Reflections on Educational Testing: Problems and Opportunities

Edward H. Haertel
Stanford University
2009

Prepared for the Carnegie Corporation of New York-Institute for Advanced Study Commission on Mathematics and Science Education

Carnegie Corporation of New York
437 Madison Avenue
New York, NY 10022

Institute for Advanced Study
Einstein Drive
Princeton, NJ 05840

Assessment is woven into the fabric of educational practice in the United States. Individual assessments help determine the classifications of students as gifted, learning disabled, English Learners, or ADHD. The quizzes, unit tests, and final exams that teachers create or choose help determine the pacing of classroom instruction, instructional grouping, and marks and grades, as well as informing students about expectations for learning and about their success in meeting those expectations. Advanced Placement and International Baccalaureate tests define ambitious curricula for respected high school courses. The SAT and the ACT are central to the sorting and selecting process at the point of college admissions. High school exit examinations are viewed as a form of quality assurance, but also stand as significant barriers to graduation for substantial numbers of students. State testing systems mandated under the No Child Left Behind Act of 2001 (NCLB) define school-level success or failure, and a range of sanctions are imposed if scores repeatedly fall short of targeted levels. Teachers may be required to pass tests of basic skills and/or subject matter knowledge for initial licensure, and over 70,000 accomplished teachers have passed rigorous examinations set by the National Board for Professional Teaching Standards. There is increasing interest in and use of standardized assessments at the post-secondary level.

This brief paper is intended to offer some guidance to the Carnegie-IAS Commission on Mathematics and Science Education as to possible assessment-related initiatives to improve U.S. mathematics and science learning. It is necessarily selective, taking as its focus (1) classroom assessments and (2) assessments for accountability. These forms of testing have both positive and negative effects on teaching and learning. There is certainly room for improvement in the tests themselves, but before turning to matters of test format or test content, it will be useful to consider some of the ways tests function within the larger education system.

The paper first touches upon some major intended and unintended influences of achievement testing and of test-based accountability on student learning. In the light of that discussion, it then offers some specific near-term and longer-range recommendations.

Test Uses, Supporting Rationales, and Unintended Consequences

It is generally understood that “validity” inheres not in a test itself, but in a use or interpretation of test scores. A test valid for one purpose may be invalid for another. Test validation can be thought of as evaluation of the interpretive argument for using a specific test in a specific way — a weighing of theoretical rationales and empirical evidence for a particular testing application. Unfortunately, once a test is created it may be appropriated for uses unforeseen when it was built. When this happens, validity must be examined afresh. This section addresses some rationales and unintended consequences pertaining to the validity of achievement testing in general, then turns to issues specific to high-stakes accountability testing.

The Logic of Achievement Testing

Mehan (2008, p. 46) nicely summarizes the traditional argument for achievement tests as a basis for grading, promotion, and college admissions: "In the traditional view of schooling in the United States, [educational opportunity] is defined in meritocratic, individualistic, and competitive terms. Students … are placed in environments where they can achieve through their effort and hard work. They have the opportunity to compete with peers for precious resources. They are judged on the basis of their individual performance on presumably objective measures such as tests. Athletic metaphors abound in this tradition: Students engage in 'competitions' and 'races for success.' … The achievement ideology undergirding the meritocratic thesis defines educational success as a matter of individual effort and hard work. The corollary of this proposition is that academic failure or difficulty stems from a lack of effort and hard work. That is, placement in the lower rungs of the economic hierarchy is the fault of the individual who did not try hard enough."

The relevant points here are that students are supposed to work at learning ("earning" grades), and tests are supposed to reveal what they have accomplished. Tests are fair because they are objective and because each student answers the same questions under the same conditions, alone and unaided. Their content communicates to students what is important to learn. ("Is it going to be on the test?") "Achievement tests" motivate and reward effort by providing students with opportunities to demonstrate their learning "achievements."

Achievement tests also benefit curriculum and instruction. Writing test questions helps keep teachers focused on measurable learning goals. Students' test performances, individually and collectively, give teachers feedback on the effectiveness of their instruction, guiding lesson planning, instructional pacing, and the organization of individualized or small-group instruction. Low-stakes "formative" assessments assist both students and teachers with ongoing monitoring of student learning, enabling timely intervention when understanding falters.

Unintended Consequences

There are, of course, competing accounts of the ways testing functions in our educational system. As Mehan (2008) goes on to discuss, the fairness and objectivity of educational tests become less clear when differences in educational opportunity are considered. If educational success is substantially determined by factors *other than* individual aptitude and effort, then sorting and selecting based on test performance may be regarded as quite unfair. In short, achievement reflects both individual effort and educational opportunity. Educational opportunity, in turn, comprises both within-school and out-of-school factors. Within-school factors, including access to highly qualified teachers and other resources, are unequally distributed. Out-of-school factors, including home and community resources, are also unequal. The simplified logic of a meritocracy in which students compete on an equal basis ignores both in-school and out-of-school differences in opportunity to learn. The simplified logic of school accountability based on test scores ignores out-of-school differences in opportunity to learn, and subsumes (average level of) individual effort together with curriculum and instruction as matters under the school's control. Debates on the coachability of the

SAT or the correlation of SAT scores with family income touch on the same theme. The "myth of the meritocracy" may mask structural inequalities that tests help to perpetuate.

In addition, tests that students must complete alone and unaided, in competition with others, comport well with a view of knowledge as an individual possession, carried inside the heads of learners. Adherence to that view may create a gulf between conceptions of mathematics or science in the classroom versus the contexts of professional practice. Some contemporary theorists instead locate knowledge in the *interaction of individuals with their environments.* "Knowing" in this view is knowing how to participate meaningfully in a range of settings and activities. Along with the view of knowledge as an individual possession, our accustomed testing practices fit comfortably with a "knowledge transmission" model of schooling, in which the teacher and textbook are sources of knowledge and students are its (more or less passive) recipients. (That said, conventional modes of assessment by no means preclude students' active engagement with the subject matter.)

These broad critiques find various expressions. Over fifty years ago, the *Taxonomy of Educational Objectives* (Bloom, et al., 1956) drew attention to the tendency for item writers to focus on low-level skills rather than "higher-order thinking." Frederiksen (1984) noted that a focus on objectivity leads to tests posing "well-structured problems" with a single right answer of a pre-determined form, with clear criteria for distinguishing that answer from incorrect ones, and solution procedures guaranteed to reach the right answer if executed correctly. In contrast, real-world "ill-structured problems" involve ambiguities and tradeoffs, with a range of solutions that may be judged better or worse along different dimensions. Bransford and Schwartz (1999) characterized traditional test taking as "sequestered problem solving" (SPS) in contrast to real-world problem solving with access to a range of resources, including other people. Instead of high test scores, they framed academic success in the language of "preparation for future learning" (PFL). In this view, schooling should equip students to approach new problems and figure out what they would need to learn in order to solve them. This is a sophisticated version of "learning to learn" as a goal of schooling. Critiques from other quarters have called for assessments of workplace skills, including effective group participation and collaborative problem solving, that are largely missing from achievement testing (SPS) as now practiced. Cognitive psychologists and scholars in the learning sciences might add that metacognitive awareness (thinking about thinking, self monitoring during learning and problem solving, strategic formulation of subgoals) are also ignored. In summary, current testing practices may reinforce a flawed, narrow view of the subject matters of science and mathematics as static domains of received knowledge to be memorized, together with procedures for solving routine problems. And, learning to do well on tests may be poor preparation for later participation in "communities of practice" employing mathematics or the sciences.

In addition to distorting students' conceptions of the subject matter, conventional testing practices may reinforce dysfunctional student identities as learners. In a long line of studies, Carol Dweck and her students and collaborators have shown that students may adopt a "mastery orientation" or a "performance orientation" toward school subject

matter. The "mastery orientation" is much preferred. Students with this view find their reward in gaining new knowledge and skills. They tend to believe that intelligence is malleable (we get smarter the more we learn) and that failure is part of learning. They set appropriately challenging goals for themselves and persist in the face of difficulties. Students with the contrasting "performance orientation" find their rewards in the system of social comparisons within the classroom. Their academic identity is defined by how smart they are relative to other students, and their aim is to appear as smart as possible. Classroom life becomes a performance. They tend to view intelligence as fixed and unalterable (some students are just smarter than others), avoid challenges they may be unable to meet, and give up quickly in the face of learning difficulties. Bright students who fall into a performance orientation may do well in early grades, but may fail to persist in the middle grades or high school when the subject matter becomes more challenging. It is certainly not inevitable that achievement testing will promote a performance orientation, but competition and social comparison are closely linked, and great care must be taken in the use of tests and the communication of test results, lest the goals and satisfactions of excelling on the test and surpassing one's peers supplant the goals and satisfactions of learning. (Or, lest the humiliation of doing poorly on the test overshadow any satisfaction that might otherwise be found in learning.)

The Logic of Test-Based Accountability

Test-based accountability systems add another layer of complexity to the interpretive arguments for intended testing applications, with a corresponding potential for additional unintended consequences. Although "high-stakes testing" is often discussed without further elaboration, the term actually encompasses several distinct testing uses. For present purposes, the most important of these are as follows.

First, tests aligned with "content standards" are intended to monitor and enforce adherence to a prescribed curriculum. The theory goes that if curriculum and instruction are aligned with content standards (e.g., a state's curriculum frameworks) and if the high-stakes test is also aligned with the content standards (as required under NCLB), then that which is tested will be that which was supposed to be taught.

Second, in the same vein, "performance assessments" have been used to encourage attention to different kinds of learning outcomes. In the early 1990s, policy makers seized upon performance assessments as a way to encourage higher-order thinking and problem solving in the classroom. The hope was to harness the power of measurement-driven instruction for good rather than for ill with "tests that we would *want* teachers to teach to" (Haertel, 1999, p. 663). Whereas alignment to academic content standards under NCLB is concerned primarily with the content aspect of curriculum, performance assessment as implemented during the early 1990s in the context of large-scale, externally-mandated assessments generally placed more emphasis on the process aspect.

Third, standardized tests inform citizens and elected officials about the effectiveness of the public schools they are supporting. The public reporting of test performance school

-by-school is enough to make a test "high stakes." Under NCLB, public reporting is extended to the level of numerically significant subgroups within schools, so that the performance of students in different racial/ethnic groups and of English Learners, students with disabilities, and students from low-income families are all publicly available. While NCLB imposes sanctions if any of these separate groups score too low, reporting per se is seen as a way to bring public pressure to bear for needed school improvement. The law requires that parents be notified individually, by letter, if a school is found to be "in need of improvement," and after a certain point they must be offered the option of transferring their children out of the school if there is an available place in a school not in need of improvement.

Fourth, related to the point above, various school choice initiatives have sought to bring about school reform by creating markets in which different education providers would offer consumers alternatives. For any such market to function as intended, consumers (parents) must have information about the relative quality of different educational offerings. It would be simplistic to suggest that all consumers should prefer the available school with the highest average test score, but under these choice models, standardized tests do provide one important source of comparative information.

Finally, today's most prominent test-based reform effort, NCLB, prescribes a complex procedure for determining whether schools and districts have met Annual Measurable Objectives (AMOs) for the school as a whole and numerically significant subgroups, or if they have otherwise (e.g., via the law's "safe harbor" provision or via "margin of error" adjustments added in regulations) met the requirements for "Adequate Yearly Progress" (AYP) to avoid the label "in need of improvement." This standards-based accountability model is intended to work in conjunction with standards-aligned assessments, public reporting, and various choice and other provisions to bring about school improvement, so that 100 percent of students will be "proficient" (by the varying standards of the different states) by the year 2014.

Other rationales beyond the scope of this paper apply to tests with stakes for individual students (e.g., high school exit examinations), tests used to evaluate or compare instructional methods or curricula (as with the What Works Clearinghouse), tests for teachers, tests for college admissions, and standardized tests used at the post-secondary level.

Unintended Consequences

Five mechanisms of influence were just described: (1) promote adherence to prescribed curriculum, (2) promote teaching of more complex skills, (3) promote transparency and accountability, (4) provide needed information for efficient educational markets, and (5) hold schools accountable by establishing clear performance standards. Each of the mechanisms just described has potential problems.

Promoting adherence to prescribed curriculum. The primary purpose of comprehensive

academic content standards is to guide curriculum and instruction, not to guide the building of tests. Test specifications must be derived from the content standards, and a lot of learning expectations must be set aside in that process. Thus, "alignment" turns out to mean that everything on the test can be found in (or at least related to) the content standards. It does *not* mean that everything in the content standards can be found on the test. This does not merely mean that some content standards are dropped. The larger problem is that objective test items are often pale reflections of the standards they are "aligned" with. A standard for chemistry might call for students to "select and use appropriate tools and technology" but a corresponding test item might just require selecting "a pH probe" from the list of four options in a multiple-choice question. That is very different from actually selecting (let alone using) the right tools and technology in a real-world situation. As another example, a (released) science item from the 2005 California Standards Test in biology asks, "Which of these organisms would be *most* likely to be found at the top of an energy pyramid? A clams; B sardines; C sharks; D kelp." The corresponding standard is as follows: "_Students know_ at each link in a food web some energy is stored in newly made structures but much energy is dissipated into the environment as heat. This dissipation may be represented in an energy pyramid." Note again that while the item is clearly related to the standard, answering the item correctly requires far less than the standard calls for. The phrase "energy pyramid" appears in the question stem, but the item can be solved without any knowledge of energy storage in newly made structures or of energy dissipation into the environment. The phrase "food chain" would have served as well.

There is certainly evidence that high-stakes tests encourage an instructional focus on the content tested (e.g., see Koretz, 2008). The problem is that this is just a sample from some larger domain that the test was originally intended to represent. If the sampled content receives special emphasis in the classroom, then one can no longer generalize from students' performance on that sampled content to their performance across the larger domain. In the examples given, preparing students to answer questions like the ones cited would be much simpler than teaching them the concepts embodied in the corresponding standards. Comparisons of score trends over time on high-stakes tests versus "audit" (low-stakes) tests administered concurrently typically show larger gains on the high-stakes test — sometimes much larger (Koretz, 2008). Rotating tested standards from year to year or matrix sampling can increase domain coverage somewhat, but cannot remedy the problem that arises if critical elements of a given standard are omitted from all of the items available to assess that standard. (As described later in this paper, matrix sampling can be of much greater benefit for improving domain coverage if it is used to enable inclusion of more complex kinds of assessment tasks, beyond multiple-choice. There may also be technical advantages to matrix sampling of multiple-choice items, e.g., to improve accuracy of year-to-year equating of test forms or to field test new items under realistic administration conditions.)

Promoting teaching of more complex skills. Turning to the second mechanism of influence, performance assessments have been touted as a mechanism for pushing curriculum and instruction away from superficial learning of masses of disconnected facts toward "real-world" reasoning and problem solving. The "alignment" logic is similar, but when many states embraced performance assessments during the early 1990s, the

focus tended to be more on process than on content. The theory was (and is) compelling (see, e.g., Resnick & Resnick, 1992), but for various reasons, implementation fell far short (Baxter & Glaser, 1988; Haertel, 1999; Shavelson, Baxter, & Pine, 1992). Neither the states nor their testing contractors had developed the expertise to build sufficient numbers of technically sound performance assessments within the time and cost constraints imposed. In addition, competing objectives (teacher in-service development, curriculum reform) may have led to compromises with respect to reliability and validity (e.g., see Kirst & Mazzeo, 1996).

These first two mechanisms of influence seek to exploit the reactive character of high-stakes testing. Actors in a social system will change their behavior in response to test-based rewards and sanctions. So, the theory goes, what gets tested gets taught. The next two mechanisms are more traditional, returning to the fundamental purpose of tests, which is simply to provide information.

Promoting transparency and accountability; providing information for efficient markets. Unintended consequences also arise from the uses of tests to provide feedback on how well schools are doing or to inform consumers in educational markets: Users of test information draw conclusions that cannot be supported by the data. There are multiple determinants of educational success, both in-school and out-of-school. Parents paying a premium to live in a neighborhood with high-scoring schools may not care whether students score high because of students' family backgrounds, community resources, student peer culture, or instructional quality. But policy makers trying to evaluate teacher or school quality must work at disentangling these factors.

Complex statistical models have been devised to try to separate "school effects" from factors beyond schools' control. All such models are quite imperfect, but among the most promising are "value-added" models incorporating prior test scores for individual students. While value-added methodologies have proven useful in research to model the effects of education policies or to quantify the factors influencing schooling outcomes, most technical experts are reluctant to endorse their use as a basis for reaching consequential decisions about the effectiveness of individual teachers or of schools (Braun, 2005; Briggs & Wiley, 2008; McCaffrey, et al., 2003). Even with strong data, there is substantial error and variability in teacher or school rankings across classes taught, across subjects and grades, and from year to year. Available achievement data may lack "vertical scales" needed to calculate meaningful year-to-year gain scores. State data systems may not enable reliable tracking of students over two or more years and may not enable reliable linkage of students to specific teachers. These data limitations in turn limit the range of statistical models that can be used. Clearly, further research is called for.

Standards-based school reform. The fifth and final mechanism listed in the previous section was standards-based reform: the use of complex decision rules to determine which schools were meeting annual targets for "Adequate Yearly Progress" for the school as a whole and all numerically significant subgroups, and which were "in need of improvement." Research has documented various unintended consequences that have compro-

mised the effectiveness of NCLB in meeting its goals, and discussion of specifics is deferred to a later section of this paper. The fundamental problems, though, may lie in the nature of the tests used and in the clumsy treatment of achievement-related factors beyond the schools' control. Forthright analysis of out-of-school factors is hampered by the "myth of the meritocracy" described by Mehan (2008) and others. There is an understandable commitment to setting uniform achievement expectations for all students and, by extension, for all schools. Accountability systems implemented at the state level must treat schools in a uniform manner, and cannot incorporate nuanced interpretations of local conditions. This commitment to uniformity (clear, objective rules, consistently enforced), coupled with the politically expedient but wildly unrealistic goal of 100% proficiency by 2014, has resulted in school performance targets under NCLB that are not even remotely close to reasonable. Such targets lead only to discouragement, cynicism, and efforts to game the system, not to sincere and constructive efforts at improving learning.

Recommendations

As stated at the outset, assessment is woven into the fabric of educational practice. In thinking through promising assessment-related initiatives, "better tests" per se may not be the best place to start. Tests do need to be improved, but if tests alone are changed, then new forms of tests will just be assimilated into old schooling patterns and not much will really change. (The rise and fall of performance assessment as a policy tool could be read in this light.) Competition will remain a fact of life for students and for schools. Objective tests with clear right answers will continue to be regarded as more fair and more appropriate than complex, ill-structured assessments that are unreliable and expensive to administer and to score. Whatever tests looks like, if high stakes are attached, then direct efforts to improve test scores will displace other learning goals to some extent. Regardless of the form of any new high-stakes test, scores may be expected to rise rapidly the first few years it is in use, then level off. Patterns of group differences as defined by race/ethnicity, socioeconomic status, language background or disability status are quite likely to be about the same on new tests as on old tests. The search for tests that would erase these differences has gone on for decades, without much success.

It follows that the most successful initiatives are likely to require simultaneous changes in tests and in the rules by which they are used or interpreted. This may sometimes require decoupling the multiple purposes for which some tests are used. In addition, it would be worthwhile to push for technical changes to remedy serious problems with the current federally mandated accountability system. In this section, five initiatives are described:

- Portfolio-Based School Accountability—Incorporate student- or classroom-level math and science portfolios into school accountability systems
- Performance Assessment Component for School Accountability_—Incorporate matrix-sampled school-level performance assessments into school accountability systems
- Classroom Assessment for Learning—Improve the structure of curriculum embed-

ded formative assessments

- Better High-Stakes Tests—Offer guidance on improvement of high-stakes tests in current accountability systems

Better Decision Rules for Evaluating School-Level Assessment Results—Offer guidance on technical improvements to NCLB implementation

Portfolio-Based School Accountability

Over the past sixty years, beginning with mandated evaluations of post-Sputnik NSF-sponsored curriculum projects and continuing with mandated evaluations of Head Start and other Great Society programs, there has been a shift from process-based to outcomes-based evaluation of social programs. In education especially, test scores have offered a seductive metric for quantifying "results," and direct measurement of educational processes has declined. Schools face many compliance requirements, of course, but the core teaching and learning activities in classrooms receive scant attention in educational accountability systems. Measurement of classroom process could be a powerful means for influencing teaching and learning. Also, direct classroom process measures might be less subject to the confounding influences of out-of-school factors. Thus, the Commission might consider an initiative to incorporate modest process measures of mathematics and science teaching, to supplement (not supplant) direct measures of student learning outcomes.

These could take the form of math and/or science "portfolios" at either the student level or (preferably) the classroom level. The portfolios would document the kinds of work students were expected to do. They would include student work samples and might also incorporate teacher logs. At the secondary level, they might be supplemented with school reports of course taking, enabling estimation of the proportion of students participating in the kinds of work documented in different classrooms. The portfolios would be evaluated primarily on the kinds of activities students were engaged in, *not* the quality of student work per se. Portfolio scoring rubrics would credit evidence of productive collaborative work, evidence that students were confronting complex problems that might not have clear answers, and evidence that all students in the classroom were participating meaningfully. Where appropriate, evidence of actual lab work would also be expected.

Implementation would be incremental, with scaffolding to assist teachers unaccustomed to new instructional approaches. A state or district might begin phasing in a portfolio-based classroom process assessment by requiring a single learning unit (of some minimum specified duration), and offering teacher professional development as well as course-specific templates to guide both instruction and portfolio creation. Once in place, this requirement would remain unchanged for several years. If an evaluation indicated that it was well received and was achieving its aims, then the weight of the proc-

ess measure in the accountability formula might be increased and/or additional units of the same kind might be required. Unit content might be more or less tightly constrained. If it were tightly constrained, then corresponding student learning outcomes might be incorporated into accountability tests. But, this would be unnecessary and might actually be counterproductive. (The goal might be seen as freeing up some space in the curriculum for teaching and learning *not* driven by concerns over student test scores.) Evaluation of the multi-year initiative would be designed to determine whether there was any decrement in performance on pre-existing learning outcome measures and whether students had a better appreciation of what mathematicians/scientists do, more interest in these subjects as potential careers, and a more sophisticated understanding of the nature of mathematics and science as disciplines (e.g., shaking students loose from simplistic ideas about "the scientific method"). The fundamental intent of this initiative would be to create a space for teaching toward important learning outcomes that are now entirely absent from school accountability systems. The theory here is that some kinds of learning outcomes might be better promoted by careful attention to *classroom process* than by direct measurement of *student outcomes.*

Performance Assessment Component for School Accountability

It is generally accepted that some important learning outcomes, especially in the sciences, are better assessed using performance assessments than multiple-choice or even constructed-response questions. Tasks that require students to decide what steps to follow in using physical apparatus to solve a problem can engage different kinds of reasoning processes as students interact with the equipment provided to solve the problem posed. Performance assessments that require mathematical modeling may tap into skills that a paper-and-pencil test cannot. Andy diSessa (2000, pp. 32-33) has observed that the formal notation of algebra does not distinguish "among motion (d = rt), converting meters to inches (i = 39.37 x m), defining coordinates of a straight line (y = mx), or a host of other conceptually varied situations" (quoted in Gee, 2008, pp. 87-88). Mapping a concrete problem to a symbolic representation is one of the skills performance assessments could tap.

The benefits of performance assessments will not be realized unless they are carefully designed and validated (cf. Baxter & Glaser, 1998). Validation should include "cognitive labs" in which students are observed while conducting the assessments (possibly using think-aloud protocols) and then debriefed, to confirm that the intended cognitive processes are engaged. The primary goal must be sound assessment, *not* (as has happened in the past) the modeling of potentially useful and engaging classroom activities. The use of performance assessments in externally mandated ("on-demand") testing (versus classroom instruction) poses additional challenges (Haertel, 1999). Such assessments must be self-contained, with essentially all required materials delivered to schools as a package. Administrator training may pose another challenge. In addition, if performance assessments are to be used on an annual basis, the problem of designing alternate forms must be thought through in advance of the first administration. "Design templates" might be used to specify classes of assessment tasks so that forms designed in successive years can be made parallel. The time frame of an externally mandated

assessment also limits the kinds of tasks that can be posed. These challenges are not insurmountable. Experience with science performance assessments in the context of the National Assessment of Educational Progress (NAEP) can offer valuable guidance. The framework for the next NAEP science assessment offers guidance for further improvements. Also, computer-based simulations might be used to engage the reasoning processes called for in performance assessment while avoiding the logistical complexities and scoring costs that actual hands-on performance assessments would entail.

A sensible design, which has been proposed far more often than it has been implemented, calls for a small set of performance assessments (perhaps five to ten), each administered to only a fraction of the students in each school. (This is an application of matrix sampling.) Five stations might be set up in a school lunch room, for example, and a small group of students might work at each station for an entire period. Over the course of a day, six groups would have completed each of five assessments. (It would not be necessary for any individual student to take more than one assessment.) Scores would be reported at the school level only. No individual score reports would be generated. Thus, if small-group work were more appropriate than individual work for some tasks, that could easily be accommodated. Because students would be randomly chosen to complete each task, statistical generalization to the entire student population within the school would be straightforward.

In addition to providing information about critical mathematics and science learning outcomes now largely ignored, high-stakes performance assessments of this kind would also influence schools to adopt more hands-on activities in their ongoing mathematics and science instruction, so that their students were better prepared for the assessments that "mattered."

Classroom Assessment for Learning

By far the largest proportion of the tests students take are routine, low-stakes classroom assessments—the quizzes, exams, graded homework assignments, unit tests, midterms, and finals given by teachers. While these tests comprise the largest share of assessments, they have received the least research attention. There is evidence that improving classroom assessment has huge potential to improve learning (Black & William, 1998). Because classroom assessments are so decentralized, improving these tests and the ways they are used will require a lot of resources and will take a long time. Measurement researchers will need to get smarter about life in classrooms, and preservice teacher preparation will need to become more sophisticated (Stiggins, 2001). Here, just one point of entry is proposed: guidelines for curriculum publishers to help and encourage them to provide better curriculum-embedded assessments. If this were done well (perhaps with some "seal of approval" that textbooks or textbook series could earn), market forces could propel rapid adoption.

Curriculum-embedded assessments are among the varieties of ancillary materials provided by publishers in conjunction with textbooks. Guidelines for better formative as-

sessments might require that publishers provide tests in distinct categories as appropriate, perhaps including (1) problems for use during instruction, (2) problems for routine classroom use to check understanding, (3) problems requiring near and remote transfer, and (4) secure problems for more formal assessment/evaluation. The first of these categories would include problems to be directly taught. The second would include similar, non-secure problems for use following instruction and independent practice. The third category would require application of new learning in different contexts. Teachers would be strongly encouraged *not* to reference these different contexts in the preceding instruction, so that students were required to engage in genuine problem solving and transfer. Finally, the fourth category would include secure assessments (perhaps updated annually and delivered to qualified users via the internet), which would be used for unit tests, midterms, or final examinations. The design of these formative assessments would not begin to approach the complexity or rigor of large-scale standardized test construction, but publishers might be expected to produce simple technical manuals that would at least provide test specifications ("blueprints"), document that the assessments had been piloted, and tout any special features (e.g., test creation informed by design principles derived from the learning sciences, including PFL and not just SPS).

There are some old ideas here as well as new. Good teachers have been using assessments in these ways for a long time. The new ideas are (1) greatly increasing the number and quality of assessments provided by publishers, (2) explicit classification of these assessments according to intended instructional use, (3) explicit guidance for teachers concerning these intended uses, and (4) assessment design informed by contemporary research and theory.

Curriculum-embedded assessments today are notoriously bad. There is much room for improvement. A first step might be for the Commission to call for creation of a guidance document for publishers as well as an accompanying public quality assurance / certification mechanism. Options for a related kind of oversight mechanism were discussed about twenty years ago, in the final report of a project funded by the Ford Foundation and led by Professor George F. Madaus, of Boston College. That report, *From Gatekeeper to Gateway: Transforming Testing in America* (National Commission on Testing and Public Policy, 1990), called for "greater public accountability" around test quality and test use. Their comments on potential mechanisms were as follows:

While we cannot recommend one particular mechanism as the single best way to ensure accountability, several approaches are possible. First, some form of governmental scrutiny might be considered. In the same way that the federal Food and Drug Administration helps to protect the public against unsafe and ineffective drugs, a federal test bureau might help to protect against faulty tests or flawed uses of tests. At the same time, we recognize the limits of government regulation…. A second option, then, is some form of independent quality control, perhaps modeled on the practices of the Consumers' Union or the Underwriters Laboratory, to evaluate the technical quality of tests and the ways they are used.

Promulgation of voluntary standards for textbooks and ancillaries appears to be an easier problem than monitoring tests and test uses. There is already an elaborate state-level textbook review process. The textbook adoption criteria in the big states (primarily California and Texas) have substantial power to drive textbook content nationally. The Carnegie/IAS Commission might sponsor the creation of a guidelines document targeted to state boards and/or state departments of education, with the aim of persuading states to require adherence to these criteria for textbook adoption. Marginal costs beyond the current textbook adoption process would not be large. The process might be corrupted, of course, but it would probably be at least as transparent as any feasible process invented from scratch. If this route proved politically or practically infeasible, an alternative would be to work with professional societies (e.g., the National Council of Teachers of Mathematics, the National Association for Research on Science Teaching) and/or to commission reviews of curriculum-embedded assessments to inform textbook adoption decisions at the school or district level.

Better High-Stakes Tests

This and the following final recommendation describe possible improvements to the States' current, federally mandated school accountability systems. Change is in the air, and a lot of advice has come from different quarters concerning the reauthorization of NCLB. For reasons set forth above, better tests alone will probably have limited impact. That said, tests can be improved, and better tests would be a good idea.

The technical quality of current state-mandated high-stakes tests under NCLB, defined in narrowly psychometric terms, is actually quite good. The NCLB peer review guidance addresses matters of reliability, validity, alignment, testing of special populations, and so forth. Simply pushing harder on these kinds of criteria alone is unlikely to bring much further improvement. Also, the NCLB legislation already provides grants to states for enhanced assessment instruments, using multiple measures from multiple sources, including "development of comprehensive academic assessment instruments, such as performance and technology-based academic assessments." Vague calls for increased investment in alternative forms of assessment, absent specifics and absent changes in the incentives created by current legislation, are unlikely to have much positive effect.

That said, current accountability rules may also work to diminish test validity. Despite calls for multiple measures, the NCLB legislation does not provide any real incentives to include performance or technology-based assessments. Because scores need to be reported at the individual level, matrix sampling can be used only in limited ways. (There is no provision in NCLB for assessments at the school level that do not yield scores reportable to parents.) Requirements that all students in grades 3 through 8 be tested annually and that their scores be reportable at the individual level may make performance assessment prohibitively expensive. If NCLB rules were changed, that could free up resources enabling states to make material improvements in high-stakes tests, especially the incorporation of alternative formats.

It would be useful to promote greater public awareness that the quality assurance offered through mandated "alignment" of tests with "challenging State academic achievement standards" means much less than it appears to. Two sample items from one state's science tests were described above, along with their corresponding content standards. These examples each took about one minute to locate using the World Wide Web. Hundreds or thousands more could be cited. Local discussions, newspaper articles and editorials, academic articles, and independent reviews and critiques to shed light on this problem would be valuable. If the Commission were able to highlight the problem, that could help to arouse greater public concern and awareness for the need to (1) change current testing requirements and (2) move beyond the present, nearly exclusive reliance on multiple choice tests for school accountability.

This is an area where practice is heavily constrained, and change will be difficult. Better tests might sample less content but require deeper engagement with that content and more complex reasoning. This would probably require use of performance assessments and/or technology-based simulations to tap learning objectives not amenable to multiple-choice testing. Such significant changes in assessments would require substantial lead time for planning, followed by phased implementation. Ideally, they would be accompanied by investments in revised instructional materials and in teacher professional development, so that students were well prepared to respond to the new assessments as they became operational. Such testing changes would also afford an ideal occasion for rethinking the accountability decision rules and performance targets intended to raise achievement and reduce achievement disparities among demographic groups. This last topic is addressed in the following section.

Better Decision Rules for Evaluating School-Level Assessment Results

Despite a rhetorical commitment to educational practice guided by "scientifically based research," the NCLB legislation itself set forth detailed accountability design specifications with essentially no empirical evidence. At the heart of the accountability formula are "academic achievement standards," operationalized as cut scores on multiple-choice tests. The primary accountability statistic is the proportion of students scoring at or above the "Proficient" cut score within a school or within a student subgroup in a school. These cut scores are the "standards" of "standards-based reform." The idea of holding all students to the same absolute standard of accomplishment has immediate appeal. Norm-referenced tests doom about half the population to the label "below average," but in principle, all might be "proficient." Also, percent-proficient statistics direct attention to the proportions succeeding, even within low-achieving groups. Reporting of group means may instead reinforce an association of all group members with the group's average score, shifting attention away from within-group variation.

These advantages notwithstanding, the "percent proficient" metric has extremely poor statistical properties (Holland, 2002). Whether an "achievement gap" between two groups of students appears to be growing or shrinking over time is almost entirely an artifact of the cut score chosen (Ho, 2008). Standards-based reporting might nonethe-

less be defended if "proficient" cut scores had some defensible absolute meaning, but they simply do not. Study after study, report after report, evaluation after evaluation have found that judgmental standard setting methods cannot be relied upon to produce meaningful results (Haertel & Lorié, 2004; Linn, 2003). In addition to being statistically flawed, standards-based reporting encourages an allocation of instructional resources away from students far below or far above the standard toward the "bubble kids" judged likely to reach the cut point with additional work (e.g., see Diamond & Cooper, 2007). One consequence of judgmentally based definitions of "proficient" and a mandate to escalate annual measurable objectives to 100 percent "proficient" by 2014 has been the setting of wildly unrealistic performance expectations (Koretz, 2008). Unattainable targets cannot serve as effective spurs to improvement. Moreover, under NCLB, the problem of unrealistic achievement targets is further compounded by a conjunctive decision rule under which a school fails if even one student subgroup falls short of either its testing participation requirement or its achievement target either in reading or in mathematics. Stories about exemplary schools "in need of improvement" have become so commonplace that they are no longer make the news.

Purely statistical considerations would suggest a return to the reporting of means rather than percent proficient or other "percent-above-cut" (PAC) statistics, but the intuitive appeal of standards and of labels like "proficient" is strong. Two realistic proposals for change are as follows. First, increase the salience of *multiple cut points,* not just a single cut point. Second, establish more meaningful cut points by relying on *benchmarking* instead of judgmental standard setting.

Multiple cut points. NCLB requires each state to establish at least three cut points defining at least four performance levels, called Below Basic, Basic, Proficient, and Advanced. However, only the percent at or above the Proficient cut score figures into the accountability formula in the original legislation. Through waivers to various states, the Department of Education has permitted some use of "indexing," whereby schools receive credit for students' attainment of lower cut scores, even if they have not yet reached the Proficient level. This is a step in the right direction, but implementation has been tortuous and labored because the rhetorical commitment to 100 percent "Proficient" by 2014 has remained inviolate. A simpler scheme, crediting improvements across the achievement spectrum, would be easy to devise. An accountability index can be designed, as was done for California's Academic Performance Index, so that more credit is given for score gains by lower-achieving students than by higher-achieving students. Such a scheme creates an incentive to allocate more instructional resources toward students with lower scores.

Benchmarking. The term "benchmarking" refers to the simple idea of using actual performance, from some place and time, as the "benchmark" against which other students or schools are judged. Two distinct methods of benchmarking have been used, over time, in the Trends in International Mathematics and Science Study (TIMSS), for example. In the 1995 and 1999 TIMSS assessments, an international achievement score distribution was constructed across all participating nations, and benchmarks were established at the 25th, 50th, 75th, and 90th percentile points of this distribution. (A National

Academy of Education panel once suggested that the combined score distribution for the four top-performing countries in the world might be used in a similar fashion.) In the context of a single state, one could take a band of schools near the median in terms of demographics, and take, say, the 90th percentile of the student-level score distribution for students in those schools, and call that "proficient."

In 2003, TIMSS moved to a different form of benchmarking, called scale anchoring. Because percentile ranks change over time, TIMSS sought a method for defining cut scores that were independent of any particular group's score distribution. They turned to a scale-anchoring method that had been used earlier for NAEP. In this method, items are positioned along the achievement scale using item response theory (IRT) methods. Then, a series of cut points are chosen arbitrarily. (For TIMSS in 2003, cut points were chosen to approximate the locations of the earlier, percentile-based benchmarks.) It is then possible to identify the particular set of test items that distinguish students at a given cut point from those at lower cut points. Content experts review that set of items and write substantive descriptions of the kinds of knowledge and skills that characterize the cut point. This results in cut points with a defensible, empirically established relationship to concrete descriptions of what students at those levels know and are able to do (cf. Haertel & Lorié, 2004).

An additional problem with the NCLB legislation, also easily remedied, concerns the treatment of certain student subgroups. Subgroup reporting is generally regarded as a positive feature of the legislation. Poor performance within one demographic group can no longer be masked by the high performance of other groups. But the law treats targets and growth expectations for all groups in the same fashion, and for two groups in particular, this does not make sense. The first problematical group is English learners. Unlike race/ethnicity, a student's English proficiency status is itself defined in terms of a degree of educational attainment. When students gain proficiency in English, they are "redesignated" and are no longer part of the English learner group. For this reason, it is not sensible to expect 100 percent of the English learner group to reach proficiency. It is still a good idea to call out the performance of English learners. Ideally, individual students might be tracked over time to assure steady progress toward proficiency. Performance targets based on individual progress of students within the EL group would be more helpful than the current Annual Measurable Objectives defined in terms of percent proficient. The second problematical group is students with disabilities. Problems of definition, classification, and testing accommodations are technically challenging, and some may have no good solutions (Koretz, 2008). Once more, however, one obvious problem with a straightforward solution is the unrealistic legal requirement for 100 percent proficiency by 2014.

Summary

Tests are used in different ways to guide educational decisions and as tools of educational policy. They influence curriculum and instruction, students' self concepts as learners, students' conceptions of school subjects and the associated academic disciplines, and public perceptions of schools and of public education. Discussions of educational testing, especially in the United States, are too often dominated by psychometric

concerns of reliability and standard errors, differential item functioning, form-to-form equivalence, and so forth. Wise testing reforms must be fashioned and undertaken with due attention to the tacit assumptions embodied in our testing practices (e.g., the meritocracy), the messages they convey (e.g., about sequestered problem solving), and the dimensions of student proficiency left unexamined (e.g., cooperative group work to solve ill-structured problems). This paper has suggested some ways of thinking about educational assessment, as well as some specific directions for reforms. Much work remains.

References

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

Briggs, D. C., & Wiley, E. (2008). Causes and effects. In K Ryan & L. Shepard (Eds.), *The future of test-based educational accountability.* New York: Routledge.

Black, P., & William, D. (1998). Assessment and classroom learning. *Educational Assessment: Principles, Policy and Practice,* 5(1), 7-74.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (Eds.). (1956). *Taxonomy of educational objectives. Handbook 1: Cognitive domain.* New York: David McKay Company.

Bransford, J.D. & Schwartz, D.L. (1999). Rethinking transfer: A simple proposal with multiple implications. In A. Iran-Nejad & P.D. Pearson (Eds.), *Review of Research in Education* (vol. 24, pp. 61-101). Washington: American Educational Research Association.

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models.* Princeton, NJ: Educational Testing Service.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfield, F. D., & York, R. L. (1966). *Equality of educational opportunity.* Washington, DC: U. S. Government Printing Office.

Diamond, J. B., & Cooper, K. (2007). The uses of testing data in urban elementary schools: some lessons from Chicago. In P. A. Moss (Ed.), *Evidence and decision making* (106th Yearbook of the National Society for the Study of Education, Part 1, pp. 241-263). Malden, MA: Blackwell Publishing.

diSessa, A. A. (2000). *Changing minds: Computers, learning, and literacy.* Cambridge: MIT Press.

Frederiksen, N. (1984). The real test bias: influences of testing on teaching and learning. *American Psychologist, 33*(3), 193-202.

Gee, J. P. (2008). A sociocultural perspective on opportunity to learn. In P. A. Moss, D. C. Pullin, J. P. Gee, E. H. Haertel, & L. J. Young (Eds.), *Assessment, equity, and opportunity to learn* (pp. 76-108). New York: Cambridge University Press.

Haertel, E. H. (1999). Performance assessment and education reform. *Phi Delta Kappan, 80*, 662-666.

Haertel, E. H., & Lorié, W. A. (2004). Validating standards-based score interpretations. *Measurement: Interdisciplinary Research and Perspectives, 2*(2), 61-103.

Ho, A. D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher, 37*(6), 351-360.

Holland, P. (2002). Two measures of change in the gaps between the CDFs of test-score distributions. *Journal of Educational and Behavioral Statistics, 27,* 3-17.

Kirst, M. W., & Mazzeo, C. (1996). The rise, fall, and rise of state assessment in California: 1993-1996. Phi Delta Kappan, 78__(4), 319-323.

Koretz, D. (2008). *Measuring Up: What Educational Testing Really Tells Us.* Cambridge, MA: Harvard University Press.

Linn, R. L. (2003b). Performance standards: Utility for different uses of assessments. *Education Policy Analysis Archives, 11*(31). Retrieved December 19, 2008 from http://epaa.asu.edu/epaa/v11n31/.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability* (Report No. MG-158-EDU). Santa Monica, CA: The RAND Corporation.

Mehan, H. (2008). A sociological perspective on opportunity to learn and assessment. In P. A. Moss, D. C. Pullin, J. P. Gee, E. H. Haertel, & L. J. Young (Eds.), *Assessment, equity, and opportunity to learn* (pp. 42-75). New York: Cambridge University Press.

Mullis, I. V. S., Erberber, E., & Preuschoff, C. (2008). The TIMSS 2007 International Benchmarks of Student Achievement in Mathematics and Science. In J. F. Olson, M. O. Martin, & I. V. S. Mullis (Eds.), *TIMSS 2007 technical report* (pp. 339-347). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. (Downloaded from http://timss.bc.edu/timss2007/PDF/T07_TR_Chapter13.pdf on January 11, 2009)

National Commission on Testing and Public Policy. (1990). *From Gatekeeper to Gateway: Transforming Testing in America.* Chestnut Hill, MA: National Commission on Testing and Public Policy, Boston College.

Ogbu, J. U. (1992). Understanding cultural diversity and learning. *Educational Researcher, 21*(8), 5-14, 24.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New Tools for Educational Reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: alternative views of aptitude, achievement and instruction* (pp. 37-75). Boston: Kluwer.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992, May). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22-27.

Stiggins, R. J. (2001). The unfulfilled promise of classroom assessment. *Educational Measurement: Issues and Practice, 20*(3), 5-15.