Journal of Teacher Education

The Near Impossibility of Testing for Teacher Quality

May/June 2005

Journal Article

David Berliner

Regents' Professor

Arizona State University

THE NEAR IMPOSSIBILITY OF TESTING FOR TEACHER QUALITY

David C. Berliner
Arizona State University

The federal No Child Left Behind Act of 2001 (NCLB) mandated that a highly qualified teacher be in all our nation's classrooms by academic year 2005-2006. To accomplish this laudable goal, each state must define what it means by a *highly qualified* teacher. States are permitted to use teacher licensure tests to demonstrate to the federal government that their teachers are highly qualified, that is, capable, competent, skilled, trained, practiced, and so forth. The theory of action behind the policy is that if America's teachers were of sufficiently high quality, then education would improve.

But all this appears to be political spectacle (Smith, 2004); pure theater with no other purpose than to look like something positive is happening, whereas it is not. The way to recognize spectacle is through analysis of the slogans and policies promulgated by politicians. For example, although the call is to ensure highly qualified teachers, there is no evidence to suggest that teachers, as a group, are not now highly qualified. Perhaps this legislation is designed in part to scare ordinary citizens into thinking that millions of unqualified teachers are in charge of their children, although the 3 million teachers we now have in our schools appear to be overwhelmingly qualified to be teaching. How do we know that America's teachers as a group are competent, skilled, and qualified to teach?

One way we know is from the evidence we collect on how well our children are learning, because the ability to promote learning in children must be part of the criteria applied when we designate a teacher as highly qualified. Evidence of learning is obtained from the National Assessment of Educational Progress (NAEP), from which there is evidence of growth for all racial and ethnic groups among 9-, 13-, and 17-year-olds in all three of the subject areas regularly tested: reading, mathematics, and science. In fact, on the NAEP reading tests, scores for 9-year-old African Americans place them approximately 2 years ahead of where their parents scored when the tests were first introduced (Berliner, 2004)

The quality of our teachers can also be inferred through the performance of American 9-year-old students in reading. On Progress in International Reading Literacy Study (PIRLS) tests, U.S. students performed remarkably well. Despite the Bush administration's criticism of reading instruction in the United States, PIRLS shows White American students performing better than those of the highest performing nation in the world; all U.S. students, combined, placed about third in the world statistically; and even America's minority public school students scored above the international average (Berliner, 2004). These results raise questions about the claim that we lack qualified teachers of reading (for more on how the premise of a failing school system cannot be supported, see Berliner, 2004).

Certainly America's poor and minority students do not do well in most national and international assessments. In particular, it is these students who need access to more qualified teachers in their schools. However, it is hard to make the case that America's teachers in general are not qualified, as implied by the NCLB legislation.

Additional examination of the federal requirement for highly qualified teachers reveals that no means are offered to accomplish the desired goal, suggesting that nothing is

Journal of Teacher Education, Vol. 56, No. 3, May/June 2005 205-213
DOI: 10.1177/0022487105275904

expected to happen. Moreover, the language used to rally the politically faithful is kept purposely ambiguous, with the term *highly qualified* providing no concrete referents for anyone to understand what is so ardently being promoted. Furthermore, demands for highly qualified teachers must take into account the relationship between teacher quality and the pay and status of teachers in our society. Attracting "high-quality" teachers—whatever that might turn out to be—may be more difficult than imagined by legislators, given the economic and social status of teachers in our society.

Political spectacle is also evident when serious problems are ignored and lesser ones addressed. For example, there appears to be a much more pressing need in our nation than highly qualified teachers. That is the need for highly qualified pharmacists and physicians in hospitals. According to the Institute of Medicine, part of the National Academy of Sciences, their errors are killing somewhere between 50,000 and 100,000 patients a year (Kohn, Corrigan, & Donaldson, 2000). More recent data, based on a bigger and more representative sample (HealthGrades, 2004), estimate that an average of almost 200,000 people a year are dying from such errors. This makes physician and pharmacist errors the 6th leading cause of death, costing immeasurable personal grief as well as tens of billions of dollars annually. High-quality teachers are desirable, of course, but were I to mandate ways to ensure quality in service to the public, I would turn my regulatory eye first to physicians and pharmacists. Although unqualified teachers can do damage, unqualified physicians and pharmacists are killing us at rates of about 550 per day!

Nevertheless, the federal law demands that we eventually have highly qualified teachers in every classroom, in every state. That will surely result in 50 different definitions of *quality*, with each definition intertwined with and perhaps inseparable from the hiring needs of states and districts (teacher shortages or surpluses in a state will influence the definition each state chooses). We are also likely to see no mention of university training as a requirement for beginning teachers to promote alternative teacher education programs; accommodate those who believe a bachelor's degree is sufficient to teach; and placate those wanting prospective teachers protected from the "liberal agenda" of university teacher education programs. We will also see the promotion of paper-and-pencil tests to determine if teachers know enough about their profession to be called qualified teachers. The rest of this article focuses on the testing of teachers to assure quality.

DEFINING QUALITY TEACHING

Defining quality in teaching is unusually difficult. Were anyone serious about this issue, they would soon realize that quality is an ineffable concept, as the best-selling book by Pirsig (1974) made clear. Defining *quality* always requires value judgments about which disagreements abound. Studying teaching cross-culturally makes this evident (Alexander, 2000). A high-quality teacher in India does not allow questioning by the students. Students simply listen for hours on end. The opposite is true in many American classes, where students are expected to raise questions during class. Alexander (2000) found that maintaining discipline is not part of any definition of *quality* in Russia or India because there are almost no discipline problems in their schools. But in the organizationally complex world of American and British schools, with individualization of some activities, promotion of collaboration and negotiation, and a concern for students' feelings, there is a greater incidence of behavior problems. Thus, American and British teachers of high quality must have classroom management skills that are unnecessary in Russia or India.

In the United States, we see quality teaching taking on different characteristics in programs such as Success for All for inner-city youngsters, in contrast to the schooling offered advantaged students in middle-class suburbs. Quality in reading and mathematics instruction has been vigorously fought over for decades, as has the nature of quality kindergarten instruction.

Under the best of circumstances, it would be difficult to define a *quality teacher*; under political mandate to do so, it is likely to lead to silly

and costly compliance-oriented actions by each of the states. The discernment of quality, an integral part of the identification of a highly qualified individual, always requires keen insight and good judgment (Fenstermacher & Richardson, 2005). It is unlikely that any federal law can mandate the employment of keen insight and good judgment.

So what should we do? Let us approach the issue by thinking of quality teaching as consisting of two conceptually separate parts—what I once referred to as good teaching and effective teaching (Berliner, 1987). Good teaching occurs when the standards of the field are upheld. If you are a physician or a waiter, good practice includes washing your hands frequently. If you are a teacher, good practice may include greeting students warmly at the classroom door. Good is normative. It is what is expected of people in a position. In contrast, effective teaching is about reaching achievement goals. It is about students learning what they are supposed to in a particular class, grade, or subject. A high-quality teacher shows evidence of both good and effective teaching.

Fenstermacher and Richardson (2005) referred to these two qualities as good and successful teaching, arguing that quality teaching means high marks on both dimensions:

> By "good teaching" we mean that the content taught accords with disciplinary standards of adequacy and completeness, and that the methods employed are age-appropriate, morally defensible, and undertaken with the intention of enhancing the learner's competence with respect to the content. . . . By "successful teaching" we mean that the learner actually acquires, to some reasonable and acceptable level of proficiency, what the teacher is engaged in teaching.

Fenstermacher and Richardson (2005) went on to point out

> that not all instances of good teaching are successful, nor are all instances of successful teaching good teaching. Indeed, considerations of successful teaching took us into the domain of learning, where it became apparent that successful learning (in the context of schooling) requires more than teaching of a certain kind. Learning also requires willingness and effort on the part of the learner, a supportive [school and community] social surround, and opportunity to learn through the provision of time, facilities, and resources. These features of learning add greatly to the probability that teaching will be successful. When teaching is both successful and good, we can speak of quality teaching.

Following Fenstermacher and Richardson's (2005) analysis, we see that good teaching is normative and made up of at least three components: the logical acts of teaching (defining, demonstrating, modeling, explaining, correcting, etc.); the psychological acts of teaching (caring, motivating, encouraging, rewarding, punishing, planning, evaluating, etc.); and the moral acts of teaching (showing honesty, courage, tolerance, compassion, respect, fairness, etc.). When coupled with demonstrations of student learning, we have a start toward a definition of *quality* in teaching.

Highly qualified teachers, then, provide evidence that certain qualities of teaching are frequently present in the everyday experiences of their students. The teacher's competence, proficiency, ability, and talent—the many synonyms for having qualifications—are demonstrated in the logical, psychological, and moral acts of teaching, along with evidence that desirable kinds of learning are taking place.

This is a reasonable start to defining the almost indescribable concept of quality in teaching and a reasonable place to ask the question: Can a paper-and-pencil test of teachers' professional knowledge come close to capturing these dimensions of quality in teaching? (I will not comment here on that part of an assessment of teacher quality that examines subject matter competency. Valid tests of subject matter competency can be produced, or teachers' credentials can be evaluated for making judgments of subject matter competency, although not all states do this.)

TESTING TEACHER QUALITY

It should first be noted that successful teaching (evidence of student learning) is not part of the assessment of beginning teachers. So half of what it means to be highly qualified is ignored at the start of one's career. Moreover, after teachers are more experienced, measurement of their success in promoting learning through "pay for

performance" or "value-added" assessments is so filled with psychometric problems that no current system is acceptable for assessing this dimension of teacher quality. The accountability demands of NCLB appear to make it more likely that teachers will be evaluated by their students' performance, but logical and methodological problems associated with inferring quality from student performance do not go away with legislation. So this half of the criteria by which we might judge teacher quality cannot now be measured satisfactorily.

Let us then look at the other half of quality—judgments derived from assessment of the logical, psychological, and moral dimensions of teaching. We can disregard the moral dimension because that cannot be adequately assessed through paper-and-pencil tests. To reliably assess honesty, courage, tolerance, compassion, respect, fairness, and so forth would require unique abilities for discernment by classroom observers during the course of many days and over long periods of time. The moral dimension of teacher quality is simply too difficult to assess, given the costs and time that would be needed, and probably could not be validly assessed for beginning teachers at all.

The psychological dimension of quality teaching is equally difficult to assess. Here we ask if teachers are exemplary in their demonstrations of caring, motivating, encouraging, rewarding, punishing, planning, evaluating, and so forth. It also needs to be judged by discerning observers during the passage of lengthy periods of time in real classes. Although multiple-choice and constructed response items can be built with these dimensions in mind, there is no evidence that they ever predicted the behavior of teachers in classrooms.

For example, despite many attempts by the Educational Testing Service to demonstrate the validity of the National Teacher Examination in predicting ratings of teaching competency and/or student achievements, no predictive validity could be found (Haney, Madaus, & Kreitzer, 1987). Although the tests are made up of items that look like they are related to quality classroom teaching, there is no evidence that this is so. If we genuinely want a highly qualified teacher in every classroom, we should not confuse a highly qualified taker of tests about teaching with a highly qualified classroom teacher.

The construct we measure with a paper-and-pencil test may be quite different than that which is measured through classroom observations of actual teaching, despite the apparent face validity of the items in the paper-and-pencil test. The high reliability and low cost of these tests matters little if the construct of quality teaching is distorted or unmeasured with such tests.

To assess what we really want will require highly discerning observers who spend their time watching teachers teach. This form of assessment is too costly in time and money and might yield reliability estimates that a state would find difficult to defend were it to use the information in summative rather than formative evaluations. But the alternative, tests that measure the wrong construct or that cannot predict quality in teaching, may be worse. Such tests may be used only to calm the public's fears while serving no genuine purpose. If this is indeed the case, then political spectacle is being substituted for a sincere response to public concerns about teacher quality.

The other dimension of quality, the logical dimension, requires the assessment of teachers' skills at defining, demonstrating, modeling, explaining, correcting, and so forth. As with the psychological dimension, this too is hard to assess outside of the classroom setting. The words of two measurement experts who support using paper-and-pencil tests for many educational purposes are important to ponder. In their review of the kinds of tests we are discussing here, they said, "Passing a multiple-choice test does not ensure that one will be a good teacher—or necessarily even a minimally competent one" (Madaus & Mehrens, 1990, p. 260).

Nonetheless, the behaviors encompassed by the logical and psychological dimensions of teacher quality are currently the basis for the professional knowledge tests of teacher quality. I now turn to examples from two paper-and-pencil tests of teacher quality.

TESTING TEACHERS' PROFESSIONAL KNOWLEDGE

Arizona, like many states, has a test to ensure that new teachers are of high quality. Recently, 95% of all applicants who took this multiple-choice test of teachers' professional knowledge and basic skills passed the test the first time they took it. Almost all those who originally fail the test pass it if they take it a second or third time. Arizona has a perpetual tight supply of teachers, approximately 1.1 teachers available for each teaching job, and actual shortages in mathematics, science, special education, bilingual education, and areas that serve poor children. Under such conditions, too rigorous a test of quality is not possible for it would exacerbate the shortage problem, and that takes precedence over the issue of quality. So, although it is never discussed openly, the passing or cut score on the test of teacher quality appears to be synchronized with the labor market realities of my state. This provides more evidence that the test is political spectacle rather than a sincere attempt to improve teacher quality.

Because the overwhelming majority of candidates pass the test, quality among our newest educators appears to be very high. On the other hand, it is more likely that the cut score is set very low because of the demands of the market. The point on the scale that allows for discrimination between those who pass the test and those who do not is always arbitrary and, thus, negotiable.

The company that profits from this testing program is National Evaluation Systems (NES), which has similar contracts with other states. But with such high pass rates and clear adverse impact on minorities and nonnative language speakers, it is not clear that the tests aid the state of Arizona in finding highly qualified teachers for its children. In fact, given the earlier discussion of what a highly qualified teacher might look like, it seems unlikely that any paper-and-pencil test of professional knowledge, mostly of the multiple-choice variety, could identify quality in teaching. That is the problem we address next.

MULTIPLE-CHOICE TEST ITEMS OF PROFESSIONAL KNOWLEDGE

Items parallel to those on the actual Arizona test are given in the test guide to help teachers prepare for the test (NES, 2003). Here is one such item, with an unusually long set up, for the novice teacher to ponder before the actual question is asked:

> Mr. Rivera's fourth-grade class has started a health unit that emphasizes the obligations of individuals and societies to protect the environment. In this unit, the class has been discussing the town's landfill crisis. One day the students return from lunch commenting on the amount of waste they saw in the cafeteria that day and noting that all the garbage generated by the school is contributing to the landfill problem. One student, Kahlil, remarks, "if they'd feed us stuff we like better, maybe there'd be less to throw out." Other students join in, talking about how wasteful it is to serve food that nobody likes and wondering what might be done about the waste. (NES, 2003, p. 21)

Once the class settles down, the teacher remarks that the students have made some very interesting observations and poses the following questions to the class:

- Is the amount of food you saw wasted today typical?
- Kahlil has suggested that if the school served lunches that students like, there would be less waste. Do you agree or disagree with Kahlil's suggestion, and why?
- What kinds of information could you collect to support your opinions? (NES, 2003, p. 21)

The first thing to be concerned about when such lengthy stimulus materials are part of a test question is whether *construct irrelevant variance* is being introduced. Questions that rely on lengthy stimulus materials require high levels of verbal ability and good short-term memory skills, thus, introducing sources of variance that may be irrelevant to the construct one is hoping to measure. Items like these may correlate very highly with measures of verbal intelligence, suggesting that a construct that is not of primary interest is being measured by this test of teacher quality and suggesting also that this test might have an adverse impact on nonnative language speakers. Which it does.

After the description of the classroom situation, a multiple-choice question is provided: "The primary role Mr. Rivera has taken in the instructional process so far has been to . . . :" (NES, 2003, p. 21). Four choices are then offered as possible answers, only one of which is correct. Choice A suggests that Mr. Rivera wants to "encourage students to generate questions about issues that are meaningful to them" (NES, 2003, p. 21). That is a perfectly reasonable answer because Jerome Bruner (1961), among others, has pointed out that the raising of questions is much more important a skill to go out into the world with, than are a set of answers. So I could be persuaded to pick Choice A.

Choice B asks us if Mr. Rivera's role has been to "prompt students to assess their own understanding of instructional content" (NES, 2003, p. 21). Educational textbooks talk of questions such as Mr. Rivera's as prompts for examining self-knowledge and as reviews of content learned. The cognitive processes students engage in as a response to such questions induce learning. So I could also be persuaded to pick Choice B.

Choice C is that Mr. Rivera's role is to "facilitate students' use of higher-order thinking in a real-world context" (NES, 2003, p. 21). Evaluating Kahlil's suggestion requires processes that Benjamin Bloom (1956) and others would identify as the highest in the taxonomy of cognitive processes that can be induced through questions. The other questions asked by this teacher also require cognitive processes that are "higher" than mere use of memory. So I could be persuaded to pick Choice C.

Choice D suggests that the teacher's role is to "provide students with information that can serve as a basis for future learning" (NES, 2003, p. 21). By modeling questions worth asking about the events following lunch, and relating them to the classroom unit about public health, Mr. Rivera provides the children with the information they need to participate as citizens in the debates of their community. This teacher's modeling of high-quality questions derived from informal conversations represents a very Vygotskian perspective on how students learn, helping to turn the social interactions to personal, psychological forms of knowledge that students can use in the future. So I think I could be persuaded to pick Choice D. Of course, only one of these answers is correct. Perhaps others may have an easier time making a choice, but I find it impossible to defend a test of professional knowledge with items and distractors such as these. Many other items on this test were equally confusing to me.

The latest attempt to test teacher quality through paper-and-pencil tests was put together by the American Board for Certification of Teacher Excellence (ABCTE). This is a subgroup of the Education Leaders Council (ELC), founded by conservative school superintendents who hired Lisa Keegan as their director. Keegan, a vocal foe of university training for teachers and a supporter of allowing anyone with a bachelor's degree a chance to teach, was state superintendent in Arizona. She has argued that paper-and-pencil tests are sufficient to distinguish between novice teachers of higher and lower quality, apparently ignoring the fact that the test she helped to build in Arizona allows almost every nonminority candidate to pass. When moving to a national level, Keegan went from concern with mere quality teaching to ensuring *excellence* in teaching, another word that suggests spectacle and not substance.

The ELC (2001) stated that it wants "certification to mean quality" and that "teachers certified by ABCTE will demonstrate academic excellence and help their students achieve" (p. 2). But no research evidence I could find supports their claim that quality, excellence, or student achievement has been evaluated. In fact, one study designed to look at those issues was cancelled by the ELC.

The test designed by the ABCTE is called the Passport to Teaching and was developed with an initial unsolicited grant of US$5 million from the Department of Education. Perhaps such leniency and generosity on the part of the government may have been due to then Secretary of Education Paige's expectations that ABCTE would further his agenda of curtailing university teacher education. Some of the funds were apparently illegally spent (Archer, 2004), and some were used to develop the test. An actual

item from a preliminary version of the test follows:

> To involve parents or guardians in a student's homework, the teacher should understand which of the following is strongly related to student achievement?
> **A**. Income of parents or guardians
> **B**. Amount of time invested by parents or guardians
> **C**. Education level of parents or guardians
> **D**. Employment of parents or guardians.

I am unable to choose one correct answer from this set of alternatives. It may be an easier item for other educators, but I can defend each choice.

Another item from this test gets at teacher-parent relations. The stem is as follows: "At conference time, parents express concern that their child is having difficulty mastering basic addition and subtraction. The teacher should advise the parents to. . . ." Alternative choices are then provided, one of which is "hire a tutor for their child." This strikes me as sensible advice for a good many parents who are middle and upper-middle class and can afford to do so. It is also possible now, under NCLB legislation, for a parent in a school identified as failing to demand a tutor for their child. So this is good advice for some parents, and it also informs other parents of their rights under recent federal laws.

A second alternative for this Passport to Teaching question is "show patience regarding the child's development." This too strikes me as reasonable advice because many parents put a good deal of pressure on their children, forgetting that academic achievement develops at different times for different children. A third alternative is "buy a calculator for the child." Although this is probably not the most desirable alternative to choose, a calculator is perfect for children to use for checking the computations that they do by hand, ensuring that they can do self-checking and thereby learn when they have made a mistake. If we value the development of metacognition, we might choose this response. So this alternative is not clearly wrong. The fourth alternative is "practice math with their child with the use of flash cards daily." I am guessing that this may be the "correct" answer to this question. But some very good math educators, although not denying the use of drill in the learning of math facts, see such activities as secondary to the use of objects, materials, and manipulatives for the learning of math concepts. These educators might not find this an appealing alternative and pick another.

My point in presenting these few examples out of dozens to be found is to suggest that it is a near impossibility to adequately assess quality in teaching through paper-and-pencil tests of professional knowledge. Such tests usually fail to adequately measure the construct of genuine interest, which is quality in teaching and, thus, they fail to identify for the public the promised highly qualified teachers. These tests fail in part because of the complexity of classroom environments and the near impossibility of capturing that reality in paper-and-pencil formats. They also fail because they rely on one correct answer to questions for which many answers are appropriate.

COGNITIVE THEORY AND TEST DEVELOPMENT

These tests also fail to assess what teachers really know because there is no mechanism to follow up answers with teachers, inquiring of them what they mean when they answer test items correctly or incorrectly. It has long been known that one way to improve the validity of test items is through such microanalyses of respondents' thinking (Messick, 1989). But that is not done for these tests. An example of this kind of inquiry was recently reported with students in New Zealand who were questioned about their responses to the science tests that were used in the Third International Mathematics and Science Study (TIMSS; Harlow, 2003). The students' thinking behind the answers given to 24 multiple-choice and free-response items was assessed. Probing of what students understood when they answered a paper-and-pencil test item a certain way revealed that the mean scores obtained from 14 items went up; that is, on 58% of the items, students appeared to know a great deal more than they were able to demonstrate on the test. In addition, the mean scores from 7 items went down. So for 29% of the items, students initially deemed successful were rejudged as not having complete understanding of the

concepts being assessed. The very well-designed TIMSS items were judged to "not necessarily represent what students know" (Harlow, 2003, p. 14):

> An important purpose of assessment is not only to determine [how much and] what people know, but also to assess how, when, and whether they use what they know. . . . Assessment of cognitive structures and reasoning processes generally require more complex tasks that reveal information about thinking patterns, reasoning strategies, and growth in understanding over time. (Pellegrino, Chudowsky, & Glaser, 2001, pp. 62-63)

This is not a plea for performance measures of teacher knowledge, because those too can be invalid. As Messick (1994) pointed out a decade ago, we often confuse task-centered and construct-centered approaches to assessment. The focus in a task-centered approach to learning is the performance of examinees on tasks that seem to be authentic. And this makes sense when the goal is to judge diving, figure skating, and art. Under these circumstances, the task performed is considered to be representative of the performers' skills, and competent judges can discern better and worse performance. But when we evaluate quality in teaching, we are categorically not interested in a single performance. We are interested in the underlying competencies that enable performance on the task at hand and related activities. In this case, a construct-centered approach is needed. Keeping our eyes on the construct probably requires us to use multiple discerning observers on multiple occasions to adequately assess highly qualified teaching.

CONCLUSION

Because "quality" eludes us, it is not surprising that a close examination of some current tests of teacher quality reveals that they are simply inadequate. Under the usual constraints of time and money, the testing of teacher quality may be nearly impossible to do satisfactorily. What is abundantly clear to me is that political spectacle has taken precedence over the public's genuine concerns about quality in teaching. As a result, many teachers are being forced to take tests that do not assess the constructs on which they claim to be based. This demeans and cheapens the teaching profession. It leads, paradoxically, to the possibility that inadequate and inappropriate testing for teacher quality may lower the quality of those who choose to enter the profession. Public education is not well served by bad tests of teacher quality. We should either pursue a genuine program to assess teacher quality or stop the charade.

REFERENCES

Alexander, R. (2000). *Culture and pedagogy: International comparisons in primary education*. Oxford, UK: Basil Blackwell.

Archer, J. (2004, January 21). Leaders group faces shortcomings. *Education Week*. Retrieved September 10, 2004, from http://www.edweek.org/ew/ewstory.cfm?slug=19ELC.h23

Berliner, D. C. (1987). Simple views of effective teaching and a simple theory of classroom instruction. In D. C. Berliner & B. Rosenshine (Eds.), *Talks to teachers* (pp. 93-110). New York: Random House.

Berliner, D. C. (2004). If the underlying premise for No Child Left Behind is false, how can that act solve our problems? In K. Goodman, P. Shannon, Y. Goodman, & R. Rapoport (Eds.), *Saving our schools*. Berkeley, CA: RDR Books.

Bloom, B., Englehart, M., Furst, E., Holl, W., & Krathwohl, D. (1956). *Taxonomy of education objectives: The classification of educational goals. Handbook 1. Cognitive domain*. New York: Longman, Green.

Brewer, J. (1961). Threat of discovery. *Harvard Education Review, 31*, 21-32.

Education Leaders Council. (2001). *Weekly policy update, October 5, 2001*. Retrieved September 10, 2004 from http://www.educationleaders.org/elc/issues/update/ELC_Weekly_Policy_Update_2001-10-05.pdf

Fenstermacher, G. D., & Richardson, V. (2005). On making determinations of quality in teaching. *Teachers College Record, 107*(1), 186-215.

Haney, W., Madaus, G., & Kreitzer, A. (1987). Charms talismanic: Testing teachers for the improvement of American education. In E. Z. Rothkopf (Ed.), *Review of Research in Education* (Vol. 14, pp. 169-238). Washington, DC: American Educational Research Association.

Harlow, A. (2003, July). *Why students answer TIMSS science test items the way they do*. Paper presented at the meeting of the Australian Science Education Research Association, Melbourne, Australia.

HealthGrades. (2004). *In-hospital deaths from medical errors at 195,000 per year, HealthGrades' study finds: Little progress seen since 1999 IOM report on medical errors*. Retrieved September 11, 2004, from http://www.healthgrades.com/aboutus/index.cfm?fuseaction=

mod&modtype=content&modact=Media_PressRelease_Detail&&press_id=135
Kohn, L. T., Corrigan, J. M., & Donaldson, M. S. (Eds.). (2000). *To err is human: Building a safer health system*. Washington, DC: Committee on Quality of Health Care in America, Institute of Medicine, National Academy Press.
Madaus, G., & Mehrens, W. A. (1990). Conventional tests for licensure. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 257-277). Newbury Park, CA: Sage.
Messick, S. (1989). Meaning and values in test validation. The science and ethics of assessment. *Educational Researcher*, *18*, 5-11.
Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13-23.
National Evaluation Systems. (2003). *Arizona educator proficiency assessments: Study guide* (Vol. 1). Amherst, MA: Author. Retrieved March 11, 2005, from http://www.aepa.nesinc.com/PDFs/AZ_vol1_frontmatter.pdf
No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).
Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: Board on Testing and Assessment, National Research Council, National Academy Press.
Pirsig, R. M. (1974). *Zen and the art of motorcycle maintenance*. New York: William Morrow.
Smith, M. L. (2004). *Political spectacle and the fate of American schools*. New York: Routledge Farmer.

David C. Berliner *is Regents' Professor of Education at Arizona State University and can be reached at Berliner@asu.edu. He is a member of the National Academy of Education and a past president of both the American Educational Research Association (AERA) and the Division of Educational Psychology of the American Psychological Association (APA). His research interests are in teaching, teacher education, and educational policy.*