

EVIDENCE OF TEST SCORE USE IN VALIDITY

Evidence of Test Score Use in Validity: Roles and Responsibilities

Paul D. Nichols

Natasha Williams

Pearson

Paper presented at the annual meeting of the National Council on Measurement in
Education, March, 2008, New York, New York

Abstract

This paper has three goals. The first goal is to clarify the role that the consequences of test score use play in validity judgments by reviewing the role that modern writers on validity have ascribed for consequences in supporting validity judgments. The second goal is to summarize current views on who is responsible for collecting evidence of test score use consequences by attempting to separate the responsibilities of the test developer and the test user. The last goal is to offer a framework that attempts to prescribe the conditions under which the responsibility for collecting evidence of consequences falls to the test developer or to the test user.

Evidence of Test Score Use in Validity: Roles and Responsibilities

The concept of validity has evolved since Guilford claimed that “a test is valid for anything with which it correlates” (1946, p. 429). The evolution of the concept of validity has been well documented (Cronbach, 1988; Kane, 2001; Messick, 1989). Currently, the field of educational measurement appears to have reached broad consensus that validity is a judgment of the degree to which arguments support the interpretations and uses of test scores (Kane, 2006). However, the field of educational measurement appears to disagree on the role that the consequences of test score use play in judgments concerning validity (Greene, 1997; Mehrens, 1997). Yet, the consequences of test score use take on increasing importance in the current era in which educators are attempting to leverage the information in test scores to improve student learning (Perie, Marion & Gong, 2007).

This paper sets out to accomplish three goals. The first goal is to clarify the role that the consequences of test score use play in validity judgments in educational measurement. In pursuing this goal, this paper recounts the role that modern writers on validity have ascribed for consequences in supporting validity judgments. The second goal of this paper is to summarize current views on who is responsible for collecting evidence of the consequences of test score use. This paper addresses that goal by attempting to separate the responsibilities of the test developer and the test user for collecting evidence of test score use consequences. The last goal of this paper is to offer a framework that attempts to prescribe the conditions under which the responsibility for collecting evidence of consequences falls to the test developer or to the test user.

The Role of Consequences in Supporting the Interpretation and Use of Test Scores

This section provides an overview of the role of consequences of test score use proposed by modern writers on validity. The inclusion of consequences of test score use in evidence to support the soundness of test score use and interpretation by writers in the early 1950s to late 1970s is discussed first. Then the role of the consequences of test score use in writings of more current authors on validity is summarized.

Early Frameworks

Writers on validity beginning near the early 1950s and continuing to the late 1970s divided validity arguments—the lines of argument to support the soundness of the interpretation and use of test scores (Cronbach, 1988; Messick, 1989)—into three or four categories based on the nature of the argument advanced in support of an interpretation of test scores. For example, a commonly used classification categorized validity evidence into criterion, content and construct validity. This division had become so widely embraced that Guion (1980) referred to it as “something of a holy trinity representing three different roads to psychometric salvation” (p. 386). This quip reflects the view of validity argument as a toolkit. As Kane (2001) notes, “Between the early 1950s and the mid to late 1970s, the practice developed of using the different models as a sort of toolkit, with each model to be employed as needed in the validation of educational and psychological tests,” (Kane, 2001, p. 323). The problem with such an approach that Kane (2001; 2006) documents is the lack of a requirement for a coherent or even reasonable argument.

Within this toolbox, no place is reserved for the consequences of test score use. Test score interpretation and test score use are treated as distinct issues. The consequences of test score use have no relevance for arguments supporting test score interpretation. Under this view, the consequences of test score use are issues for policymakers to consider rather than part of the body of evidence relevant to score interpretation.

However, Shepard (1997; see also Kane, 2001) argues that consequences of test score use were a part of the validity framework for decades before Messick's (1989) landmark chapter. Consequences were included as part of validity under the guise of discussing the soundness of test-based decisions. The inclusion of consequences as an aspect of validity evidence, "was only made to seem a major departure because of Messick's use of a new term and a new set of conceptual categories" (Shepard, 1997, p. 6). But Kane (2006) notes that the consequences of test score use were an implicit aspect of validity before Messick (1988, 1989) made consequences an explicit and prominent component of validity evidence.

Recent Frameworks

More recently, writers on validity have categorized validity arguments by the type of evidence used to support an interpretation of test scores. The number of categories is arbitrary, but Messick (1989, 1995) distinguishes the following six aspects that function as standards for construct validity: content relevance and representativeness; substantive theories, process models and process engagement; scoring models; generalizability; convergent and discriminant correlations; and consequences. Taken together, these six

aspects address the many and interrelated questions that need to be answered to justify score interpretation and use.

Within this framework, current writers on validity explicitly assign a role for the consequences of test score use (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999; Kane, 2001, 2006; Messick, 1989, 1995). Though validity is maintained as a unitary construct, consequences of test score use serve both as evidence for construct interpretation and as predictions based on construct theories. Consequences serve as evidence for construct interpretation when empirical findings on test score use either confirm or question construct theories. Consequences are predicted by construct theories when judgments of relevance or utility are based on inferences of the attributes and processes assessed, as they always are.

These writers argue that validity is called into question when the consequences of test score use can be linked to a flaw in the conceptualization of test score interpretation and use. For example, the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999) explains, “Rather, judgments of validity or invalidity in the light of testing consequences depend on a more searching inquiry into the sources of these consequences” (p. 16). This flaw in the conceptualization of test score interpretation and use may be due to construct under-representation or inclusion of sources of construct irrelevant variance. An example of the negative consequences of test score use linked to construct under-representation was offered by Shepard (1997). Medical college admissions officers were concerned that

using the MCAT to admit students to medical school was prompting pre-med students to concentrate on taking science courses. The consequence was that deans and admissions officers were concerned that students admitted to medical school were too narrowly educated. The deans and admissions officers were concerned that the interpretation of MCAT scores as level of preparation for medical practice and the use of the scores to admit students to medical school was an example of construct under-representation.

The potential threat of construct-irrelevant variance to validity has been documented by Haladyna and Downing (2004). An example of the negative consequences of test score use linked to construct-irrelevant variance is the possible presence of rater stringency when scoring typed essays compared to scoring hand-written essays. Research suggests a tendency for raters to score typed essays more stringently than hand-written essays (Hollenbeck, Tindal, Stieber, & Harniss, 1999; Powers, Fowles, Farnum, & Ramsey, 1994). A number of possible scoring biases may contribute to rater stringency including the tendency for typed essays to appear shorter than identical hand-written responses and more obvious writing errors because of the greater ease of reading typed responses compared to hand-written responses.

Lingering Controversy

But the inclusion of consequences in the current discussion of validity evidence remains controversial. Both Green (1998) and Reckase (1998) argue that imposing the responsibility for collecting evidence of test score use consequences on test developers burdens them with an impossible task. First, the test developer of a new testing program simply has no consequences yet of test score use to collect as evidence. And certainly

evidence with regard to unanticipated consequences cannot be collected for a new testing program since the effects are, by definition, unanticipated. The example provided by Reckase (1998) is the development of professional coaching for students taking the ACT Assessment, a consequence that might lead to some higher test scores from improved test taking skills.

Furthermore, even after a testing program has been in place, Green (1998) and Reckase (1998) argue that evidence of a cause-and-effect relationship is impossible to collect under the conditions of an operational testing program. Random assignment of students to a treatment, a necessary condition to establish causality, is unlikely to be approved (Reckase, 1998). Similarly, the maintenance of long-term experimental and control groups, another necessary condition to establish causality, is just as unlikely (Green, 1998).

But Green (1998) and Reckase (1998) may be setting up a straw man argument when they argue that the difficulty of random assignment and the maintenance of long-term experimental and control groups precludes the inclusion of consequences in the discussion of validity evidence. Writers who advocate consequences as validity evidence describe validity as judgments of the degree to which arguments support the interpretations and uses of test scores. But these writers do not require long-term studies incorporating random assignment to treatment to employ the effects of test score use as validity evidence. Such studies may strengthen arguments but are not required to construct convincing arguments. The results from alternative research methods, such as the case studies recounted but rejected by Green (1998), also may be used in validity arguments.

In contrast to arguments that collecting evidence of test score use consequences is too burdensome, Mehrens (1997) rejection of consequences in a discussion of validity evidence is straightforward. According to Mehrens (1997), “One can investigate the validity of the inference that a score is a reasonable indicator of the amount of a construct possessed independent of any specific use of the score” (p. 17). Consequently, Mehrens (1997) rejects the argument that consequences of test score use can have relevance for arguments concerning the accuracy of test score use and argues that the analyses of the effects of test score use should not be included as validity evidence. Furthermore, Mehrens (1997) argues that the psychometric community should narrow the use of the term validity to evidence of the accuracy of inferences regarding test scores.

But Mehrens’ (1997) wall separating the consequences of test score use from other evidence relevant to the validity of test score interpretation is artificial. The consequences of test score use can have implications for score interpretation as the earlier example from Shepard (1997) demonstrates. The consequences of test score use are evidence relevant to test score interpretation when the consequences can be linked to construct under-representation or construct irrelevant variance. Consequences of test score use that can be linked to construct under-representation or construct irrelevant variance can be excluded by fiat from evidence relevant to the inference that a score is a reasonable indicator of the amount of a construct possessed. But such a wall is fragile and easily breached.

The Assignment of Responsibility for Collecting Evidence of the Consequences of Test
Score Use

Despite the misgivings of some writers on validity, consensus has developed that the consequences of test use are an important source of validity evidence (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999; Kane, 2006). But who is responsible for collecting evidence of the consequences of test score use—the test developer or the test user? The roles of both the test developer and the test user in investigating the consequences of test score use are poorly defined. Writers disagree on the extent of test developer responsibility as compared to test user responsibility. Are test developers responsible for investigating distal effects of test score use such as possible narrowing of the curriculum? Or is the responsibility of test developers limited to more proximal effects of test score use such as class placement?

This section attempts to summarize current views on who is responsible for collecting evidence of test score use consequences. Initially, this section offers definitions of test developer and test user while recognizing that a clear delineation of those roles may not always exist. Next, this section reviews arguments for the test developers' responsibilities in providing evidence concerning the consequences of test score use. Finally, this section reviews arguments for the test users' responsibilities in providing evidence concerning the consequences of test score use.

Role Definition

Any discussion of the responsibilities of the test developer and the test user for collecting evidence of test score use consequences must be predicated on clear delineations of those roles. The Standards for Educational and Psychological Testing

(American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999) offers descriptions that attempt to mark boundaries between the roles of test developer and test user. But as the Standards note, the intermingling of roles makes the precise assignment of responsibility difficult.

Under the Standards, the test developers appear to be those individuals and organizations that fall into any of the following categories: those who prepare and design the tests, those who market the tests, those who sponsor tests and those who mandate the development and administration of tests. These individuals may include teachers, educational administrators, business representatives and state and federal officials. The organizations may include schools, school districts, businesses and state and federal agencies.

Test users appear to be those individuals and organizations who actively participate in the administration, interpretation and use of test results. The group of individuals and organizations that may be classified as test users includes a high proportion of members that may also be classified as test developers. Individuals who may be classified as test users include parents, teachers, educational administrators and state and federal officials. Organizations that may be classified as test users include schools, school districts, and state and federal agencies. The Standards explicitly excludes test takers themselves from the group of test users but this exclusion may not be reasonable in the context of educational measurement.

But distinguishing test users from test developers is more challenging than might first appear (Linn, 1998). For example, the No Child Left Behind Act of 2001 (NCLB; Public Law 107-110) requires, among other things, that all schools implement testing in

Grades 3 through 8 in mathematics and English/language arts by the 2005-2006 school year. For a grade 5 mathematics test under NCLB, the role of test user is shared across a number of parties. The federal government is arguably a test user because of the federal requirement that students make Adequate Yearly Progress in state test performance, with the eventual goal of all students reaching proficiency by the 2013-2014 school year. The state legislature, the state board of education and the state department of education also are arguably test users. On the obverse, each of these test users is also reasonably defined as a test developer; the federal government has passed federal legislation specifying features of the test, the state legislature has passed state legislation specifying additional features of the test and the state board of education may have approved the test design that the state education agency proposed.

Test Developer

Both Shepard (1997) and Kane (2001; 2006) argue that the test developer plays a circumscribed role in investigating the consequences of test score use. For example, Shepard (1997) argues that intended effects and likely side effects are clearly within the responsibility of the test developer. Furthermore, persistent unanticipated effects are also the responsibility of the test developer. But test developers are not responsible for negative consequences following test score misuse or for distal consequences such as real estate prices.

However, Moss (1998) suggests greater responsibility for the test developer and argues that considerations of test consequences should encompass the anticipated uses of test scores. Test developers are obligated to attempt to maximize positive consequences

and minimize negative consequences. But Moss goes further and argues that test developers should consider the consequences of testing in general rather than the immediate consequences of using scores from a specific test. For example, Moss argues that testing is reactive with test takers and test users. The administration of a test in a school changes the school, whether information from scores are intentionally used or ignored.

A more definitive treatment of test developers' responsibility to collect evidence to support test score use was provided by Kane (2001; 2006). The significant role for consequences in validity argument is confirmed by Kane's treatment. As Kane (2006) notes, consequences are the "bottom line" in evaluating decision procedures. Consequences include how well a decision procedure achieves its goals as well as the immediate negative consequences. Kane's argument draws a relatively clear boundary between the responsibilities of the test developer and the test user. The test developer has circumscribed responsibility for collecting evidence to support test score use as opposed to responsibility to support test score descriptive interpretation.

The interpretive argument supporting validity is divided into two parts by Kane (2001): a descriptive part supporting descriptive statements about individuals and a prescriptive part supporting decisions concerning treatments of those same individuals. Under this dichotomous framework, Kane proposes that the test developer is responsible for collecting evidence for supporting descriptive statements about individuals. In contrast, the responsibility for collecting evidence supporting decisions about individuals is divided between test developer and test user.

Under this division of responsibility, the test developer is responsible for collecting evidence to support promoted uses of test scores, e.g., using formative test scores to make decisions with regard to instructional treatment. Furthermore, the test developer is responsible for collecting evidence concerning reasonably anticipated but not promoted uses of test scores. For example, many schools administer benchmark tests throughout the school year to track student progress between more formal test administrations. Often tests developed for other purposes are adopted for use as benchmark tests, test developers may not have planned for their assessments to be used in this way but can anticipate that this is a possibility. Finally, Kane suggests that test developers are responsible for monitoring the use of their tests and the subsequent consequences. But Kane (2006) clearly defines the social consequences of test score use, those consequences that are distal to the decision procedures, as outside of the test developers' responsibilities.

Test User

According to Kane (2006), the test user is responsible for collecting evidence to support uses of test scores proposed by the test user that are outside of reasonably anticipated uses. A judgment must be made as to what uses are reasonable or not reasonable to anticipate. To whom the onus for that judgment will fall is unclear. But the test developer does have responsibility to monitor uses of test scores proposed by the test user that are outside of what the test developer anticipated.

The Standards are more demanding of test users and charges test users with greater responsibility with respect to evidence supporting uses of test scores. The

Standards discards judgments of reasonably anticipated uses in favor of clearly separating test uses that have been validated from test uses that have not been validated. Standard 1.4 states: “If a test is used in a way that has not been validated, it is incumbent on the user to justify the new use, collecting new evidence if necessary” (p. 18).

The need to assign responsibility to collect evidence to support test score use becomes unavoidable when the situation conflates the roles of test developer and test user. But sometimes the responsibility may be confused and the result is that evidence of test use consequences is uncollected. Policymakers may provide an example of such a result. As noted, policymakers such as federal and state governments often serve as both test developer and test user. But Linn (1998) observes that in such a situation the ownership of responsibility to gather evidence of the effect of test score use may be unclear.

For example, under NCLB, states are required to submit information about their student assessments to the federal government for peer review. Peer review guidance (U.S. Department of Education Office of Elementary and Secondary Education, 2007) emphasizes the importance of validity by requiring submission of evidence outlining the validity of the assessment under review. Of particular interest for this paper is the requirement for evidence related to the consequences associated with assessment use. An excerpt of peer review critical element 4.1 covering Technical Quality is given below:

For each assessment, including alternate assessment(s), has the State documented the issue of validity (in addition to the alignment of the assessment with the content standards), as described in the Standards for Educational and

Psychological Testing (AERA/APA/NCME, 1999), with respect to all of the following categories:

- a) Has the State specified the purposes of the assessments, delineating the types of uses and decisions most appropriate to each? and...
- f) Has the State ascertained that the decisions based on the results of its assessment are consistent with the purposes for which the assessments were designed? and
- g) Has the State ascertained whether the assessment produces intended and unintended consequences?

This excerpt shows that the federal government clearly places responsibility for providing evidence of consequences of test score use on the State with the federal government taking on the role of test user. What remains unclear is whether the State has taken on the role of test user, test developer, or some combination of the two. The State may also assume the role of test user and place responsibility for providing evidence of test score use consequences on another party, such as the vendor. In this situation, both the federal government and the State may avoid responsibility for collecting this specifically requested evidence by assigning responsibility to another party who might reasonably object to that assignment. We suggest in the next section that the ownership of responsibility for collecting evidence to test score use consequences must be negotiated in situations such as this that are characterized by a confusion of roles.

A Prescription for the Conditions Under Which the Responsibility for Collecting Evidence of Consequences Falls to the Test Developer or to the Test User

The last section of this paper offers a framework that attempts to prescribe the conditions under which the responsibility for collecting evidence of consequences falls to the test developer or to the test user. The framework, shown in Figure 1, attempts to capture the fluid nature of ownership of responsibility for evidence of test score use consequence. Ownership is not static but evolves over time. To capture this fluid quality, the framework is organized around three dimensions: breadth of construct, distance from intended score use, and time from test publication. This section describes each element of the framework including the three dimensions and how the dimensions interact.

Insert Figure 1 Here

The upper right corner of the framework presents the test developer's responsibility whereas the lower left corner presents the test user's responsibility. A zone of negotiated responsibility exists between the test developer's responsibility and the test user's responsibility. This zone is in recognition that the ownership of responsibility is not clearly delineated. At times, the zone may be relatively narrow in recognition that the ownership of responsibility is well understood. At other times, the zone of responsibility may be relatively wide in recognition that the ownership of responsibility is controversial or shared. The figure attempts to portray how each of the three dimensions in the figure interacts with ownership of responsibility for collecting validity evidence.

Breadth of Responsibility

The first dimension running across the top of the figure is breadth of construct. The test developer's responsibility tends to broaden as the advertised definition of the construct broadens. An example of a narrowly defined construct is proficiency in two-digit addition. An example of a broadly defined construct is high school mathematics achievement. For broad constructs such as high school mathematics achievement, the test developer assumes responsibility to collect evidence of the consequences of test score uses that may be characterized as distal, i.e., situated away from the intended test score use. This is because the use of test scores reflecting broad constructs affects a wider range of activities than does the use of test scores reflecting a narrow construct.

Note that the effects of test score use are not a property of the test scores but reflect the breadth of test score interpretation that motivate test score use. Furthermore, the interpretation attributed to a set of test scores may evolve over time. For example, Shepard (1987) illustrates how test scores initially interpreted as level of science achievement can evolve to be interpreted as preparation for professional practice.

Distance from Intended Test Score Use

The second dimension running down the left side of the figure is distance from intended test score use. Test scores cannot be used without interpreting the meaning of those scores. As Messick noted (1989, p. 21): "To interpret a test is to use it, and all other test uses involve interpretations either explicitly or tacitly." Test scores may be used in ways that involve interpretations close to the targeted construct. These uses may be characterized as proximal to the intended test score use. Conversely, test scores may be used in ways that involve interpretations that stray from the targeted construct. These

uses may be characterized as distal from the intended test score use. Student scores that are used to determine teacher salaries would be an example of test score use that was not the original purpose of the test.

Time from Test Publication

The third dimension implicit in the figure is that of time from test publication. The initial element of the figure represents the distribution of responsibility immediately after test publication. The second element of the figure represents the distribution of responsibility some time following test publication. The figure represents only two points in time, but time is essentially a continuous variable. The figure shows a shift in more responsibility falling to the test developer as time goes on, but this shift could also be in the opposite direction. The test developer's responsibilities to collect evidence of test score use may expand (or shrink) over time as experience with test score use increases.

Test developer's responsibilities to collect test score use evidence might expand to encompass a test score use that may have been unusual when the test was first released but becomes increasingly common over time. For example, a personality test may have been originally developed to help individuals identify the sort of employment in which they would be most comfortable. Decades later, the personality test may be frequently used to increase collaboration and improve the productivity of workplace teams. In this situation, the test developers' responsibilities would expand to collect evidence of test score use for improving workplace team performance.

Conversely, test developer's responsibilities to collect test score use evidence might contract to reflect test score uses that over time the community of practitioners have restrained. For example, the use of intelligence test scores has become more restrained as practitioners' interpretation of scores from intelligence tests has become more contextualized (Sternberg, 1996). In this situation, the test developers' responsibilities would shrink to reflect the more restricted use of intelligence test scores.

Example of Framework Application

An example serves to illustrate the application of this framework to assigning ownership of responsibility for collecting test score use evidence. In this example, a test developer has advertised that test scores from a mathematics assessment are useful for providing feedback on students' daily practice, differentiating instruction to address student's individual needs, and monitoring the performance of students. Soon after publication, teachers adopt the mathematics assessment for formative purposes in their classrooms. In the situation where test scores are used as advertised, the construct definition is relatively narrow, test scores are used in ways near the intended use, and little time has passed since publication the responsibility for collecting evidence to support test score use is squarely that of the test developer.

Following hard upon publication, a research organization recommends use of the mathematics assessment test scores to determine the effectiveness of teaching methods and the success of classrooms, programs, and the school as a whole. In this situation, the construct definition has been broadened somewhat and test scores are used in ways removed from the intended use. Using the framework, the responsibility for collecting

evidence to support test score use would appear to fall to the test user. The test user here is the third party making the recommendations to use scores to determine the effectiveness of teaching methods and the success of classrooms, programs, and schools.

To extend this example, educators follow the recommendations of the research organization and use the mathematics assessment test scores to evaluate classrooms, programs and schools. This use by educators continues for several years with the knowledge of the test publisher. In this situation, the construct definition has been broadened and test scores are used in ways removed from the intended use but this particular test score use has been continuing across time. Using the framework, the responsibility for collecting evidence to support test scores used in this way would shift over time back towards the test developer. However, ownership of responsibility is not clearly delineated and may be controversial or shared. Ownership of responsibility might fall into the zone of negotiated responsibility that exists between the test developer and the test user.

Conclusions

The authors' first goal was to clarify the role that the consequences of test score use play in judgments concerning validity. This paper documents how consensus has developed that the consequences of test use are an important source of validity evidence. Consequences of test score use serve as validity evidence when these consequences can be linked to a flaw in the conceptualization of test score interpretation and use.

A summary of the current views on whether responsibility for collecting evidence of the consequences of test score use should fall to the test developer or the test user was

the second goal of the paper. This summary shows that assignment of responsibility for collecting consequential validity evidence is a fluid rather than static process. And is very much dependent on the circumstances surrounding the assessment in question.

The final goal of this paper was to outline a framework that models the shifting responsibility between test user and test developer for collecting consequential validity evidence. This framework shows that the relative burden of responsibility shifts as the context of test score shifts along three dimension: the breadth of the construct underlying score interpretation, the distance score use is from the advertised score use, and the extent to which test use has evolved over time. Furthermore, responsibility for collecting evidence of the consequences of test score use may be clearly that of the test user or test developer but there is also a set of test uses for which responsibility is unclear and still being negotiated.

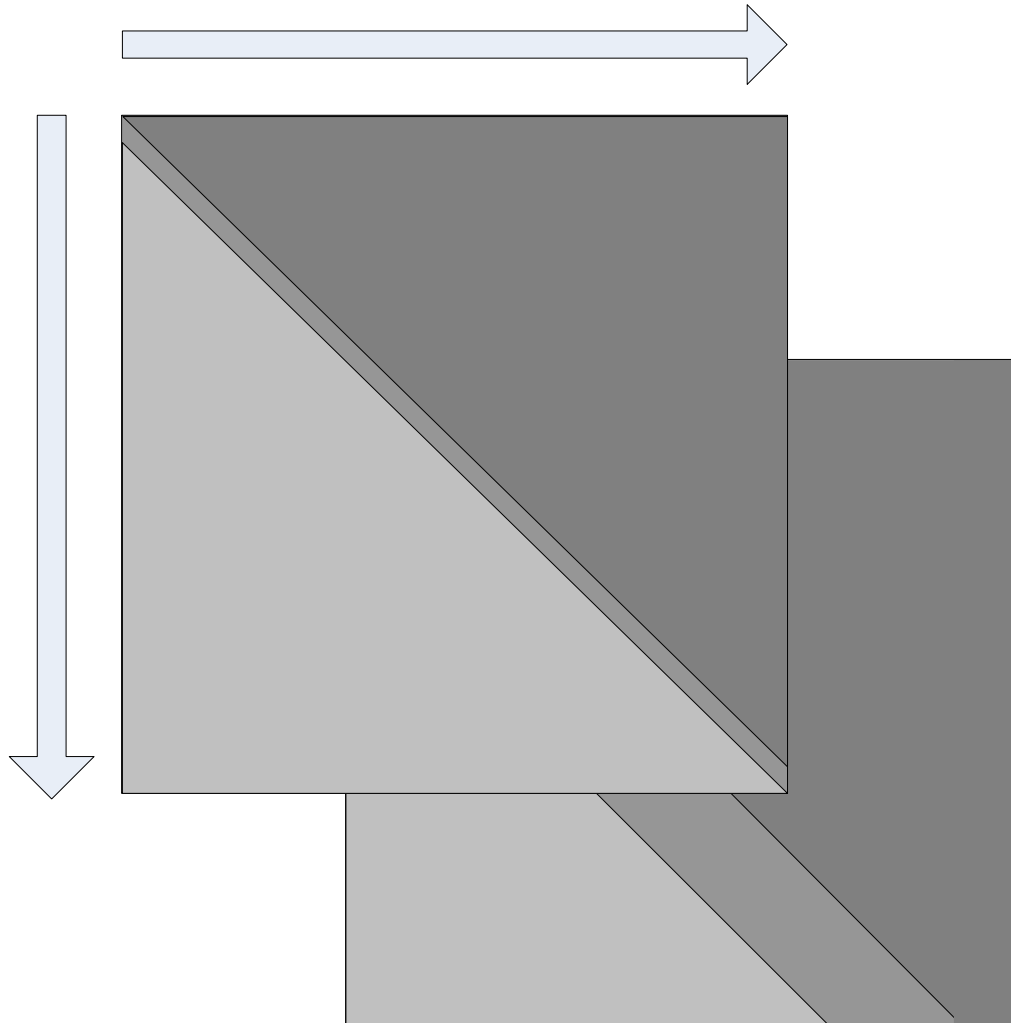
In an era of unprecedented publicity for test results, when test scores are used for everything from evaluating student progress to promoting real estate sales, the consequences of test score use are all around us. The understanding of the role test score use consequences play in validity arguments has evolved to reflect the changing uses of test scores. As Anastasi (1986) has noted, the concept of validity continues to evolve. Similarly, the role that the consequences of test score use play in judgments concerning validity will continue to evolve as the uses of test scores in society expand, contract and change.

REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.). *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.
- Green, D. R. (1998). Consequential aspects of the validity of achievement tests: A publisher's point of view. *Educational Measurement: Issues and Practice*, 17(2), 16-19, 34).
- Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, 6, 427-439.
- Guion, R. M. (1980). On trinitarian conceptions of validity. *Professional Psychology*, 11, 385-398.
- Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, 23 (1), 17-27.
- Hollenbeck, K., Tindal, G., Stieber, S., & Harniss, M.(1999). Handwritten versus word processed statewide compositions: do judges rate them differently? *Applied Measurement in Education*, 7, 255-278
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38, 319-342
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Washington, DC: The National Council on Measurement in Education & the American Council on Education.
- Linn, R. L. (1998). Partitioning responsibility for the evaluation of the consequences of assessment programs. *Educational Measurement: Issues and Practice*, 17(2), 28-30.
- Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, 16(2), 16-18.
- Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: Erlbaum.

- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Washington, DC: The American Council on Education & the National Council on Measurement in Education.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23, 13-24.
- Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, 14(4), 5-8.
- Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice*, 17(2), 6-12.
- No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).
- Perie, M, Marion, S., & Gong, B. (June, 2007). *A framework for considering interim assessments*. Paper presented at the National Conference on Large-scale Assessment Conference sponsored by the Council of Chief State School Officers, Nashville, Tennessee.
- Powers, D. E., Fowles, M. E., Farnum, M., & Ramsey, P.(1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement*, *31*, 220-233.
- Reckase, M. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice*, 17(2), 13-16.
- Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16(2), 5-8, 13, 24.
- Sternberg, R. J. (1996). Myths, countermyths, and truths about intelligence. *Educational Researcher*, 25 (2), 11-16.
- U.S. Department of Education Office of Elementary and Secondary Education. (2007). *Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001, revised July 25, 2007 to include modified academic achievement standards*, DRAFT, Washington, D.C. 20202

TOP



NAF

Figure 1

PROXIMAL

Figure Captions

Figure 1. A framework summarizing the conditions under which the responsibility for collecting evidence of consequences falls to the test developer or to the test use.