VALIDATING STANDARDS-BASED TEST SCORE INTERPRETATIONS

EDWARD H. HAERTEL
STANFORD UNIVERSITY

WILLIAM A. LORIÉ[1]
CTB/MCGRAW-HILL

Abstract

Standards-based score reports interpret test performance with reference to cut scores defining categories like "below basic" or "proficient" or "master." This paper first develops a conceptual framework for validity arguments supporting such interpretations, then presents three applications. Two of these are serve to introduce new standard-setting methods.

The conceptual framework lays out the logic of validity arguments in support of standards-based score interpretations, focusing on requirements that the performance standard (i.e., the characterization of examinees who surpass the cut score) be defensible both as a description and as a normative judgment, and that the cut score accurately operationalize that performance standard.

The three applications illustrate performance standards that differ in the breadth of the claims they set forth. The first, a "criterion-referenced testing" application, features a narrow performance standard that corresponds closely to performance on the test itself. The second, "minimum competency testing," introduces a new standard-setting method that might be used when there is a weaker linkage between the test and the performance standard. The third, a contemporary standards-based testing application, proposes a new procedure whereby the performance standard would be derived directly from the specification for the test itself.

[1] Support for this research from The Spencer Foundation is gratefully acknowledged. The ideas and positions taken are the authors' own; no endorsement by the Spencer Foundation should be inferred. An earlier version of this paper was presented as "Standards for Student Work Versus Standards for Students: Conceptual Confusions in Standards-Based Reporting" in a symposium on "Educational Standards, Assessment, and the Search for 'Consensus'" at the April 2000 annual meeting of the American Educational Research Association. Helpful comments and suggestions by the editors and anonymous reviewers are gratefully acknowledged.

Validating Standards-Based Test Score Interpretations

Edward H. Haertel
Stanford University

William A. Lorié
CTB/McGraw-Hill

The Appeal of Standards-Based Reporting

Standards-based reporting has come into common use for educational assessments in the United States. Cut scores define performance levels for individual students, and the proportions of students at or above successive levels are reported for schools, states, and the entire nation. The broad appeal of such reporting is not surprising. Labels like "Basic," "Proficient," or "Advanced" seem to convey whether students are making satisfactory progress, whether schools are doing well enough, and in what subject areas students are most in need of improvement. In this standards-based era, it no longer seems sufficient to know whether annual scores are up or down; reporting in terms of quantified goals is called for to say how much better would be good enough. In the name of "ending social promotion," score thresholds are set to indicate which children need remediation and which are ready for the next grade level. And, at the individual student level, dozens of states have enacted requirements for high school exit examinations. Many of these standards have serious consequences; the potential for mischief is great if they are set capriciously. Because these cut scores define the decision rules according to which test scores are interpreted and used, the Standards for Educational and Psychological Testing state that "the validity of test interpretations may hinge on the cut scores" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 53).

Unfortunately, standards-based score reporting and student certification may be yet another case of political expectations outpacing measurement realities. It is arguable whether many such score interpretations are technically defensible. And while enthusiasm continues unabated, current standard-setting methods have been strenuously criticized (e.g., Berk, 1995; Glass, 1978; Jaeger, Mullis, Bourque, & Shakrani, 1996; National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment, 1993; Pellegrino, Jones, & Mitchell, 1999).

This paper is divided into two major sections, the first presenting a conceptual framework and the second applying that framework to three standard-setting problems. After introducing the idea of validity argument, the first section takes up the related requirements for (1) a performance standard that describes some quality or capability of examinees that the test in fact measures (defensible criterion-referenced score interpretation); (2) a performance standard that also embodies a reasonable and defensible expectation as to the proficiency level examinees should attain (defensible normative judgment); and (3) an appropriate process for determining a cut score. The second major section presents three applications. Of these, the first is "criterion-referenced testing" as that term was used in the 1970s, which offers perhaps a best case for some current standard-setting methods. The second is minimum-competency testing and the third is a contemporary standards-based application. These two final cases are used to develop alternative standard-setting methods intended to remedy some deficiencies we identify in existing methods.

The Conceptual Basis of Standards-Based Reporting and Standard Setting

Our conceptual framework begins by distinguishing the *performance standard* from the *cut score*. As these terms are used here, the performance standard is a description of what "meeting the standard" is supposed to mean. It is typically set forth in a brief text, perhaps a paragraph, saying what "basic," "proficient," "master," or some such designation is intended to signify. Thus, the performance standard accurately characterizes some examinees and not others. Performance standards acquire additional meaning whenever meeting the standard is the basis for some inference or action, for example, partial fulfillment of the requirements for a high school diploma or denial of promotion to the next grade level. If meeting the standard is required for high school graduation, for example, then "having sufficient (tested) skills to merit a diploma" becomes part of the description. The performance standard may be stated in the language of constructs (e.g., a requisite level of reading proficiency), but must reference observable performances of some kind, at least implicitly (e.g., somehow demonstrating adequate comprehension of texts at a given reading level, or being ready for the next unit of instruction).

The cut score is the operationalization of the performance standard (Kane, 1994, p. 426). It defines the minimum test score required for an examinee to be classified as satisfying the performance standard. If the test is scored by tallying the number correct, then the cut score is some minimum number correct. If the test is scored in some more complicated way, the cut score may be defined as a point on some numerical scale. The problem of standard setting is sometimes viewed as no more than the problem of choosing a cut score, with scant attention to the performance standard. It should be clear, however, that a standards-based score interpretation is not defensible unless the cut score and the performance standard correspond to one another. By and large, the performance standard should accurately describe examinees scoring at or above the cut score. Conversely, examinees who score below the cut score should in general fall short of the capabilities set forth in the performance standard. (Note that this distinction between the performance standard and the cut score is in contrast to the common practice of referring to a cut score itself, or to a location on a score scale, as a performance standard.)

Arguments and procedures supporting a performance standard and cut score may differ according to the breadth of the claim the performance standard sets forth. Trivially, if the performance standard referred to no more than the ability to score in a certain range on the test itself (or on parallel forms of the test), then nearly any cut score might be justified by little more than evidence that the test was reliable (i.e., that it would classify examinees as masters versus nonmasters with an acceptable degree of consistency). In practice, though, the performance standard always embodies a stronger claim, pertaining to capabilities for performance in nontest settings. The range of nontest settings addressed by the performance standard is referred to here as the *outcome domain* (Haertel, 1985). The performance standard describes some acceptable degree of proficiency with respect to tasks in the outcome domain. The domain of knowledge and skills sampled by the test is referred to as the *content standards*, which may be described in a curriculum framework or similar document. The earlier and more modest term *content domain* is still used for much the same notion. If the performance standard is clear and defensible, then in principle, the determination of the appropriate cut score is a technical problem, not a judgmental task (Reckase, 2000a). A score of X is deemed good enough *because examinees scoring at or above X can do Y* (with an acceptably high probability). Clearly, the closer the correspondence between the content standards and the outcome domain, the easier it will be to argue that a cut score on the test distinguishes between examinees who have versus have not met the performance standard.

In this paper, validity and validation for standards-based score reports and for standard-setting methods are framed using *validity arguments*. A validity argument is a set of related propositions that, taken together, form an argument in support of an intended use or interpretation. Validation of a test interpretation may be approached as construction and evaluation of a validity argument, investigating both its strengths and its weaknesses (Cronbach, 1988). Some parts of the argument may be well established or noncontroversial. Others may be subject to challenge. Because the strength of the argument as a whole depends on all of its parts, the validity argument helps to guide the allocation of resources for test validation. Additional evidence for a proposition already well established is of little value, especially if some other part of the argument is poorly supported (Cronbach, 1988; Crooks, Kane, & Cohen, 1996; Shepard, 1993).[2]

An Illustrative Validity Argument

Suppose that each spring, a school district administers a standardized reading test to all of its fourth-grade students. Only those scoring at or above the cut score are deemed to have met the district's minimum expectation for end-of-fourth-grade reading proficiency. A validity argument in support of this intended standards-based score interpretation would include the following five propositions. Each is given a brief label for ease of reference in the discussion to follow.

1. Content Standards. Elementary reading instruction in the district is guided by a clear description of the intended content of the reading curriculum. The description might be found in a state or local curriculum framework, for example.

2. Alignment. The content of the reading test is well aligned with the content standards.

3. Accuracy and Precision. Reading test scores accurately reflect examinees' reading proficiency. The test is sufficiently reliable and scores are not unduly influenced by motivation, test wiseness, or other extraneous factors.

[2] Kane (1992, 1994) distinguishes between the set of propositions that constitutes a proposed test interpretation, which he terms the *interpretive argument*, and the evidence and arguments warranting the propositions in the interpretive argument, which are the concern of test validation. We have chosen the term *validity argument* following Cronbach (1988) to refer to both the interpretive argument itself and warrants for the propositions it comprises. Shepard (1993), citing House (1980), has used *evaluation argument* in the same way as we use *validity argument*.

4. Performance Standard. There is an agreed upon conception of the minimum acceptable level of reading proficiency for students at the end of fourth grade. This is set forth in a paragraph describing an outcome domain (i.e., the kinds of materials children are expected to be able to read, for different purposes and in different contexts) and an expected level of proficiency with respect to reading tasks in that outcome domain (i.e., some ways students should be able to demonstrate comprehension of what has been read). The outcome domain closely matches the content standards.

5. Cut Score. The cut score is established so that passing (i.e., scoring at or above the cut score) generally indicates that a student is accurately described by the performance standard, and conversely.

The first two propositions, Content Standards and Alignment, imply that the test is properly designed to measure a clearly defined construct. The Content Standards proposition implies not only that the intended coverage of the test is well defined, but also that it is appropriate for the intended testing application. The Alignment proposition implies that the test is an adequate reflection of the content standards. Alignment addresses primarily concerns about construct underrepresentation (Messick, 1989). These propositions would be supported by a clear description of the domain of reading skills the test was intended to measure, evidence that that domain matched the curriculum framework used by the district, documentation of a sound test development process, and an evaluation that the development was successful.

The third proposition, Accuracy and Precision, addresses test reliability, as well as many factors that could distort the meaning of test scores. These are sources of construct-irrelevant variance (Messick, 1989), which need not be unpacked for present purposes. This proposition will hold only if the test is administered following standardized procedures and is adequately proctored; if examinees are motivated to show their best work; and if they are not excessively anxious. It further implies that the test is not unduly speeded; that item formats are familiar to examinees but the actual reading passages and test items are not; that items are not amenable to strategic guessing by test-wise examinees; that the test's vocabulary is at an appropriate difficulty level; that performance is not unduly influenced by students' prior background knowledge; and that the test is free from any evident bias for or against members of identifiable demographic groups. This list could be extended. Evidence from both theory-based analyses and empirical studies would be required to address these potential concerns.

The performance standard referred to in the fourth proposition states that there is "an agreed upon conception of the minimum acceptable level of reading proficiency for students at the end of fourth grade." Note that the performance standard embodies both a normative judgment and a substantive description.[3] The level of proficiency it describes is to be regarded as an appropriate minimum expectation. Anyone who accepts as valid a report that a fourth-grader has or has not met the standard implicitly accepts that judgment. The performance standard also describes what a proficient fourth-grade reader is able to do, and at least implies some range of situations

[3] We reserve the term "standards-based reporting" for test score interpretations that are based on cut scores *and* that imply some evaluation of the adequacy or acceptability of performance.

in which these proficiencies would be manifested. In this illustrative case, it is critical that these proficiencies be drawn from the content standards and be measured by the test.

These two components, normative and substantive, are present in any performance standard, and give rise to two broad requirements for any valid standard-setting procedure. First, the standard setting must be defensible as a legitimate exercise of authority by a duly elected or appointed individual or by a duly constituted group.[4] Following Kane (1994), we refer to this as a requirement for *procedural evidence of validity*. Second, the test score scale must be *criterion-referenced*.[5] It must be possible to describe what some given level of test performance implies about examinees' capabilities. This is just another way of stating that the performance standard describes what examinees above a certain point on the score scale know or are able to do. A criterion-referenced score interpretation is embodied in any performance standard.

Note that in this example, the performance standard defines only a minimum expectation. Classroom teachers, parents, or anyone else familiar with the particular histories and circumstances of individual children may expect higher performance of some than of others, depending, for example, upon their performance at the end of the previous year. Defining a minimum achievement standard is problematical (Glass, 1978), but is not logically inconsistent with the idea of higher expectations for some, perhaps for nearly all, pupils. The performance standard establishes a floor for the range of acceptable individual expectations.[6] It does not necessarily entail any reduction in the variability of learning outcomes.

The Cut Score proposition holds that the cut score has been set at a point where the large majority of pupils earning passing scores meet the performance standard, and the large majority of those scoring lower do not. A passing score by an examinee who does not meet the performance standard is referred to as a *false positive*, and a failing score by an examinee who does meet the standard is referred to as a *false negative*. The *false positive rate* is the conditional probability that an examinee will score above the cut score, given that that examinee in fact fails to meet the performance standard, and the *false negative rate* is defined similarly. Predicted performance on virtually any criterion measure is likely to increase gradually as a function of test score (Shepard, 1980). There will be little difference in the average performance of those scoring just below versus just above the cut score, wherever it is set. Nonetheless, it is meaningful to ask whether a given examinee is correctly classified, and to define the proper cut score as one that minimizes the probability of misclassification, or yields some desired ratio of

[4] The group empowered to set the standard is referred to in this paper as the *policy body*. The analysis would be the same if a single policy maker, rather than a policy body, wrote the performance standard. Most contemporary standard-setting methods also employ a group of *panelists*, who carry out a judgment task that informs the setting of the cut score.

[5] Following standard usage, we employ the term "criterion referencing" in two different, though related, ways. The more general usage refers to the interpretation of an examinee's score as showing directly what that examinee knows or can do, without reference to the performance of other examinees (i.e., in a manner that is not norm-referenced). Later, in presenting cases of standards-based score interpretation, one of the situations we analyze is "criterion-referenced testing," referring to a type of measurement-driven instructional management system popular in the 1970s and early 1980s.

[6] Cohen and Haney (1980) argue that minimum standards may be socially acceptable precisely because the majority perceive them as levels that can easily be exceeded. Thus, setting minimum standards for educational achievement does not threaten the relative advantage of higher-achieving pupils.

false positive to false negative misclassifications. The cut score proposition implies these misclassification rates are acceptably low.

False positives and false negatives may arise for several reasons. First, due to measurement error, observed scores give imperfect information about examinees' actual proficiency with respect to the construct the test measures. The third proposition, Accuracy and Precision, addresses measurement error and the linkage of test scores to the intended construct. If the third proposition is satisfied, then measurement error should be acceptably small. Methods for quantifying this source of error and analyzing its consequences are well established (e.g., Hambleton & Novick, 1973; Huynh, 1976). Second, there may be some mismatch between the construct measured by the test and the proficiencies described by the performance standard. If the outcome domain does not reach beyond the coverage of the content standards, and if the test is aligned to the content standards, then it should be possible for the test to show whether an examinee has satisfied the performance standard. These are the concerns of the first, second, and fourth propositions, concerning content standards, alignment, and the performance standard itself. Finally, even if the test is accurately measuring the proficiencies the performance standard describes, the cut score may be set too high or too low. This is addressed by the fifth proposition, concerning the cut score.

These five propositions, Content Standards, Alignment, Accuracy and Precision, Performance Standard, and Cut Score, would appear in some form in any validity argument we constructed for a standards-based interpretation of achievement test scores. Each of the five might be elaborated with a list of more specific propositions. These elaborations would vary from one testing application to another. In the latter half of this paper, such elaborations are illustrated for the Performance Standard and Cut Score propositions, in the context of three hypothetical examples.

We next turn to a more detailed examination of performance standards and of cut scores, taking up first, the requirement for a warranted criterion-referenced score interpretation; second, the requirement for procedural evidence of validity; and third, methods to establish defensible cut scores.

Performance Standards and Criterion-Referenced Score Interpretations

Performance standards describe some actual capabilities of examinees who meet the standard. If examinees above some cut score on a test meet the standard and those below do not, then the performance standard embodies a criterion-referenced score interpretation. As argued by Messick (1995), it is logically impossible that performance at or above some given cut score attest to a quality neither measured by the test nor related to test scores through established correlation. Because standards-based score reports can convey no more information about examinees than is available from their actual scores, it follows that defensible performance standards are logically dependent on a satisfactory criterion-referenced score scale (cf. Nitko, 1980, pp. 464-465).

Item Domains and Outcome Domains

Typically, the items on a test may be regarded as a sample from some *item domain*, defined as the (hypothetical or actual) set of all items that could be included on alternate forms of the test. This domain may be ordered or unordered, unidimensional or faceted. It may be described by an assessment framework, a test specification, or occasionally, by an item generation algorithm. In any case, it should be possible in principle to generalize statistically from performance on the test to performance on alternate tests constructed in the same way from other items in the domain. With few exceptions, however, intended interpretations of test scores go beyond proficiency in taking tests:

> Test developers may show great ingenuity in devising items that simulate the nontest situation of interest, but the intended universe of generalization nonetheless transcends the universe of admissible observations (Cronbach, [Gleser, Nanda, & Rajaratnam], 1972). Items showing soup cans of different sizes and prices, or referring to a facsimile of a driver's license application, do not automatically warrant any inference as to examinee behavior in the supermarket or at the Department of Motor Vehicles. Tests are settings for structured observations designed to provide an efficient source of information about attributes of examinees. Often, these are attributes that cannot be observed directly. The necessity of making inferences to a broader domain than the test directly samples brings a need for some deeper theoretical basis for linking test and criterion. (Haertel, 1985, p. 25)

The broader domain of performances to which the standards-based interpretation is intended to pertain is referred to here as the *outcome domain*. By definition, the item domain is a subset of the outcome domain. That is to say, each potential item may be regarded as a task, to be administered under specified conditions, that elicits some portion of the knowledge or skills the performance standard describes.[7] The outcome domain includes both test and nontest tasks that call for the same knowledge and skills. Unlike inferences from test performance to performance on the item domain, warrants for generalization from the item domain to the broader outcome domain cannot be purely statistical. They may be both logical and empirical. Psychological theory or analyses of task performances may indicate that test items and other tasks in the outcome domain should entail the same skills. Correlations between test scores and ratings of performance on nontest tasks representing the outcome domain, comparisons both of test scores and nontest outcomes for instructed versus uninstructed groups, or other patterns of evidence may support the inference that test scores are indicative of ability to perform in the range of settings comprised by a broader outcome domain.

Validity Arguments for Criterion-Referenced Score Interpretations

Four distinct bases for criterion-referenced interpretation are presented in this section. The basis may be the proportion of test items an examinee can solve, the kinds of items the examinee can

[7] This generalization must be qualified if there are multiple cut points. On NAEP, for example, there could be items that "advanced" but not "proficient" examinees were expected to answer correctly. Such items would be part of the item domain but not part of the outcome domain associated with the "proficient" achievement level description.

or cannot solve, the quality of the examinee's performance, or statistical associations between test scores and ratings of performance in nontest settings.

Nitko (1980) discussed the importance for criterion-referenced score interpretations of beginning with a well-defined item domain. An item domain is well defined if, given a description of the domain, it is clear whether any given test item is or is not included. Given a well-defined domain, the most straightforward criterion-referenced interpretation may be a *domain score* estimate. The proportion-correct score on the test is interpreted as an estimate of the proportion of items in the domain that an examinee could answer correctly. Domain score interpretations support claims like "a score of X on this test indicates that the student can spell Y percent of all English words under conditions corresponding to the test administration." This form of criterion-referenced interpretation is little used, not only because it says nothing directly about performance beyond answering test questions, but also because it relies on the assumption that the items on the test are a representative sample of the domain as a whole. For that reason, is best suited to item domains that can be exhaustively specified, or where items can be generated by some algorithm.[8]

Nitko (1980) further divided well-defined item domains into those that are ordered versus unordered, although that distinction is not always clear. Domain score estimates can be obtained with unordered domains. If the item domain has some structure, it is usually possible to arrive at richer characterizations than just the proportion of items an examinee can solve. Also, criterion-referenced interpretations with ordered domains may not require the assumption that the items in the test are statistically representative of the entire domain. Nitko (1980, p. 466) set forth a number of bases for ordering domains, such as "ordering based on prerequisite sequences for acquiring an intellectual or psychomotor skill." Ordered item domains are of particular interest in connection with standards-based score interpretations, because they may offer a basis for distinguishing the knowledge and skills required to answer items in different regions of a score scale. With ordered item domains, performance standards may indicate that students at successive achievement levels are able to solve increasingly more challenging kinds of items, or have mastered successively larger areas of knowledge and skill.

A difficulty that arises with this second form of criterion-referenced interpretation may be clarified by introducing a distinction between *substantive domain orderings* and *empirical item orderings*. Substantive domain orderings are based on characteristics of the items themselves, without any direct reference to actual examinee responses. If items can be classified into distinct varieties, there may be some rationale for ordering those varieties according to the sequence in

[8] Typically, test items would be a simple random sample or a stratified random sample of all possible items. In the case of spelling, one could create rules for generating test items from English words, and could stipulate the corpus of English words eligible for inclusion. Thus, there would be a one-to-one correspondence between words and potential test items, making statistical sampling straightforward. In the case of simple arithmetic problems, there might be some stratification by arithmetic operation, number size, etc., but with suitable rules, the item domain would be finite and again, rigorous sampling would be straightforward. Note that even if one could unambiguously classify any proposed item as inside versus outside of a domain, it would not necessarily follow that one could determine whether the set of items on a test was statistically representative of that domain. Unless either (a) items are generated algorithmically, as in these examples, or (b) the item domain is defined to consist of a finite pool of existing items, a rigorous justification for a domain-referenced score interpretation is problematical.

which learners are expected to master them, or according to their expected difficulty. Achievement test specifications ("test blueprints") may employ categories that define substantive domain orderings, for example. Characteristics such as reading vocabulary level or the number of steps required in the solution of a mathematical problem might also induce substantive domain orderings.

An empirical item ordering may be determined in any of several ways by patterns of examinee responses. In classical test theory, items may be ordered according to their difficulties (*p-values*, i.e., the proportions of examinees answering each item correctly). Using item response theory (IRT), items may be ordered according to their difficulty parameters, according to the level of examinee proficiency (theta) at which the correct-response probability reaches some threshold (e.g., a probability of 65 percent, referred to as an item's *RP 65* value), according to their conditional difficulty given some fixed proficiency level, or according to their average difficulty for some standard population. (Except in special cases, these various criteria will define different orderings.) Note that a substantive domain ordering is defined over all of the actual or potential items in the domain, whereas an empirical item ordering can be rigorously defined only for a set of actual items to which examinees have responded.[9]

Qualitative distinctions among examinees at successive performance levels may be based on substantive domain orderings, but the probabilities that a given examinee can actually solve various items will be more strongly related to an empirical item ordering. Mislevy (1998, pp. 52-53) observes that in the context of standard setting, when a rater's ordering of items according to their substantive characteristics differs from the items' empirically determined difficulty ordering, "this phenomenon is referred to as *rater inconsistency* ... implicitly pointing an accusing finger at the rater. However, ... [t]here is no logical necessity for these orderings to agree, and it need not signal problems with the judge's ratings if they do not." Possible explanations for such disagreements include discrepancies between actual and desired curricula.

We may suppose that a substantive domain ordering partitions the test items into *subdomains* ordered by difficulty, and that a performance standard states which kinds of items those meeting the standard can solve. A college mathematics placement test might include items on arithmetic, prealgebra, algebra, and calculus, for example (Kane, 1994, p. 434). In principle, it might be possible to find cut scores that distinguished between examinees probably able versus probably unable to solve successively more difficult kinds of items. Typically, though, the distributions of item difficulties for different subdomains will overlap. To the extent that items in subdomains corresponding to advanced skills prove easy or conversely, the accuracy with which the performance standard distinguishes those above versus those below any possible cut score will be diminished. Empirical item orderings are not fully determined by substantive domain orderings, although standard setting would be easier if they were.

[9] However, Mislevy, Sheehan, and Wingersky (1993) discuss cases where information about item characteristics may be quantified before items are actually administered. There is always some uncertainty about empirical item orderings, and that uncertainty is a matter of degree.

It is interesting to note in this connection that Nitko (1980, p. 466) considered "ordering [item domains] based on an empirically defined latent trait," but pointedly excluded IRT scaling as a sole basis for supporting criterion-referenced score interpretations:

> Crucial to the whole process is that the latent trait must link the test scores to the behavior domain in a way that permits a statement, not so much about the numerical value of an examinee's latent-ability parameter, but about the *behavior* that the numerical value represents. (Nitko, 1980, p.471, italics in original)

Wilson and Bock (1985) demonstrated the construction of a substantive domain ordering that began with an empirical item ordering for spelling words, then derived an equation predicting spellability based on observable features of the items. Their approach did yield a description of "the behavior that the numerical value represents," in that it could be used to predict an examinee's likely success in spelling words not on the test. Other approaches to deriving substantive domain orderings from empirical item orderings are described by Fischer (1973), Embretson (1993), Tatsuoka (1990), Zwick, Senturk, Wang, & Loomis (2001), and others.

It is an empirical question whether any given substantive domain ordering defines item subdomains that are sufficiently separated by difficulty to support useful standards-based score interpretations. On broad achievement tests we have examined, it has been difficult to discern subdomains corresponding to different ranges of item difficulty. This could be due to an insufficient number of items, or to inadequate subdomain descriptions, or to items that call upon too many bits of knowledge or skills to be described in a straightforward way. It may be that only careful test design, with planful groupings of items under subdomain descriptions, will permit justifiable criterion-referenced score interpretations of this kind.[10] We consider this form of criterion-referencing further in the final example in this paper.

A third distinct form of criterion referencing is based not on the proportion of tasks examinees can perform or on the kinds of tasks they can versus cannot perform, but on the quality of their performance of a fixed set or representative sample of tasks. Nitko (1980, pp. 467-469) gives numerous examples of "criterion-referenced achievement scales" including early examples of scale books for rating handwriting or drawing. Contemporary examples would include scoring rubrics for brief essays written in response to standardized prompts. The "Body of Work" standard-setting method (Kingston, Kahl, Sweeney, & Bay, 2001), described later in this paper, is based on judgments of this kind. Note that like the domain score estimate, this form of interpretation is most useful when the test tasks themselves are representative of the outcome domain. (For example, a sample of handwriting shows directly the quality of the respondent's handwriting; generalization to nontest settings appears relatively unproblematical.)

[10] If a test were being designed solely to support a standards-based interpretation with a single cut score, it would make little sense to include items that persons not meeting the performance standard were expected to answer correctly. Ideally, all of the items on such a test would elicit the knowledge and skills constituting the performance standard. In practice, for various reasons, performance standards typically imply different expectations for different items. There may be more than one cut score, with different expectations at different mastery levels. A test may be designed for uses in addition to standards-based reporting. A single performance standard may hold that examinees should be able to answer correctly nearly all of one kind of item but only a fraction of some other kind of item.

A final class of criterion-referenced score interpretations may be based on the empirical association between test scores and performance on nontest tasks representative of the outcome domain. This form of criterion referencing may make no direct reference to the item domain, and does not even require a well-defined item domain in Nitko's (1980) sense.[11] For example, the standard-setting method proposed later in this paper for a minimum competency test relates test scores to examinees' performance of specific real-world benchmark tasks. Other standard-setting methods, the "contrasting groups" method for example, rely on more global judgments of examinees' proficiency by persons familiar with their work, without specifying particular nontest outcome-domain tasks on which to base such judgments.

In the context of standards-based score interpretation, this form of criterion referencing is attractive, because judges may be better able to bring their values, knowledge, and experience to bear in evaluating nontest performances than test performances. However, empirical methods also have their problems. Because tests are not perfectly reliable and because nontest performances are influenced by unmeasured constructs, the joint distribution of test scores and nontest performance may differ across examinee subgroups or across settings. An empirically established association between test and nontest performance may also change over time, especially if stakes are attached to test performance. Judges may find more examinees proficient with respect to one relevant nontest performance than another. Finally, it may be nearly impossible in practice to define an outcome domain, to sample from that domain, and to devise criteria for evaluating performance on the sampled tasks (Davis, 1998, pp. 86-92).

Procedural Evidence for Validity

As already noted, a performance standard conveys both a substantive description of proficiency and a normative judgment that that is the level of proficiency that ought to be expected or required. Procedural evidence of validity addresses primarily the legitimacy of the normative judgment. This discussion of procedural evidence is organized around (1) the legitimacy and authority of those establishing performance standards; (2) the appropriateness of the performance standard for a clearly stated intended use; (3) the logic and coherence of the judgment process through which both performance standard and cut score are established; and (4) the appearance of fairness and objectivity.

Legitimacy and Authority of Responsible Parties

Standard-setting processes for large-scale educational testing applications typically involve three groups of actors. First is some policy body, which may initiate the standard-setting process and which typically retains the authority to adopt, modify, or reject the results of a standard-setting study. Second are technical experts, who are charged with designing and managing the standard-setting procedure (primarily, the determination of cut scores). Third are panelists, whose judgments about items or examinees are used to determine the cut scores that best represent the performance standards. Standards are most likely to receive broad endorsement if the standard-

[11] In other testing applications, e.g., employment testing, the relation between test scores and ratings of criterion performance may be summarized in "expectancy tables" giving the conditional distributions of the criterion score for different ranges of test scores.

setting process is initiated by duly elected or appointed officials or by an established professional organization; and if the standard-setting procedure is managed by an appropriate governmental body or by a disinterested organization. Panelists should be representative of stakeholder groups. The process whereby they are chosen as well as their individual qualifications should be documented. If any of the actors in the process are viewed as strongly partisan, then it should be clear that partisans of alternative viewpoints are equally represented. Legitimacy will be further enhanced if members of the public and representatives of other interested constituencies and organizations are given meaningful opportunities to present their views before the standards are adopted (Haertel, 2002).

Appropriateness of Performance Standard to Explicit Intended Use

The description or decision rule embodied in a performance standard and its associated cut score will not be regarded as legitimate if it appears unreasonable or has unreasonable consequences. If the characterization provided by the performance standard appears either trivial or overly stringent, or if failure rates are unacceptably high or unacceptably low, then the standards-based score interpretation will be called into question. For example, defensible performance standards that all students must meet in order to receive a high school diploma will differ from "world class" standards intended to motivate striving toward a distant goal. The tests on which those standards are set must also differ. World-class standards cannot be represented by cut scores, even high cut scores, on tests that do not adequately sample challenging content. Conversely, a fair and reasonable high school exit examination must include "only the specific or generalized content and skills that students have had an opportunity to learn" (American Educational Research Association, et al., 1999, p. 146).

It follows that before a performance standard is established, its intended use should be clear. Ideally, the range of potential consequences for individual test takers would be specified in advance, as would the timeline for preparing students to meet the standard. If standards-based score reports are to be used to judge schools or districts, it should be clear in advance how test scores will be aggregated and how performance relative to standards will be evaluated (e.g., by stating the proportion of a school's students expected to meet the standard each year). In practice, standards established for one purpose have been adopted for other uses. Low-stakes programs sometimes evolve into high-stakes programs, and standards on new tests may be based on standards from other tests, established for other purposes. Clear statements of the intended uses of score-based interpretations might help to prevent future misuse.

Logic and Coherence of the Judgment Process

Establishing performance standards and setting corresponding cut scores requires normative judgments and technical procedures. The performance standard embodies judgments about the degree of proficiency that merits some descriptor in the context of an intended test use. Setting cut scores usually involves technical procedures, but may also require interpreting the language of the performance standard. Finally, it may be necessary to weigh independent evidence bearing on the reasonableness or appropriateness of performance standards and cut scores. The manner in which these different judgments are organized, delegated, and integrated is referred to here as the judgment process.

Consider the following scenario. First, a policy body articulates a performance standard that is ambitious but vague. On close inspection, it says no more than that students ought to have solid mastery of the content they should have learned, at a level sufficient to enable (perhaps even assure) their future success in further schooling or in the workplace. Next, panelists employing a modified Angoff procedure (described below) rely on this standard in reaching judgments about the proportions of minimally proficient examinees who would answer each item correctly. These judgments are used by technical experts to derive a recommended cut score, which turns out to entail a very high failure rate. They report the recommended score and the projected failure rate back to the policy body. The policy makers weigh the consequences of such a high failure rate. They conclude that it is inconsistent with other evidence of students' academic success, and therefore is not reasonable. Faced with the problem of reconciling the performance standard, the results of the standard-setting procedure, and the independent evidence concerning reasonableness and appropriateness, the policy makers look for a technical fix. The experts inform them that they could make a post hoc adjustment to the cut score, reducing it by 1.96 times the standard error of measurement, leaving the language of the performance standard unchanged and avoiding any direct criticism of the standard-setting procedure itself. The policy body makes the recommended adjustment, and all breathe a sigh of relief.

This process is flawed for several reasons. First, there is a questionable connection between the language of the performance standard and the construct measured by the test. If the performance standard is not based on a sound criterion-referenced score interpretation, then the standards-based score interpretation cannot be defended. Second, the performance standard gives insufficient guidance to the panelists. It may be easier for a policy body to reach agreement on a vague performance standard than on one that makes visible some selection among alternatives, but a lack of clarity in the performance standard increases the cognitive complexity of the panelists' work, and in effect delegates responsibility for policy making to the panelists. With clearer guidance, panelists would carry out a more constrained, more narrowly technical task. Third, the process sketched in the preceding paragraph precludes an integrated consideration of outside evidence and practical constraints that bear on the reasonableness and appropriateness of the cut scores for a clearly stated intended use. Finally, post-hoc modification of the cut score without some corresponding amendment to its intended meaning further compromises the already strained argument for the intended standards-based interpretation. Policy deliberations are messy, involving competing interests and concerns. Rather than arriving at a standard and then seeking to reconcile that standard with other concerns, a logical and coherent judgment process would enable an integrated deliberation (Haertel, 2002).

Appearance of Fairness and Objectivity

In addition to the legitimacy and authority of responsible parties, the appropriateness of the standard to its intended use, and the logic and coherence of the judgment process, procedural evidence of validity encompasses any additional issues affecting perceptions of the standard-setting process. Ideally, the entire procedure will be publicly specified in advance, to avoid any appearance that the rules are being made up or changed as the process unfolds. If the policy body initiating the process retains authority to accept, modify, or reject the results of the standard-setting study, that should be made clear from the outset. Post-hoc adjustments will be more readily accepted if the possibility of such adjustments is spelled out in advance. Similarly, if there is a possibility that outliers among panelists' judgments may be modified or set aside, rules for doing so should be specified in advance.

The standard-setting procedure should be insulated from political manipulation. In addition to assuring that panelists are representative of all major viewpoints, the instructions and materials they are given should be balanced, not intended to manipulate them into choosing a higher or lower cut point than they might otherwise.

Finally, the implementation of the standard-setting procedure should be amply documented. The deliberations of the policy body should be recorded in sufficient detail to demonstrate that they gave due consideration to relevant information. Panelists should be given opportunities to report their degree of understanding of the process, their comfort with the tasks they are asked to carry out, and their level of confidence in the judgments they provide.

Setting Cut Scores

The published literature on standard setting methods focuses almost exclusively on methods of determining cut scores, with scant attention to the prerequisite performance standard. In the first brief description of the Angoff standard-setting method, for example, almost nothing was said about a performance standard existing apart from the cut score (Angoff, 1971, pp. 514-515). The framework developed in this paper suggests that this emphasis is misplaced. Finding an appropriate cut score is only possible if prior propositions in the validity argument are satisfied. The first question to ask about setting cut scores is whether there exists any cut score at all for the given test that will warrant the intended use or interpretation. If the performance standards is clear and sound, referring to an outcome domain aligned with the test's content standards, then determining the cut score, while still requiring judgment, will be a more nearly technical process, calling for the implementation rather than the determination of policy (Reckase, 2000a).

The Judgment Task

Numerous methods for determining cut scores have appeared in the published literature. Helpful reviews have been published by Shepard (1980), Livingston and Zieky (1982), Berk (1986), Jaeger (1989), and Reckase (2000a, 2000b). Nearly all of these methods involve a group of panelists who make some sort of judgments. The results of their judgments are then summarized to obtain an initial recommended passing score, which may be subject to further adjustments. At the heart of these methods is the atomic judgment process carried out by an individual panelist, which we refer to as the *judgment task*. Typically, multiple executions of the judgment task by each of multiple panelists provide the quantitative information from which initial cut scores are derived. Both the judgment task and the algorithm by which resulting judgments are combined to produce cut scores are defined as part of the standard-setting procedure. Additional elements of these procedures include definition of panelist qualifications and procedures for panelist selection and training. The instructions to panelists and methods for preparation of any exemplars ("benchmarks") they are to consult must also be specified. There may be rules or guidelines for multiple iterations of the judgment task, often with different kinds of feedback provided to panelists between rounds. Finally, the complete cut score determination process may include further adjustments to the initial cut scores derived from panelists' judgments.

In the judgment task itself, a panelist examines some entity, focusing on particular attributes of that entity, and compares it to a fixed standard. The result of that comparison may be a classification of the entity into one of a small number of categories, or an estimate of a probability or proportion. This judgment task is repeated for a collection of entities, each of which is compared to the same fixed standard. The judgment task can be described in terms of (a) the *judgment locus* (the pertinent attribute(s) of the entities judged); (b) the *judgment reference* (the fixed standard to which the judgment locus is compared or against which it is evaluated), and (c) the *judgment procedure* (how the judgment locus and the judgment reference are related and the form in which their judged relation is expressed). Table 1 presents the judgment locus, reference, and procedure for four typical standard-setting methods. Note that the products of these judgment tasks take different forms. They may indicate which items a minimally proficient examinee could answer correctly. They may quantify the conditional probability for each item of a correct response given borderline proficiency. They may classify examinees according to whether each meets the performance standard. Or, they may classify student test papers according to whether each does or does not demonstrate sufficient proficiency.

Insert Table 1 about here

Each execution of the judgment task yields a separate evaluation. Results of these evaluations are then combined according to some formula to arrive at a cut score. That process may require further decisions, such as the relative costs to be assigned to false positive versus false negative misclassifications. Judgments involved in selecting benchmark responses may be highly influential (Moss & Schutz, 1999).

Standard-setting methods are often classified according to the judgment locus. Thus, the Angoff or Modified Angoff methods, in which the judgment locus is a perceived property of an item, are now referred to as test-centered, item-centered or item-judgment methods, and the Contrasting Groups method, where the judgment locus is an attribute of an examinee, is referred to as an examinee-centered or person-judgment method (Kane, 1998; Livingston & Zieky, 1982, 1989; Jaeger, 1989)[12]. Methods in which the locus of judgment is an examinee response or pattern of responses form a third category, which may be termed performance-centered. The Body of Work method is performance-centered, as are Wilson and Draney's (1998, 2002; Draney & Wilson, 1998) "construct-referencing" approach, as well as methods in which panelists judge actual samples of student writing. Note that for these performance-centered methods, the criterion-referenced interpretation is based on an ordered domain of examinee responses to fixed tasks, rather than an ordering of tasks.

[12] In some earlier literature (e.g., Berk, 1986; Hambleton, 1980) what are now termed "item-centered" methods were referred to as "judgmental" and what are now termed "examinee-centered" methods were referred to as "empirical." Jaeger (1989, p. 493) argued in favor of the more recent terminology used here.

By way of illustration, the four methods in Table 1 are briefly described to ground the discussion to follow. The Bookmark standard setting method is also briefly introduced.

The Original and Modified Angoff Methods

William H. Angoff's original proposal for a standard-setting method[13] was as follows:

> A systematic procedure for deciding on the minimum raw scores for passing and honors might be developed as follows: keeping the hypothetical "minimally acceptable person" in mind, one could go through the test item by item and decide whether such a person could answer correctly each item under consideration. If a score of one is given for each item answered correctly by the hypothetical person and a score of zero is given for each item answered incorrectly by that person, the sum of the item scores will equal the raw score earned by the "minimally acceptable person." A similar procedure could be followed for the hypothetical "lowest honors person."
>
> With a number of judges independently making these judgments it would be possible to decide by consensus on the [cut scores] without actually administering the test. If desired, the results of this consensus could later be compared with the number and percentage of examinees who actually earned passing and honors grades. (Angoff, 1971, pp. 514-515)

In a footnote to the first of the two paragraphs just quoted, Angoff suggested that rather than saying which items the "minimally acceptable person" would answer correctly, judges could state the probability that each item would be answered correctly. This suggestion is incorporated into most procedures now referred to as "modified Angoff" methods.

The Angoff method has evolved since it was first proposed. Some of the modifications featured in one or another standard-setting exercise include elaboration of panelists' (judges') training, changes in panelists' instructions, incorporation of multiple rounds of ratings with feedback and discussion between rounds, segmentation of panelists into subpanels to provide replicated cut score estimates so that standard errors can be calculated, and the addition of post-task corrections, such as corrections for guessing when tests are multiple choice. Nonetheless, the judgment locus – perceived skill requirement of single test items – remains an essential aspect of the Angoff method.

Even though the original and modified Angoff judgment tasks may be experienced as quite similar by panelists, they differ sharply in the complexity of the judgment procedure by which the locus and reference are to be related. On its face, the original Angoff task is far simpler. A binary decision is required, as opposed to an estimate of a conditional probability involving hypothetical test takers. Cognitive psychological research on probability judgments gives little

[13] Livingston and Zieky (1989, p. 122, footnote) report that "Angoff ... attributed this method to Ledyard R. Tucker."

cause for optimism that panelists can accurately reach such judgments. That research also offers abundant evidence that panelists' expressed confidence in their judgments does not assure their accuracy (e.g., Tversky & Kahneman, 1993). In fact, panelists report little difficulty in carrying out such tasks, and generally do express confidence in their ratings (Hambleton, et al., 2000), with some exceptions (Pellegrino, Jones, & Mitchell, 1999, p. 168).

The Contrasting Groups Method

The contrasting groups method (Livingston & Zieky, 1982) requires panelists to classify examinees according to whether they have or have not met the performance standard. These judgments are made independent of any knowledge of examinees' test scores. Next, students are grouped by test score, and the proportions judged proficient versus not proficient are determined for each successive score level. The cut score is set at the level where roughly half the students are judged proficient, or at a level determined by the relative costs of false positive versus false negative misclassifications. Clearly, the contrasting groups method requires that panelists have some basis for judging examinees' proficiency, apart from performance on the test for which the cut score is to be determined. In a typical application, teachers might judge the proficiency of their own students. A potential strength of examinee-centered methods is that panelists' judgments can be based on direct observations of examinees' work on nontest tasks in the outcome domain.

The Body of Work Method

The Body of Work method (Kingston, Kahl, Sweeney, & Bay, 2001) requires panelists to evaluate entire booklets of answers to constructed-response questions. Thus, panelists can examine a more substantial body of an individual student's work than is possible with item-centered methods. Not only are multiple responses examined together, but each of these extended written responses is more informative than a response to a multiple-choice question. Each booklet is judged as to whether, taken as a whole, it shows that the student has met the performance standard. Standard setting proceeds in two phases. For each phase, panelists examine booklets purposively chosen according to their IRT scale scores. In the first phase, panelists examine a heterogeneous sample of booklets spanning a broad range of proficiency levels and classify them by proficiency level. The results of these classifications determine the quarter-logit or half-logit range within which each cut score is located. (Logits are the units of an IRT score scale.) In the second phase, panelists examine a much more homogeneous sample of booklets within the narrow range where a cut point is to be located. The proportion of these booklets classified as above versus below the cut point determines the cut point's precise location.

The Bookmark Method

In the Bookmark standard-setting method (Mitzel, Lewis, Patz, & Green, 2001), items are arranged according to an empirical item ordering, and panelists are asked to determine the point in the series that separates those items a borderline examinee should versus should not have mastered. This method might be classified as either item-centered or performance-centered, depending on how the judgment task is conceptualized. If the task is framed as one of examining each item in turn and comparing its apparent requirements to the proficiencies implied by the performance standard, then the method is item-centered. As discussed later in this paper, however, the present authors prefer to frame the Bookmark judgment task as one of selecting among an ordered series of hypothetical profiles of performance on the entire set of items in the test. This description casts the Bookmark method as performance-centered. Note that the bookmark method requires specifying what probability of a correct response to an item signifies "mastery" of that item. Reliance on this arbitrary specification complicates the judgment task, and has been cited as a weakness of the Bookmark standard-setting method (Lorié, 2001).

Selected Applications

Before considering validity arguments to justify current standards-based score interpretations, it will be helpful to begin with earlier and simpler applications of standards-based reporting. The discussion here is idealized. It should not be taken to imply that standards for earlier applications were in fact established in the ways described, nor that they were better justified than current applications (Glass, 1978).

Three hypothetical standard-setting applications are presented. The first case, that of standard setting for a criterion-referenced test (CRT), is perhaps a best-case scenario for familiar item-centered, examinee-centered, and performance-centered approaches. Note that the term "CRT" is used here to refer to a particular kind of test popular in the 1970s, in distinction to the more general usage in connection with criterion-referenced test interpretations. The CRT application is low-stakes, and features a narrowly defined outcome domain of tasks closely resembling the task of responding to the CRT itself. The second case, standard setting for a minimum-competency test (MCT), carries high stakes for individual examinees, and poses the problem of an ill-defined outcome domain. A new standard-setting process is proposed for this application, in which outcome-domain proficiency is operationalized by a set of representative benchmark tasks linked empirically to MCT performance. The third case is a contemporary standards-based testing application, which might also carry high stakes for individuals. Compared to the CRT application, the outcome domain is much broader. Compared to the MCT application, the outcome domain is more narrowly focused on academic performance and is better described by formally adopted content standards. Another new standard-setting process is proposed for this contemporary application, in which a set of alternative possible performance standards is derived from the item domain, each of which corresponds to a possible cut score. Information is organized to enable policy makers to make an informed selection among these alternative warranted performance standards, and its corresponding cut score follows immediately.

Criterion-Referenced Testing

The term "criterion-referenced test" (CRT) was introduced by Glaser (1963; Glaser and Klaus, 1962) for tests designed to support what are now termed criterion-referenced (as opposed to norm-referenced) score interpretations. Although Glaser's original conception of criterion-referenced testing did not give a prominent role to cut scores (Glass, 1978), within a few years the term "criterion-referenced test" came to be applied to tests supporting a form of measurement-driven instructional planning in which the knowledge and skills to be learned were analyzed into narrow, carefully sequenced units, and each child was tested to assure mastery of all prerequisite units before progressing to a new unit (e.g., Bloom, 1976). Cut scores defining "mastery" on CRTs guided decisions about the pacing of instruction. This instructional approach, rooted in behaviorist psychology, was controversial even at the height of its popularity (Rutherford, 1979; Schmidt, 1982), and is no longer generally accepted (Resnick & Resnick, 1992; Shepard, 2000). Nonetheless, it can serve as a useful illustration.

Consider a criterion-referenced test of, say, two-column subtraction without "borrowing" (regrouping). The test might consist of twenty problems like "76 - 35 = ?". This test might be used at the end of a brief learning unit to determine which children were ready to proceed to the next unit, on two-column subtraction with borrowing, learning to solve problems like "75 - 36 = ?".

Validity Argument

Returning to the five propositions of the illustrative validity argument, the Content Standards (or content domain) for the CRT would be stated or implied by the curriculum. It would comprise the algorithm to be taught and learned for solving two-column subtraction problems without borrowing. Assuming that the problems on the CRT were a representative sample of the roughly 3000 arithmetic problems of this kind, the Alignment proposition would be satisfied. The Accuracy and Precision of the test would depend on adequate test length, appropriate administration conditions, and of course, avoidance of test questions identical to specific subtraction problems that were directly taught.

The Performance Standard is implied by the intended use. A passing score is intended to signify that a student is ready for the next unit of instruction, whereas a failing score should imply at least a moderate chance of difficulties during the next unit due to deficiency in the tested prerequisite skill. The outcome domain, instructional tasks to be encountered during the next learning unit, comprises activities similar to test taking itself. Thus, there is little need for extrapolation from test performance to performance in nonacademic settings. The intended use makes this a low-stakes test. The consequences of misclassifications are minor, and test-based decisions are easily revisited. Passing merely signifies readiness for the next instructional unit, and failing merely signifies a need for more practice with a particular, narrow skill. For these reasons, procedural evidence of validity is relatively unimportant.

Determining a Cut Score

The cut score might be determined by any of several methods, built around alternative judgment tasks. An *examinee-centered* approach might proceed as follows. First, the test would be administered to a representative group of students at the point in the instructional sequence where the CRT was to be used. Next, these students would be taught the subsequent learning unit. Teachers would be unaware of students' CRT scores. Finally, teachers would classify students as successful or not successful in mastering the subsequent unit. A cut score would then be chosen to maximize agreement between CRT-based pass-fail classifications and teacher judgments.

Using an *item-centered* approach, teachers familiar with the students and the curriculum at the grade level tested could examine each test item and judge the probability that a child who had barely satisfied the performance standard would respond correctly. Equivalently, they might be asked to judge the proportion of correct responses to the item by a group of such children. These judgments of conditional probabilities would be aggregated across items and across panelists to determine the cut score. Note that this judgment task requires teachers to inspect a test item and accurately predict the performance of a borderline examinee (or a group of such examinees) on that item. They must be presumed to be able to express their predictions as unbiased numerical judgments of the probabilities of a correct responses by borderline examinees.

As already discussed, this judgment task seems unduly complex. In this particular situation, however, there may be an alternative argument in support of the original Angoff judgment task, involving binary classifications of items rather than estimates of conditional probabilities. Classroom teachers are likely to have had direct experience with children's performance on actual tests and test-like classroom activities; and the intended use of the CRT implies an outcome domain largely limited to such activities. Arithmetic problems are likely to be constructed-response, so modifications of the cut score to account for guessing are not necessary. Moreover, the performance standard is not constrained by the actions or statements of any policy body. It does not directly describe the skills of examinees showing mastery, but instead defines mastery indirectly, by reference to success in the subsequent learning unit. Thus, teachers carrying out the Angoff judgment task might not actually need to first form a conception of a borderline examinee and then mentally simulate such an examinee's performance on each item. Instead, the conception of the minimally proficient examinee might in effect be built up through the process of reaching item-by-item judgments. Under this alternative interpretation, teachers setting a cut score for a CRT may be a best case for Angoff's (1971) original method. They might simply examine each item and decide whether failing that item would indicate weaknesses serious enough to cause difficulties in subsequent instruction. If so, then passing that item would be required for mastery. The total number of items so identified would equal the cut score.[14]

A *performance-centered* approach might have teachers examine test papers completed by a representative group of students after appropriate instruction. They would decide whether each test paper demonstrated sufficient proficiency for the examinee to proceed to the next unit of instruction. In reaching these judgments, teachers might attend to more than just the number correct; different kinds of errors might imply different instructional remedies. One paper might indicate a partial failure to master still earlier prerequisite skills (e.g., systematic errors with certain simple subtraction facts). Another paper might show a pattern of general carelessness or inattention. Still another might show that a child performed well on the first part of the test, but then ran out of time. A paper with wildly incorrect errors might indicate a more fundamental

[14] It might be possible to extend this argument to cover judgment tasks in which teachers provided probabilities, as opposed to binary decisions, but the procedure for deriving a cut score from teachers' judgments would be more complex than the one currently used. If teachers were able to judge the probability that students unable to solve a given item were below the threshold of adequate preparation, then Bayes Theorem might be used to derive the probabilities that students below that threshold would fail to solve each item. Subtracting these latter probabilities from one would give the probabilities that students below the threshold would succeed in solving each item. These values could then be summed to obtain a lower bound to the cut score.

lack of understanding. Thus, distributions of number-correct scores might overlap for the groups of students a teacher classified as not ready versus ready to proceed to the next unit. The number-correct cut score best distinguishing between these two groups of papers would be taken as that teacher's preferred cut score, and the median of these preferred cut scores across teachers could be taken as the performance standard.

Minimum Competency Testing

By the mid-1970s, the popularity of criterion-referenced testing was waning, and the minimum competency testing movement had begun (Linn, 2000, p. 6). Beginning in the early 1970s and continuing into the 1980s, most states enacted legislation requiring public-school students to demonstrate mastery of basic skills prior to high school graduation, usually by passing a Minimum Competency Test (MCT). Thus, failure to pass the MCT could result in denial of a high school diploma, even if all other graduation requirements were satisfied. The validity argument proposed here suggests a procedure for determining and validating MCT performance standards and cut scores that is unlike any process historically used for that purpose by a state or district. It might be possible to develop sound arguments in support of other procedures.

Justifying standards for MCTs is more challenging than for CRTs, for three reasons. First, the stakes are higher, so procedural evidence of validity assumes greater importance. Second, the outcome domain is not well defined. Legislation might allude to basic skills; basic competencies; basic reading and mathematics proficiencies required for success in adult life; reading at the eighth-grade level; consumer mathematics; or ability to compute with whole numbers, fractions, and decimals, for example, but would typically not provide any clear specification of either the precise skills to be covered or the range of situations in which those skills were to be proficiently applied. Third is the related problem that the (typically) multiple-choice MCT itself is less representative of the intended outcome domain than was the case for the CRT.

Standard-setting methods for the CRT application capitalized on the facts that the outcome domain, which consisted of instructional activities in the next curricular unit, was well defined and that the activities included in the outcome domain were much like the activity of answering CRT test questions. This meant that the requirement for a criterion-referenced score interpretation was easy to satisfy--The items of the CRT were like a sample of actual tasks in the outcome domain. Thus, it was plausible, for example, that teachers could reach sound judgments about the readiness of children who missed particular test questions to proceed to the next learning unit.

For the MCT application, a different argument is proposed, which constructs a criterion-referenced score interpretation by investigating empirically the relation between MCT scores and performance on benchmark tasks representative of the outcome domain. As part of the process of standard-setting, the outcome domain is more clearly defined by the selection of those benchmark tasks. The performance standard is defined in the course of deliberations about the criteria of adequate performance on each benchmark task and about the proportion of benchmark tasks minimally competent examinees should be able to complete successfully. By providing a forum for deliberation about the meaning of minimum competency in real-world settings, the proposed procedure develops procedural evidence of validity.

Validity Argument

The Content Standards for the MCT would be based on state curriculum frameworks, but might comprise material that was to be taught at earlier grade levels. By law, the MCT could not include material test takers had not had an adequate opportunity to learn.[15] Alignment of the test to the content standards would be established through appropriate (and appropriately documented) test construction. Accuracy and Precision are discussed at greater length following the presentation of the proposed standard-setting procedure. Because the proposed procedure would generate information about the correlation between MCT performance and benchmark task performance, it would be possible to quantify both misclassifications due to measurement error and misclassifications due to the imperfect correspondence between the construct measured by the MCT and the demands of performances in the outcome domain. The Performance Standard and Cut Score propositions are addressed in tandem. Under the proposed procedure, once the performance standard is chosen, the corresponding cut score follows immediately.

Benchmark Task Standard-Setting Procedure

The outcome domain for the MCT includes a variety of real-world applications of basic academic skills. It would be expected that different examinees would experience varying degrees of difficulty with these tasks. Some might do relatively better on one set of tasks and others on another set. Nonetheless, agreement might be reached on a definition of minimum competency with respect to the entire domain, as follows. First, a set of benchmark tasks would be specified, which taken together could be considered representative of the domain. Second, for each of these tasks, a rubric would be developed for rating performance as satisfactory versus unsatisfactory. Finally, the policy body would determine what proportion of successful benchmark task completions should be taken to define minimum competency. This proposed procedure shifts the locus of judgment from success at answering multiple-choice questions to success in performing concrete, real-world tasks. It replaces the one large problem of defining proficiency with respect to a diffuse, ill-defined outcome domain with the many smaller problems of first clarifying the domain definition by selecting the set of benchmark tasks, next defining success on each benchmark task, and finally deciding on the proportion of tasks that must be completed successfully. Each of these smaller component tasks should enable judges to bring their practical knowledge, experience, and values to bear in a principled fashion.

Once benchmark tasks were identified and pass/fail rubrics were created for each task, studies would be undertaken in which examinees took the MCT and also performed one or more of the benchmark tasks. Logistic regression or some other statistical procedure would be used to model the probability of successful task performance as a function of MCT score. It would then be straightforward to model the expected proportion of all benchmark tasks completed successfully as a function of MCT score. Thus, once the policy body determined the proportion of benchmark tasks defining mastery, the MCT score predicting that degree of success would be

[15] Debra P v. Turlington, 644 F. 2d 397 (5th Cir. 1981).

adopted as the cut score. It would never be necessary for any one examinee to perform all of the benchmark tasks. Once standard setting was complete, pass-fail judgments would be based on MCT scores alone.

Identification of benchmark tasks might proceed as follows. Representatives of stakeholder groups including employers and the general public would be surveyed to identify actual out-of-school tasks requiring the kinds of basic reading and mathematics proficiency they would expect of any high school graduate. Technical experts would assist in selecting a representative subset of the proposed tasks. Task selection would proceed iteratively, with revisions guided by feedback from policy makers and stakeholder representatives until consensus was reached.

Next, a performance assessment would be derived from each benchmark task. Assessment materials, instructions for task administration, instructions to examinees, and scoring criteria would be developed. Scoring rubrics would only specify the minimum requirements for satisfactory performance; there would be no need to differentiate further among those passing or among those failing. It might be that stakeholders proposing particular benchmark tasks would take a leading role in defining adequate performance on the corresponding assessment tasks. Scoring rubrics would emphasize the successful application of basic skills in benchmark task performance. The resulting catalog of benchmark assessments might exist in two versions, one secure, including specific materials to be used in each task (e.g., specific text passages), and the other public, in which specific materials were replaced with sufficiently general descriptions to maintain test security.

The third step would relate performance on each assessment task to performance on the MCT. Groups of high school students representing a broad range of proficiency levels would be asked to perform each task, and their performances would be videotaped or otherwise recorded for later rating as satisfactory or unsatisfactory. Participating students would also take the MCT. Task performance ratings might have considerable error, although reliability could be improved by having multiple raters judge the adequacy of each performance. However, high accuracy at the level of individual task judgments would not be critical, because these judgments would be used only to establish the empirical relation between MCT scores and the probability of successful task completion. The results of these empirical studies might be summarized in a table with a row for each possible MCT score and a column for each benchmark task. Each table entry would be the estimated probability that a student with that row's MCT score would successfully perform that column's benchmark task. The row sums would give the number of benchmark tasks a student with the corresponding MCT score would be expected to complete successfully.

The final step would be to specify the performance standard, expressed as a proportion of outcome-domain assessment tasks representing minimum competency. At this point, all relevant information would be brought together, enabling a rational and coherent deliberation. The meaning of possible performance standards would be clarified by reference to the public version of the catalog of assessment tasks; members of the policy body might also examine the secure version. In addition to information relating MCT performance to outcome domain proficiency, the policy body would also be provided with information derived from the actual distribution of MCT scores from a representative sample of examinees. Thus, for each possible cut score, in addition to benchmark task performance, they would also know the likely passing rate, both overall and for major demographic subgroups. The policy body might also be provided with information concerning expected misclassification probabilities at each possible cut score, as described below. The policy body's decision would simultaneously establish both the performance standard and the cut score.

Accuracy and Precision

The third proposition of the general validity argument concerns the accuracy and precision of score-based pass/fail classifications. Classification accuracy will depend on several factors. Obviously, the reliability of the MCT itself is one consideration. A second is the strength of the correlation between MCT scores and the expected number of outcome-domain benchmark tasks completed successfully (i.e., the validity coefficient). If predicted outcome domain performance changes only gradually with increasing MCT scores, then misclassifications will occur more frequently. Finally, classification accuracy will depend on the base rate (proportion of masters in the tested population) and the passing rate (proportion of examinees who pass the MCT). The passing rate should be set equal to the base rate only if the relative costs of false positive and false negative classifications are equal. With the MCT, the cost of denying a diploma to a minimally competent examinee may be judged to exceed the cost of granting a diploma to an examinee whose basic skills fall short of the performance standard. For that reason, the passing rate may be chosen to exceed the base rate.

Table 2 illustrates the accuracy of an MCT under several scenarios.[16] In these illustrations, it is assumed that the observed test score and the outcome domain criterion variable have a bivariate normal distribution, with a correlation given by the validity coefficient for each row. The validity coefficient is less than one due to both unreliability in the MCT and the imperfect correlation between the construct measured by the MCT and the corresponding proficiencies required in benchmark tasks. In the top half of the table, it is assumed that the lowest two percent on the proficiency continuum fell below the performance standard (i.e., the base rate equals 98 percent actually meeting the performance standard). The cut score on the MCT results in a 98 percent passing rate. With these assumptions, statistics are presented for MCTs with validity coefficients of 0, .50, .80, .90, and .95. Note that even with a validity coefficient of zero, a base rate of 98 percent would result in over 96 percent of the examinees being classified accurately as masters versus nonmasters. Also, as shown in the column labeled P(Fail | Truly Proficient), only one in fifty examinees who deserved to pass would be denied a diploma unjustly. On the other hand, the column showing P(Truly Proficient | Fail) shows that of those who failed the MCT, 98 percent would deserve to pass. With a validity coefficient of 0.5, 96.7 percent would be classified accurately, but of those who failed, almost five sixths would deserve to pass. Even with a validity coefficient of .95, almost a third of the students denied a diploma would deserve to pass, and the same fraction of true nonmasters would receive passing scores.

[16] Calculations of areas under the bivariate normal distribution were carried out using Mathematica® version 3.0. Note that these calculations are for a single test administration. They do not take account of opportunities for repeated testing. Retakes would increase false positives and decrease false negatives.

Insert Table 2 about here

It is unlikely in practice that an MCT would be sufficiently reliable and sufficiently aligned with outcome-domain proficiency to yield a validity coefficient higher than .80. Under the scenario in the top half of Table 2, with a validity coefficient of .80 it would be expected that more than half of those students denied a diploma in fact met the performance standard. The lower half of Table 2 presents results for the same calculations assuming a base rate of 95 percent and a passing rate of 98 percent. That is to say, even though five percent of examinees are assumed to fall below the standard, only the bottom two percent are to be denied diplomas. Now, with a validity coefficient of .80, over two-thirds of the examinees failing the MCT in fact fall below the proficiency standard. However, this improvement is purchased at the cost of a much higher false-positive probability. Only about a quarter of those truly not proficient would fail the MCT. Note that with a base rate of 95 percent and a failure rate of two percent, the maximum possible correct classification rate is only 97 percent.

MCT scores might correlate with outcome domain performance not only due to common underlying skills, but also due to common dependence of MCT scores and benchmark task performance on other unmeasured variables. Thus, as schools made direct efforts to improve test performance, the empirical relationships between test scores and criterion task performances might change over time. Even if such a costly standard-setting effort were undertaken, the meanings established for passing MCT scores might change within a few years. However, the fact that MCT validity would erode is a criticism of the MCT as a policy, not a defect of the proposed standard-setting method. The library of performance assessments created in this standard-setting exercise could be used in future research to quantify any such erosion of validity.

Contemporary Standards-Based Testing Applications

The current standards-based reform movement in the United States may be dated from three events in the late 1980s and early 1990s. First was the release by the National Council of Teachers of Mathematics (NCTM) of the *Curriculum and Evaluation Standards for School Mathematics* (NCTM, 1989), which popularized the idea of standards as instruments of educational reform. Second was the creation of Basic, Proficient, and Advanced achievement levels for the National Assessment of Educational Progress (NAEP), by the National Assessment Governing Board (NAGB). Achievement levels were established first in reading and mathematics and later in five additional content areas at the fourth, eighth, and twelfth grade levels. NAGB policy made the achievement levels the primary vehicle for reporting of NAEP results (Vinovskis, 1998). Third, following the creation of the National Education Goals Panel, a National Council on Education Standards and Testing (NCEST) was created by an act of Congress and charged with advising "on the desirability and feasibility of national standards and tests," and recommending "long-term policies, structures, and mechanisms for setting voluntary education standards and planning an appropriate system of tests" (NCEST, 1992, p. 1). The NCEST report popularized the terminology of content standards, performance standards, and opportunity-to-learn standards, and recommended the establishment of national tests with performance standards indicating how good was good enough.[17]

In the decade since, the rhetoric of educational standards, standards-based reform, and standards-based testing has become so commonplace in the United States that it may be surprising to recall how recently it first was heard. In addition to the NAEP, nearly all state testing programs are standards-based. Over half the states have implemented or are in the process of implementing high school exit examinations, with promises that unlike the MCTs of twenty-five years ago, these new tests will demand students' mastery of challenging academic content. The recently enacted "No Child Left Behind Act" (P.L. 107-110) links receipt of certain federal education funds to states' compliance with testing requirements, including annual standards-based testing in reading and mathematics of all public-school students in grades three through eight.

Most of these new standards-based initiatives are moving away from basic skills toward high or "world-class" standards. At the same time, compared to MCTs, they are focusing more narrowly on academic achievement and preparation for further schooling. There is less talk of real-world applications of skills or of workplace readiness than there was for MCTs. The intended content of the new standards-based tests is set forth in states' academic content standards. It is generally understood that content standards should be developed first, and that new tests should then be constructed in alignment with them. This ideal may be very imperfectly realized, but developing new tests is preferable to retrofitting standards to existing tests.

As with Minimum Competency Testing, the central problem in standard setting for these recent applications is developing a defensible criterion-referenced interpretation of test scores. In the proposed MCT method, this was done by relating test scores to benchmark performances representing the outcome domain. For recent standards-based applications, an approach is proposed that derives criterion-referenced score interpretations from an analysis of the item pool itself, with no reference to any separate outcome domain. This approach capitalizes on the fact that the outcome domain for these applications is intended to match the content standards on which the test is based. In other words, the item domain and the outcome domain are derived from the same source. Like the MCT approach, this proposed standard-setting method features informed selection from among a series of possible performance standards. Each possible performance standard provides a defensible description of test performance at or above a specific cut score. Thus, when a performance standard is chosen, the corresponding cut score is entailed. Unlike the widely used standard-setting methods discussed earlier, this method does not require a separate panel to carry out a judgment task.

Validity Argument

Following the logic of the 1992 NCEST report, standards-based reform is commonly understood to begin with the development of content standards, which guide curriculum and instruction, test content, and the language of performance standards. In practice, however, even if performance standards and tests are based on the same content standards, the performance standards may

[17] The NCEST usage of "performance standards" encompassed both performance standards and their associated cut scores, as the terms are used in this paper.

describe proficiencies the test does not actually measure. This is because content standards tend to be broad in their coverage, are often imprecise in their descriptions of expected knowledge and skills, and include material not amenable to measurement within the constraints of large-scale testing programs. Thus, the Alignment proposition is only partially satisfied. Although everything in the test may be found in the content standards, the converse is unlikely to hold.

Rather than relying on the content standards to link the test and the performance standard, the method described here instead derives a criterion-referenced score interpretation from the test itself. First, descriptions are generated of the proficiency demonstrated at each successive raw score. Then, one of these descriptions is chosen as the performance standard, thereby determining the cut score, as well. The performance standards developed by this approach describe proficiency in terms of the proportions of items of different types that examinees at a given level would be expected to solve correctly. This form of description affords a rigorous treatment of accuracy and precision, using IRT.

Continuum of Defensible Performance Standards

In order to develop criterion-referenced interpretations of test scores, the proposed procedure begins with the construction of a *continuum of performance* (Lorié, 2001). This consists of a series of performance descriptions akin to the anchor level descriptions used in reporting NAEP results roughly between 1986 and 1992. NAEP anchor level descriptions were created by first choosing an arbitrary series of numerical score levels (*anchor levels*) on a NAEP proficiency scale. Next, all items were identified that appeared to discriminate between examinees at one anchor level (e.g., 300) and the next lower level (e.g., 250), based on the estimated difficulties of the items for examinees at each level. Content experts then examined the identified items and wrote brief characterizations of the distinguishing features of performance at each successive level.

The procedure for constructing a continuum of performance is more formal, and yields a larger number of distinct descriptions -- one for each possible number-correct score. It begins with the mapping of each item onto a specific location on an IRT-based proficiency continuum. Next, items are classified into categories (*item types*) according to the kinds of knowledge and skill each appears to require.[18] Finally, potential performance standards are stated in terms of the proportions of items of each type that a borderline examinee should be able to solve correctly.

Linking Item Types to a Continuum of Performance

In thinking through the challenge of mapping item types to regions on a continuum of justified performance standards, Lorié (2001) reasoned that the problem can be reduced to that of mapping items onto a test's latent trait scale. The way that such item mapping has been done in the past, however, is not fully satisfactory. Using IRT, examinees' abilities can be represented by locations on a latent trait scale, and the probability of a correct response to any given item can

[18] Although the definition of item types may be informed by psychological theory, classifications of items into types are based on surface features. It is not necessary to develop accurate analyses of the knowledge and skills different item types actually require, and it is not necessary to assume that different item types measure nonoverlapping sets of skills.

be predicted for each possible latent trait scale location. The problem is that the predicted probability of a correct response changes smoothly as a function of scale location. Thus, in order to identify an item with a specific point on the latent trait scale, item mastery must be defined in terms of some correct response probability. Each item might be mapped to the point where examinees have a 50 percent chance of answering it correctly, for example. Alternatively, each item might be mapped to the scale location where examinees have an 65 percent chance of answering it correctly (the item's *RP 65* location), or to some other probability value. This is referred to as the problem of mastery probability indeterminacy (Kolstad, 1996; Kolstad, Cohen, Baldi, Chan, DeFur, & Angeles, 1998). Arguments have been made for the appropriateness of different mastery probability conventions, ranging from .5 to 1, but no argument is fully convincing. As a practical matter, adopting 1 as the mastery probability convention makes mapping individual items using IRT impossible, since that convention would assign all of the items a scale location of plus infinity.

Instead of mapping individual items, Lorié (2001) constructed a specific *pseudoresponse pattern* corresponding to each item, and then mapped these pseudoresponse patterns onto the latent trait scale in a manner that did not require adopting any arbitrary mastery probability convention. To find the location of a given item (the reference item) on the latent trait scale, a response pattern is constructed in which that item and all easier items are answered correctly, while all more difficult items are answered incorrectly or at the chance level. This is referred to as the canonical pseudoresponse (CPR) pattern for the reference item.[19] The estimated ability of a hypothetical examinee who produced that pattern of responses is taken as the canonical pseudoresponse scaling (CPRS) location of the reference item. CPRS locations depend to some extent on the set of other items included with the reference item, and on the empirical item ordering chosen to sort the items by difficulty. However, Lorié (2001) found that the variability of item CPRS locations across item ordering schemes is far smaller and less systematic than the variability in item scale locations when the current mapping method is used with alternative mastery probability conventions. The validity of performance standards developed using continuums of performance depends, in part, on the stability of item mappings. Thus, CPRS seems to hold promise for strengthening the validity arguments that support such performance standards.

Determining the Performance Standard and Cut Score

Lorié (2001) noted that since any possible cut score is just above the CPR location of some test item, a performance standard can be constructed for any given cut score by stating the percentage of items of each type that a “hypothetical borderline examinee” at that cut score could answer correctly. On Kane's (1994) hypothetical college mathematics placement test, for example, some possible cut score might indicate that an examinee was expected to answer correctly 88% of arithmetic items, 75% of prealgebra items, 30% of algebra items, and 6% of calculus items. Higher cut scores would correspond to higher proportions for each category, and

[19] The CPR pattern for the n^{th} item is a vector that contains ones for item *n* and easier items, and zero or some chance probability for items harder than *n*. An actual response pattern would contain only ones and zeros. The inclusion of fractional values representing performance at the chance level requires a modification to the equation for the maximum-likelihood estimate of ability, as proposed by Lord (1974) in connection with the scoring of omitted responses. It is for this reason that the term *pseudoresponse pattern* is used.

conversely. The performance standard is chosen from the list of such descriptions for successive cut scores. When the CPRS distributions of items within the respective types overlap to a great extent, this kind of performance standard may indicate little more than the overall percentage of test items mastered by the hypothetical borderline examinee. When item CPRS distributions for the different item types do not overlap much, however, such performance standards which distinct kinds of items examinees at different score levels can solve. An extreme (and highly unlikely) case is when item type distributions do not overlap at all. Then, performance standards can be found that indicate complete mastery over successively more difficult item types.

Accuracy and Precision

CPRS provides an unambiguous characterization of an item's latent trait location, within the context of a given test. For example, in a slight modification to the Bookmark standard-setting method (Mitzel, Lewis, Patz, & Green, 2001), panelists could be asked to consider a series of hypothetical examinees, each of whom was able to answer all of the items up to a certain point, and none of the remaining items. Once a panelist determined which of these hypothetical examinees was at the borderline of acceptable proficiency, the corresponding latent trait location could be calculated without any use of an arbitrary mastery probability convention. However, the proportions of items of each type answered correctly by the hypothetical borderline examinee would not, in general, match the expected proportions correct by type for all examinees at that (borderline) ability level. Those expected proportions, as well as the expected variability of borderline examinee performance for each item type, would be readily obtainable using item response theory (e.g., Lord, 1980, p. 45, eqs. 4-2 and 4-3). Thus, in addition to linked performance standards and cut scores, CPRS also provides a framework for quantifying the accuracy with which its possible performance standards characterize borderline examinees.

Standard Setting Using CPRS

Criterion referencing through continuums of performance transforms standard setting as it is typically conceived. The familiar practice of first specifying a performance standard and then having a panel use that standard in reaching judgments that determine a cut score does not cohere for tests with an established continuum of performance standards. That continuum already provides a set of justified performance standards, and a technical argument links each of these standards to a (latent trait) test score. Any test score chosen as a cut score has an associated performance standard. With CPRS, the one-to-one relationship between performance standards and cut scores, defined by a continuum of performance, implies that if a performance standard is chosen then its corresponding cut score is entailed, and conversely.

What might standard setting look like in this situation? Here is one possible scenario. The series of performance standards derived from the continuum of performance would be assembled in a briefing book for policy makers. Each possible performance standard would be accompanied by information about the corresponding projected passing rate, overall and for significant demographic subgroups. If appropriate, the projected distribution of passing rates at the school level might also be presented. For some standards-based testing applications, especially at the high school level, expectancy tables relating standards-based test performance to patterns of course taking or to scores on college entrance, advanced placement, or other examinations might also be provided.[20] Policy makers would deliberate concerning the tradeoffs for different performance standards, in the context of a procedure, determined beforehand, for arriving at a performance standard / cut score combination.

The legitimacy of the resulting standard for tests with preestablished continuums of performance may turn out to depend greatly on the expertise of those who put together the continuum of performance, and on evaluations of its intelligibility. Lorié (2001) outlined and implemented a partial evaluation for a continuum of performance developed using 1992 and 1996 NAEP mathematics items. A central question of his investigation was whether potential audiences for standards-based reports generated using continuums of performance would understand what the item types meant. To support the validity of any continuum of performance, the author of the item type descriptions and members of potential standards-based report audiences should agree on their classifications of test items into different types. That is, descriptions of item types should be clear enough that one can tell which type each item on the test represents.

Summary and Conclusions

Performance standards defining categories like "proficient," and the cut scores determining examinees' membership in those categories, give rise to score interpretations and support score-based decision rules beyond those afforded by raw-score or other derived score scales. Validity arguments supporting these standards-based uses and interpretations must address the adequacy and appropriateness of performance standards as well as the accuracy of examinee classifications based on the corresponding cut scores.

In this paper, we have presented a conceptual framework for the analysis of standard setting and standards-based score interpretations, and have used that framework to develop illustrative validity arguments for three applications. Because every performance standard embodies a criterion-referenced score interpretation, its corresponding cut score cannot be warranted unless the validity argument supports that interpretation. By and large, the performance standard should accurately describe only those examinees who meet or exceed the cut score. Performance standards also have a normative component, embodying conceptions of adequacy or of appropriate expectation. Thus, procedural evidence for validity is also necessary. This evidence comprises arguments for the legitimacy and authority of the body setting the performance standard, the appropriateness of the standard for its intended use, the logic and coherence of the judgment process, and its fairness and objectivity. We briefly described some common methods for setting cut scores, and analyzed their atomic judgment tasks in terms of a judgment locus, judgment reference, and judgment procedure. Analyses of alternative judgment tasks revealed weaknesses in validity arguments for some of these methods.

We then applied our framework to analyze CRT, MCT, and current standards-based testing applications. These cases showed some of the ways appropriate standard-setting methods and validity arguments might vary across applications. The CRT application was chosen as a "best

[20] The joint distribution of scores on a standards-based test and another examination might change over time as a function of changes in curriculum and instruction or special test preparation, especially if high stakes are attached to test performance. The use or interpretation of such collateral information is not considered further in this paper.

case" for current standard-setting methods. It is low-stakes, and relies on the judgments of classroom teachers in reaching instructional decisions, affording procedural evidence of validity. The performance standard, namely sufficient mastery of prerequisite skills to enable success on the next learning unit, is one that teacher panelists are well qualified to understand and interpret. The item domain and the outcome domain are closely aligned, which increases the plausibility of the claim that there exists a cut score that can accurately distinguish between examinees meeting the performance standard and those falling short. In connection with Angoff methods for standard setting in the CRT application, an alternative interpretation of the modified Angoff judgment task was presented, which suggested a new method of using panelist judgments to derive a cut score. Showing how the justification for conventional standard-setting methods depends on particular features of the CRT application calls into question the suitability of these methods for applications where criterion-referenced score interpretations are less well defined or where item domains and outcome domains are less well aligned.

The MCT application highlighted the challenge of defining a performance standard when the outcome domain is less well defined. A new standard-setting method was proposed, which was designed to build strong procedural evidence of validity and to construct a performance standard in a way that enabled empirical linkage between that performance standard and an MCT cut score. The proposed method appears feasible, but much more time-consuming and expensive than methods actually used. One key feature of the proposed method was its shifting of judgments from performance on test items, which have a tenuous connection to the language of the performance standard, to performance on benchmark tasks, which make visible the meaning of the performance standard in the real world.

We have come to regard as central the problem of constructing performance standards that embody defensible criterion-referenced score interpretations. For the MCT application, that problem was addressed by constructing a rigorous performance standard referring to a real-world outcome domain, then empirically relating test performance to outcome domain performance. For the contemporary application, the same problem was addressed from another direction. Beginning with plausible patterns of performance on the test itself, a series of justified performance standards was constructed. A standard-setting procedure was then described in which a policy body was constrained to choose the actual performance standard from among that series. The proposed procedure in some ways resembles the Bookmark method, but unlike the Bookmark method, it assures that the performance standard is stated in terms of what the test actually measures. Also unlike the Bookmark method, the proposed procedure does not require the adoption of any arbitrary mastery probability convention. Note that, because the CPRS method defines a cut score on a latent trait continuum, that cut score can be used with alternate test forms constructed from the same calibrated item bank.

Our major conclusions should come as no surprise. Tests that support the criterion-referenced interpretations entailed by performance standards are critical to rational and defensible standard setting. The cut scores that operationalize these standards must accurately distinguish those who meet the performance standard from those who do not. It follows that not all standard-setting methods are equally defensible. Validity arguments are more straightforward and more plausible for methods in which the locus of judgment is concrete performances illustrative of the outcome domain. As the judgment task becomes more abstract or hypothetical, arguments become less direct and more difficult to support, and panelists (if used) are less able to bring their knowledge, experience, and values to bear in a principled fashion. Currently used standard-setting methods may be better suited to earlier, lower-stakes applications than to contemporary standards-based score interpretations. For contemporary applications, defensible standards-based reporting might best be accomplished via systematic construction of performance standards that hew closely to whatever it is the test actually measures.

It is our hope that the kind of analysis undertaken in this paper will lead to more critical consideration of the validity arguments in support of standards-based score interpretations and uses. The conceptual model may be useful in determining what standard-setting methods are defensible for a given purpose and context. The illustrations show that absent necessary preconditions for existing standard-setting methods, new methods may be developed. Absent such development, however, contemporary standards-based score interpretations may promise more than they can possibly deliver.

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.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Berk, R. A.(1986). A Consumer's Guide to Setting Performance Standards on Criterion-Referenced Tests. Review of Educational Research, 56, 137-172.

Berk, R. A. (1995). Standard setting -- The next generation. Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments, Volume II (pp. 161-181). Washington, DC: National Assessment Governing Board, National Center for Education Statistics.

Bloom, B. S. (1976). Human characteristics and school learning. New York: McGraw-Hill.

Cohen, D. K., & Haney, W. (1980). Minimums, competency testing, and social policy. In R. M. Jaeger & C. K. Tittle (Eds.), Minimum competency achievement testing: Motives, models, measures, and consequences (pp. 5-22). Berkeley: McCutchan.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I Braun (Eds.), Test Validity (pp. 3-17). Hillsdale, NJ: Erlbaum.

Cronbach, L. J., Gleser, G. C., Nanda, H, & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: John Wiley & Sons.

Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to the valid use of assessments. Assessment in Education, 3, 265-285.

Davis, A. (1998). Special Issue: The Limits of Educational Assessment. Journal of Philosophy of Education, 32(1), 1-159. (Whole Issue.)

Draney, K., & Wilson, M. (1998, June). Creating composite scores and setting performance levels: Use of a scaling procedure. Council of Chief State School Officers National Conference on Large Scale Assessment, Colorado Springs, CO.

Embretson, S. (1993). Psychometric models for learning and cognitive processes. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), Test theory for a new generation of tests (pp. 125-150). Hillsdale, NJ: Erlbaum.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. Acta Psychologica, 37, 359-374.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 18, 519-521.

Glaser, R., & Klaus, D. J. (1962). Proficiency measurement: Assessing human performance. In R. M. Gagné (Ed.), Psychological principles in systems development. New York: Holt, Rinehart, & Winston.

Glass, G. V. (1978). Standards and criteria. Journal of Educational Measurement, 15, 237-261.

Haertel, E. H. (1985). Construct validity and criterion-referenced testing. Review of Educational Research, 55, 23-46.

Haertel, E. H. (2002). Standard setting as a participatory process: Implications for validation of standards-based accountability programs. Educational Measurement: Issues and Practice, 21(1), 16-22.

Hambleton, R. K. (1980). Test score validity and standard-setting methods. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art (pp. 80-123). Baltimore: The Johns Hopkins University Press.

Hambleton, R. K., Brennan, R. L., Brown, W., Dodd, B., Forsyth, R. A., Mehrens, W. A., Nellhaus, J., Reckase, M., Rindone, D., van der Linden, W. J., & Zwick, R. (2000). A response to "Setting Reasonable and Useful Performance Standards" in the National Academy of Sciences' Grading the Nation's Report Card. Educational Measurement: Issues and Practice, 19(2), 5-14.

Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 10, 159-170.

House, E. R. (1980). Evaluating with validity. Beverly Hills, CA: Sage.

Huynh, H. (1976). Statistical considerations of mastery scores. Psychometrika, 41, 65-78.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn, Ed., Educational Measurement (3rd ed., pp. 485-514). New York: Macmillan.

Jaeger, R. M., Mullis, I. V. S., Bourque, M. L., & Shakrani, S. (1996). Setting performance standards for performance assessments: Some fundamental issues, current practice, and technical dilemmas. In G. W. Phillips (Ed.), Technical issues in large-scale performance assessment (Report No. NCES 96-802, pp. 79-115). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

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

Kane, M. (1994). Validating the performance standards associated with passing scores. Review of Educational Research, 64, 425-461.

Kane, M. (1998). Choosing between Examinee-Centered and Test-Centered Standard-Setting Methods. Educational Assessment, 5, 129-145.

Kingston, N. M., Kahl, S. R., Sweeney, K. P., & Bay, L. (2001). Setting performance standards using the Body of Work method. In G. J. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives (pp. 219-248). Mahwah, NJ: Erlbaum.

Kolstad, A. (1996). The response probability convention embedded in reporting prose literacy levels from the 1992 National Adult Literacy Survey. Paper presented at the 1996 annual meeting of the American Educational Research Association.

Kolstad, A., Cohen, J., Baldi, S., Chan, T., DeFur, E., & Angeles, J. (1998). The response probability convention used in reporting data from IRT assessment scales: Should NCES adopt a standard? Paper prepared for the National Center for Education Statistics, U.S. Department of Education.

Linn, R. L. (2000). Assessments and accountability. Educational Researcher, 29(2), 4-16.

Livingston, S. A., & Zieky, M. J. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service.

Livingston, S. A., & Zieky, M. J. (1989). A comparative study of standard-setting methods. Applied Measurement in Education, 2, 121-141.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 39, 247-264.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lorié, W. A. (2001). Setting defensible cut scores: Canonical pseudoresponses, item types, and performance standards. Unpublished doctoral dissertation, Stanford University, Stanford, CA.

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

Messick, S. (1995). Standards-based score interpretation. Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments, Volume II (pp. 291-309). Washington, DC: National Assessment Governing Board, National Center for Education Statistics.

Mislevy, R. J. (1998). Implications of market-basket reporting for achievement-level setting. Applied Measurement in Education, 11, 49-63.

Mislevy, R. J., Sheehan, K. M., & Wingersky, M. (1993). How to equate tests with little or no data. Journal of Educational Measurement, 30, 55-78.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives (pp. 249-281). Mahwah, NJ: Erlbaum.

Moss, P. A., & Schutz, A. (1999). Risking frankness in educational assessment. Phi Delta Kappan, 80, 680-687.

National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment. (1993). Setting performance standards for student achievement: A report of the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment: An Evaluation of the 1992 Achievement Levels. Stanford, CA: National Academy of Education.

National Council on Education Standards and Testing [NCEST]. (1992). Raising standards for American education. Washington, D.C.: U.S. Government Printing Office.

National Council of Teachers of Mathematics [NCTM]. (1989). Curriculum and evaluation standards for school mathematics (Report of the Working Groups of the Commission on Standards for School Mathematics). Reston, VA: The Council, c1989.

Nitko, A. J. (1980). Distinguishing the many varieties of criterion-referenced tests. Review of Educational Research, 50, 461-485.

Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (Eds.) Committee on the Evaluation of National and State Assessments of Educational Progress, Board on Testing and Assessment, Commission on Behavioral and Social Sciences and Education, National Research Council. (1999). Grading the Nation's Report Card: Evaluating NAEP and transforming the assessment of educational progress. Washington, DC: National Academy Press.

Reckase, M. D. (2000a). The evolution of the NAEP Achievement Levels setting process: A summary of the research and development efforts conducted by ACT. Iowa City, IA: ACT, Inc.

Reckase, M. D. (2000b). A survey and evaluation of recently developed procedures for setting standards on educational tests. In M. L. Bourque & S. Byrd (Eds.), Student performance standards on the National Assessment of Educational Progress: Affirmations and improvements (pp. 41-69). Washington, DC: National Assessment Governing Board.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), Changing assessments: Alternative views of aptitude, achievement, and instruction (pp. 37-75). Boston: Kluwer.

Rutherford, W. L. (1979). Criterion-referenced testing programmes: The missing element. Curriculum Studies, 11, 47-52.

Schmidt, G. N. (1982). Chicago Mastery Reading: A case against a skills-based reading curriculum. Learning, 11(4), 36-37, 39-40.

Shepard, L. (1980). Standard setting issues and methods. Applied Psychological Measurement, 4, 447-467.

Shepard, L. A. (1993). Evaluating test validity. Review of Research in Education, 19, 405-450.

Shepard, L. A. (2000). The role of assessment in a learning culture. Educational Researcher, 29(7), 4-14.

Tatsuoka, K. K. (1990). Toward an integration of item response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), Diagnostic monitoring of skill and knowledge acquisition (pp. 453-488). Hillsdale, NJ: Erlbaum.

Tversky, A., & Kahneman, D. (1993). Probabilistic reasoning. In A. I. Goldman (Ed.), Readings in Philosophy and Cognitive Science (pp. 43-68). Cambridge, MA: MIT Press.

Vinovskis, M. A. (1998). Overseeing the Nation's Report Card: The creation and evolution of the National Assessment Governing Board (NAGB). Washington, DC: National Assessment Governing Board. (Available on the World Wide Web at www.nagb.org.)

Wilson, M. , & Bock, R. D. (1985). Spellability: A linearly ordered content domain. American Educational Research Journal, 22, 297-307.

Wilson, M., & Draney, K. (1998, June). Creating composite scores and setting performance levels: Comparison of raw score and scaling procedures. Council of Chief State School Officers National Conference on Large Scale Assessment, Colorado Springs, CO.

Wilson, M., & Draney, K. (2002). A technique for setting standards and maintaining them over time. In S. Nishisato, Y. Baba, H. Bozdogan, & K. Kanefugi (Eds.), Measurement and multivariate analysis (Proceedings of the International Conference on Measurement and Multivariate Analysis, Banff, Canada, May 12-14, 2000). Tokyo: Springer-Verlag.

Zwick, R., Senturk, D., Wang, J., & Loomis, S. C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. Educational Measurement: Issues and Practice, 20(2), 15-25.

Table 1. Judgment locus, reference, and procedure for four typical standard setting methods.

Method	Judgment Locus	Judgment Reference	Judgment Procedure
Angoff, as originally proposed	Proficiency an item appears to require	Proficiency of examinee just at Performance Standard	Binary judgment of whether a borderline examinee would or would not be expected to respond correctly
Modified Angoff, as typically implemented	Proficiency an item appears to require	Proficiency of examinee just at Performance Standard	Quantitative judgment of probability of correct response by borderline examinee, or of proportion of such examinees who would respond correctly
Contrasting Groups	Proficiency shown, in nontest settings, by an actual examinee	Proficiency, in nontest settings, that is just at Performance Standard	Binary judgment of whether the examinee meets the performance standard
Body of Work	Proficiency shown in responses to a set of constructed-response items	Proficiency shown in responses of examinee just at Performance Standard	Binary judgment of whether the examinee has met the performance standard

Table 2. Classification accuracy for hypothetical MCT as a function of base rate and validity coefficient, assuming 98% passing rate.

Base Rate	Validity Coefficient	P(Fail | Truly Proficient)	P(Pass | Truly Not Proficient)	P(Truly Proficient | Fail)	P(Truly Not Proficient | Pass)	P(Correctly Classified)
98%	.00	2.00%	98.00%	98.00%	2.00%	96.08%
	.50	1.70%	83.06%	83.06%	1.70%	96.68%
	.80	1.18%	57.74%	57.74%	1.18%	97.69%
	.90	0.86%	42.02%	42.02%	0.86%	98.32%
	.95	0.61%	30.13%	30.13%	0.61%	98.79%
95%	.00	2.00%	98.00%	95.00%	5.00%	93.20%
	.50	1.45%	87.57%	68.94%	4.47%	94.24%
	.80	0.70%	73.27%	33.18%	3.74%	95.67%
	.90	0.32%	66.07%	15.18%	3.37%	96.39%
	.95	0.11%	62.06%	5.16%	3.17%	96.79%