DOCUMENT RESUME

ED 380 501 TM 022 864

AUTHOR Messick, Samuel
TITLE Standards-Based Score Interpretation: Establishing Valid Grounds for Valid Inferences. Research Report.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-94-57
PUB DATE Dec 94
NOTE 31p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Construct Validity; *Educational Assessment; *Inferences; Knowledge Level; Measurement Techniques; Scores; Scoring; *Standards; *Test Interpretation; *Validity
IDENTIFIERS *Performance Based Evaluation

ABSTRACT

The construct validity of content standards is addressed in terms of their representative coverage of a construct domain and their alignment with the students' cognitive level of developing expertise in the subject matter. The construct validity of performance standards is addressed in terms of the extent to which they reflect increasing levels of construct complexity as opposed to construct-irrelevant difficulty. Also critical is the extent to which performance standards characterize the knowledge and skills operative at each level both to accredit specific accomplishment and to serve as goals for further learning. All of this depends on construct-valid assessment attuned to the content standards and the development of dependable scoring rubrics and measurement scales for representing the performance standards. (Contains 29 references.) (Author)

RR-94-57

RESEARCH REPORT

STANDARDS-BASED SCORE INTERPRETATION: ESTABLISHING VALID GROUNDS FOR VALID INFERENCES

Samuel Messick

Educational Testing Service
Princeton, New Jersey
December 1994

STANDARDS-BASED SCORE INTERPRETATION: ESTABLISHING VALID GROUNDS FOR VALID INFERENCES

Samuel Messick
Educational Testing Service

Standards-Based Score Interpretation: Establishing Valid Grounds for Valid Inferences

Samuel Messick
Educational Testing Service

The construct validity of content standards is addressed in terms of their representative coverage of a construct domain and their alignment with the students' cognitive level of developing expertise in the subject matter. The construct validity of performance standards is addressed in terms of the extent to which they reflect increasing levels of construct complexity as opposed to construct-irrelevant difficulty. Also critical is the extent to which performance standards characterize the knowledge and skills operative at each level both to accredit specific accomplishment and to serve as goals for further learning. All of this depends on construct-valid assessment attuned to the content standards and the development of dependable scoring rubrics and measurement scales for representing the performance standards.

STANDARDS-BASED SCORE INTERPRETATION: ESTABLISHING VALID GROUNDS FOR VALID INFERENCES

Samuel Messick[1]
Educational Testing Service

In standards-based education reform, a lot depends on the establishment of valid standards specifying both the critical content and the desired performance levels of student accomplishment, because these two kinds of standards comprise the driving force that energizes the reform movement. Basically, content standards specify what students should know and be able to do; performance standards specify the level and quality of that knowledge and skill that is deemed acceptable. Appraising whether or not assessed student competence meets a performance standard requires that the two be compared as points, as it were, on the same measurement scale. As a consequence, the validity of these standards cannot be separated from the validity of the assessment itself. That is, a construct-valid measurement scale of some sort is needed because without it there is not only no assessed content competence, there is also no performance standard.

Hence, in order to address the construct validity of both content standards and performance standards, we must turn to the same framework of validity criteria and forms of evidence needed to appraise the construct validity of assessed student competence. This is so because the construct validity of the content standards and the construct validity of the measurement scales go hand in hand. Moreover, to be meaningfully interpreted and reported, both the assessed competence and the performance standard need to be described in the same *construct* terms. That is, a performance standard

---

[1] This paper was commissioned by the National Assessment Governing Board and The National Center for Education Statistics and a briefer version was delivered at the Joint Conference on Standard Setting for Large-Scale Assessments, Washington, DC, October, 1994. I gratefully acknowledge helpful comments on the manuscript provided by Ann Jungeblut, Robert Mislevy, and Michael Zieky.

has two critical aspects: One is its location on the measurement scale; the other is its meaning in terms of the nature or quality of the knowledge and skill characterizing proficiency at that level. As a consequence, the construct validity of the meaning of the performance standard as well as that of the assessed competence, both being interpreted points on the same measurement scale, must be evaluated in the same ***evidential*** terms. Because the location of the performance standard is fundamentally a matter of value judgment, its validity must be addressed in terms of the reasonableness of the procedures used for determining it.

Next, we briefly review the criteria or standards of validity as well as the forms of evidence pertinent to the construct validation of any assessment, including performance assessment. Then we apply these general validity principles to a consideration of the construct validity of both content standards and performance standards, drawing implications as well for evaluating the standard-setting process whereby performance standards are determined.

STANDARDS OF VALIDITY

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. Hence, validity becomes a unified concept that integrates multiple supplementary forms of convergent and discriminant evidence.

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 test 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.

Aspects of Construct Validity

In particular, six distinguishable aspects of construct validity are highlighted as a means of addressing central issues implicit in the notion of validity as a unified concept. These are content, substantive, structural, generalizability, external, and consequential aspects of construct validity. In effect, these six aspects function as general validity criteria or standards for all educational and psychological measurement (Messick, 1989, 1994b). They are briefly characterized as follows:

- 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 fidelity of the scoring structure to 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).

In one way or another, these six aspects seek evidence and arguments to discount the two major threats to construct validity -- namely, construct underrepresentation and construct-irrelevant variance -- as well as to evaluate the action implications of score meaning. In construct underrepresentation, the assessment is too narrow and fails to include important dimensions or facets of the construct. In construct-irrelevant variance, the assessment is too broad, containing excess reliable variance that is irrelevant to the interpreted construct. Both threats are operative in all assessment. Hence a primary validation concern is the extent to which the same assessment might underrepresent the focal construct while simultaneously contaminating the scores with construct-irrelevant variance.

Validity As Integrative Summary

The 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. In previous writings I maintained that it is "the relation between the evidence and the inferences drawn that should determine the validation focus" (Messick, 1989. p. 16). This relation is embodied in theoretical rationales or persuasive arguments that the obtained evidence both supports the preferred inferences and undercuts plausible rival inferences. From this perspective, as Cronbach (1988) concluded, validation is evaluation argument. That is, as stipulated earlier, validation is empirical evaluation of the meaning and consequences of measurement. The term "empirical evaluation" is meant to

convey that the validation process is scientific as well as rhetorical and requires both evidence and argument.

By focussing on the argument or rationale employed to support the assumptions and inferences invoked in the score-based interpretations and actions of a particular test use, one can prioritize the forms of validity evidence needed in terms of the important points in the argument that require justification or support (Kane, 1992; Shepard, 1993). Helpful as this may be, there still remain problems in setting priorities for needed evidence because the argument may be incomplete or off target, not all the assumptions may be addressed, and the need to discount alternative arguments evokes multiple priorities. This is one reason that Cronbach (1989) stressed cross-argument criteria for assigning priority to a line of inquiry, such as the degree of prior uncertainty, information yield, cost, and leverage in achieving consensus.

The point here is that the six aspects of construct validity afford a means of checking that the theoretical rationale or persuasive argument linking the evidence to the inferences drawn touches the important bases and, if not, requiring that an argument be provided that such omissions are defensible. These six aspects 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 presumed or documented 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.

VALIDITY OF STANDARDS

With these fundamental aspects of construct validity in mind, let us now take up the question of the validity of both content and performance standards. The construct validity of content standards is addressed in terms of their representative coverage of a construct domain and their alignment with students' cognitive levels of developing expertise in the subject matter. In this treatment of the validity of content standards, special emphasis is given to the content, substantive, generalizability, and consequential aspects of construct validity.

The construct validity of performance standards as interpreted score levels is addressed in terms of the extent to which they reflect increasing construct complexity as opposed to construct-irrelevant difficulty. Special emphasis is given to the structural, generalizability, external, and consequential aspects of construct validity.

Content Standards as Blueprints for Teaching and Testing

Content standards specify what students should know and be able to do in a subject area at their particular level of developing expertise in the field. Concern is with what students should know and be able to do in different years of study in a discipline, as in fourth, eighth, and twelfth grade mathematics or language arts. Within-grade levels of proficiency, as we shall see, are captured by performance standards.

There is thus a temporal dimension to content standards. They specify not only *what* knowledge and skills should be attained, but *when*. Hence, the construct validation of content standards needs to address not only the relevance and representativeness of the *what* of the subject matter or construct domain, but also the appropriateness of the *when* to the students' cognitive levels of developing expertise. Judgments of what and when are usually combined by setting distinct content standards for different grade levels in a discipline, but these also need to be coordinated across grades to reflect an appropriate course of academic development. Ideally, both the substantive and the temporal aspects of content standards should be addressed in appraising their construct validity.

By specifying what students are expected to learn, content standards provide blueprints for what is important to teach as well as to test. However, content standards typically refer to generic constructs: For example, some content standards in grade-eight mathematics require that students should "understand the process of gathering and organizing data;" furthermore, they should "be able to calculate, evaluate, and communicate results" (Mullis, Dossey, Owen, & Phillips, 1993, p. 51). Learning exercises and assessment tasks are then selected or created to embody these generic

processes. The validity of the content standards as blueprints for teaching and testing depends on the extent to which the learning exercises engender these processes and the assessment tasks tap them. Thus, the construct validity of the content standards and the construct validity of the assessment tasks are inseparable.

Although a separate topic not to be elaborated here, the construct validity of opportunity-to-learn standards and that of the learning exercises are similarly inseparable. That is, the validity of opportunity-to-learn standards depends on exposure to learning experiences that are construct valid in the sense that they actually engender or facilitate the development of the knowledge and skills specified in the content standards.

The linking of content standards to assessment tasks shown to engage the specified processes bears on the substantive aspect of construct validity. In regard to the content aspect, a key issue is the extent to which the content standards circumscribe the boundaries and reflect the structure of the subject-matter construct domain. The major concern is to minimize construct underrepresentation. To be valid, the content standards and their associated assessment tasks should be both relevant to and representative of the construct domain. Hence, the content standards should specify (and the associated assessment tasks should sample) domain processes in terms of their functional importance.

A major problem is sorting out evidence of domain processes in complex tasks and especially disentangling focal construct processes from ancillary processes involved in task performance (Wiley, 1991). This is serious because ancillary processes, which are ordinarily numerous in complex task performance, are a potential source of construct-irrelevant variance.

Functionally important knowledge and skill in a subject-matter or construct domain may be addressed from at least two perspectives. One is in terms of what is actually done in the performance domain, for example, as revealed through techniques akin to job analysis. The other is in terms of what differentiates and characterizes developing expertise in the domain, which would usually emphasize different tasks and processes. The former perspective addresses the substantive aspect of content standards and the latter addresses the temporal aspect.

In effect, the content standards specify the constructs that are not only to be taught but are also to be assessed in standards-based performance assessment. Furthermore, as I have maintained elsewhere, "the meaning of the construct is tied to the range of tasks and situations that it generalizes and transfers to" (Messick, 1994a, p. 15). This brings us to the generalizability aspect of construct validity and, in particular, to the distinction between generalizability as consistency or reliability and generalizability as transfer.

In the previous presentation at this conference, Brennan discussed generalizability across judges, occasions, and tasks, which are topics typically subsumed under the heading of reliability, as well as generalizability across standard-setting methods. Also of concern are generalizability across measurement methods and scoring rubrics, which we will comment upon further in the next section on performance standards. Generalizability as reliability refers to the consistency of performance across the judges, occasions, and tasks of a particular assessment, which might be quite limited in scope. For example, we have all been concerned that some assessments with a narrow set of tasks might attain higher reliability in

the form of cross-task consistency, but at the expense of construct validity.

In contrast, generalizability as transfer requires consistency of performance across tasks that are representative of the broader construct domain. That is, transfer refers to the range of tasks that performance on the assessed tasks facilitates the learning of or, more generally, is predictive of (Ferguson, 1956.) Thus, generalizability as transfer depends not only on generalizability theory but also on domain theory, that is, on the construct theory of the subject-matter domain. In essence, then, generalizability evidence is an aspect of construct validity because it establishes boundaries on the meaning of the construct scores.

Content standards are at the heart of standards-based education reform because they are presumed to have positive consequences for teaching and learning. Evidence documenting such positive outcomes bears on the consequential aspect of construct validity, which for its full appraisal also requires attention to the possibility of unintended adverse side effects. For example, establishing common content standards for all students is a selective process that privileges certain knowledge and skills over other possibilities. This might inadvertently lead, as Coffman (1993) reminds us, to limitations on the development of those other skills and, hence, to unintended restrictions on the diversity of talent.

In effect, content standards for all students constitute a common denominator. To be sure, the impact of such a common denominator is most insidious when students are held to low standards, as in minimum-competency testing, rather than to high standards. Hence, the levels that students are challenged to reach become important as educational goals, which brings us directly to the topic of performance standards.

Performance Standards as Challenges or Hurdles

Performance standards refer to the level of competence a student should attain in the knowledge and skills specified by the content standards as well as the form or forms of performance that are appropriate to be evaluated against the standards. To take into account the differential complexity of information-processing requirements in different years of study in a discipline, which is attuned to the students' levels of developing expertise, standards of performance considered to be "good enough" are typically set separately by grade level. For example, in the National Assessment of Educational Progress (NAEP), performance standards for basic, proficient, and advanced levels are set separately for grades 4, 8, and 12.

By specifying the form or forms of performance that are appropriate to evaluate, performance standards essentially circumscribe the nature of the evidence relevant to deciding whether the standards have been met, for example, whether the evidence should be an essay, project, demonstration, mathematical proof, scientific experiment, or some combination of these. An important issue is whether standards-based score interpretation and reporting can legitimately be formulated in terms of the generic constructs of knowledge and skill specified in the content standards or whether it needs to be specific to the method of measurement, that is, specific to knowledge and skill exhibited via a particular method.

If the latter specificity holds, interpretation is limited to construct-method units, which implies that there are not only distinct performance standards but also distinct content standards for each method of measurement. To attain the power of the former interpretation in terms of generic constructs requires evidence of generalizability across measurement methods.

This is a fundamental issue because the content standards evoke generic constructs of knowledge and skill which, if attained at the levels specified by the performance standards, are deemed to be relevant to a range of diverse problems and applications. That is, the content standards specify knowledge and skill that is considered to be important precisely because it is generalizable and transferable across problems and situational contexts, including measurement contexts. Hence, the degree of generalizability of the construct scores across measurement methods bears directly on the meaning of the constructs.

Moreover, attention should be paid not just to convergent evidence of consistency across methods but also to discriminant evidence of the distinctness of constructs within method. Such multiconstruct-multimethod comparisons are needed to help disentangle construct-relevant variance from construct-irrelevant method variance. But more about this later.

As ordinarily conceptualized, a performance standard is a point on a measurement scale or a set of points on a profile of scales or a region in a multidimensional space. A "softer" version of performance standards is associated not with cut-points but with utility functions in a decision-theoretic approach to standard-setting (van der Linden, 1994). For simplicity, I will characterize a performance standard only as a point on a scale because the argument about the centrality of the measurement scale in standard-setting applies equally well to profiles and other multidimensional representations as well as to utility functions.

A measurement scale of some sort is critical to the setting and use of performance standards for at least two reasons. First, without a measurement scale, there can be no points on the scale and hence no performance standards.

This is so because the notion of performance standards implies an ordering of tasks (or of performances on a particular task) such that some of the performances are considered to be good enough and others not good enough. Such an ordering constitutes a rudimentary measurement scale. By taking into account the structure of interrelations among task or performance scores, more powerful model-based measurement scales may be fit to the data, such as the IRT-based scales developed for NAEP.

A second reason that not just a measurement scale but an interpreted measurement scale is critical is that meeting a performance standard should not just attest that the assessed performance is good enough. To be educationally useful, performance standards should also characterize the nature of the knowledge and skill entailed at that level as well as point to what needs be accomplished for further mastery. One implication of this is that the measurement scale should extend beyond the level of the performance standard. Another implication is that various levels on the scale, especially the performance-standard levels, should be tied to process descriptions of what constitutes proficiency at each level.

The development of such process descriptions would be facilitated if the various levels were benchmarked by tasks for which students scoring at each level have a high probability of success while students at lower levels have less likelihood of performing well. This is important because the interpretation and reporting of scores relative to performance standards require evidence, first, that tasks at a given scale level actually engage the knowledge and skill attributed to proficiency at this level and, second, that the performance of students at this level is validly characterized by the

process description. Thus, the construct validity of the performance standards and the construct validity of the measurement scale are inseparable.

The construct validity of the performance standards as well as of the measurement scale is vulnerable to threats of both construct underrepresentation and construct-irrelevant variance. For example, if acknowledged masters fail to meet the standard, one would suspect construct irrelevancy in the measure. Alternatively, if acknowledged nonmasters do meet the standard, one would suspect construct underrepresentation, that is, the nonmasters may be proficient in the assessed part of a sparsely covered domain but less proficient in the unassessed part. This latter situation is the bane of selection testing and of criterion prediction more generally, that is, some individuals do well on the domain processes covered in the predictor tests but perform poorly on unmeasured processes important in criterion performance. This problem of construct underrepresentation is critical in standards-based educational assessment. Even NAEP, with BIB spiralling, has trouble covering the important bases.

The identification of benchmark tasks and process descriptions is facilitated by development of the more powerful model-based measurement scales, to be sure, but much of performance assessment is limited to rudimentary scales that order performances in a small number of categories, such as the four- to six-point range typical of most scoring rubrics. Many scoring rubrics employ at least partly evaluative as opposed to descriptive labels for the performance categories, such as "undeveloped response" or "extensively elaborated response." In such cases, the rubric embodies a kind of primitive performance standard, at least for the task being evaluated.

In effect, the scoring rubric provides a score scale for evaluating task performance and, hence, a basis for setting performance standards for the particular task. At issue is whether or not the scoring categories have the same meaning across tasks, especially in the face of variations in task difficulty. Whether the same scoring rubric can be meaningfully applied to different tasks to generate a cross-task measurement scale depends on evidence of generalizability of the scoring rubric across tasks. Moreover, because the particular scoring rubric is usually only one among several that might just as well have been formulated, we should also inquire into generalizability across scoring rubrics.

These issues of generalizability are especially important for score interpretation and reporting because scoring rubrics are typically task-based rather than construct-based. That is, more often than not, scoring rubrics refer to aspects of a student's response or product, such as degree of elaboration or coherence, rather than to aspects of process or skill. Going from a task-specific interpretation to a construct interpretation of some generality and power requires evidence of generalizability.

For performance standards to be valid, the increasing achievement levels characterized by such terms as "basic," "proficient," and "advanced" -- as well as the tasks that benchmark these levels -- should reflect increases in complexity of the construct specified in the content standards and not increasing sources of construct-irrelevant difficulty. However, what constitutes construct-irrelevant variance is a tricky and contentious issue (Messick, 1994a,1994b). This is especially true of performance assessments, which typically invoke constructs that are higher-order and complex in the sense of subsuming or organizing multiple processes.

For example, skill in communicating mathematical ideas might well be considered irrelevant variance in the assessment of mathematical knowledge (although not necessarily vice versa). But both communication skill and mathematical knowledge are considered relevant parts of the higher-order construct of mathematical power according to the content standards delineated by the National Council of Teachers of Mathematics. The problem, as was previously mentioned, is to separate evidence of the operation of the focal construct from that of ancillary skills involved in task performance serving as potential sources of construct-irrelevant difficulty.

The concept of construct-irrelevant variance is important in all educational and psychological measurement, especially in richly contextualized assessments and so-called "authentic" simulations of real-world tasks. This is the case because, "paradoxically, the complexity of context is made manageable by contextual clues" (Wiggins, 1993, p. 208). And it matters whether the contextual clues that are responded to are construct-relevant or represent construct-irrelevant difficulty. Everything depends on how compelling the evidence and arguments are that the particular source of variance is a relevant part of the focal construct as opposed to affording a plausible rival hypothesis to account for the observed performance regularities and relationships with other variables.

To disentangle construct-relevant from construct-irrelevant variance, one must turn to the construct theory of the subject-matter domain, that is, to the best available integration of scientific evidence about 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 and other assessment formats.

If the theoretical sources of task difficulty are actually used as a guide for test construction, the resulting exercises or tasks should have some critical properties. In particular, their ordering and approximate placement on the measurement scales should be predictable. Empirical evidence that the actual scale placement of these tasks is predicted by theory-based indices of task difficulty then provides strong support for the construct validity of both the theory and the measurement scale -- for example, as was done for the prose, document, and quantitative scales in the National Adult Literacy Survey (Kirsch, Jungeblut, & Mosenthal, 1994).

Performance standards are central to standards-based education reform because they are thought to transform educational assessments into worthwhile educational experiences serving to motivate and direct learning. That is, because performance standards specify the nature and level of knowledge and skill a student should attain, the criteria of good performance should become clear to them, in terms of both how the performance is to be scored and what steps might be taken to improve performance. In this sense, the criteria of successful performance are transparent or demystified and hence should be more readily internalized by students as self-directive goals (Baron, 1991; Wiggins, 1993).

Evidence needs to be accrued, of course, that the performance standards are understood by students and teachers and that they indeed facilitate learning, because the meaningfulness or transparency of performance standards cannot be taken for granted. In particular, the meaningfulness of the

performance standards as applied to the assessment tasks should be appraised. Such evidence bears on the consequential aspect of construct validity.

Also of consequence is the possibility that common performance standards for all students may not uniformly serve as challenges for further growth. For some students they may represent hurdles or artificial barriers that channel educational experiences in ways that are not personally fulfilling, thereby limiting development in line with personal interests and values (Coffman, 1993). Those who learn different things at different rates may be consigned to failure because they do not learn the common things at the expected rate. Such potential adverse side effects need to be appraised because they bear on the very meaning of the performance standards as well as on their implications for educational policy.

VALIDITY OF STANDARD-SETTING

The meaning of content and performance standards also depends in large measure on the credibility of procedures used in setting the standards. Because standard-setting inevitably involves human judgment, a central issue is *who* is to make these judgments, that is, whose values are to be embodied in the standards. From the discussion thus far, it seems clear that informed judgments about content standards require knowledge of the subject-matter domain as well as of the students' levels of developing expertise. Hence, the group of judges should certainly include teachers and curriculum specialists, who are also appropriate for setting performance standards. An important question in a pluralistic society is who else should participate in the standard-setting process?

The more diverse the group of judges, of course, the less consistency should be expected in their judgments and the more difficulty in reaching consensus. With a heterogeneous group of judges, one should anticipate a range of disagreement around the consensus, which can be reduced in refined standard-setting procedures by feedback and discussion among the judges. This range of disagreement has been called a "consensus distribution" (Phillips, Herriot, & Burkett, 1994), which should be robust in being replicable over a variety of settings with the same mix of judges' backgrounds.

An important issue is not just the extensiveness of this distribution, but also whether it represents random variation around the consensus as opposed to consistently different value perspectives or points of view. If the latter, some means of accommodating diverse viewpoints needs to be considered to make consensus meaningful under conditions of pluralism (Messick, 1985).

Much of our discussion of the construct validity of performance standards has highlighted the need for convergent and discriminant evidence supporting the meaning of the measurement scale and, in particular, the nature of the cognitive processes entailed at each performance-standard level. Whether the levels themselves are set at the proper points is a most contentious issue and depends on the defensibility of the procedures used for determining them. That is, because setting these levels is inherently judgmental, their validity depends on the reasonableness of the standard-setting process and of its outcome and consequences, not the least of which are passing rates and classification errors.

For example, consider the reasonableness of the widely used Angoff (1971) method of standard-setting. In this procedure, expert judges are asked to

estimate the probability that a minimally competent respondent would answer each item correctly. The average estimate for each item provides a kind of minimum passing level for the item. These estimates are summed to determine a passing or cut-score for the test. Modified versions of the Angoff method are typically used to set nonminimum standards such as the basic, proficient, and advanced levels of NAEP. The reasonableness of such judgments clearly depends on the expertise of the judges, that is, they should be knowledgeable not only about the subject-matter domain but also about the performance of persons exhibiting various levels of proficiency in the field.

Other aspects of the reasonableness of this standard-setting process can also be addressed. For example, one could appraise the logical or internal consistency of the process by comparing the Angoff probability estimates for each item with the proportion of minimally competent respondents who get the item correct, such respondents being defined as those scoring at or just above the cut-score for the test (Kane, 1984). In one such appraisal, the results were modest but encouraging, as witness a correlation of .71 between the mean Angoff probability estimates for the judges and the mean performance of minimally competent respondents (DeMauro & Powers, 1993). However, for correlations between estimated and observed item difficulties, both by item and by judge, medians were in the low forties, which is consistent with other studies of subject-matter experts' only modest ability to estimate item difficulty and discrimination (Bejar, 1983).

One line of development at this point pursues methods to improve the precision and consistency of the judgmental estimates (e.g., Kane, 1987). Another line might be to identify vulnerabilities in the judgmental process and attempt to overcome them. For example, a major weakness of item-level

judgmental procedures such as the Angoff method occurs precisely because judgments are made at the item level for each item separately. When each item is considered in isolation, item-specific variance looms large compared with construct variance. This tends to distort probability estimates that are supposed to reflect minimal or whatever level of *construct* competence. This distortion might be reduced by requiring judgments of the probability of success on small sets of items where the construct variance would be more salient because it cumulates across items while the item-specific variance does not.

Another problem with item-by-item judgments is that they do not capitalize on the structure of interrelations among the items as does IRT scaling or other model-based approaches to developing measurement scales. Indeed, once a measurement scale is constructed, especially if exercises are benchmarked along the scale and validated process descriptions are formulated for various scale levels, standards could be set directly as points on the scale. This would involve judgments as to what level of process complexity (and of associated scaled exercises) is appropriate to performance at minimal, basic, proficient, or advanced levels.

Another weakness of standards based on item-level methods such as Angoff's is that they may not hold if the items are changed, although extrapolations are generally defensible to equated tests or item sets. In contrast, if standards are set as points on a measurement scale such as those based on item-response theory (either directly as just described or by combining judges' probability estimates for items calibrated to the scale), the standards should remain relatively invariant when calibrated items are

added or dropped from the set. As a consequence, such scale-level standard-setting is amenable to use with computer-adaptive as well as linear tests.

Moreover, if a well-developed theory of the sources of construct-relevant difficulty has guided test construction and if the resultant exercises fall on the scale in their predicted order and approximate expected placement, it may be possible, as was done in the National Adult Literacy Survey (Kirsch et al., 1994), to empirically delineate regions of the scale where construct processes emerge, differentiate, compound, hierarchically integrate, or otherwise become more complex. The empirical delineation of such scale regions then provides a rational basis for judges to set standards in terms of desired levels of process complexity for different grade levels and degrees of expertise.

Finally, the measurement scale can be elaborated by projecting onto it a variety of other behaviors, scores, and real-world tasks (Messick, Beaton, & Lord, 1983; Phillips et al., 1994). These might include ACT, SAT, Advanced Placement, or Regents Examination scores; achievement in math and science; and, skill in reading *TV Guide* or *New York Times* editorials as well as high school or college textbooks. In this procedure, which I refer to as behavioral anchoring, ***nonassessment*** tasks and scores are projected onto the scale. This is in contrast to benchmarking, in which ***assessment*** tasks mark particular points on the scale.

By adding behavioral anchoring to validated process descriptions, scale levels can be related to a variety of accepted norms, thereby giving policy makers and laypersons alike a better sense of what is implied by the scale levels and hence by standards set as points on the scale. With such information in hand, it becomes possible to open up the standard-setting process beyond the group of experts needed to make item-level judgments

(Messick, 1985). By focussing not on isolated items but on the ordered set of benchmark exercises, on the associated process descriptions, and on the implications of behavioral anchoring, meaningful standards judgments could be obtained from policymakers, parents, businesspeople, representatives of minority groups, and other stakeholders in education in a pluralistic society.

OVERVIEW

Because content standards specify what is important to teach and to learn, they provide blueprints for standards-based educational assessment. Because performance standards specify accredited levels of student accomplishment, they require a measurement scale of some type to characterize the location and meaning of those levels. Hence, the validity of content and performance standards cannot be separated from the validity of the assessment itself or of the measurement scale. Therefore, the validity of standards must be addressed in terms of the same criteria needed to appraise the validity of assessments generally. These include content, substantive, structural, generalizability, external, and consequential aspects of construct validity.

With these fundamental aspects of construct validity in mind, the validity of content standards was addressed in terms of their representative coverage of the construct domain and their alignment with students' cognitive levels of developing expertise in the subject matter. The construct validity of performance standards as interpreted score levels was addressed in terms of the extent to which they reflect increasing construct complexity as opposed to construct-irrelevant difficulty.

Operationally, performance standards are interpreted points on a measurement scale. At issue are both the proper placement of those points and their meaning in terms of the knowledge and skill entailed in performance at those levels. The meaning of the performance standards depends on the construct validity of the measurement scale. The appropriateness of their placement depends on the reasonableness of procedures used for setting them. The advantages of standard-setting based on scale-level judgments as opposed to the compounding of item-level judgments were explored, especially as they bear on opening up the standard-setting process to a pluralism of stakeholders beyond subject-matter experts.

In sum, it may seem that providing valid grounds for valid inferences in standards-based educational assessment is a costly and complicated enterprise. But when the consequences of the assessment affect accountability decisions and educational policy, this needs to be weighed against the costs of uninformed or invalid inferences.

REFERENCES

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 508-600). Washington, DC: American Council on Education.

Baron, J. B. (1991). Strategies for the development of effective performance exercises. *Applied Measurement in Education*, 4, 305-318.

Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, *7*, 303-310.

Coffman, W. E. (1993). A king over Egypt, which knew not Joseph. *Educational Measurement: Issues and Practice*, *12*(2), 5-8.

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

Cronbach, L. J. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity*. Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147-171).

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.

Ferguson, G. A. (1956). On transfer and the abilities of man. *Canadian Journal of Psychology*, *10*, 121-131.

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. (1984, April). *Strategies in validating licensure examinations*. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Kane, M. T. (1987). On the use of IRT models with judgmental standards setting procedures. *Journal of Educational Measurement*, *24*, 333-345.

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

Kirsch, I. S., Jungeblut, A., & Mosenthal, P. B. (1994). Moving toward the measurement of adult literacy. In *Technical report on the 1992 National Adult Literacy Survey*.

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. (1985). Progress toward standards as standards for progress: A potential role for the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice*, *4*(4), 16-19.

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

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

Messick, S. (1994b). *Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning* (ETS RR-94-45). Princeton, NJ: Educational Testing Service.

Messick, S., Beaton, A., & Lord, F. (1983). *National Assessment of* Educational Progress reconsidered: A new design for a new era (NAEP Report 83-1). Princeton, NJ: National Assessment of Educational Progress.

Mullis, I. V. S., Dossey, J. A., Owen, E. H., & Phillips, G. W. (1993). *NAEP 1992 mathematics report card for the nation and the states* (Report No. 23-ST02). Washington, DC: National Center for Education Statistics.

Phillips, G., Herriot, R., & Burkett, J. (1994). Issues in establishing technical guidelines for standards-based reporting (Draft). Washington, DC: National Center for Education Statistics.

Shepard, L. A. (1993). Evaluating test validity. *Review of research in education*, *19*, 405-450.

van der Linden, W. J. (1994, October). Standards for standard setting in large-scale assessment. Paper presented at the Joint Conference on Standard Setting for Large-Scale Assessments, Washington, DC.

Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappan*, *83*, 200-214.

Wiley, D. E. (1991). Test validity and invalidity reconsidered. In R. E. Snow & D. E. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach* (pp. 75-107). Hillsdale, NJ: Erlbaum.