Can Today's Standardized Achievement Tests Yield Instructionally Useful Data?

Challenges, Promises, and the State of the Art

W. James Popham

University of California, Los Angeles

David C. Berliner

Arizona State University, Tempe, Arizona

Neal M. Kingston

University of Kansas, Lawrence

Susan H. Fuhrman and Madhabi Chatterji

Teachers College, Columbia University, New York, New York

Steven M. Ladd

Elk Grove Unified School District, Elk Grove, California

Jeffrey Charbonneau

Zillah High School, Zillah, Washington

Author Note

W. James Popham is Professor Emeritus in the Graduate School of Education and Information Studies at the University of California, Los Angeles.

David C. Berliner is Regents' Professor Emeritus of Education at Arizona State University. He is coauthor of *The Manufactured Crisis*, *Collateral Damage* and, most recently, *50 Myths and Lies That Threaten America's Public Schools*.

Neal M. Kingston is Professor of Psychology and Research in Education, and Director of Achievement and Assessment Institute at the University of Kansas.

Susan H. Fuhrman is the President of Teachers College, Columbia University, the founding Director and Chair of the Management Committee of the Consortium for Policy Research in Education (CPRE), and a past-President of the National Academy of Education.

Steven M. Ladd is the Superintendent of schools at the Elk Grove Unified School District, California.

Jeffrey Charbonneau is a National Board Certified Teacher who teaches science at Zillah High School in the Zillah District School in the state of Washington. He is the 2013 US National Teacher of the Year.

Madhabi Chatterji is Associate Professor of Measurement, Evaluation, and Education, and the founding Director of the Assessment and Evaluation Research Initiative at Teachers College, Columbia University (AERI@TC).

Correspondence concerning this article should be addressed to Madhabi Chatterji, Teachers College, Columbia University. Email: mb1434@tc.columbia.edu

This article was published in 2014 in *Quality Assurance in Education, Special Issue: Assessment, Accountability and Quality, (22)*4, 303–318. http://dx.doi.org/10.1108/QAE-07-2014-0033

Abstract

***Purpose*:** Against a backdrop of high-stakes assessment policies in the USA, this paper explores the challenges, promises, and the "state of the art" with regard to designing standardized achievement tests and educational assessment systems that are instructionally useful. Authors deliberate on the consequences of using inappropriately designed tests, and in particular tests that are insensitive to instruction, for teacher and/or school evaluation purposes.

***Methodology/approach*:** The method used is a "moderated policy discussion." The six invited commentaries represent voices of leading education scholars and measurement experts, juxtaposed against views of a prominent leader and nationally recognized teacher from two American education systems. The discussion is moderated with introductory and concluding remarks from the guest editor, and is excerpted from a recent blog published by *Education Week.* References and author biographies are presented at the end of the article. ***Findings*:** In the education assessment profession, there is a promising movement toward more research and development on standardized assessment systems that are instructionally sensitive and useful for classroom teaching. However, the distinctions among different types of tests vis-à-vis their purposes are often unclear to policymakers, educators and other test users, leading to test misuses. The authors underscore issues related to validity, ethics, and consequences when inappropriately designed tests are used in high-stakes policy contexts, offering recommendations for the design of instructionally sensitive tests and more comprehensive assessment systems that can serve a broader set of educational evaluation needs. As instructionally informative tests are developed and formalized, their psychometric quality and utility in school and teacher evaluation models must also be evaluated. ***Originality/value*:** Featuring perspectives of scholars, measurement experts and educators "on the ground," this article presents an open and balanced

exchange of technical, applied and policy issues surrounding “instructionally sensitive” test design and use, along with other types of assessments needed to create comprehensive educational evaluation systems.

Keywords: assessment for instruction, norm-referenced testing, criterion-referenced testing, standardized tests, teacher evaluation, school evaluation, educational accountability

Paper type: Technical paper

Introduction

Madhabi Chatterji, Guest Editor, Teachers College, Columbia University

Standardized achievement tests—tools that enjoy widespread use in decision-making and research contexts at all levels in education around the world—are typically designed by test developers to have particular properties, and are best-suited for serving particular purposes. For example, tests could be developed to help spread out and rank-order individual test-takers on the tested domains. In applied contexts, this type of an instrument (called a *norm-referenced test, NRT*) is intended to help decision makers compare, separate, and select a few examinees from the overall test score distribution, usually for admission into specialized programs or for job placements. Another widespread use of information from NRTs is for describing the general performance of large groups of test-takers in aggregate form at given points in time. In American school systems, NRTs have been historically used for school evaluation purposes where policymakers are interested in tracking student achievement trends annually (Crocker & Algina, 2006; AERA, APA, & NCME, 1999).

In the same vein, educational tests could also be designed to generate information that facilitates domain-referenced interpretations of learning and mastery. For example, a *criterion-referenced* or *domain-referenced* test (*CRT*) is built deliberately to reveal areas of student strength and weakness in ways that are useful for informing instruction or evaluating the quality of instruction in classroom teaching contexts (Crocker & Algina, 2006; AERA et al., 1999). The methodology and validation procedures for designing CRTs versus NRTs differ on several counts, and the types of validity evidence we need to support interpretations and uses of scores for one purpose versus the other also vary. It stands to reason, then, that a given type of test should not be used loosely or freely for purposes it was not designed (Chatterji, 2013a, 2013b).

In educational practice and policy contexts, the close inter-dependencies between test design and a test's purposes are not always clear to all concerned players. Information on the technical demarcations between NRTs and CRTs, for example, is not always accessible to educators and policymakers when they make decisions on how test-based information will be employed for educational accountability purposes. In such high-stakes policy contexts, test misuse can result in serious and sometimes negative consequences for stakeholders at schools (Chatterji, 2014).

A critical issue in assessment policy contexts in the USA deals with the *lack* of instructional sensitivity of currently-popular NRTs. How this deficiency affects teacher and school evaluations is not clear. Another concern is whether standardized tests today have the properties necessary for evaluating schools and teachers at all, with more suitable tools becoming available in the near future.

To examine these issues more deeply, QAE presents its first "moderated policy discussion," titled: *Can Today's Standardized Achievement Tests Yield Instructionally Useful Data? Challenges, Promises, and the State of the Art.* The dialogue features the thinking of leading scholars in measurement and education policy (W. James Popham, David C. Berliner, Neal M. Kingston, and Susan H. Fuhrman, listed in presentation order) counter-balanced against views of a prominent leader and a nationally recognized teacher from two education systems (Steven M. Ladd and Jeff Charbonneau, also listed in order of presentation). The discussion is excerpted from a recent blog published in *Education Week*, co-facilitated by James Harvey of the National Superintendents Roundtable and myself (see http://blogs.edweek.org/edweek/assessing_the_assessments). In conclusion, I summarize authors' recommendations with a few thoughts of my own.

Correcting a Harmful Misuse of Students' Test Scores

James W. Popham, University of California, Los Angeles

A widespread, yet unsound use of students' standardized achievement test results is currently being countenanced in the educational measurement community in the USA. I refer to the role that student test results play in evaluating the quality of schools and, more recently, the effectiveness of teachers, with tests that lack instructional sensitivity. Depending on the tests being used, this practice is—at best—wrongheaded and—at worst—unfair. The *evaluative* use of students' performance results from most of today's *instructionally insensitive* tests violates the single, most important precept of educational assessment: namely, validity.

To use students' test results to evaluate schools and teachers, it is necessary to validate the interpretations and uses of test scores (Kane, 2013). To satisfy this sacrosanct validity cannon of educational testing, *evidence* must be available that allows test-users to answer two questions:

Q1. How accurate are score-based inferences about test-takers apt to be?

Q2. How appropriate will be the real-world uses to which those inferences will be put, such as evaluating the effectiveness of schools and teachers?

The higher the stakes associated with the use of an educational test's results, the greater should be the scrutiny given to both the accuracy of score-based interpretations and to the appropriate usage of the test's results. Chatterji (2013b, 2013b) and others recently addressed a range of such validity concerns.

Accordingly, if students' performances on educational tests are being used to evaluate the success of *schools*, it becomes imperative that those tests are accompanied by evidence supporting the legitimacy of using test results for accountability purposes. So, for example, if a state's tests are intended to measure students' mastery of official, state-approved curricular aims,

then evidence should accompany those tests indicating that test-based inferences about students' mastery status regarding the state's curricular aims are, indeed, valid (i.e. accurate). Moreover, if the test results are also used to rank the state's schools according to their levels of instructional success, then evidence must also be on hand indicating that the tests can actually differentiate among schools according to their relative effectiveness. Similarly, if students' test scores are to be used in evaluating the quality of *teachers*, then we need to not only have evidence showing that the tests measure what a particular teacher is supposed to be teaching; evidence is also needed indicating that results from the specific test being used can distinguish between well-taught and poorly taught students.

During the past decade or so, attempts have been made to determine if educational tests used for high-stakes accountability purposes measure the right curricular aims—typically by using so-called "alignment" studies in which systematic judgments are made about the degree to which a test's items address the knowledge and skills that students are supposed to learn. Yet, essentially no serious efforts have been made to indicate whether the *tests* being used to evaluate schools or teachers are actually up to that job. And this is a huge omission.

What are instructionally sensitive tests? *Instructional sensitivity* refers to a test's ability to provide results that allow us to tell how well test-takers were taught. Although minor definitional differences exist, most educational measurement specialists who deal with this topic accept a definition along the following lines:

> Instructional sensitivity is the degree to which students' performances on a test accurately reflect the quality of instruction specifically provided to promote students' mastery of what is being assessed (Popham, 2013, p. 64).

This conception of instructional sensitivity has an unabashed focus on the *quality of instruction.* Thus, if the inference to be made from a test's results centers on the effectiveness of instruction provided to students by an individual teacher or by a school's entire faculty, the validity of those inferences about instructional quality from an instructionally *insensitive* test would clearly be suspect.

Because of recent federal incentives, almost every state in the USA has adopted new teacher evaluation programs in which students' test scores must play a prominent role. Regrettably, almost all of the tests currently being used in the USA to evaluate school quality or teacher quality have been created according to traditional testing practices of providing test scores that sort and rank test-takers best.

Tests that rank are not built to be instructionally sensitive. For almost a full century, creators of America's standardized tests have been preoccupied with constructing tests that permit *comparative score interpretations* among test-takers. There is an educational need for such tests, particularly for admitting candidates into fixed-quota settings when there are more applicants than openings. But the need for comparative score interpretations disappears when we use a test's results to evaluate the quality of instruction given by teachers or schools.

As it turns out, many of the test-development procedures that are most effective in creating traditional, comparatively oriented tests are likely to *diminish* a test's instructional sensitivity. For example, suppose a state's teachers have, because of a strong emphasis from state authorities, done a crackerjack instructional job during the last few years in promoting students' mastery of, say, Skill X. Well, during the process of designing a test for making comparisons among student performances, it is quite likely that items on the test measuring the well-taught Skill X will be deleted from the test. This is because too many students will be scoring well on

Skill-X items. As a consequence, those items do not contribute to spreading out students' total-test scores—a necessary property if the test is going to do its comparative-interpretation and ranking job well.

Consequences of using the wrong tests for teacher and school evaluations. It is bad enough when traditional, instructionally *insensitive* tests are used to evaluate the quality of our nation's schools. The insensitivity of those tests surely leads to unsound decisions about schooling, and many students get a far less effective education than they should. That is surely reprehensible. But now, because of the above-mentioned federal incentives, most of our states have installed teacher evaluation programs in which students' test scores play a major role. If the tests being used are instructionally insensitive, then the evaluative judgments made about many teachers will be flat-out wrong. Effective teachers, misjudged, will be forced out of the profession. Ineffective teachers, misjudged, will be regarded as acceptable and, therefore, will continue to supply less than sterling instruction to their students.

And all these adverse consequences flow from at least one patently fixable flaw—the instructional sensitivity of tests. When we evaluate schools or teachers with students' scores on tests whose suitability for those tasks has not been demonstrated, we not only violate validity fundamentals, we also violate fundamentals of fairness.

In November 2013, the First International Conference on Instructional Sensitivity was held at Lawrence, Kansas, directed by Neal Kingston and colleagues, where both judgmental and empirical methods of determining a test's instructional sensitivity were described (see also Kingston, this issue). To be candid, these recently proposed procedures for enhancing a test's instructional sensitivity still need to be tried out and scrutinized with care. But such scrutiny will

never transpire if we do not first recognize the existence of this insidious shortcoming in our current assessment arsenal for evaluating schools and teachers.

Morality, Validity and the Design of Instructionally Sensitive Tests

David C. Berliner, Arizona State University

Moral reasons for using appropriate tests to evaluate teachers and schools. The first reason for caring about how sensitive our standardized tests are to instruction is moral. If the tests we use to judge the effects of instruction on student learning are not sensitive to differences in the instructional skills of teachers, then teachers will be seen as less powerful than they might actually be in affecting student achievement. This would not be fair. Thus, *instructionally insensitive* tests give rise to concerns about fairness in using these tests for evaluating teachers, a moral issue.

Additionally, we need to be concerned about whether the scores obtained on *instructionally insensitive* tests have been misinterpreted with adverse consequences when used, for example, to judge a teacher's performance with the possibility of the teacher being fired or rewarded. If that is the case, then we move from the moral issue of fairness in trying to assess the contributions of teachers to improving student achievement, to the psychometric issue of test validity: What inferences can we make about teachers from the scores students get on a typical, norm-referenced standardized test?

Validity reasons for using appropriate tests to evaluate teachers and schools. What does a change in a student's test score over the course of a year actually mean? To whom or to what do we attribute the changes that occur? If the standardized tests we use are not sensitive to instruction by teachers, but still show growth in achievement over a year, the likely causes of such growth will be attributed to other influences on our nations' students. These could be school

factors *other* than teachers—such as, say, the qualities of the peer group, or the textbook, or the principal's leadership. Or such changes might be attributed to *outside-of-school factors*, such as parental involvement in schooling and homework, income, and social class of the neighborhood in which the child lives and so forth.

Currently, all the evidence we have shows that teachers are not particularly powerful sources of influence on aggregate measures of student achievement, such as mean scores of classrooms on standardized tests. Certainly, teachers do, *occasionally and to some extent,* affect the test scores of everyone in a class (Pedersen et al., 1978; Barone, 2001). And teachers can make a school or a district look like a great success based on average student test scores (Casanova, 2010; Kirp, 2013). But exceptions do not negate the rule.

Teachers account for only a little variance in students' test scores. Teachers are not powerful forces in accounting for the variance we see in the achievement test scores of students in classrooms, grades, schools, districts, states, and nations. Teachers, it turns out, affect *individuals* a lot more than they affect *aggregate* test scores, for example the means of classrooms, schools or districts.

A consensus is that *outside-of-school factors* account for about 60 per cent of the variance in student test scores, while schools account for about 20 per cent of that variance (Haertel, 2013; Borman & Dowling, 2012; Coleman et al., 1966). Further, about half of the variance accounted for by schools is attributed to teachers. So, on tests that may be insensitive to instruction, teachers appear to account for about 10 per cent of the variance we see in student achievement test scores (American Statistical Association, 2014). Thus, outside-of-school factors appear six times more powerful than teachers in affecting student achievement.

How instructionally sensitive tests might help. What would teacher effects on student achievement test scores be were tests designed differently? We cannot tell yet because we have no information about the sensitivity of the tests currently in use to detect teacher differences in instructional competence. Only with instructionally sensitive tests can we begin to be fair to teachers and make valid inferences about their contributions to student growth.

It might be helpful if teachers judged as excellent were able to screen items for instructional sensitivity during test design. Even better, I think, might be cognitive laboratories, in which teachers judged to be excellent could provide instruction to students on curriculum units appropriate for a grade. The test items showing pre-post gains—or items found to be sensitive to instruction empirically—could be chosen for the tests, while less sensitive items would be rejected.

Would the per cent of variance attributable to teachers be greater if the tests used to judge teachers were more sensitive to instruction? I think so. Would the variance accounted for by teachers be *a lot* greater? I doubt that. But even if the variance accounted for by teachers went up from 10 to 15 per cent, then teacher effects would be estimated to be about 50 per cent greater than in the current systems over one year. Over a period of say 12 years, teachers can be shown to play an influential role more clearly on aggregated student data, while continuing to be a powerful force on their students individually.

Hammering Out Better Testing Programs: A Call for Action

Neal M. Kingston, University of Kansas

> I suppose it is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail. (Maslow, 1966, p. 15)

In the educational assessment profession, we are surrounded by experts and practitioners wielding hammers. Ask the experts to create an assessment to be used for accountability purposes and they will take out their hammer and bang out a test that is highly standardized, using methods with philosophical bases set between 1930 and 1970. Students will need to all answer the same questions (mostly multiple-choice) on the same days (typically in one week at the end of the school year). Test specifications will be tightly controlled so that everyone is tested with the same amount of each type of content. Statistical data will be analyzed using the same biserial (Pearson, 1909) and point-biserial (Richardson & Stalnaker, 1933) correlations, or the more recent item response theory (IRT) parameters (Lord, 1952). Items will be selected to maximize the internal consistency reliability of test scores and, often, sacrifice the validity of inferences based on those scores for the intended purposes. Much care will be spent on statistical issues like test equating, reliability, and comparability. These standardized tools, once thought important to help us reach our goal of better student learning, have become goals in their own right.

Ask a testing company to create an assessment that can be used to inform instructional decisions and they will take out the same hammer and create a test that looks and feels pretty much like the accountability test. It will be developed in the same way using the same statistical analyses. They may use more modern methods like IRT-based adaptive testing (developed circa 1970), but to what end? Perhaps the tests will be a bit shorter or cover a narrower range of content, but there is no evidence that they will support student learning any better than the accountability test. The largest difference is likely to be an increased number of unreliable sub-scores for which there is no evidence of utility. And why should we be surprised? They were created using the same hammer.

Meanwhile, we have tens of thousands of teachers to whom we have provided hammers and to whom we have shown how to hit nails, and perhaps, to whom we have lied by suggesting that that screws are like nails and that they should hammer them down […] But, perhaps I have taken this metaphor far enough.

No one test can serve multiple purposes equally well. Worse yet, in the face of tests embedded in accountability systems, other tests have little power to support instructional improvement. We need to choose the most important goal for a testing program and focus on achieving it. Secondary considerations should get secondary priorities. We need to identify the evidence we need to support the inferences we must make and support the actions we must take. Rather than develop tests that first and foremost support accountability decisions and then, as an afterthought, try to figure out how to extract a few morsels of instructionally useful information from these inappropriate tests, we should develop assessments that are designed to help teachers make good instructional decisions. We then need to figure out how (and if) those tests can be used to reasonably support accountability decisions.

High on the list of lessons learned from high-stakes accountability systems, including No Child Left Behind (NCLB, 2002), is that when a test is used in a high-stakes accountability system, teachers will teach to the test (Stufflebeam, 1997, p. 56). So, it behooves us to have a test worth teaching to (Kifer, 2001, p. 26).

This means designing the test from the ground up based on how students learn (including especially, how *different* students learn *differently*), modeling good instructional activities, and providing information that supports decision-making during instruction. Once we start there, and without allowing ourselves to compromise those desiderata, we can develop and apply appropriate psychometric models to address the secondary goals. We are not talking about tests

driving instruction; we are talking about a design where instruction drives assessment, but with a statistical underpinning so that it does so efficiently and effectively.

Tests need to be instructionally embedded, relevant, and sensitive. To support the goal of better student learning, tests must be:

- instructionally embedded;
- instructionally relevant; and
- instructionally sensitive.

They must be instructionally embedded because otherwise information will be gathered at a point that is too remote for anything to be done with it. Useful feedback must occur while learning is occurring or wrong ideas will be encoded and right ideas will not be reinforced (Hattie & Timperly, 2007). Using an end-of-year summative assessment to try to improve student learning is too late to be useful; it is like house-training a puppy after letting it do its business all over the house for one year.

Testing must be instructionally relevant. Otherwise, when teachers teach to the test they will teach in ways that are either known to be inferior, or do damage to the curriculum by narrowing it (Madaus et al., 2009, pp. 142–150) and to instruction by replacing teaching higher forms of cognition with memorization of facts (Shepard & Cutts-Dougherty, 1991). Instructionally relevant tests should model what we know about good instruction to the extent that master teachers would want to use similar tasks purely as part of ongoing instruction (Kingston & Reidy, 1997, p. 192).

To provide useful feedback, let alone use test scores as a formal part of a teacher or school evaluation systems, tests should be instructionally sensitive (Kosecoff & Klein, 1974; Muthen et al., 1991). An instructionally sensitive test is one where, on average, students do better if they

have received instruction on a topic covered by the test and do worse if they have not received such instruction. While it might seem obvious that instruction on a topic should make a difference on student performance, we have only a few studies on this issue. These studies show that most items are not instructionally sensitive (Kao, 1990; Niemi et al., 2007; Chen & Kingston, 2013). If teaching a topic does not make a statistically significant difference on results based on student achievement tests then why should we believe that better teachers will produce better student performance? And if lower scores do not indicate poorer teaching then why should teachers and schools whose students have lower scores be punished?

Supporting teachers with cognitively based assessment systems. Teachers have a huge amount of work to do. Not all teachers can be equally expert at all aspects of teaching, including the elicitation of feedback they need to optimally instruct their students. Most teachers will benefit from tools designed to support them, especially a structured assessment tool based on sound cognitive science. These tools should support teacher decision-making but not attempt to replace teachers as decision makers.

The way in which we have chosen to over-standardize the testing experience is part of the problem. There have long been better approaches to similar problems in other fields. We are just not developing or using them as part of large-scale educational assessment. The concept of mass customization—the well-structured designing of flexibility into a product to allow it to meet individual needs—needs to be applied to educational assessment. We need to do this to design tests that are instructionally embedded, instructionally relevant and instructionally sensitive. We need to replace our hammer with the right tool and we need to do so now.

A next generation of assessments is on the horizon: assessments that apply the power of statistical modeling in a way that is driven by instruction, assessments that provide pinpoint

diagnoses in support of teacher personalized prescription, and assessments that are sensitive to instruction because they are part of instruction. Cognitively Based Assessment of, for, and as Learning (CBAL, www.ets.org/research/topics/cbal/initiative/) is one research initiative trying to apply some of these principles, but CBAL is an initiative not a product. In fall 2014, the Dynamic Learning Maps Alternate Assessment (DLM, http://dynamiclearningmaps.org/), developed at the University of Kansas, will be the first operational assessment system to embody these principles to serve the learning needs of individual students. To facilitate this much needed movement, the Achievement and Assessment Institute at the University of Kansas held a conference on instructional sensitivity in November 2013. It is expected that this was only the first of a series of conferences on instructionally useful large-scale assessment that the institute will hold to encourage greater progress. With these new assessments as models and the conferences as one dissemination mechanism, perhaps soon we will put down our hammers and start focusing on teaching and learning again.

Varied Measures for Varied Purposes

Susan H. Fuhrman, Teachers College, Columbia University

Just about everyone agrees there is too much stress on testing in American schools. Ever since No Child Left Behind in 2001 (NCLB, 2002), schools have been testing every student in English/Language Arts and Mathematics in Grades 3–8 and Grade 10. The same state standardized tests have been used for multiple purposes: to monitor student learning, to hold schools accountable and to evaluate teacher performance. Putting so much weight on these tests has meant that they drive instruction, forcing teachers to attend to the specific content likely to be tested rather than the whole curriculum, and crowding out important subjects like social studies, physical education, and the arts.

New developments in research and practice open up the possibility of varying the measures used for what I consider to be the main purposes of assessment in public education:

- tracking student progress and guiding the improvement of student learning;
- getting information about school performance and progress for incorporation into accountability systems; and
- evaluating teachers.

Building a hybrid system of assessment would alleviate placing undue pressure on a single yearly state test.

Using assessment and other data from students and teachers. Student performance should be regularly tracked by teachers, who are positioned to be the best assessors of student learning, through in-class tests, midterms, finals and grades for homework and class work. Excellent curricular approaches, like the Teachers College Reading and Writing Project (see http://readingandwritingproject.com/), incorporate formative assessment processes. In addition, some states and districts are developing item banks of good test questions that teachers can use in classrooms. Eventually, new learning software and games used by students in some school subjects will track student learning and supplement teacher in-class tests with embedded measures of student progress. New, increasingly ubiquitous information systems compile all these data so that educators can observe the progress of every child. Parents and students can also access versions of these reports through what may be considered a learning management system (Watson & Watson, 2007; National Research Council [NRC], 2014).

Hopefully, additional research and investment will enable us to design better measures that are embedded in the curriculum and help us aggregate and standardize the results of such assessments to measure school, classroom or group progress with validity. Until that day, schools

and districts, and even states, could include some common measures among all the data they collect for the purpose of ensuring that all students are learning essential material. It is hard to see why standardized state testing of each individual student would need to take place in as many grades as presently—if at all—given all this accumulating student-level information (NRC, 2014).

Matrix sampling for school evaluations. For school-level information, we should consider a matrix sampling approach. Assessment experts have long argued that one way for tests to cover the whole curriculum and not narrow in on a few areas that can be covered in a single test session is to give different students different questions. Matrix sampling, used by the National Assessment for Educational Progress (NAEP) in the USA (www.nces.ed.gov/nationsreportcard/about), enables coverage of many more areas of content than giving everyone the exact same test (as we do now, in an approach called standardized, census testing). If we separate the need to track individual students from the need for school-level information, it is possible that we can introduce sampling for the latter purpose and mitigate the narrowing that comes from teachers having to focus on the limited content most state-mandated tests can cover. With some common questions, in a partial matrix sampling approach, some individual-level scores might be available.

California successfully petitioned the federal government to accept a sampled approach in lieu of traditional NCLB testing this spring while it pilots new tests from the Smarter Balanced Assessment Consortium (see www.smarterbalanced.org). A few smaller states have also opted to use sampled field tests in lieu of traditional census testing. Perhaps a policy window has opened.

I hope to collaborate in some research that focuses on how a sampling approach can satisfy policymaker needs for school-level and subgroup information. While the consortia working on

assessments aligned to Common Core State Standards Initiative are designing tests that provide individual scores (see www.corestandards.org), the cost and burden of such assessments over time would fall on the states. If a sampling approach, perhaps at benchmark grades, provides good information on school progress, it might be an attractive alternative for state policymakers.

We also have to figure out what to do about teacher evaluation. Despite growing resistance, many states are using state test measures to indicate how much difference in student learning can be traced to individual teachers. Moving to a matrix sample design for school accountability, as just discussed, will not provide sufficient information to link standardized test scores to individual teachers and is resisted by those who think a single test best captures teacher influence. Fortunately, at least one major research effort has demonstrated that there are valid ways to examine teacher contributions through the work they assign and that their students complete, through student surveys and through observations of practice (Mihaly et al., 2013). Putting more effort into developing efficient systems using such measures seems warranted.

Duct Tape Won't Fix Current Assessments

Steven M. Ladd, Elk Grove Unified School District, California

Beside the hammer mentioned by Kingston (this issue), the other widely used, and multi-faceted tool in most toolboxes is "duct tape." When standardized test results for individual students are taped together—figuratively speaking—as a means of examining student performance and growth, educators in school systems receive a macro-level view. Looking at mean grade-level scores of students is helpful as a way to reveal aggregate student trends. But aggregate data reports do not help teachers know what individual students need. Commentaries by Popham, Berliner, Fuhrman, and Kingston (this issue) note that current standardized tests are misguided as tools for informing instruction. They are right on point.

How students are assessed fundamentally drives the way in which teachers plan and deliver instruction. If we accept the fact that one of the most significant aspects of the new Common Core State Standards (CCSS) reform movement is assessment (see www.corestandards.org), then we must acknowledge the critical need to ensure that assessments add value to the teaching and learning process.

Teachers are now incorporating elements of the CCSS into their lessons, re-designing how they will present content and changing how they will assess their students' academic performance. If they are to continue to differentiate instruction in the classroom based on variable needs of students, then teacher-developed assessments will become even more essential.

It is, therefore, critically important to understand that while there should be, and likely will be, professional development to support teachers in the development of assessments aligned to CCSS, that is only the beginning and not the end of the task of ensuring the success of the Common Core reforms. If the federal and state governments and foundations that have supported the Common Core want to protect their investments, they must understand that more professional learning will be necessary to ensure that the assessments help instruction. Our teachers deserve a lot of support as they learn how to tie assessment to their daily work.

We need new and better assessments to inform instruction. Newly developed assessments must be constructed in a manner that informs instruction. Kingston sheds light on how new assessments may be developed with the use of the Educational Testing Service's (ETS) CBAL as one research initiative. The Dynamic Learning Maps (DLM) Alternate Assessment developed at the University of Kansas represents another exciting new approach that uses the principles of ETS' CBAL project to deliver assessments aimed at informing teaching and learning, and at the level of the individual student.

Educators have a right to expect that Common Core assessments included in professionally produced curriculum materials (such as textbooks, supplemental materials and standardized tests) incorporate design elements such as those of DLM at the University of Kansas and CBAL at ETS. Professional materials may evolve to include supports that would serve both to define what students did (and did not) understand and to identify strategies that teachers could use to re-introduce and fortify student knowledge of specific content. Such an approach, far from replacing a teacher, would work as a tool to strengthen assessment. In the end, the shift in design would add value to learning.

Assessments should be designed afresh to make formative and summative decisions in schools. This is one area where hammers and duct tape will not suffice. There is no benefit to hammering out or taping together standardized tests that have not been designed to guide instruction.

New challenges provide new opportunities for improving assessment and accountability systems. The opportunity for improvement should not be lost by those developing the new reform standards. The same holds true for implementation of CCSS reforms by educators. It should also be true for those developing assessments and corresponding evaluation systems for teachers and schools.

Using Multiple Assessments to Mimic the Classroom

Jeffrey Charbonneau, Zillah High School, Zillah, Washington

As a teacher, I use multiple assessment strategies—ranging from laboratory investigations to group discussions to projects and assignments—to modify my instruction and meet student needs in a timely fashion. But I also give more formal assessments, such as tests and examinations, to assign grades. My examinations serve as culminating confirmations of learning. The

examinations evaluate an aspect of learning that the projects cannot. The projects do the same. The final grade is a combination of all forms of evidence gathered from all my assessments. It is very possible for students to have low examination scores but do well on all other parts of the class and still receive a passing grade. But to receive the best overall grade, students must perform well in all areas.

My classroom assessment system. Imagine for a moment that you are a student in my high school physics class. The class is demanding. The content and challenge high enough that the course counts for college credit, even though it is taught in high school. On the first day of class, you (or any student) learn about the grading scheme. There are three options:

1. No homework. No quizzes. No Labs. In fact, the only grade that counts is the final examination.
2. No tests. Not one. Your grade will be determined by daily homework assignments only.
3. Your grade will be a combination of homework assignments, labs, quizzes, chapter tests, and a final examination.

Which one would you want? Which one would you choose for your child to be evaluated upon?

If we used Option 1, I am fairly certain that I would have parents seeking a meeting with the school principal, if not the school board, for inappropriate grading practices. There would be arguments that the stakes were too high, the stress too great, and the student evaluation unduly influenced by whether their child was having a good or bad day at school on the day of the examination.

If we used Option 2, I am fairly certain that I would lose the agreement I have made with the regional four-year university to offer my course for college credit. With no culminating

examination, no equivalent to the on campus/university experience, how could they ensure that the rigor and expectations of the course were being met? If the course no longer counts for college credit, how many students do you see lining up to take such a challenging class again?

And then there is Option 3, which is my preferred option. Students are allowed to showcase their knowledge through multiple pathways—from writing essays to projects to solving problems. They create a wealth of evidence chronicling their learning over time. The secret to the third option being popular among students, parents, and teachers? Option 3 works. We know that using multiple measures and techniques to evaluate student-learning works. We have known this for a very long time.

So my question is: If we know that multiple measures work best for showing comprehensively how students are growing in the classroom, why do we not mimic the classroom when it comes to using data for evaluating schools and teachers? Fuhrman (this issue) suggests in “Varied Measures for Varied Purposes” that:

> Student performance should be regularly tracked by teachers, who are positioned to be the best assessors of student learning, through in class tests, midterms, finals, and grades for homework and classwork.

I could not agree more!

Using a parallel system for evaluating teachers and schools. From one perspective, the current form of statewide testing in the USA does fill a need. It can be used to help find holes in district-wide curriculum, explore inequities between districts that are not otherwise realized, and help identify schools that have successful programs. This can be a great wealth of useful information. Currently used state tests have the most use at the district and state levels, where sample sizes are large enough to see trends.

However, the issue is that state testing should be *PART* of a larger picture in school-wide assessment and evaluation. Statewide testing does not tell me anything about the quality of a district's music program, after-school programs, anti-bullying strategies, and gives only very limited information on support for students with special needs. The list goes on. As a parent, I want to know so much more about my child's school than simply test scores.

We already know what it takes to create a good evaluation of learning in the classroom. So, let us start there. Let us start by mimicking a good classroom. Let us create a system of school evaluation that looks at the entirety of the student experience. One that uses tests as a small fraction of the whole, and does a much better job at showing what it is really like to be educated in our public schools.

Conclusion: Promising Directions, Daunting Challenges

Madhabi Chatterji, Guest Editor, Teachers College, Columbia University

This article provided an open and balanced exchange of technical, applied and policy issues surrounding "instructionally sensitive" test design and use, featuring perspectives of scholars, measurement experts, and educators "on the ground." We also saw recommendations for other types of assessments needed to create more comprehensive evaluation systems for schools and teachers. There is consensus among authors on the need to prioritize tests that are useful in instructional contexts, particularly in helping teachers and learners.

Some authors called for better ways to design instructionally sensitive standardized tests and assessment systems that could serve improved systems of teacher evaluation (see Popham; Berliner; Kingston, this issue). Others offered more comprehensive visions of school evaluation with multiple kinds of assessments and data sources, highlighting the traditional place held by classroom assessments—created, controlled and managed by teachers to serve diverse student

and classroom needs (see Charbonneau; Fuhrman, this issue). Others endorsed the need for more professional development of school-based educators in assessment (Ladd, this issue).

I concur completely with the validity issues raised by authors, as well as the moral and ethical arguments presented here against the use of current student achievement tests (mostly NRTs) for evaluating teachers and schools (Popham; Berliner; Kingston, this issue). But, would instructionally sensitive tests be better for evaluating schools and teachers in high-stakes contexts? I worry that designing instructionally sensitive tests would only be a first step—albeit an important first step—in our quest to develop better teacher and school evaluation systems.

Changing test design parameters would not automatically produce a more valid teacher or school evaluation system overall. This would particularly be true if high-stakes rewards or sanctions are tied to test scores, and lead again to a "teaching to the test" predicament (i.e. where teachers teach to the new "instructionally sensitive test" instead of the broader curriculum goals so as to get better results). Formal validation studies must be therefore be undertaken to ensure validity in interpretations of results within the larger accountability policy context where schools and teachers are evaluated (Chatterji, 2013a, 2013b).

The movement toward more research and development on instructionally sensitive standardized assessment systems, useful for classroom teaching purposes, is a most promising change. Specific purposes and limitations of different types of tests are often unclear to policymakers, educators and other public test users, leading to various levels of test misuse in applied settings. As better systems of school and teacher evaluation are designed and implemented, the need for educating stakeholders on relevant issues must be recognized and continually addressed by the assessment profession.

References

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

American Statistical Association. (2014). *ASA statement on using value-added models for educational assessment*. Alexandria, VA: Author.

Barone, T. (2001). *Touching eternity: The enduring outcomes of teaching*. New York, NY: Teachers College Press.

Borman, G. D., & Dowling, M. (2010). Schools and inequality: A multilevel analysis of Coleman's equality of educational opportunity data. *Teachers College Record, 112*(5), 1201–1246.

Casanova, U. (2010). *¡Sí se puede! Learning from a school that beats the odds*. New York, NY: Teachers College Press.

Chatterji, M. (Ed.). (2013a). *Validity and test use: An international dialogue on educational assessment, accountability, and equity*. Bingley, England: Emerald Group Publishing.

Chatterji, M. (Ed.). (2013b). When education measures go public: Stakeholder perspectives on how and why validity breaks down [Special issue]. *Teachers College Record, (115)*9.

Chatterji, M. (2014, March 17). Validity, test use, and consequences: Pre-empting a persistent problem. *Education Week*. Retrieved from http://blogs.edweek.org/edweek/assessing_the_assessments/2014/03/validity_test_use_and_consequencespre-empting_a_persistent_problem.html

Chen, J., & Kingston, N. M. (2013, November). *A comparison of empirical and judgmental approaches for detecting instructionally sensitive items*. Paper presented at the Instructional Sensitivity Conference, Lawrence, KS.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D. & York, R. L. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Department of Health, Education and Welfare.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. New York, NY: Wadsworth Publishers.

Haertel, E. H. (2013, March). *Reliability and validity of inferences about teachers based on student test scores*. The 14th William H. Angoff Memorial Lecture presented at The National Press Club, Washington, DC. Princeton, NJ: ETS.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. doi: 10.3102/003465430298487

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. doi: 0.1111/jedm.12000

Kao, C. F. (1990). *An investigation of instructional sensitivity in mathematics achievement test items for U.S. eighth-grade students*. Los Angeles: University of California.

Kifer, E. (2001). *Large-scale assessment: Dimensions, dilemmas, and policy*. Thousand Oaks, CA: Corwin Publishers.

Kingston, N. M., & Reidy, E. (1997). Kentucky's accountability and assessment systems. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* Thousand Oaks, CA: Corwin Publishers.

Kirp, D. L. (2013). *Improbable scholars: The rebirth of a great American school system, and a strategy for America's schools*. New York, NY: Oxford University Press. doi: 10.1093/acprof:oso/9780199987498.001.0001

Kosecoff, J. B., & Klein, S. P. (1974, April). *Instructional sensitivity statistics appropriate for objective-based test item.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Lord, F. M. (1952). *A theory of test scores* (Psychometric Monograph, No. 7). Richmond, VA: Psychometric Corporation.

Madaus, G., Russell, M., & Higgins, J. (2009). *The paradoxes of high stakes testing: How they affect students, their parents, teachers, principals, schools, and society*. Charlotte, NC: Information Age Publishing.

Maslow, A. H. (1966). *The psychology of science*. New York, NY: Joanna Cotler Books.

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). *A composite estimator of effective teaching* (MET Project research paper). Bill and Melinda Gates Foundation. Retrieved from http://k12education.gatesfoundation.org/wp-content/uploads/2015/12/MET_Composite_Estimator_of_Effective_Teaching_Research_Paper.pdf

Muthén, B. O., Kao, C.-F., & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement, 28*(1), 1–22. doi: 10.1111/j.1745-3984.1991.tb00340.x

National Research Council. (2014). *Developing assessments for the next generation science standards*. Washington, DC: The National Academies Press.

Niemi, D., Wang, J., Steinberg, D. H., Baker, E. L., & Wang, H. (2007). Instructional sensitivity of a complex language arts performance assessment. *Educational Assessment, 12*(3/4), 215–237.

No Child Left Behind Act of 2001. Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

Pearson, K. (1909). On a new method of determining a correlation between a measured character of A and a character of B, of which only the percentage of cases wherein B exceeds intensity is recorded for each grade of A. *Biometrika, 7*(1/2), 96–105. doi: http://www.jstor.org/stable/i315989

Pedersen, E., Faucher, T. A., & Eaton, W. (1978). A new perspective on the effects of first-grade teachers on children's subsequent adult status. *Harvard Educational Review, 48*(1), 1–31. doi: http://dx.doi.org/10.17763/haer.48.1.t6612555444420vg

Popham, W. J. (2013). *Evaluating America's teachers: Mission possible*. Thousand Oaks, CA: Corwin Publishers.

Richardson, M. W., & Stalnaker, J. M. (1933). A note on the use of bi-serial r in test research. *Journal of General Psychology, 8*(2), 463–465. doi: http://dx.doi.org/10.1080/00221309.1933.9713200

Shepard, L. A., & Cutts-Dougherty, K. (1991, April). *Effects of high-stakes testing on instruction*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Stufflebeam, D. L. (1997). Oregon teacher work sample methodology: Educational policy review. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* Thousand Oaks, CA: Corwin Publishers.

Watson, W. R., & Watson, S. L. (2007). An argument for clarity: What are learning management systems, what are they not, and what should they become? *TechTrends, 51*(2), 28–34. doi: doi:10.1007/s11528-007-0023-y