Perils of Standardized Achievement Testing

by Thomas M. Haladyna

Abstract

This article argues that the validity of standardized achievement test-score interpretation and use is problematic; consequently, confidence and trust in such test scores may often be unwarranted. The problem is particularly severe in high-stakes situations. This essay provides a context for understanding standardized achievement testing, then presents and discusses threats to validity, many of which are currently unaddressed. The public and several constituencies support standardized achievement testing. Many educators, however, especially educators in testing, have argued consistently that test-score interpretations and uses are inadequately validated. Standardized achievement test scores provide one valid source of information about student learning if they corroborate other information about student learning. Unfortunately, so many factors undermine the validity of test scores that we should be very careful in the way we interpret and use them.

Richard Phelps (2006) correctly points out many facts about standardized achievement testing: the public wants it, other nations may do it better, and many critics offer no viable alternatives to it. As he concedes, though, standardized achievement tests will never be perfect. Given this state of affairs, what should we do about standardized achievement tests in America? Do we reject the messenger of student achievement? Should we have higher standards for tests when the outcomes of test scores are high stakes? Should we exercise some caution when we use such test scores?

The thesis of this article is that using standardized achievement test scores for high-stakes purposes is perilous because many threats to test validity have gone unaddressed. Part I of the article provides a context

for discussing these perils. Part II identifies and discusses the ways in which such perils threaten validity. Part III then attempts to answer the questions stated in the first paragraph.

PART I: THE CONTEXT

The aims of Part I are explaining basic concepts; describing what students learn; presenting a model for student learning; discussing the role of testing in this model; and finally, considering the role of validity and validation in standardized achievement testing.

Basic Terms

The jargon that permeates education constitutes one difficulty in discussing standardized achievement testing and student learning. Several terms are defined here.

Student *achievement,* distinguished from *intelligence,* is cognitive behavior changed by learning experiences. Intelligence and the cognitive abilities that make up intelligence are less subject to such change. Generally, achievement and intelligence are highly correlated. In most states a set of content standards, which reflect what students should know and can do, defines student achievement.

Test. A test is a measuring instrument. An achievement test, if validated, measures student achievement. Without validation, it is hard to make and justify a claim that an achievement test measures student achievement.

High-stakes test. Some uses of test scores have significant consequences for students, teachers, schools, and school districts. The term "high stakes" designates a test-score use with such consequences, including graduation or promotion; school accountability, such as the federal No Child Left Behind (NCLB) legislation requires; merit pay or continued employment based on test scores in schools or school districts; and intervention in schools or school districts due to chronically low achievement test scores.

A teacher undertakes *assessment* after collecting information about student learning. Assessment, a judgment about student learning, helps plan future instruction. A standardized achievement test score is one piece of information useful for the assessment. We often mistakenly equate the terms "assessment" and "test," but in this context the test is clearly what we use to help us make an assessment.

Accountability. In the past, accountability meant providing information to policymakers to aid their decisions about instructional programs and resources for students. A newer interpretation of accountability—holding people responsible for student learning—is simplistic in its logic: teachers are seldom fully equipped to deal with student learning, and they often have little control over resources needed to help students learn. In test-based accountability, the test score becomes the only basis for assessing a student or a group of students in a classroom, school, school district, or state. One test, however, should never be the sole basis of assessment (AERA 2000); other information should corroborate the test score and better inform us about student learning.

Validity. Validity refers to the adequacy of any test-score interpretation. Let's say Bob, the fastest runner in our high school as a freshman, refused to run hard during a time trial. He finished a 100-yard dash in 15.2 seconds. In other words, his standardized test score was 15.2. Would the coach's assessment of his running speed based on that result be valid?

What Students Learn

A graduate of the University of California at Berkeley has stated that what got him through "Cal" in the 1950s were the three Rs: read, remember, and regurgitate. What he regurgitated was knowledge at the lowest cognitive level: recall. Most of us have experienced that kind of learning. Recall is still part of learning, but understanding and using knowledge are now also widely recognized as dimensions of student learning. Both

can be conceived of as domains that contain many tasks. A test is a sample of the tasks from a domain.

The first domain of student learning consists of *knowledge* and *skills*. Knowledge exists as facts, concepts, principles, or procedures. Knowledge can be recalled, understood, or applied. Spelling and punctuation are examples of writing skills. Most student learning involves knowledge and skills in the domains of reading, writing, mathematics, science, and social studies. Each subject has a large domain of multiple-choice test items representing knowledge and skills at all levels of proficiency. The multiple-choice format, which has proved effective and efficient in measuring knowledge and skills (Downing and Haladyna 2006; Haladyna 2004), should continue in use as a valid measurement of the knowledge and skills domain.

The second and newer domain of student learning draws from cognitive psychology and the persistent belief that learning entails more than simply learning knowledge and skills. *Cognitive ability* is another name for what this domain represents: a mental capacity for achieving an end through complex use of knowledge and skills. Different writers apply different names to this capacity: *developing ability* (Messick 1984); *fluid ability* (Lohman 1993); and *learned ability* (Sternberg 1998). Each cognitive ability is easily recognizable: reading, writing, speaking, listening, mathematical and scientific problem solving, and critical thinking. The acquired knowledge and skills of the first domain are put to use in this second domain. Tests of cognitive ability require that students apply and not simply regurgitate knowledge. Skills are used in unique and complex ways. Standardized achievement tests were not designed to measure cognitive abilities. The advent of state content standards motivated by the federal No Child Left Behind legislation makes clear that future tests will have to address how we use knowledge and skills. Students will be learning knowledge and skills they can apply in their own lives. The performance-test format is best suited for measuring cognitive abilities such as writing.

A Model for Student Learning

Carroll's classic generic model for student learning (1963) holds that educators present their students with clear learning outcomes; aligned instruction; aligned measurement of student learning that provides a valid basis for assessment; and re-instruction where needed to achieve student-learning goals. This model for student learning has not been altered through the years, but the standards for aligned instruction and testing have greatly improved. Not only must today's standardized achievement test be aligned to our content standards, but the alignment of instruction both to content standards and to assessment tests must

also be demonstrated. All students must be provided opportunities for learning and re-learning until they meet desirable performance levels. Both NCLB and AERA (2000) promote this model for student learning and provide guidelines.

The Role of Testing in This Model

Two main uses of test scores are helping teachers improve future instruction by assessing student learning, and providing responsible parties (i.e., both district and state school boards; state and federal legislators; and the public) with information about student learning. Many of those constituencies need such information to formulate policy and allocate resources to schools. In some circumstances, accountability use includes graduation or promotion testing.

Any and all uses of test information must be validated (AERA, APA, and NCME 1999). Without validation by a test-score interpretation, the information culled from a standardized achievement test is dubious.

Validity and Validation

Messick (1984, 1989, 1994, 1995a, 1995b) and Kane (in press) identify validity as the most important goal in testing. Validation is the investigative process that appraises validity for test-score interpretation. *Standards for Educational and Psychological Testing* (AERA, APA, and NCME 1999) is clear about how to validate test-score uses. The process of validation has many steps: first, defining the content being tested; then proposing the interpretation of a test score (the test developer must argue that the test is created to measure this trait validly); and later, gathering evidence to support the claim of validity (Haladyna and Olsen 2006). Validation is a long-term study of a test; the goal is to improve validity.

PART II: PERILS OF STANDARDIZED ACHIEVEMENT TESTING

Haladyna and Downing (2004) show that the many factors threatening validity fall into two main categories: 1) content irrelevance—factors that incorrectly and systematically increase or decrease test scores for some students, and 2) content underrepresentation—flaws in the design of the test that fail to evaluate its full range of content and cognitive behaviors.

Those factors weaken, undermine, or destroy validity. One obvious factor is cheating, which inflates a test score inaccurately. We need to investigate standardized achievement test scores before endorsing and accepting them as unquestionable truth. By studying each factor, we can

reduce or eliminate a threat before using test scores as desired. Eliminating or reducing threats increases score validity. Presenting this information to the public should provide proof that test results can be trusted.

The following discussion focuses on high-stakes uses of standardized achievement test scores. (With low-stakes uses, the need for validity is important, but not to the same extent.) The relevant factors include

- students
- instruction
- test preparation
- cheating
- test development
- test administration
- test scoring
- standard setting

Some of these factors are more serious than others, but all undermine validity to some extent and all have been documented in American testing.

Students

Students themselves are one major source of contamination in testing. Students who cannot read the test material adequately tend to stop taking the test or to mark answers aimlessly. Our inference may be that they have not learned, when a more fundamental problem exists: they cannot read. Nonresponse and omitted responses are more prevalent with English language learners (Haladyna, Osborn Popp, and Weiss 2005). Other factors seldom assessed when students take standardized achievement tests include motivation and fatigue; varying incentives among schools and school districts for performance on the test (Haladyna, Nolen, and Haas 1991); and the motivation level of different students—some students are highly motivated, whereas others seem not to care. Random marks or large blocks of unanswered items are scored as "wrong" when they were in fact omitted from the test by the student. The resulting test scores are inaccurate.

Instruction

NCLB and the new version of accountability have ensured that teachers align instruction to the state's curriculum. The transition from a relaxed, enlightened selection of subject matter to a more rigid system requiring grade-appropriate goals in each subject matter can fulfill the model of student learning presented earlier in this article. Other nations'

unified curricula may make their alignment and testing more uniform and effective than does the United States (Phelps 2006), but most U.S. states have abandoned "states' rights" to follow curricular guidance provided by national organizations.

Opportunity to Learn (OTL). An outgrowth of the accountability movement has been nearly universal support for OTL. AERA and the National Council of Teachers of English, among many organizations, have espoused OTL standards that address various concerns: content standards-based instruction; the diverse ways students learn; highly qualified teachers; best classroom practices; and assessing schools and classroom learning environments. Unfortunately, few state or school district standardized-testing programs assess the conditions of classroom learning that directly address OTL. Despite accountability's premise of adequate instruction for all students, insufficient information about OTL makes it nearly impossible to evaluate student learning for high-stakes purposes such as graduation. In other words, students must first be taught before we use a test to measure what they should have learned. Further, students who are not initially successful should be given repeated opportunities to learn so that they really are not left behind.

Lack of Test Alignment. The introduction of the Stanford Achievement Test in 1923 initiated widespread evaluation of student learning by such tests, whose alignment with an idealized curriculum was regarded as an advantage. However, the trend toward state content standards has caused many states to abandon the publishers' tests in favor of tests aligned to the new standards. The publishers' tests have survived because they provide norm-referenced comparisons for millions of students. Nonetheless, norm-referenced test-score data can be vexing. Haladyna (2004) pointed out that the National Assessment of Educational Progress (NAEP) ranked one state's students fourth from the bottom nationally, whereas they performed well above average in a nationally normed standardized achievement test. One suspicion is that the latter test was the object of considerable coaching; thus the national norms were compromised and state policymakers were misinformed about their students. This tendency, originally called the "Lake Wobegon effect" in reference to humorist Garrison Keillor's mythical Minnesota town where all the children are above average, is still prevalent today.

Test Preparation

In my files is the Stan Fordnine test, a cloned version of the Stanford 9, used for test preparation in one high-scoring school district. When teachers can study a test and identify the content and specific objectives that each test item measures, there is a strong temptation to teach content that will directly affect test performance. This practice is known

pejoratively as "teaching to the test." Advocates of standardized achievement testing and test-based accountability often defend teaching to the test, claiming it is better to learn something valued than not. Teaching to the test, however, is a type of consumer fraud. Any test is only a sample from a large domain of knowledge and skills; mastering a small part of the domain that happens to be tested creates a biased test score. Students, parents, and the public think that more learning has occurred than really did.

Test preparation must be ethical (Haladyna, Nolen, and Haas 1991). Many well-documented test-preparation practices essentially "trick" test scores for the obvious advantage of making it appear that more learning has happened than was the case. The most ethical test preparation is good teaching: using the content standards, aligning instruction to these standards, assessing learning, re-teaching, and re-assessing. The teacher follows all content standards for the grade level and aligns instruction and assessment with that content. Teaching to the test is one possibility when test scores increase in peculiar patterns from one year to another. The tactic may work in test-based accountability, but only in the sense that the public is fooled.

Cheating

Cheating on tests is a pervasive problem in American education as well as throughout the world. Test scores can be badly corrupted by cheating. A Google search on the Web will yield eighty-two pages of hits on test cheating. The National Center for Fair and Open Testing (www.fairtest.org) regularly reports on cheating, and Caveon, a test-security company, provides weekly updates on test-cheating scandals worldwide (www.caveon.com). That company's recent survey of thirty-four states (Sorenson 2006) reveals that test security is a great concern: efforts to increase security are on the rise, detection methods are increasing, and more security measures are being planned. Lost or stolen booklets seem to cause the greatest concern.

Recent reports from New Jersey and Texas (www.philly.com; www.npr.org) document the most pernicious problem: educators cheating to satisfy accountability requirements. Without security audits and studies in states and school districts, cheating may go undetected unless a local newspaper investigates and exposes the problem.

Test Development

Test development is a science with a considerable technology. The *Handbook of Test Development* (Downing and Haladyna 2006) provides many instances of standards that apply to test development. Each standard represents an important source of *validity evidence*. Without this

evidence, the validity of any test-score interpretation is in question. In each step of test development, peril exists.

One of the most important categories of validity evidence is content. Is the test content matched to the intended content? Is the test content trivial or learned by rote—or does the test content ask students to apply knowledge in new situations they may encounter in life? Tests can be highly biased samples of what students need to learn. We need assurance and evidence that the content of the test is exactly that prescribed by the state.

Reliability is another important category of validity evidence. Only high reliability can maintain confidence in making high-stakes, pass-fail decisions. High reliability is difficult to attain when writing tests (Haladyna and Olsen 2006). Continued use of test scores for life-altering decisions not only runs the risk of prompting legal challenges; it also damages the students who unfairly fail. One remedy is to ensure high reliability and a small margin of error.

The quality of test items is another important category of validity evidence. Item development is a costly and lengthy process, an estimated $300 to $1,200 for a single test item, depending on the effort spent developing an item bank. Evidence that quality of test items is a high priority should be assembled and posted on state Web sites with technical documentation establishing item development and validation. Such evidence is in short supply, as a search of such sites will confirm.

In high-stakes testing, the cut score is a point on the test-score scale at which students are classified one way or another. One important cut-score determination is pass-fail. Another is using a test score to classify each student in one of four categories: highly proficient, proficient, approaching proficient, and well-below proficiency. The test-score scale should mean the same from year to year, because cut scores are standardized for pass-fail decisions and those other classifications. The goal is accurately measuring progress in student learning over several years. The science of scaling for comparability is well established for multiple-choice tests, but not for performance tests (Kolen and Brennan 2004). Scaling for performance tests can be particularly troublesome when tracing growth vertically across grade levels (Haladyna and Olsen 2006).

This discussion is not intended to suggest that test development is faulty or contributes to lower validity, but annual documentation must assure the public that the validity evidence of such high-stakes tests supports reliance on test scores. Technical reports and other indicators of validity covering this evidence should be abundant and comprehensive. Searching Web sites of major test developers, however, reveals that such evidence is scarce and usually nonexistent. A review of such reports (Ferrara and DeMauro, in press) finds them lacking in many

respects. In summary, considerable peril lurks at each step of test development. Without documentation, there is no way to determine what has happened.

Test Administration

Altering the administration of high-stakes standardized achievement tests renders any "standardized" achievement test less standardized. According to anecdotal reports, two common practices are altering the administration time of high-stakes tests for the advantage of the students and reading test items to students. Evidence of such problems in test administration (Haladyna and Downing 2004; McCallin 2006), although growing, is hard to come by because few test sponsors, states, and school districts consistently monitor test administration with care. The most decisive strategy to combat the peril would be hiring professional test administrators. Computer-based tests offer hope of standardizing test administration and reducing this category's threat to validity.

Test Scoring

A serious source of invalidity is test scoring. The National Center for Fair and Open Testing and other sources, mainly in the media, have found that scoring errors, not to mention monitors sanitizing answer sheets by reviewing them and cleaning up bad erasures, are not unusual. Such practices are unfair when some schools and districts employ them and others don't.

The scoring of performance tests that require subjective judgments poses many threats to validity, which typically go unacknowledged (Haladyna and Olsen 2006). Results close to the cut score should be rescored when the stakes are very high, as in graduation or promotion.

The most noteworthy scoring error to date occurred with the Scholastic Assessment Test, the highly respected college-admission test of the Educational Testing Service, one of America's foremost test companies (National Center for Fair and Open Testing, May 2006): excessive moisture in answer sheets resulted in misscoring. Whether the problem has occurred in the past is unknown. ETS also had to pay $11.1 million to settle a lawsuit involving 4,100 teachers who received erroneous scores on a teacher-licensing test.

Such incidents are not limited to ETS; the problem of test-scoring errors is widespread. If a test score does not seem representative of a student's true ability, scoring error is a potential culprit. The major point is that using a single, flawed score for a high-stakes purpose, such as college admission or licensing a teacher, has legal consequences. We need to be smarter in quality control and use of test scores when such errors seem more likely than ever.

Standard Setting

As discussed earlier, high-stakes tests often employ a cut score to designate students as passing or failing. Most commonly, a committee of subject-matter experts reviews items and makes judgments that are aggregated to form the recommended cut score. Another committee may accept the recommendation or change it. Do different methods of development produce different cut scores? Is one cut score more valid than another? Which is the most valid? How does one decide? Those are imponderable questions. Cut scores are set and decisions are made based on arbitrary criteria; the labels used to identify students or groups of students (e.g., proficient, basic) are social conventions, not true categories. The difficulty of tests and the position of the cut scores are sure to vary from state to state. Studies on the validity of cut scores are typically not reported. Consequently, we have very little information about the validity of cut scores used for high-stakes purposes.

Consequences

The benefits and deficits of a high-stakes testing program are its consequences. AERA (2000) argues that negative consequences of test-score uses should be made public. A recent study by Warren, Jenkins, and Kulick (2006) of high-stakes testing's impact on state graduation rates from 1975 to 2002 is not positive. Graduation rates are negatively correlated with high-stakes testing, and the rate of GED testing is increasing in high-stakes states. Although other factors may contribute, those findings support a plausible hypothesis that high-stakes uses of the tests have negative consequences. The public needs to be informed about any negative consequences of standardized achievement testing.

Summary

Threats to the validity of high-stakes standardized achievement test scores are well documented in many sources, both scholarly and popular. This section has sought to point out the often-questionable validity of standardized achievement test score interpretations. Documentation that could assure critics and the public of the scores' accuracy is scarce; in many instances scores are simply inaccurate.

PART III: CONCLUSIONS

The public demands standardized achievement testing, and the uses of the scores increasingly involve high stakes. Elected representatives have responded to the public's demand for testing. The federal government has long maintained a national testing program, the National Assessment of Educational Progress, that measures student achievement over time. Standardized achievement testing programs in most states and

virtually all school districts provide information to help teachers assess student learning, plan better instruction, and inform policymakers and their constituencies. Their goals for student learning include appropriate curriculum, aligned instruction, and assessment based on multiple sources of valid information. But in most circumstances, the basis of assessment and accountability is a single source of information: the standardized achievement test. National education organizations have argued that test-based accountability is shortsighted, narrow, and inadequate. Given the frequent tendency for such scores to be inaccurate, we may in fact be doing more damage than good to our students. To continue using standardized achievement testing, we need to assure the public that our interpretation and each intended use are valid, not flawed or contaminated, as they often seem.

The many threats to the validity of standardized achievement tests this article has pointed out should concern us all. We need to evaluate these threats honestly and minimize or eliminate them. Without documentation or research that dismisses or qualifies such threats, it is hard to justify the public's longstanding confidence in standardized achievement test scores.

Perfection in test development and validation is unattainable (Phelps 2006). I agree. However, high validity standards are mandatory, particularly when test scores affect students' lives and their teachers' careers. If the messenger of student learning is so badly flawed, where is the truth in the message? The perils need more attention than they have received in the past.

References

American Educational Research Association (AERA). 2000. "Position Statement of the American Educational Research Association Concerning High-stakes Testing in Pre-K–12 Education." *Educational Researcher* 29: 24–25.

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). 1999. *Standards for Educational and Psychological Testing.* Washington, D.C.: American Educational Research Association.

Carroll, J. B. 1963. "A Model for School Learning." *Teachers College Record* 64: 723–733.

Downing, S. M., and T. M. Haladyna, eds. 2006. *Handbook of Test Development*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Ferrara, S., and G. E. DeMauro. In press. "Standardized Assessment of Individual Achievement in K–12." In *Educational Measurement,* ed. R. L. Brennan, 4th ed. Westport, Conn.: American Council on Education/Praeger.

Haladyna, T. M. 2004. "The Conditions of Assessment of Student Learning in Arizona: 2004." In *The Conditions of Pre-K–12 Education in Arizona: 2004*, ed. A. Molnar. Tempe, Ariz.: Education Policy Studies Laboratory at Arizona State University.

Haladyna, T. M., and S. M. Downing. 2004. "Construct-irrelevant Variance in High-stakes Testing." *Educational Measurement: Issues and Practice* 23(1): 17–27.

Haladyna, T. M., S. B. Nolen, and N. S. Haas. 1991. "Raising Standardized Achievement Test Scores and the Origins of Test Score Pollution." *Educational Researcher* 20: 2–7.

Haladyna, T. M., and R. M. Olsen. 2006. Threats to validity in large-scale writing performance tests: What are these threats and what should be done about it? Paper presented at the annual CCSSO Large-Scale Assessment Conference, San Francisco, Calif.

Haladyna, T. M., S. Osborn Popp, and M. Weiss. 2005. Nonresponse in large-scale assessment. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Kane, M. T. In press. "Validation." In *Educational Measurement*, ed. R. L. Brennan, 4th ed. Westport, Conn.: American Council on Education/Praeger.

————. 2006. "Content-related Validity Evidence." In *Handbook of Test Development*, eds. S. M. Downing and T. M. Haladyna, 131–154. Mahwah, N.J.: Lawrence Erlbaum Associates.

Kolen, M. J., and R. L. Brennan. 2004. *Test Equating, Scaling, and Linking: Methods and Practices.* 2nd ed. New York: Springer-Verlag.

Lohman, D. F. 1993. "Teaching and Testing to Develop Fluid Abilities." *Educational Researcher* 22: 12–23.

McCallin, R. 2006. "Test Administration." In *Handbook of Test Development*, eds. S. M. Downing and T. M. Haladyna, 625–652. Mahwah, N.J.: Lawrence Erlbaum Associates.

Messick, S. 1984. "The Psychology of Educational Measurement." *Journal of Educational Measurement* 21: 215–237.

————. 1989. "Validity." In *Educational Measurement*, ed. R. L. Linn, 3rd ed., 13–104. New York: American Council on Education and Macmillan.

————. 1994. "The Interplay of Evidence and Consequences in the Validation of Performance Assessments." *Educational Measurement: Issues and Practices* 23(2): 13–23.

————. 1995a. "Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning." *American Psychologist* 50: 741–749.

————. 1995b. "Standards of Validity and the Validity of Standards in Performance Assessment." *Educational Measurement: Issues and Practice* 14(4): 5–8.

Phelps, R. 2006. "Characteristics of an Effective Student Testing System." *Educational Horizons* 85(1).

Sorenson, D. 2006. *2006 State Education Test Security Results*. Midvale, Utah: Caveon Test Security.

Sternberg, R. J. 1998. "Abilities Are Forms of Developing Expertise." *Educational Researcher* 27(3): 11–20.

Warren, J. R., K. N. Jenkins, and R. B. Kulick. 2006. "High School Exit Examination and State-level Completion and GED Raters, 1975 through 2002." *Educational Evaluation and Policy Analysis* 28(2): 131–152.

Tom Haladyna, Professor of Educational Psychology in the College of Teacher Education and Leadership at Arizona State University, has been an elementary school teacher, a test developer, a test researcher, and a teacher educator. He has published thirteen books, more than seventy articles and chapters, and hundreds of technical reports, validity studies, and conference papers. His primary interests are planning and validating testing programs, item development, and validity studies. He has worked in different capacities in more than fifty testing programs. He is co-editor, with Steve Downing, of the Handbook of Test Development *(Lawrence Erlbaum Associates 2006).*