be available in high schools, in after-school programs, in community colleges, or in universities. I expect that Web delivery, with its entrepreneurial brainpower and finance, will be a powerful and welcome source of creative Qualifications. Public and private institutions, such as museums, businesses, governments, and advocacy, private, and service organizations, can help, too.

How to begin? Our international colleagues have wonderful Qualifications that we can adapt. If we want to take this reform path, it will be a national challenge of commitment. At the very least, the expectations and rhythms of secondary school will change. Although Qualifications will need to be understood, energized, and partly managed by teachers, expertise outside school will matter, too.

Is this rebalancing of secondary school curricula feasible? Yes, if four benchmarks are met within a reasonable time.

First, the framework(s) for Qualifications must reside in one or more politically credible national or state organizations. Achieve, Inc., has a great start in its American Diploma Project, with more than half of the states participating. Or see North Carolina's Department of Public Instruction for another example. Second, quality control will be necessary. What evidence of quality is needed for a Qualifications adoption? Forms of evidence could include empirical comparisons and expert judgments. Third, acceptance by businesses and universities in their selection processes will be essential if Qualifications are to matter to students. Finally—and this is extraordinarily important—Qualifications must be made a part of state and federal accountability systems and linked to other politically powerful structures. With that function added, we need clear protections against phony performance, plagiarism, and overzealous parental help.

How fast? As a lover of 10-year funded research grants, I'm sorry to say our timelines in this area must be unusually brief. No years of trials and pilots, but concurrent validity studies. By starting with examples and data from other countries or adopting existing Qualifications from our own business and arts sectors, we can have a pool of 40 Qualifications available by 2008, and only as a start. We should use systematic R&D models from other national ventures and invest in parallel development of Qualifications in different areas. The federal government, along with the private sector, can be a financial catalyst. A modified No Child Left Behind Act should phase in Qualifications now.

Unless we begin right now to fix educational balance, with a strong focus on secondary schools, we will have stark questions to answer. Who decided to triage large numbers, and particular groups, of current secondary school students? Why don't we care that many will not find a life of contribution and meaning? A goal of balance in the revitalization of secondary schooling will develop generations of students to be far better prepared than we are.

Research Agenda for Qualifications

The research agenda for the Qualifications recommendation is long and complex, and at this hour our attention is short. Yet opportunities traverse theories of learning and validity; explorations of equity and individual and group differences; definitions of adaptability; and studies of transfer within and across fields, of expertise, of narrative, efficacy, self-management, and motivation. For both students *and* teachers. We can look at second or third chances, mobilizing community and business interests, and of course, costs. We will need to consider many questions if we are serious about exploring a system of Qualifications.

- How large should the initial pool of Qualifications be?
- How many approved Qualifications should a student accumulate, and how much time would be allowed for the completion of requirements?
- Who, other than teachers, can certify performance? How does performance-based, as opposed to test-based, certification occur?
- Can Qualifications be obtained from existing Web enterprises? What is the quality control mechanism? What level of evidence must be available from those presenting Qualification options for consideration?
- How much flexibility should schools or students have in their choices?
- How can businesses help provide Qualifications?
- What can be borrowed, adapted, or copied from other countries, and how well will it work for various learners?
- When should the collection of Qualifications start? In what grade? Should they be subject based?
- How much out-of-school or free within-school time will students have to collect certifications? How will they be supervised?
- Who will pay the costs?
- What teacher development will be needed?
- Can Qualifications be made available on game platforms, PDAs, and phones in addition to computers?
- What are the various possible plans for phasing in Qualifications, and what trade-offs are there in the existing accountability models?
- How will change in Qualification acquisitions be monitored over time? At what point will they affect the formal system?
- Will Qualifications be conjunctive (added to requirements) or compensatory (substituted for some)?
- How will standards be set?
- How will businesses and colleges help define preferences for Qualifications and use them in their induction processes? Must they be in different tracks?
- Will there be preference differences related to student background?
- How much balance among different areas (e.g., the arts, technology), if any, is desirable?
- Should Qualifications be attained by groups of students providing support for one another, by individuals, or in other ways?
- What is the impact of Qualifications on students' ability to learn? To retain knowledge? To transfer learning? On their satisfaction and retention in school? On their self-efficacy and collaborative skills?
- What impact do Qualifications have on how instruction occurs in other subjects?

I'll let my betters specify questions constructing indexes or profiles and the details of time spans and longitudinal analyses. I believe the focus on Qualifications at secondary school will add a new vector to accountability systems, one that values real accomplishments rather than seat time or test scores. Consider the notion of Qualifications as but one example. Don't reject the larger idea if you don't like the particulars of my proposal.

The studies fall into categories related to feasibility, quality assurance, impact on schools, and costs. The persisting problem (pointed out by McDonnell in her 2004 book) is how to muster the political will and the practical smarts to produce, for all groups, the learning required for future success. For now, I want to discuss the nature of requirements for testing, of either conventional or qualifications types, to underscore current limitations that must be corrected.

We should agree that unless we find something tangible, beyond a test score, that engages and fulfills students and teachers, education will continue to shrink and shrivel, and with it, our chances to balance our hopes and aspirations with the comfort of accomplished learners.

So to return to the beginning, can we find ways to get balance into the schools and promote a different quality of learning—within the essential framework of accountability? Can we focus on need and take on the high school substantively? Let's cross over to a new path, built on previous research, and reinvest in learning where accomplishments come with validity and the balance is redressed between what *we* think high school students need and what *they* think they need. The path of Qualifications shifts attention from schoolwork to usable and compelling skills, from school life to real life. Choice becomes personal rather that structural. With pride, our students can assemble their unique collection of Qualifications to show to their families, to adults in university and workforce, and to themselves. With collaboration of both the international community and our own communities, we can enable education to prepare our students far better for the future. As you know, when we are balanced, each of us is able to move comfortably in a range of directions. With balance, and with help from the world community, our students will succeed and fulfill their not-yet-imagined promise.

REFERENCES

Aguirre-Muñoz, Z., & Baker, E. L. (1997). *Improving the equity and validity of assessment-based information systems* (CSE Rep. No. 462). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Aleven, V., & Koedinger, K. (2002) An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science*, *26*, 147–179.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research*, *70*(2), 181–214.

Atkinson, R. K., Renkl, A., & Merrill, M. (2003). Transitioning from studying examples to solving problems: Effects of self-explanation prompts and fading worked out examples. *Journal of Educational Psychology*, *95*(4), 774–783.

Baker, E. L. (1994). Learning-based assessments of history understanding [Special issue]. *Educational Psychologist*, *29*(2), 97–106.

Baker, E. L. (2003). Multiple measures: Toward tiered systems. *Educational Measurement: Issues and Practice*, *22*(2), 13–17.

Baker, E. L., & O'Neil, H. F. (2006). Evaluating Web-based learning environments. In H. F. O'Neil & R. S. Perez (Eds.), *Web-based learning: Theory, research, and practice* (pp. 3–20). Mahwah, NJ: Lawrence Erlbaum.

Baker, E. L., & O'Neil, H. F., Jr. (2003). Technological fluency: Needed skills for the future. In H. F. O'Neil, Jr., & R. Perez (Eds.), *Technology applications in education: A learning view* (pp. 245–265). Mahwah, NJ: Lawrence Erlbaum.

Baker, E. L., & Saloutos, A. (1974). *Evaluating instructional programs*. Washington, DC: U.S. Department of Health, Education, and Welfare.

Bassok, M., & Holyoak, K. J. (1989). Interdomain transfer between isomorphic topics in algebra and physics. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *15*, 153–166.

Baxter, G. P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice*, *17*(3), 37–45.

Baxter, G. P., & Mislevy, R. (2005). *The case for an integrated design framework for assessing science inquiry* (PADI Tech. Rep. No. 5). Menlo Park, CA: SRI International.

Beaton, A. E., Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Smith, T. A., & Kelly, D. L. (1996). *Science achievement in the middle school years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: Boston College, Center for the Study of Testing, Evaluation, and Educational Policy.

Beaton, A. E., Mullis, I., Martin, M., Gonzalez, E., Kelly, D., & Smith, T. (1996). *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: TIMSS International Study Center, Boston College, Center for the Study of Testing, Evaluation, and Educational Policy.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice*, *5*(1), 7–74.

Bruner, J. (1996). *The culture of education*. Cambridge, MA: Harvard University Press.

Catrambone, R. (1998). The subgoal learning model: Creating better examples so that students can solve novel problems. *Journal of Experimental Psychology: General*, *127*(4), 355–376.

Catrambone, R., & Holyoak, K. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *15,* 1147–1156.

Chi, M. T. H. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology* (pp. 161–238). Hillsdale, NJ: Lawrence Erlbaum.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*, 145–182.

Chung, G. K. W. K., Delacruz, G. C., & Bewley, W. L. (2006). *Performance assessment models and tools for complex tasks* (CSE Rep. No. 682). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Chung, G. K. W. K., Delacruz, G. C., Dionne, G. B., & Bewley, W. L. (2006). *Linking assessment and instruction using ontologies* (CSE Rep. No. 693). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Clare, L., & Aschbacher, P. (2001). Exploring the technical quality of using assignments and student work as indicators of classroom practice. *Educational Assessment*, *7*, 39–59.

Commission on Instructionally Supportive Assessment. (2001). *Building tests to support instruction and accountability* (W. J. Popham, Chair). Washington, DC: National Education Association. Retrieved August 7, 2007, from http://www.nea.org/accountability/buildingtests.html

Conati, C., & Van Lehn, K. (1999). Toward computer-based support of meta-cognitive skills: A computational framework to coach self-explanation. *International Journal of Artificial Intelligence in Education*, *11*, 398–415.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives*, *8*(1). Available at http://epaa.asu.edu/epaa/v8n1/

Donovan, M. S., Bransford, J. D., & Pellegrino, J. W. (Eds.). (1999). *How people learn: Bridging research and practice*. Committee on Learning Research and Educational Practice, Commission on Behavioral and Social Sciences and Education, National Research Council. Washington, DC: National Academy Press.

Finland Ministry of Education. (2007). *The Finnish matriculation examination*. Retrieved August 7, 2007, from http://www.ylioppilastutkinto.fi/english.html

Fletcher, J. D., & Tobias, S. (2007, April). *What research has to say (thus far) about designing computer games for learning*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology*, *95*(2), 393–408.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15,* 1–38.

Glaser, R. (1977). *Adaptive education: Individual diversity and learning*. New York: Holt, Rinehart and Winston.

Glaser, R. (1996). Changing the agency for learning: Acquiring expert performance. In K. A. Ericsson (Ed.), *The road to excellence: The acquisition of expert performance in the arts and sciences, sports and games* (pp. 303–311). Mahwah, NJ: Lawrence Erlbaum.

Goldschmidt, P., & Martinez-Fernandez, J.-F. (2004). *The relationship between school quality and the probability of passing standards-based high-stakes performance assessments* (CSE Rep. No. 644). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Hasegawa-Johnson, M., & Alwan, A. (2002). Speech coding: Fundamentals and applications. In J. Proakis (Ed.), *Wiley encyclopedia of telecommunications* (Vol. 5, pp. 2340–2359). New York: John Wiley.

Heritage, H. M., & Niemi, D. (2007). Toward a framework for using student mathematical representations as formative assessments. *Educational Assessment*, *11*, 265–282.

Herl, H. E. (1995). *Construct validation of an approach to modeling cognitive structure of experts' and novices' U.S. history knowledge.* Unpublished doctoral dissertation, University of California, Los Angeles.

Herman, J. L., Osmundson, E., Ayala, C., Schneider, S., & Timms, M. (2006). *The nature and impact of teachers' formative assessment practices* (CSE Tech. Rep. No. 703). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Hong Kong Examinations and Assessment Authority. (2007). *HKALE [Hong Kong Advanced Level Examination]*. Retrieved August 7, 2007, from http://eant01.hkeaa.edu.hk/hkea/topper_hkcee.asp?p_coverdown=hkcee_1.html

Improving America's Schools Act of 1994, Pub. L. No. 103–382, 108 Stat. 3518 (1994).

James, M. (1998). *Using assessment for school improvement*. Oxford, UK: Heinemann.

Klein, S. P., Kuh, G., Chun, M., Hamilton, L., & Shavelson, R. (2005). An approach to measuring cognitive outcomes across higher education institutions. *Research in Higher Education*, *46*(3), 251–276.

Koretz, D. (2002, April). *Using multiple measures to address perverse incentives and score inflation*. Paper presented at Multiple Perspectives on Multiple Measures symposium at annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Lindquist, E. F. (Ed.). (1951). *Educational measurement*. Washington, DC: American Council on Education.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20*(8), 15–21. (ERIC Document Reproduction Service No. EJ 436 999)

Lumsdaine, A. A. (1965). Assessing the effectiveness of instructional programs. In R. Glaser (Ed.), *Teaching machines and programmed learning: Data and directions* (pp. 267–320). Washington, DC: National Education Association of the United States.

Markle, S. M. (1967). Empirical testing of programs. In P. C. Lange (Ed.), *Programmed instruction: Sixty-sixth yearbook of the National Society for the Study of Education* (Part 2, pp. 104–140). Chicago: National Society for the Study of Education.

McDonnell, L. M. (2004). *Politics, persuasion, and educational testing*. Cambridge, MA: Harvard University Press.

McGaw, B. (2007, April). *Internationalizing conceptions of quality*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

McGaw, B. (in press). The role of the OECD in international comparative studies of achievement. *Assessment in Education: Principles, Policy & Practice*.

Mislevy, R. J., Chudowsky, N., Draney, K., Fried, R., Gaffney, T., Haertel, G., et al. (2003). *Design patterns for assessing science inquiry: Principled Assessment Designs for Inquiry* (PADI Tech. Rep. No. 1). Menlo Park, CA: SRI International. Retrieved July 8, 2003, from http://padi.sri.com/downloads/TR1_Design_Patterns.pdf

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (1999). A cognitive task analysis with implications for designing simulation-based performance assessment. *Computers in Human Behavior*, *15*, 335–374.

National Council on Education Standards and Testing. (1992). *Raising standards for American education: A report to Congress, the Secretary of Education, the National Education Goals Panel, and the American people*. Washington, DC: U.S. Government Printing Office.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment. J. Pelligrino, N. Chudowsky, & R. Glaser (Eds.). Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

New American Dream. (n.d.). *Survey confirms that Americans overworked, overspent and rethinking the American dream.* Takoma Park, MD: Author. Retrieved June 11, 2007, from http://www.newdream.org/about/PollRelease.pdf

New Zealand Ministry of Education. (2007). *Schools*. Retrieved August 7, 2007, from http://www.minedu.govt.nz/index.cfm?layout=index&indexid=1072&indexparentid=1000

New Zealand Qualifications Authority. (2007). *New Zealand qualifications*. Retrieved August 7, 2007, from http://www.nzqa.govt.nz/

Niemi, D. (1996). Assessing conceptual understanding in mathematics: Representations, problem solutions, justifications, and explanations. *Journal of Educational Research*, *89*, 351–363.

Niemi, D., Wang, J., Steinberg, D. H., Baker, E. L., & Wang, H. (in press). Instructional sensitivity of a complex language arts performance assessment. *Educational Assessment*, *12*.

Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 510–520.

Novick, L. R., & Holyoak, K. J. (1991). Mathematical problem solving by analogy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 398–415.

Nyquist, J. B. (2003). *The benefits of reconstructing feedback as a larger system of formative assessment: A meta-analysis*. Unpublished master's thesis, Vanderbilt University, Nashville, TN.

Organisation for Economic Co-operation and Development. (2007). *Programme for international student assessment (PISA)*. Retrieved August 7, 2007, from http://www.oecd.org/topic/0,3355,en_2649_35845621_1_1_1_1_1,00.html

Paas, F., & van Merrienboer, J. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology*, *86*, 122–133.

Perez, R. S., Cherniavsky, J., & Hamilton, E. R. (2006). Summary and conclusions. In H. F. O'Neil, & R. S. Perez (Eds.), *Web-based learning: Theory, research, and practice* (pp. 393–404). Mahwah, NJ: Lawrence Erlbaum.

Popham, W. J., & Baker, E. L. (1970). *Systematic instruction*. Englewood Cliffs, NJ: Prentice Hall.

Quilici, J. L., & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology*, *88*, 144–161.

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255.

Schleicher, A. (2007, April). *Impact of the Programme for International Student Assessment (PISA) on educational quality*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*, 207–217.

Shanker, A. (1988, November). Reforming the reform movement. *Educational Administration Quarterly*, *24*, 366–373.

Shavelson, R. J., Gao, X., & Baxter, G. (1993). *Sampling variability of performance assessments* (CSE Rep. No. 361). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Shermis, M. D., & Burstein, J. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Hillsdale, NJ: Lawrence Erlbaum.

Siegler, R. S. (2002). Microgenetic studies of self-explanations. In N. Granott & J. Parziale (Eds.), *Microdevelopment: Transition processes in development and learning* (pp. 31–58). New York: Cambridge University Press.

Simon, H. A. (1997). *What is research?* Video presented at the 50th anniversary conference of the American Institutes for Research, Washington, DC.

Skinner, B. F. (1953). *Science and human behavior*. New York: MacMillan.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*, 257–285.

Sweller, J., van Merrienboer, J., & Paas, F. (1988). Cognitive architecture and instructional design. *Journal of Experimental Psychology: General*, *112*, 639–661.

Tyler, R. W. (1949). *Basic principles of curriculum and instruction*. Chicago: University of Chicago Press.

van Merrienboer, J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, *17*(2), 147–177.

Vendlinski, T. P., Niemi, D., & Wang, J. (2005, March 1–5). *Learning assessment by designing assessments*. Paper presented at the Society for Information Technology and Teacher Education (SITE) 16th international conference, Phoenix, AZ.

Vendlinski, T. P., Niemi, D., Wang, J., & Monempour, S. (2005). Improving formative assessment practice with educational information technology. *Third International Conference on Education and Information Systems, Technologies and Applications (EISTA 2005)*.

Vygotsky, L. S. (1962). *Thought and language*. New York: John Wiley.

Weinberger, D. (2007, February). The HBR List: Breakthrough ideas for 2007. *Harvard Business Review*, *85*(2), 21–36, 44–48, 50, 52, 54.

Zakaria, F. (2005, November 28). The earth's learning curve. *Newsweek*. Available at http://www.mywire.com/pubs/Newsweek/2005/11/28/1193333?extID=10037&oliID=229

Zakaria, F. (2006, May 29). Voices. *Newsweek International*, *CXLVII*(22), 28.

AUTHOR

EVA L. BAKER, the 2006–2007 President of AERA, is Distinguished Professor and Director at the University of California, Los Angeles, Center for Research on Evaluation, Standards, and Student Testing (CRESST), 300 Charles E. Young Drive North, Room 301, Los Angeles, CA 90095–1522; *baker@cse.ucla.edu*. Her research interests are the integration of learning research and assessment, including design and empirical validation of principles for developing technologically based measures of complex human performance.