Standards of Validity and the Validity of Standards in Performance Assessment

Samuel Messick
Educational Testing Service

What are six distinct aspects of construct validation? How do these aspects apply to performance assessment? Are the consequences of performance assessment on teaching and learning relevant to construct validation?

Validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *interpretations* and *actions* based on test scores or other modes of assessment (Messick, 1989). Validity is not a property of the test or assessment as such, but rather of the meaning of the test scores. These scores are a function not only of the items or stimulus conditions but also of the *persons* responding as well as the *context* of the assessment. In particular, what needs to be valid is the meaning or interpretation of the scores as well as any implications for action that this meaning entails (Cronbach, 1971). The extent to which score meaning and action implications hold across persons or population groups and across settings or contexts is a persistent and perennial empirical question. This is the main reason that validity is an evolving property and validation a continuing process.

The Value of Validity

The principles of validity apply to all assessments, whether based on tests, questionnaires, behavioral observations, work samples, or whatever. These include performance assessments which, because of their promise of positive consequences for teaching and learning, are becoming increasingly popular as purported instruments of standards-based education reform. Indeed, it is precisely because of these politically salient potential consequences that the validity of performance assessment needs to be systematically addressed, as do other basic measurement issues such as reliability, comparability, and fairness.

These issues are critical for performance assessment because validity, reliability, comparability, and fairness are not just measurement principles; they are *social values* that have meaning and force whenever evaluative judgments and decisions are made. As a salient social value, validity assumes both a scientific and a political role that can by no means be fulfilled by a simple correlation coefficient between test scores and a purported criterion (i.e., classical criterion validity) or by expert judgments that test content is relevant to the proposed test use (i.e., traditional content validity).

Indeed, broadly speaking, validity is nothing less than an evaluative summary of both the evidence for and the actual as well as potential consequences of score interpretation and use (i.e., construct validity conceived comprehensively). This comprehensive view of validity integrates considerations of content, criteria, and consequences into a construct framework for empirically testing rational hypotheses about score meaning and utility. Fundamentally, then, score validation is empirical evaluation of the meaning and consequences of measurement. As such, validation combines scientific inquiry with rational argument to justify (or nullify) score interpretation and use.

Aspects of Construct Validity

The validity issues of score meaning, relevance, utility, and social consequences are many-faceted and intertwined. They are difficult if not impossible to disentangle, which is why validity has come to be viewed as a unified concept (APA, AERA, & NCME, 1985; Messick, 1989). However, to speak of validity as a unified concept does not imply that validity cannot be usefully differentiated into distinct aspects to underscore issues and nuances that might otherwise be downplayed or overlooked, such as the social consequences of performance assessments or the role of score meaning in applied use. The intent of these distinctions is to provide a means of addressing functional aspects of validity that help disentangle some of the complexities inherent in appraising the appropriateness, meaningfulness, and usefulness of score inferences.

In particular, six distinguishable validity aspects are delineated emphasizing content, substantive, structural, generalizability, external, and consequential aspects of con-

Samuel Messick is a Vice President for Research at Educational Testing Service, Rosedale Rd., Princeton, NJ 08541. His specializations are validity and educational and psychological measurement.

struct validity (Messick, 1994, in press). In effect, these six aspects function as general validity criteria or standards for all educational and psychological measurement (Messick, 1989). Following a capsule description of these six aspects, I highlight some of the validity issues and sources of evidence bearing on each:

- The content aspect of construct validity includes evidence of content relevance, representativeness, and technical quality (Lennon, 1956; Messick, 1989).
- The substantive aspect refers to theoretical rationales for the observed consistencies in test responses, including process models of task performance (Embretson, 1983), along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks.
- The structural aspect appraises the extent to which the internal structure of the assessment reflected in the scores, including scoring rubrics as well as the underlying dimensional structure of the assessment tasks, is consistent with the structure of the construct domain at issue (Loevinger, 1957).
- The generalizability aspect examines the extent to which score properties and interpretations generalize to and across population groups, settings, and tasks (Cook & Campbell, 1979; Shulman, 1970), including validity generalization of test-criterion relationships (Hunter, Schmidt, & Jackson, 1982).
- The external aspect includes convergent and discriminant evidence from multitrait-multimethod comparisons (Campbell & Fiske, 1959), as well as evidence of criterion relevance and applied utility (Cronbach & Gleser, 1965).
- The consequential aspect appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice (Messick, 1980, 1989).

Content Relevance and Representativeness

A key issue for the content aspect of construct validity is the specification of the boundaries of the construct domain to be assessed—that is, determining the knowledge, skills, and other attributes to be revealed by the assessment tasks. The boundaries and structure of the construct domain can be addressed by means of job analysis, task analysis, curriculum analysis, and especially domain theory—that is, scientific inquiry into the nature of the domain processes and the ways in which they combine to produce effects or outcomes. A major goal of domain theory is to understand the construct-relevant sources of task difficulty, which then serves as a guide to the rational development and scoring of performance tasks. For an example of how domain theory can inform both test construction and validation, see Kirsch, Jungeblut, and Mosenthal (in press). At whatever stage of its development, then, domain theory is a primary basis for specifying the boundaries and structure of the construct to be assessed.

However, it is not sufficient merely to select tasks that are relevant to the construct domain. In addition, the assessment should assemble tasks that are representative of the domain in some sense. The intent is to ensure that all important parts of the construct domain are covered, which is usually described as selecting tasks that sample domain processes in terms of their functional importance. Both the content relevance and representativeness of assessment tasks are traditionally appraised by expert professional judgment, documentation of which serves to address the content aspect of construct validity.

In standards-based education reform, two types of assessment standards have been distinguished. One type is called *content standards*, which refers to the kinds of things a student should know and be able to do in a subject area. The other type is called *performance standards,* which refers to the level of competence a student should attain at key stages of developing expertise in the knowledge and skills specified by the content standards. Performance standards also circumscribe, either explicitly or tacitly, the form or forms of performance that are appropriate to be evaluated against the standards.

From the discussion thus far, it should be clear that not only the assessment tasks but also the content standards themselves should be relevant and representative of the construct domain. That is, the content standards should be consistent with domain theory and be reflective of the structure of the construct domain. This is the issue of the construct validity of content standards. There is also a related issue of the construct validity of performance standards. That is, increasing achievement levels or performance standards (as well as the tasks that benchmark these levels) should reflect increases in complexity of the construct under scrutiny and not increasing sources of construct-irrelevant difficulty. These and other issues related to standards-based assessment will be discussed in the subsequent article on standard setting.

Substantive Theories, Process Models, and Process Engagement

The substantive aspect of construct validity emphasizes two important points: One is the need for tasks providing appropriate sampling of domain processes in addition to traditional coverage of domain content; the other is the need to move beyond traditional professional judgment of content to accrue empirical evidence that the ostensibly sampled processes are actually engaged by respondents in task performance. Thus, the substantive aspect adds to the content aspect of construct validity the need for empirical evidence of response consistencies or performance regularities reflective of domain processes (Embretson, 1983; Loevinger, 1957; Messick, 1989).

Scoring Models as Reflective of Task and Domain Structure

According to the structural aspect of construct validity, scoring models should be rationally consistent with what is known about the structural relations inherent in behavioral manifestations of the construct in question (Loevinger, 1957; Peak, 1953). That is, the theory of the construct domain should guide not only the selection or construction of relevant assessment tasks but also the

rational development of construct-based scoring criteria and rubrics. Ideally, the manner in which behavioral instances are combined to produce a score should rest on knowledge of how the processes underlying those behaviors combine dynamically to produce effects. Thus, the internal structure of the assessment (i.e., interrelations among the scored aspects of task and subtask performance) should be consistent with what is known about the internal structure of the construct domain (Messick, 1989). This relation of the assessment structure to the domain structure has been called *structural fidelity* (Loevinger, 1957).

To the extent that different assessments (i.e., those involving different tasks or different settings or both) are geared to the same construct domain, using the same scoring model as well as scoring criteria and rubrics, the resultant scores are likely to be comparable or can be rendered comparable using equating procedures. Otherwise, score comparability is jeopardized but can be variously approximated using such techniques as statistical or social moderation (Mislevy, 1992). Score comparability is clearly important for normative or accountability purposes whenever individuals or groups are being ranked. However, score comparability is also important even when individuals are not being directly compared, but are held to a common standard. Score comparability of some type is needed to sustain the claim that two individual performances in some sense meet the same local, regional, national, or international standard. These issues are addressed more fully in the subsequent article on comparability.

Generalizability and the Boundaries of Score Meaning

The concern that a performance assessment should provide representative coverage of the content and processes of the construct domain is meant to ensure that the score interpretation not be limited to the sample of assessed tasks but be generalizable to the construct domain more broadly. Evidence of such generalizability depends on the degree of correlation of the assessed tasks with other tasks representing the construct or aspects of the construct. This issue of generalizability of score inferences across tasks and contexts goes to the very heart of score meaning. Indeed, setting the boundaries of score meaning is precisely what generalizability evidence is meant to address.

However, because of the extensive time required for the typical performance task, there is a conflict in performance assessment between time-intensive depth of examination and the breadth of domain coverage needed for generalizability of construct interpretation. This conflict between depth and breadth of coverage is often viewed as entailing a trade-off between validity and reliability (or generalizability). It might better be depicted as a trade-off between the valid description of the specifics of a complex task performance and the power of construct interpretation. In any event, as Wiggins (1993) stipulates, such a conflict signals a design problem that needs to be carefully negotiated in performance assessment.

In addition to generalizability across tasks, the limits of score meaning are also affected by the degree of generalizability across time or occasions and across observers or raters of the task performance. Such sources of measurement error associated with the sampling of tasks, occasions, and scorers underlie traditional reliability concerns; they are examined in more detail in the subsequent article on generalizability.

Convergent and Discriminant Correlations With External Variables

The external aspect of construct validity refers to the extent to which the assessment scores' relationships with other measures and nonassessment behaviors reflect the expected high, low, and interactive relations implicit in the theory of the construct being assessed. Thus, the meaning of the scores is substantiated externally by appraising the degree to which empirical relationships with other measures, or the lack thereof, is consistent with that meaning. That is, the constructs represented in the assessment should rationally account for the external pattern of correlations.

Of special importance among these external relationships are those between the assessment scores and criterion measures pertinent to selection, placement, licensure, program evaluation, or other accountability purposes in applied settings. Once again, the construct theory points to the relevance of potential relationships between the assessment scores and criterion measures, and empirical evidence of such links attests to the utility of the scores for the applied purpose.

Consequences as Validity Evidence

Because performance assessments promise potential benefits for teaching and learning, it is important to accrue evidence of such positive consequences as well as evidence that adverse consequences are minimal. In this connection, the consequential aspect of construct validity includes evidence and rationales for evaluating the intended and unintended consequences of score interpretation and use in both the short- and long-term, especially those associated with bias in scoring and interpretation or with unfairness in test use. However, this form of evidence should not be viewed in isolation as a separate type of validity, say, of *consequential validity.* Rather, because the values served in the intended and unintended outcomes of test interpretation and use both derive from and contribute to the meaning of the test scores, appraisal of social consequences of the testing is also seen to be subsumed as an aspect of construct validity (Messick, 1980).

The primary measurement concern with respect to adverse consequences is that any negative impact on individuals or groups should not derive from any source of test invalidity such as construct underrepresentation or construct-irrelevant variance (Messick, 1989, in press). That is, low scores should not occur because the assessment is missing something relevant to the focal construct that, if present, would have permitted the affected students to display their competence. Moreover, low scores should not occur because the measurement contains something irrelevant that interferes with the affected students' demonstration of competence. Positive and negative consequences of assessment, whether intended or unintended, are discussed in more depth in the subsequent article on fairness.

Validity as Integrative Summary

These six aspects of construct validity apply to all educational and psychological measurement, including performance assessments. Taken together, they provide a way of addressing the multiple and interrelated validity questions that need to be answered in justifying score interpretation and use. They are highlighted because most score-based interpretations and action inferences, as well as the elaborated rationales or arguments that attempt to legitimize them (Kane, 1992), either invoke these properties or assume them, explicitly or tacitly. That is, most score interpretations refer to relevant content and operative processes, presumed to be reflected in scores that concatenate responses in domain-appropriate ways and are generalizable across a range of tasks, settings, and occasions. Furthermore, score-based interpretations and actions are typically extrapolated beyond the test context on the basis of documented or presumed relationships with nontest behaviors and anticipated outcomes or consequences. The challenge in test validation is to link these inferences to convergent evidence supporting them as well as to discriminant evidence discounting plausible rival inferences. Evidence pertinent to all of these aspects needs to be integrated into an overall validity judgment to sustain score inferences and their action implications, or else provide compelling reasons why not, which is what is meant by validity as a unified concept.

References

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana, IL: University of Illinois Press.

Embretson (Whitely), S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197.

Hunter, J. E., Schmidt, F. L., & Jackson, C. B. (1982). *Advanced meta-analysis: Quantitative methods of cumulating research findings across studies.* San Francisco: Sage.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527–535.

Kirsch, I. S., Jungeblut, A., & Mosenthal, P. B. (in press). *Moving toward the measurement of adult literacy* (Technical report on the 1992 National Adult Literacy Survey). Washington, DC: U. S. Government Printing Office.

Lennon, R. T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement, 16,* 294–304.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3,* 635–694 (Monograph Supplement 9).

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35,* 1012–1027.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Messick, S. (in press). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist.*

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: ETS Policy Information Center.

Peak, H. (1953). Problems of observation. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences* (pp. 243–299). Hinsdale, IL: Dryden.

Shulman, L. S. (1970). Reconstruction of educational research. *Review of Educational Research, 40,* 371–396.

Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappan, 75,* 200–214.

NCME OUTREACH COMMITTEE NEEDS CONTACTS

The NCME Board has named a committee whose charge is to offer the expertise of our membership to other professional education-related organizations. This Outreach Committee will arrange for NCME members to make assessment-related presentations for our fellow organizations at their annual meetings or related forums. The presentation topics will be of the host group's choosing; presentations will be made by selected NCME volunteer members who are expert on the topic of choice. The host organizations will pay all expenses of the presenter, but no fee will be charged.

The initial step in this process is to contact prospective cooperating organizations with an offer of our services. To begin this step, NCME members who are also members of other professional organizations are asked to suggest groups with which we should be in contact. Also, if you know a person within the organization who would be best to contact (e.g., president, program chair, executive director), please let us know that information. Please do NOT contact the organization directly with any offers to assist. All such contact should be initiated through the Outreach Committee.

If you have any suggestions—or other comments concerning this committee or its mission—please contact:

Michael D. Beck,
BETA, Inc.,
35 Guion Street,
Pleasantville, NY 10570
(Telephone: 914-769-5235, fax: 914-769-4809).