2

What Is Standard Setting?

In its most essential form, standard setting refers to the process of establishing one or more cut scores on a test. As we mentioned in the previous chapter, in some arenas (e.g., licensure and certification testing programs) only a single cut score may be required to create categories such as pass/fail, or allow/deny a license, while in other contexts (e.g., K–12 student achievement testing programs) multiple cut scores on a single test may be required in order to create more than two categories of performance to connote differing degrees of attainment via-à-vis a set of specific learning targets, outcomes, or objectives. Cut scores function to separate a test score scale into two or more regions, creating categories of performance or classifications of examinees.

However, the simplicity of the definition in the preceding paragraph belies the complex nature of standard setting. For example, it is common—though inaccurate—to say that a group of standard-setting participants actually *sets* a standard. In fact, such panels derive their legitimacy from the entities that authorize them—namely, professional associations, academies, boards of education, state agencies, and so on. It is these entities that possess the authority and responsibility for *setting* standards. Thus it is more accurate to refer to the process of standard setting as one of "standard recommending" in that the role of the panels engaging in a process is technically to provide informed guidance to those actually responsible for the act of setting, approving, rejecting, adjusting, or implementing any cut scores. While we think that such a distinction is important, we also recognize that the term *standard recommending* is cumbersome and that insistent invocation of that

term swims against a strong current of popular usage. Accordingly, for the balance of this book, we continue to refer to the actions of the persons participating in the implementation of a specific method as "standard setting."

Kinds of Standards

The term *standards* is used in a variety of ways related to testing programs. For example, licensure and certification programs often have *eligibility standards* that delineate the qualifications, educational requirements, or other criteria that candidates must meet in order to sit for a credentialing examination.

Test sites—particularly those where examinations are delivered in electronic format (e.g., as a computer-based test, a computer-adaptive test, or a web-based assessment)—often have *test delivery standards* that prescribe administration conditions, security procedures, technical specifications for computer equipment, and so on.

In several locations in this book we will be referring to "the Standards" as shorthand for the full title of the reference book *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999). The *Standards* document is a compilation of guidelines that prescribe "standard" or accepted professional practices. To further complicate the issue, each of the entries in the *Standards* is referred to as "a standard."

In K–12 educational achievement testing, the concept of **content standards** has recently been introduced. In educational testing contexts, *content standards* is a term used to describe the set of outcomes, curricular objectives, or specific instructional goals that form the domain from which a test is constructed. Student test performance is designed to be interpreted in terms of the content standards that the student, given his or her test score, is expected to have attained.

Throughout the rest of this book, we focus almost exclusively on *performance standards*. As indicated previously, we will be using the term **performance standard** essentially interchangeably with terms such as **cut score**, standard, passing score, and so on. Thus when we speak of "setting performance standards" we are not referring to the abstraction described by Kane (1994b), but to concrete activity of deriving cut points along a score scale.

Definitions of Standard Setting

When defined, as we did at the beginning of this chapter, as "establishing cut scores for tests," the practical aspect of standard setting is highlighted. However, we believe that a complete understanding of the concept

of standard setting requires some familiarity with the theoretical foundations of the term. One more elaborate and theoretically grounded definition of standard setting has been suggested by Cizek (1993), who defines standard setting as "the proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states or degrees of performance" (p. 100). This definition highlights the procedural aspect of standard setting and draws on the legal framework of due process and traditional definitions of measurement.

This definition, however, suffers from at least one deficiency in that it addresses only one aspect of the legal principle known as due process. According to the relevant legal theory, important decisions about a person's life, liberty, or property must involve due process—that is, a process that is clearly articulated in advance, is applied uniformly, and includes an avenue for appeal. The theory further divides the concept of due process into two aspects: procedural due process and substantive due process. Procedural due process provides guidance regarding what elements of a procedure are necessary. Cizek's (1993) definition primarily focuses on the need for a clearly articulated, systematic, rational, and consistently implemented (i.e., not capricious) system; that is, his definition focuses on the procedural aspect of standard setting.

In contrast to the procedural aspect of due process is the substantive aspect. Substantive due process centers on the *results* of the procedure. In legal terms, the notion of substantive due process demands that the procedure lead to a decision or result that is fundamentally fair. Obviously, just as equally qualified and interested persons could disagree about whether a procedure is systematic and rational, so too might reasonable persons disagree about whether the results of any particular standard-setting process are fundamentally fair. The notion of fairness is, to some extent, subjective and necessarily calls into play persons' preferences, perspectives, biases, and values. This aspect of fundamental fairness is related to what has been called the "consequential basis of test use" in Messick's (1989, p. 84) explication of the various sources of evidence that can be tapped to provide support for the use of interpretation of a test score.

Another definition of standard setting that highlights the conceptual nature of the endeavor has been suggested by Kane (1994b). According to Kane, "It is useful to draw a distinction between the *passing score,* defined as a point on the score scale, and the *performance standard,* defined as the minimally adequate level of performance for some purpose. . . . The performance standard is the conceptual version of the desired level of competence, and the passing score is the operational version" (p. 426, emphasis in original). Figure 2-1 illustrates the relationship between these two concepts. Panel

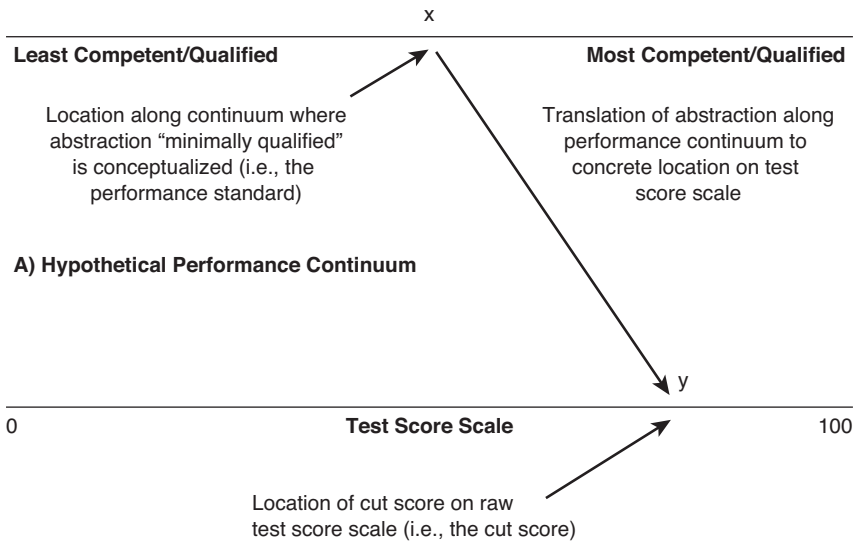

Figure 2-1 Relationship Between Performance Standard and Cut Score

A in the figure shows a hypothetical performance continuum; Panel B shows a test score scale. Participants in standard setting conceptualize a point along the performance continuum that separates acceptable from unacceptable performance for some purpose. This point is indicated in Panel A as "x." The process of setting cut scores can be thought of as one in which the abstraction (i.e., the performance standard or "x") is, via systematic, judgmental means, translated into an operationalized location on the test score scale (i.e., the cut score). This point is indicated as "y" in Panel B of the figure.

Two clarifications related to Kane's (1994b) definition of standard setting are also warranted. First, while we share Kane's desire to distinguish between the *performance standard* and the *passing score,* we think that the distinction between the two is consistently blurred. Like our own preference for use of the term *standard recommending* over *standard setting*, we recognize that the term *performance standard* is routinely used as a synonym for the terms *cut score, achievement level, standard,* and *passing score.* Thus, throughout this book and in deference to common though less-than-accurate invocation of those terms, we too use each of these terms essentially interchangeably.

Second, we think it is essential at this point to introduce the concept of **inference**, which is a key concept underlying Kane's definition. Implicit in this

definition is that the passing score creates meaningful categories that distinguish between individuals who meet some performance standard and those who do not. However, even the most carefully designed and implemented standard-setting procedures can yield, at best, defensible *inferences* about those classified. Because this notion of inference is so essential to standard setting—and indeed more fundamentally to modern notions of **validity**, we think it appropriate to elaborate on that psychometric concept at somewhat greater length.

According to the *Standards for Educational and Psychological Testing,* "validity is the most fundamental consideration in developing and evaluating tests" (AERA/APA/NCME, 1999, p. 9). The *Standards* defines validity as "the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests" (p. 9). Robert Ebel, the prominent psychometrician and namesake of a standard-setting method described later in this book, captured the special place that validity has for those involved in testing, using a memorable metaphor. He referred to validity as "one of the major deities in the pantheon of the psychometrician" (although Ebel also chastised the alacrity with which validity evidence is gathered by adding that "it [validity] is universally praised but the good works done in its name are remarkably few"; 1961, p. 640). In order to fully grasp the importance of validity as it pertains to the effects of test anxiety, we go into a bit more detail about this important testing concept.

Strictly speaking, tests and test scores cannot be said to be valid or not valid. Messick (1989) has emphasized the modern concept of validity as pertaining to the interpretation or inference that is made based on test scores. This fundamental concept was put forth by Cronbach and Meehl, who, in 1955, argued that "one does not validate a test, but only a principle for making inferences" (p. 300).

An inference is the interpretation, conclusion, or meaning that one *intends* to make about an examinee's underlying, unobserved level of knowledge, skill, or ability. From this perspective, validity refers to the accuracy of the inferences that one wishes to make about the examinee, usually based on observations of the examinee's performance—such as on a written test, in an interview, during a performance observation, and so on. Kane (2006) has refined Messick's work focus more squarely on the utility of the inference. According to Kane, establishing validity involves the development of evidence to support the proposed uses of a test or intended interpretations of scores yielded by a test. In addition, Kane suggests that validation has a second aspect: a concern for the extent to which the proposed interpretations and uses are plausible and appropriate.

Thus, for our purposes, the primacy of test purpose and the intended inference or test score interpretation are essential to understanding the

definition of standard setting. It is the accuracy of the inferences made when examinees are classified based on application of a cut score that is ultimately of greatest interest, and it is the desired score interpretations that are the target toward which validation efforts are appropriately directed.

Finally, in wrapping up our treatment of the definition of standard setting, we think it is important to note what standard setting is *not*. The definitions suggested by Cizek, Kane, and all other modern standard-setting theorists reject the conceptualization of standard setting as capable of discovering a knowable or estimable parameter. Standard setting does not seek to find some preexisting or "true" cutting score that separates real, unique categories on a continuous underlying trait (such as "competence"), though there is clearly a tendency on the part of psychometricians—steeped as they are in the language and perspectives of social science statisticians—to view it as such. For example, Jaeger has written that

> We can consider the mean standard that would be recommended by an entire population of qualified judges [i.e., standard-setting participants] to be a population parameter. The mean of the standards recommended by a sample of judges can, likewise, be regarded as an estimate of this population parameter. (1991, p. 5)

In contrast to what might be called a "parameter estimation paradigm" is the current view of standard setting as functioning to evoke and synthesize reasoned human judgment in a rational and defensible way so as to *create* those categories and partition the score scale on which a real trait is measured into meaningful and useful intervals. Jaeger appears to have embraced this view elsewhere and rejected the parameter estimation framework, stating that "a right answer [in standard setting] does not exist except, perhaps, in the minds of those providing judgment" (1989, p. 492). Shepard has made this same point and captured the way in which standard setting is now viewed by most contemporary theorists and practitioners:

> If in all the instances that we care about there is no external truth, no set of minimum competencies that are necessary and sufficient for life success, then all standard-setting is judgmental. Our empirical methods may facilitate judgment making, but they cannot be used to ferret out standards as if they existed independently of human opinions and values. (1979, p. 62)

To some degree, then, because standard setting necessarily involves human opinions and values, it can also be viewed as a nexus of technical, psychometric methods and policy making. In education contexts, social, political,

and economic forces cannot help but impinge on the standard-setting process when participants decide what level of performance on a mathematics test should be required in order to earn a high school diploma. In licensure contexts, standard-setting participants cannot help but consider the relative cost to public health and safety posed by awarding a license to an examinee who may not truly have the requisite knowledge or skill and of denying a license—perhaps even a livelihood—to an examinee who is truly competent.

The *Standards for Educational and Psychological Testing* acknowledges that standard setting "embod[ies] value judgments as well as technical and empirical considerations" (AERA/APA/NCME, 1999, p. 54). Cizek (2001b) has observed, "Standard setting is perhaps the branch of psychometrics that blends more artistic, political, and cultural ingredients into the mix of its products than any other" (p. 5). Seen in this way, standard setting can be defined as a procedure that enables participants using a specified method to bring to bear their judgments in such a way as to translate the policy positions of authorizing entities into locations on a score scale. It is these translations that create categories, and the translations are seldom, if ever, purely statistical, psychometric, impartial, apolitical, or ideologically neutral activities.

Policy Issues and Standard Setting

Whether taken into account explicitly as part of—or, better, in advance of—the actual implementation of a standard-setting method, there are many policy issues that must be considered when performance standards are established. In our experience, important policy issues are often not considered at all. However, the failure to consider such issues does not mean that decisions have not been made by default. By way of illustration, we might think of a family preparing a monthly budget, including amounts for food, housing, transportation, insurance, entertainment, and so on. Not included in the budget is any amount to be set aside for donations to charitable causes. Now, the failure to include this budget item was not purposeful; when planning the budget, this "line item" was simply not salient in the process and not even considered. However, in this context it is easy to see how failure to consider an issue actually *is,* in effect, a very clear and consequential budgetary decision. In this case, the amount budgeted is $0.

Of course, the same budgetary decision to allocate $0 might have been reached after considering how much to allocate to subsistence needs and consideration of other priorities, values, resources, and so on. Whether the $0 allocation was made because of a conscious decision or because the family's values placed greater priority on, say, political over charitable giving, or

because of any other rationale, is not necessarily germane. The decision is clearly within the family's purview; we do not intend here to make a claim about whether the decision was morally or otherwise right or wrong.

By extension, our point in this section is not to suggest the outcome or commend any particular policy position as "correct," but to insist that certain policy issues must be explicitly considered; the risk of not doing so is that the failure to consider them will result in de facto policy decisions that may be well aligned—or may conflict—with an organization's goals for setting standards in the first place. In the following paragraphs, we consider four such issues.

Scoring Models

In general, a test scoring model refers to the way in which item, subtest, or component scores are combined to arrive at a total score or overall classification decision (e.g., *Pass/Fail, Basic/Proficient/Advanced*, etc.). Perhaps the most common scoring model applied to tests is called a **compensatory** scoring model. The term *compensatory model* derives from the fact that stronger performance by an examinee on one item, subtest, area, or component of the decision-making system can compensate for weaker performance on another. The opposite of a compensatory model is called a **conjunctive** model. When a conjunctive model is used, examinees must pass or achieve a specified level of performance on each component in the decision-making system in order to be successful.

It may be helpful to illustrate the difference between a compensatory and a conjunctive model in different contexts. Suppose, for one example, that a medical board examination for ophthalmologists required candidates, among other things, to pass a 200-item multiple-choice examination. Further suppose that the written examination was developed to consist of ten 20-item subtests, each of which assessed knowledge of well-defined subareas of ophthalmic knowledge (e.g., one subtest might assess knowledge of the retina, one group of 20 items might deal with the orbit of the eye, one set of items might assess refraction and the physics of light, lenses, and so on). The entity responsible for credentialing decisions might decide that passing or failing the board examination should be determined by a candidate's total score on the total test (i.e., performance out of 200 items), irrespective of how the candidate performed on any of the 10 subareas. That is, the board explicitly decided on a compensatory scoring model. In such a case, it would be possible (depending on the cutting score chosen) that an examinee could pass the board examination without having answered correctly *any* of the items pertaining to knowledge of the retina.

This would have been possible if the candidate's knowledge of other subareas was strong enough to overcome his or her lack of knowledge with respect to the retina items. The credentialing entity would be justified in opting for a compensatory scoring model if, for example, there was evidence that the subareas were highly intercorrelated, if ophthalmologists often tended to specialize in one area (so that knowledge in all areas was not deemed essential), and so on. Regardless of the rationale, it would have been important for the credentialing entity to have explicitly articulated the rationale, investigated possible sources of evidence, and considered the implications of such a decision in advance.

For another example, suppose that a state had in place a testing program consisting of five tests—one each in mathematics, reading, writing, science, and social studies—that high school students must take in order to be eligible for a high school diploma. Let us assume that the state established what might be considered fairly "lenient" passing scores on each of the tests. Although not completely realistic, let us further suppose that the five tests measure independent **constructs**. If the state were to choose a compensatory model, a student who did not read well (or perhaps at all) could receive a diploma due largely to his or her strong performance in, say, science. If state policymakers decided that such an outcome was not desirable, a decision might have been made to use a conjunctive model instead. Use of a conjunctive model would require that a student earned a passing score on each of the five components in the system (i.e., on each of the five tests).

On the surface, this might seem like a prudent decision. Again, however, the state would be wise to explicitly consider the rationale, costs, and implications related to the choice of a conjunctive model. As we have constructed this scenario with five variables (i.e., test scores), the *real* probability of a student being eligible for a high school diploma when a conjunctive model is used can be calculated as the product of the individual, independent probabilities. For example, if the probability of passing each of five tests were .85, the product of $.85 \times .85 \times .85 \times .85 \times .85$, or approximately .44, would be the probability of passing all five tests. It is likely that in adopting a conjunctive model, policymakers may not have intended to establish a standard that would result in only approximately 44% of students being eligible to graduate. To be sure, the 44% figure is a lower bound, and the example assumes that performance on each of the tests is independent. To the extent that performance across tests is correlated, the figure would be higher. Nonetheless, the example highlights what can be an unintended consequence of adopting a conjunctive model.

We note that the preceding example is used for illustration purposes and does not take into account that the state might also permit students multiple

attempts to pass each test, that there may be strong remediation opportunities available to students, that performance on the five tests is not likely to be completely independent, and other factors. Nonetheless, the probability of obtaining a diploma based on test performance alone would still almost certainly be substantially less than the .85 the state may have mistakenly believed they were adopting when they established performance standards on the five tests that passed 85% of students on each one—and when the decision was made to implement a conjunctive scoring model.

The use of a conjunctive scoring model has other consequences as well. As Hambleton and Slater (1997) have demonstrated, the use of a conjunctive model results in slightly lower overall levels of decision consistency and decision accuracy (attributable to the impact of random errors increasing **false negative** classification errors).

We must also note that completely compensatory or conjunctive systems are not the only alternatives. Continuing with the illustration of the student assessment program consisting of separate tests in reading, mathematics, writing, science, and social studies, it would be possible for a state to adopt a partially compensatory model. Such a policy decision might, for example, include a conjunctive aspect whereby a student would be required to pass, say, separate reading and mathematics components, and a compensatory aspect whereby a student's relative strengths in his or her area of interest and coursework (e.g., science) would compensate for his or her relative weakness in an area of lesser interest or preparation (e.g., social studies).

Research on Standard Setting

It is not uncommon to encounter the phrase "standard setting study" used to describe a procedure designed to derive one or more cut scores for a test. Indeed, standard-setting procedures can be configured as studies that provide information beyond the practical need for identifying one or more points on a score scale for making classifications. Those who are responsible for setting performance standards often seek psychometric advice on standard setting from those with expertise in that area. For example, advice may be sought from consultants, standing technical advisory committees, testing companies, university-based researchers, and so on.

On the one hand, it is our experience that independent, external advisors are very valuable in that they usually offer insights, experience, ideas, and so on that may not have arisen otherwise and which usually improve the quality of the standard-setting procedures and the defensibility of the results. On the other hand, such advisors often have perspectives and goals that may not be shared by the entity responsible for setting the standards.

One such perspective that we emphasize in this section is a research orientation. Although we may be painting the contrast too sharply and the perspective as more homogeneous than it is, we believe that the research or scholarly perspective often characteristic of external advisors quite naturally compels them to recommend configuring procedures that yield information about the process, the participants, and the results that may extend beyond the entity's need. Those responsible for implementing performance standards may simply wish to have, in the end, a defensible set of cut scores.

We see the value of pursuing basic information about standard setting and the appeal of such information to those with somewhat more academic interests; we also see the value of streamlined, cost-effective, and time-efficient methods for obtaining cut scores that add little or nothing to the knowledge base of applied psychometrics. We mention the potential for differing interests in this section because we believe that the relative weighting of the two perspectives is another policy consideration best addressed well in advance of the actual standard-setting procedure. Ultimately, the board or other entity responsible for setting standards must decide which aspects of a standard-setting procedure recommended by external advisors are necessary for substantiating the validity of inferences based on application of the cut scores, and which are less germane to that goal. We recommend that explicit, a priori deliberation and consensus on a general framework regarding the extent to which research will play a part in standard-setting activities should be undertaken by the policy and decision-making entity responsible for oversight of the testing program.

Rounding

What might at first appear to be a minor issue of no consequence is the issue of rounding. The rounding we refer to here refers to the process of going from a mathematically very precise value to a value of lesser precision. The normal rounding rules indicate that, for example, when rounding to the nearest whole number, the value of 17.3 is rounded to 17, whereas a value of 17.6 would be rounded to 18. The issue, like the level in school at which students typically learn about rounding, seems elementary.

In standard setting, however, the issue is rarely without serious consequences. To illustrate, we consider a situation, increasingly common, in which a cut score derived from a standard-setting procedure is not represented (at least initially) in terms of raw score units such as number correct, but in the units of some other scale, such as the logit scale when an item response theory (IRT) ability metric is used. On the logit scale, standard setters might identify a cut score (in theta units) of, say, –1.2308. However,

because in most cases examinees' test scores are expressed as a summed number-correct value, those scores are almost always whole numbers. Further, it is highly unlikely that the ability level in theta units that standard setters indicated must be met or exceeded by examinees in order to pass, be deemed proficient, and so on will translate neatly into a whole number.

For example, let us consider the situation in which a cut score, in theta units, of –1.2308 resulted from a standard-setting procedure and was adopted by the board or other entity responsible for the license, credential, and the like. Now, suppose that a raw score of 17 corresponded to a theta value of –1.2682 and a raw score of 18 corresponded to a theta value of –1.2298. Under these circumstances, the cut score value adopted by the board lies somewhere between raw scores of 17 and 18, and a decision must be made regarding which raw score value to use for actual decision making. On the one hand, if a raw score of 17 were used as the operational cut score, some—perhaps many—examinees whose level of ability was below that deemed as necessary by both the standard-setting participants and the board would be classified as passing. If, on the other hand, a raw score of 18 were used, the operational passing score would be higher (in this case only slightly) than that adopted by the board.

Herein lies a dilemma—and the policy question—that faces the entity responsible for the testing program. How should the theta value from standard setting be rounded to obtain a value on the raw score scale? If a board adopts as a policy that the theta value resulting from the standard-setting procedure is consistently rounded to the closest whole number/number correct raw score, over the course of subsequent test administrations, the procedure will inevitably and nonsystematically result in, effectively, a lower passing standard than was approved being applied to examinees for some administrations and a higher passing standard than was approved being applied to examinees for other administrations. Alternatively, if a board adopts a policy that the theta value resulting from the standard-setting procedure must always be reached or exceeded, then over the course of subsequent test administrations, that policy decision will inevitably and systematically result in, effectively, a higher passing standard than was approved being applied to examinees at each administration—sometimes only slightly higher, though sometimes potentially very much higher. The dilemma becomes slightly more complicated when a second score conversion is used (i.e., when scaled scores are reported to examinees instead of, or in addition to, raw scores).

For the issue of rounding, our advice is again—not surprisingly—that the entity responsible for the testing program consider the issue in advance of standard setting and adopt an explicit rationale and procedure to be applied

consistently across test administrations. Simply following the "default" process of normal mathematical convention and rounding to the nearest whole is actually a policy decision that may or may not align well with the purposes of the testing program or responsibilities of the responsible entity. For example, in the case of a medical licensure examination, it may not be concordant with a mission of public protection to round to the nearest whole number when the use of such a procedure will lead to the awarding of licenses to examinees for whom there is test score evidence that they have not demonstrated the level of knowledge, skill, or ability deemed necessary by the board for safe and effective practice.

Classification Errors

The issue of rounding just described can be seen as a special case of the more general issue of classification error. As indicated previously, test scores and the Pass/Fail or other classifications resulting from application of a cut score are essentially inferences; that is, they represent best guesses based on available evidence about the "real" level of knowledge or skill possessed by an examinee, or about the examinees' "correct" classification. High-quality tests and well-conceived and implemented standard-setting procedures will result in a high proportion of correct classifications. However, because all tests, by definition, are based on limited samples of evidence and require inference, some score interpretations and classifications will, in almost all conceivable contexts, be inaccurate.

In concrete terms, it is safe to say that sometimes examinees who truly do possess the knowledge, skill, or ability required to pass, be classified as proficient, be awarded a credential, and so forth will be classified as failing, be placed in a less-than-proficient category, be retained in grade, be denied the credential or diploma they deserve, and so on. Such classification errors are referred to as **false negative** decisions. Conversely, sometimes examinees who truly lack the knowledge, skill, or ability required to pass, be classified as proficient, be awarded a credential, and so on will be classified as passing or proficient, be promoted to the next grade, be awarded a credential or diploma they do not deserve, and so on. Such classification errors are referred to as **false positive** decisions.

We introduce the concepts of false negative and false positive decisions for two reasons. First, although under usual circumstances they cannot be accurately identified (i.e., if it could be known for sure that a false positive decision was made about an examinee, we would correct it), classification errors are omnipresent and sound decision making must take them into account. Second, there are almost always differential costs or consequences

associated with each type of error, and these must be weighed against each other.

To illustrate the first point, we observe that many credentialing organizations routinely permit examinees multiple attempts to pass an examination on which licensure or certification hinges. Millman (1989) has dramatically demonstrated the effect of multiple attempts on false positive decisions: the greater the number of attempts permitted, the greater the likelihood that an examinee truly lacking the level of knowledge or skill judged to be minimally necessary will pass the examination. Figure 2-2 is taken from Millman's work. On the x-axis of the graph, the number of attempts is plotted; the y-axis shows the percentage passing. A passing standard of 70% mastery is assumed; it is also assumed that examinees' levels of knowledge, motivation, effort, and so on remain constant across repeated attempts.

Each of the four lines in the graph illustrates results for examinees of four different ability levels. For example, the lower line shows that an examinee far below the mastery standard (i.e., an examinee with only 60% mastery) has little—but some—chance of passing regardless of the number of attempts permitted. Even after 10 attempts, such an examinee has only approximately a 15% chance of passing. At the other extreme, an examinee who in truth is clearly above the standard (i.e., an examinee with 75% mastery) has a greater than 90% chance of passing on the first attempt; that percentage quickly rises to nearly 100% in as few as two attempts. An examinee exactly at the standard (i.e., an examinee with 70% mastery), has a greater than 50% chance of passing on the first attempt and dramatically increased chances with two or more attempts.

The disconcerting finding regards the false positive decisions that would be made for examinees moderately, but truly, below the standard. The line showing results for an examinee with only 65% mastery indicates that such an examinee has a fairly good chance of passing the test—close to 40%—with as few as three attempts. The examinee has a better than 50/50 chance of capitalizing on random error and passing the test in only five attempts! In summary, our first point is that classification errors are ubiquitous, affected by various policy decisions of the entity responsible for the testing program, and must be considered when those policy decisions are made.

To illustrate the second point, we consider tests with different purposes: a medical licensure test and a test to identify elementary school pupils in need of additional help with their reading skills. In the case of the medical licensure test, the consequences of awarding a license to an examinee who did not truly possess the level of knowledge or skill required for safe and effective practice may range from relatively harmless and involving minor cost (e.g., prescribing a few too many sessions of physical therapy than necessary for

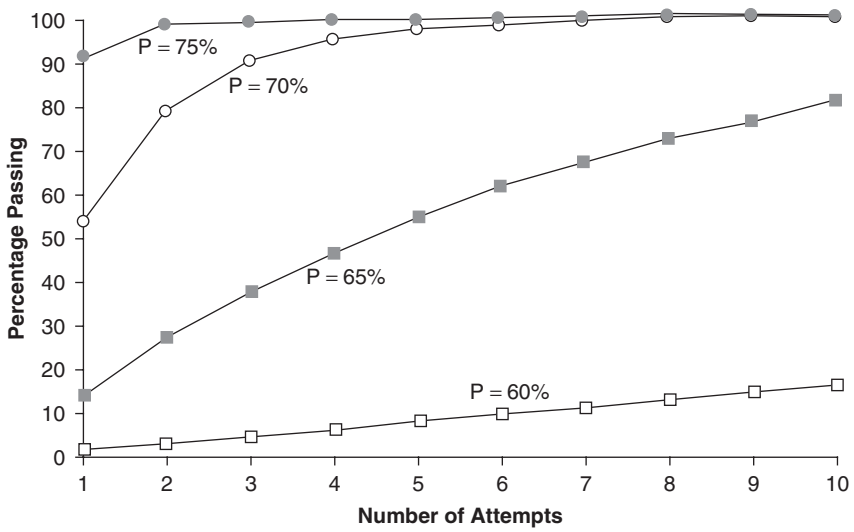

Figure 2-2 Percentage of Examinees at Four Levels of Competence Expected to Reach a 70% Passing Standard as a Function of the Number of Attempts Permitted

SOURCE: Millman (1989).

shoulder rehabilitation) to very severe and involving great cost or even loss of life (e.g., prescribing the wrong drug or dosage or incorrectly performing a surgery). In cases where the consequences or costs of false positive decisions are as serious as this, those participating in a standard-setting procedure might recommend a very high standard to preclude a large proportion of false positive decisions. And the entity responsible for the license or credential might well adopt a more stringent cut score than that recommended by the standard-setting participants to further guard against what might be viewed as a very serious false positive classification error with the potential for great harm to patients.

In some educational testing contexts, the situation might be precisely the opposite. For example, suppose a school district had a testing program at certain "gateway" grade levels (say, Grades 2, 5, and 8) to ensure that students would not simply progress through the system without acquiring a level of reading proficiency judged to be necessary for success in subsequent grades. Further, suppose that a student who failed to demonstrate the required level of reading comprehension and other skills on the Grade 2 test could be promoted to Grade 3 but would, during the summer between

Grades 2 and 3, be provided with intensive remediation and individualized instruction by reading specialists who could focus on the student's specific areas of weakness. In this case, we define false positive and false negative classifications differently than in the medical licensure example. In the medical context, failing a test was considered to be a negative decision (because of the potential economic and career effects on the physician); here we identify placement in the special remedial program as a positive (because of the potential educational benefit for the student).

As with the medical licensure example, there would be costs and consequences associated with any false positive classifications, as well as with false negative ones, as a result of applying the cut scores on the reading test. Again, in this education example, we define false positive and false negative errors in the opposite way in which they are often thought of; our use of the terms is consistent, however, in that the term *false positive* is always used to identify the inappropriate award of a credential, benefit, and so on, and the term *false negative* is consistently used to identify situations in which a reward, benefit, license, and so on is incorrectly denied. Thus, in our education example, we will define a false positive classification as occurring when a student was incorrectly identified as needing the extra remediation when in fact he or she did not, and a false negative classification would be one that identified a student as not needing the extra help when in fact he or she did.

In contrast to the medical licensure examination, a different weighing of the relative costs and consequences of the two types of errors would likely apply to the reading test context. A school board might decide that the costs and consequences associated with false positive decisions were minor. The student did suffer the loss of some free time over the summer and, during the first part of the next school year, was provided with assistance that he or she did not truly need to be successful at that grade level. The school board might also take into account the actual financial cost of false positive decisions, that is, the costs associated with salaries, benefits, instructional supplies, and so on required to provide extra reading instruction to students who did not truly need it. However, the board might weigh as more serious the costs and consequences of false negative decisions, that is, classifying a student as not needing the intervention who truly did. On that side of the ledger might be the risk of the student struggling in every subsequent grade to be successful, the risk of the student never attaining a level of reading comprehension necessary for him or her to be self-sufficient, the risk of the student dropping out of school, and so on. When faced with the relative costs of each type of classification error, the board might choose a policy

that judged false negative classification errors to be potentially far more serious than false positive decisions and budget accordingly.

Of course, it may be the case that an entity responsible for a testing program and setting performance standards might decide that both kinds of classification errors are equally serious and might set a cut score that makes the probabilities of each type of error equal (i.e., .50 and .50). In fact, the equal weighting of false positive and false negative classification errors is effectively the "default" weighting that is adopted when the issue of relative costs is *not* deliberated. As has been our point with respect to other policy issues described in this section, a policy decision is implicitly made to adopt a position related to classification errors even when no explicit deliberation of the issue occurs. Because of the potential gravity of the issue, and because of the serious consequences associated with it, we again urge that the entity responsible for setting standards give explicit attention to and document a reasoned position regarding the relative costs of classification errors in advance of implementing any cut scores.

Item Scoring Criteria and Total-Test Performance Standards

In this portion of the chapter, we seek to make an important distinction between three interrelated concepts: *performance standards, item scoring criteria,* and *performance level descriptions* (see Chapter 3 for a more detailed treatment of performance level descriptions). In a previous portion of the chapter, we noted that performance standards relate to content standards by specifying in a quantitative way how much of the content an examinee must have mastered. Performance standards refer to mastery of content standards in a global or holistic way, that is, how well the student must perform on the whole test. Somewhat similar to performance standards are **item scoring criteria**. Item scoring criteria specify how much of the content an examinee must have mastered, although in a comparatively much narrower context. Item scoring criteria specify the level of performance required in order to earn a particular score on one specific item, where the item is polytomously scored (i.e., it is not scored right/wrong, but a range of score points can be awarded based on the quality or characteristics of the response). Item scoring criteria are sometimes referred to as a scoring **rubric**, which is created and applied in conjunction with constructed-response format items or performance tasks.

Table 2-1 provides an illustration of a set of generic item scoring criteria developed for a statewide mathematics assessment. The rubric shown is

Table 2-1 Scoring Guide for Open-Ended Mathematics Items

Points	*Response Characteristics*
3	The response shows complete understanding of the problem's essential mathematical concepts. The student executes procedures completely and gives relevant responses to all parts of the task. The response contains few minor errors, if any. The response contains a clear, effective explanation detailing how the problem was solved so that the reader does not need to infer how and why decisions were made.
2	The response shows nearly complete understanding of the problem's essential mathematical concepts. The student executes nearly all procedures and gives relevant responses to most parts of the task. The response may have minor errors. The explanation detailing how the problem was solved may not be clear, causing the reader to make some inferences.
1	The response shows limited understanding of the problem's essential mathematical concepts. The response and procedures may be incomplete and/or may contain major errors. An incomplete explanation of how the problem was solved may contribute to questions as to how and why decisions were made.
0	The response shows insufficient understanding of the problem's essential mathematical concepts. The procedures, if any, contain major errors. There may be no explanation of the solution, or the reader may not be able to understand the explanation.

used as a guide to develop specific scoring guides or rubrics for each of the 4-point (i.e., 0 to 3 points possible) open-ended items that appears on the assessment, and it helps ensure that students are scored in the same way for the same demonstration of knowledge and skills regardless of the particular test question they are administered. In practice, the general rubric is augmented by development and use of extensive training sets and samples of prescored and annotated responses. It is important to note, however, that in the scoring rubric, there is no attempt to generalize to the student's overall level of proficiency in the area being assessed (i.e., mathematics).

Now, however, consider the performance level descriptions (PLDs; see Chapter 3) used for a high school graduation test in reading, presented in Table 2-2. Notice that, in contrast to specific scoring rubrics, the focus within PLDs is on the global description of competence, proficiency, or performance; there is no attempt to predict how a student at a particular

Table 2-2 Performance Level Descriptions for a Reading Test

Advanced	Students performing at the *Advanced* level typically demonstrate more abstract and sophisticated thinking in their analysis of textual information. They consistently demonstrate a firm grasp of the methods used by authors to affect the meaning and appropriateness of text. They are able to determine the meaning of unknown or complex words by using their knowledge of structural understanding and are able to discuss an author's use of figurative language.
Proficient	Students performing at the *Proficient* level can typically show an overall understanding of textual information. Students are generally able to identify and explain the various ways authors may influence text and assess the appropriateness of provided information. Students usually make appropriate choices regarding the author's use of figurative language and are able to determine the meanings of unknown or complex words using context clues or having a basic understanding of word structure.
Basic	Students performing at the *Basic* level demonstrate limited understanding and are able to make some interpretations and analytical judgments of textual information. Students generally can define unknown or complex words through context clues and can determine resources required to define or understand the more complex words.
Below Basic	Students performing at the *Below Basic* level can typically perform simple reading tasks but have not yet reached the level of Basic.

achievement might perform on a specific item. Scoring rubrics address only single items; PLDs address overall or general performance levels.

This distinction is salient for planning and conducting standard-setting activities. In standard-setting sessions, it is customary to provide as much background as possible about the test. Frequently, the panelists actually take the tests and score them using scoring keys and guides created for and used by the professional scorers, thereby gaining some familiarity with the rubrics. One tendency on the part of participants, however, is to attempt to apply scoring rubrics rather than PLDs when making the judgments required by a specific procedure chosen for setting the cut scores. In the example illustrated in Table 2-1, a participant might express the opinion that unless a student receives a score of at least 2 (or 3 or any other number) on this item, that student cannot be considered *Proficient*. If that panelist were engaged in a holistic standard-setting activity (e.g., the Body

of Work method; see Chapter 9), he or she might attempt to rescore a sampled student response and assign that student to one of the four categories on the basis of the score on this item (e.g., 3 for *Advanced*, 2 for *Proficient*, 1 for *Basic*, and 0 for *Below Basic*). That same panelist might then attempt to rescore each remaining constructed-response item, using a similar strategy, and then form an overall impression by noting which score seemed to predominate or even take the average of the individual item scores. Similarly, if that panelist were engaged in a Bookmark or other item-mapping standard-setting activity (see Chapter 10), he or she might withhold the Proficient bookmark until he or she encountered at least the first response at score point 3.

In both instances, the participant would be deviating from the specific procedures dictated by the chosen standard-setting method but, more importantly, would not be engaged in the appropriate application of expertise to the issue of overall performance standards. In a holistic standard-setting activity, the panelists should focus on how well the student performed on this item, along with all other items, and form an overall holistic impression of the student's performance level. Similarly, in a Bookmark activity, panelists should focus on the relationship between a given PLD and a given item. If that item happens to be one that calls for a constructed response, then the focus should be on the relationship between the PLD and the content of the sample response, not the score point assigned to the specific response being reviewed.

Alert and effective facilitation of the standard-setting meeting is required to aid participants in avoiding this error. In some instances one participant may describe to other participants a method he or she has discovered to make the task easier. In other instances, the facilitator may note that a panelist has written numbers on the standard-setting materials, along with calculations or other indications of attempts to summarize the numbers, a clear indication that the panelist is employing this strategy.

This point is essential for standard-setting participants to understand about item scoring criteria: The overall performance standards are numerical cut points that operationally define the PLDs and must apply to total scores rather than to scores on individual items. A student who has met the numerical criterion for *Proficient* (i.e., earned enough points to meet or exceed the cut score) may or may not do well on this or any other particular item. At least some of the *Proficient* students will perform poorly on this item (i.e., earn a low score), just as some of the *Basic* students will perform well on this item (i.e., earn higher scores).

Clearly, this distinction between item scoring criteria, performance standards, and PLDs is vital to the success of a standard-setting procedure and must be addressed effectively during the orientation and training of

participants in the standard-setting task (see Chapter 3 for more on selection and training). The introduction to the PLDs should include a detailed contrast with the rubrics and an admonition to adhere to the PLDs, rather than individual item scoring rubrics, when deciding where to set cut scores. The distinction must then be reinforced during periods of discussion between rounds of standard setting.

Conclusions

In this concluding portion of the chapter, we offer two kinds of summaries: one practical and one conceptual. On the practical side, we conclude that the choice of a scoring model, decisions about rounding rules, the emphasis to be placed on research activities, and the relative costs of classification errors are important policy decisions that ought not be left to chance. We urge those responsible for the oversight of testing programs to not allow such important decisions to be left to "default" values. Rather, these matters should be explicitly considered and decided on by the entity setting standards—in as conscientious a manner as the cut scores themselves are set.

At a more conceptual level, we conclude that standard setting lives at the intersection of art and science. Standard setting involves both thoughtful research and decisive action. We want to understand the decision-making process better, but at the same time we have to make real-time decisions with real-life consequences. Given the overt policy aspects of standard setting and the range of perspectives involved, it is no wonder that the field is replete with overlapping and sometimes contradictory terms. We have highlighted some of the key areas where confusion may lurk, and we will continue to shed additional light on these issues throughout the book.

We have attempted in this chapter to begin to shape a definition of standard setting both in terms of what it is and what it is not. Again, we note that present terminology is often the unfortunate victim of historical accident or perhaps of too many cooks spoiling the broth. We have content standards, professional standards, ethical standards, and performance standards. It is this final term, which we will also refer to as establishing cut scores (or simply as cut scores), to which we devote our attention in Chapter 3 and the following chapters of this book.