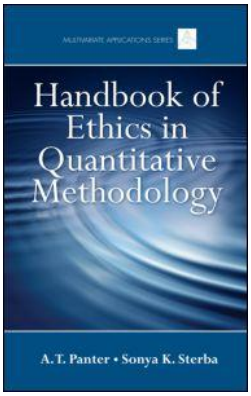

Handbook of Ethics in Quantitative Methodology

A.T. Panter, Sonya K. Sterba

Psychometric Methods and High-Stakes Assessment: Contexts and Methods for Ethical Testing Practice

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9780203840023.ch8
Gregory J. Cizek, Sharyn L. Rosenberg
Published online on: 20 Jan 2011

8

Psychometric Methods and High-Stakes Assessment: Contexts and Methods for Ethical Testing Practice

Gregory J. Cizek
University of North Carolina at Chapel Hill

Sharyn L. Rosenberg
American Institutes for Research

Psychometricians routinely use quantitative tools in test development and after test administration as part of the procedures used to evaluate the quality of the information yielded by those instruments. To some degree, nearly all those procedures play a part in ensuring that tests function in ways that promote fundamental fairness for test-takers and support the ethical use of test information by those who make decisions based on test results. In this chapter, we survey some of the quantitative methods used by testing specialists to accomplish those aims. The sections of this chapter are organized along the lines of three major phases of testing: test development, test administration, and test score reporting and use.

These topics are treated within four contexts. First, we have adopted the perspective on fairness proposed by Camilli, who has stated that "While there are many aspects of fair assessment, it is generally agreed that tests should be thoughtfully developed and that the conditions of testing should be reasonable and equitable for all students" (2006, p. 221). Further, we agree with Camilli that, although "issues of fairness involve specific techniques of analysis.… many unfair test conditions may not have a clear statistical signature" (p. 221). Thus, although the focus of this Handbook is on quantitative methods, we will occasionally allude to other methods for promoting ethical testing practice.

Second, our coverage of the psychometric methods for ethical testing practice focuses on *high-stakes* tests. Not all tests are included here, or even all standardized tests—only those tests for which important positive or negative consequences are attached. And, it is most precisely the

decisions—based in whole or in part—that are consequential and have stake associated with them, not strictly the tests themselves. However, high-stakes situations in which test data play a central role are increasingly common in education, psychology, occupational licensure and certification, and other contexts. Examples of high-stakes testing contexts include those of making clinical diagnoses of depression, judging the effectiveness of interventions for students with autism, counseling teenagers about career options, placing college first-year students in appropriate foreign language courses, awarding or withholding a license or credential for a given occupation, selecting or promoting civil servants, and numerous other situations. The common attribute is that the high-stakes test yields information that contributes to decisions that have meaningful consequences for individual persons, groups, or organizations. In each situation, quantitative methods can be used to promote fair and ethical decisions.

Third, high-stakes tests are not new. Miyazaki (1976) reports on the testing procedures associated with Chinese civil service examinations circa 200 B.C. An emphasis on ethical assessment has not always been a central focus of the testing profession (see, e.g., Gould, 1996). Within the past 40 years, however, increasing attention has been paid to ethical issues in high-stakes testing, and numerous standards and guidelines have been promulgated to provide direction for test developers and test users. Among these resources are:

- *Rights and Responsibilities of Test Takers: Guidelines and Expectations* (Joint Committee on Testing Practices, 1998)
- *Code of Professional Responsibilities in Educational Measurement* (National Council on Measurement in Education, 1995)
- *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 2004)
- *Family Educational Rights and Privacy Act* (1974)
- *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999)

Of these, the *Standards for Educational and Psychological Testing* (hereafter, *Standards*) is widely considered to be the authoritative source for best testing practices in education and psychology. The *Standards* is now in its fifth edition, a series that began with the publication of *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (American Psychological Association, 1954). In preparing this chapter, we have relied heavily on the current edition of the *Standards*, and linkages to relevant portions of the

Standards will be made throughout this chapter. We have also provided citations to specific portions of other resources where appropriate.

Finally, we have chosen a standards-referenced mathematics test required for high school graduation as a context for illustrating the application of quantitative methods to promote ethical testing practice. Several states require passage of these so-called "exit" tests or "end-of-course" examinations in subject areas such as mathematics, language arts, or science for students to be awarded a high school diploma. To be sure, the graduation decision does not hinge solely on the passage of such tests; rather, they are but one of multiple measures used. In all cases, other criteria (e.g., attendance, grades, specific courses requirements, community service hours, etc.) must also be satisfied. However, the test would still be classified as "high stakes" because failing to meet the performance standard on the test would have serious consequences for students.

Test Development

Ethical concerns arise at many junctures of the test development process, and the discipline of psychometrics has produced both qualitative and quantitative methods to promote fundamental fairness during this stage. Test development refers to "the process of producing a measure of some aspects of an individual's knowledge, skill, ability, interests, attitudes, or other characteristics," and it is "guided by the stated purpose(s) of the test and the intended inferences to be made from test scores" (AERA, APA, & NCME, 1999, p. 37). According to the *Standards*, "Tests and testing programs should be developed on a sound scientific basis. Test developers and publishers should compile and document adequate evidence bearing on test development" (p. 43).

The following subsections describe six decision points in the test development process where the application of quantitative procedures can help promote fair and ethical testing practice. The specific areas to be addressed include (a) identification of test purpose and content coverage, (b) choice of psychometric model, (c) item–task construction and evaluation, (d) test form development, (e) standard setting, and (f) validation.

Identification of Test Purpose and Content Coverage

According to the *Standards*, test development activities should be "guided by the stated purpose(s) of the test and the intended inferences to be made from test scores" (AERA, APA, & NCME, 1999, p. 37). Thus, the first step in

producing any test is to articulate a sharp focus on the construct the test is intended to measure and the test purpose(s). Construct definition and purpose may flow from theory development, clinical needs, industrial/organizational requirements, or legislative mandates. Whether in educational achievement testing or occupational testing, the first step in test development is typically to conduct a curriculum review, job analysis, role delineation study, or task survey (see Raymond & Neustel, 2006; Webb, 2006). These activities typically result in a set of *content standards*—a collection of statements that express the knowledge, skills, or abilities that are to be included in a curriculum, the focus of instruction, and assessed by an examination. Once these clusters of essential content, prerequisites, or critical job demands that will be sampled on the test have been established, the proportions or weightings for each cluster in the examination specifications must be derived, and various quantitative procedures for doing so are used (see Raymond, 1996).

In the context of a state-mandated, high-stakes exit examination in mathematics, delineating the domain to be tested and obtaining weights for subdomains are usually accomplished via judgmental procedures that seek to balance expert input, feasibility, cost, and other factors. A large and diverse panel of mathematics teachers, curriculum specialists, mathematicians, business leaders, parents, and others might be assembled to provide recommendations on decisions such as (a) the appropriate number of items or tasks for high school students to attempt; (b) the specific subareas of mathematics to be covered (e.g., algebra, geometry, probability, and statistics) and the relative proportion of the total test devoted to each of these; (c) the appropriate item formats and contexts (e.g., multiple-choice, constructed-response); (d) the acceptable level of language load of the test items or level of writing skill necessary for constructed-response items; and (e) policies for the use of calculators, and other decisions requiring knowledge of the intended test population and test content. It should be noted that, although representative panel membership is a goal of procedures to delineate domains and develop test specifications, inequities can still result (e.g., if the opinions of the mathematicians carry the most weight in panel discussions, or if practitioners in academic settings are overrepresented in job analysis survey returns). Thus, such procedures must be constantly monitored to foster equitable results.

Another ethical issue that can arise when developing test specifications is the need to ensure that the specifications reflect two characteristics. First, as in the case of the mathematics test, the instruction provided to students would need to be aligned to the test specifications. A fundamental concept in the area of test fairness is *opportunity to learn.* In this case, opportunity to learn would reflect the extent to which examinees were provided with instruction in the knowledge, skills, and abilities to be covered on the high-stakes mathematics test. Second, if the mathematics

test were used to predict success in subsequent courses or occupations, it would be necessary to collect evidence that the content specified in the test specifications was related to performance in the courses or the skills required for safe and effective practice in the occupation. Of course, at the most fundamental level, these are issues of validity, a topic addressed later in this chapter and elsewhere in this Handbook (see Carrig & Hoyle, Chapter 5, this volume).

Choice of Psychometric Model

Many aspects of test development, scoring, reliability, and validity are affected by the psychometric model that is used. There are two general classes of models used for building tests in education and psychology: classical test theory (CTT) and item response theory (IRT). CTT posits that an examinee's observed score is composed of a true component and a random error component, with the true score defined as the examinee's average score over an infinite number of parallel forms. With CTT, examinees' observed scores most often are calculated as a raw score (i.e., number correct) or percentage of items answered correctly, although more complicated scoring rules are possible (see Crocker & Algina, 1986).

An alternative set of models, IRT models, has become more widespread over the past few decades. IRT models invoke stronger assumptions than CTT models; they are more computationally complex; and they generally require larger sample sizes for successful use. IRT models posit that an observed score is an indicator of an underlying latent trait. They provide the probability of an examinee responding correctly to an item, with that probability dependent on the examinee's latent ability and the characteristics of the test item. IRT models require specialized software to compute estimates of examinees' standing on the latent trait; the software programs vary according to the estimation procedures used (e.g., joint maximum likelihood, marginal maximum likelihood) and the characteristics of the items (e.g., difficulty, discrimination, lower asymptote) that are estimated. There are several different types of IRT models, for example, the one-parameter logistic (1-PL) model, 2-PL model, 3-PL model, partial credit model, graded response model, and others. (For an introduction to IRT models, see Hambleton & Swaminathan, 1985.)

The choice of psychometric model has ethical implications. For example, the choice of one psychometric model over another may lead to different outcomes (e.g., pass–fail decisions, performance category classifications) for examinees. The choice of a psychometric model will affect the information that is gained about uncertainty (i.e., error) in examinees' scores. For example, one of the central features of IRT is the emphasis on a conditional standard error of measurement (CSEM). Unlike the CTT

standard error of measurement, which provides an overall, constant estimate of measurement error across the entire score range for a test, the CSEM provides an indication of the precision of an ability estimate at each score point on the test scale and varies across the test score range. In high-stakes contexts such as a high school graduation test, test construction efforts can enhance fairness by maximizing precision (i.e., minimizing the CSEM) in the regions of the score scale where cut scores are located and where classification decisions are made. In CTT, item discrimination indices and, in IRT, the *a*-parameter (in 2-PL and 3-PL models) also can be used to ensure that the most discriminating items contribute the most toward examinees' scores.

It is important to note, however, that the benefits of using a CTT, IRT, or other psychometric model only accrue to the extent that the model fits the data. Using a psychometric model that does not fit the data well in some parts of the score scale (particularly in the region where decisions are made) can compromise the fairness of those decisions. At minimum, procedures for assessing model–data fit (e.g., examination of residuals, assumptions, and fit statistics) should be used during field testing or after the first operational administration of an item.

The choice of a psychometric model often is driven by a combination of technical, practical, philosophical, and political considerations. For example, a 1-PL (Rasch) model may be chosen for developing a high school graduation test even before field test data are collected. This strategy is in sharp contrast to other contexts (e.g., structural equation modeling), where accepted practice involves comparing the fit of several alternative models and choosing the one that provides the best fit to the data (see McArdle, Chapter 12, this volume). Such a decision may be guided in part by philosophical considerations (e.g., the belief that additional parameters estimated in models accounting for item characteristics beyond item difficulty are only modeling error, a classic stance taken by Rasch model proponents such as Wright, 1997), or by political considerations (e.g., a concern that it would be difficult to explain to parents why items in the test are not weighted equally, and the possibility that students with the same raw scores can be assigned to different pass–fail or performance categories). Proponents of the Rasch model assert that "The data must fit, or else better data must be found" (Wright, 1997, p. 43). Regardless of approach, it is important to consider how examinees whose response patterns do not fit the prescribed model could be adversely impacted.

In summary, the process of choosing a psychometric model differs from the fitting of a structural equation model in the psychological literature. Model choice is not an area of psychometric methodology that has received wide attention to ethical issues, but the choice of test model can carry ethical implications. It is important to evaluate the reasons for and assumptions of choosing a particular CTT or IRT model. Examinee scores,

as well as subsequent decisions based on them, are directly related to the extent to which the psychometric model is appropriate for the test data. Unsatisfied assumptions or large modeling error can pose a serious threat to the inferences made about an examinee.

Item–Task Construction and Evaluation

The next step after defining the domain, producing test specifications, and identifying an appropriate psychometric model is the creation of items and/or tasks and scoring guides and rubrics that will comprise the test. According to the *Standards for Educational and Psychological Testing*: "The type of items, the response formats, scoring procedures and test administration procedures should be selected based on the purpose of the test, the domain to be measured, and the intended test takers" (AERA, APA, & NCME, 1999, p. 44).

The item writing and evaluation process can pose several ethical concerns. First, it is essential that item writers have adequate knowledge of the test content and are trained on the item writing process in a consistent manner. If item writers do not have this requisite knowledge, then they are likely to produce items that may compromise fairness by failing to adequately represent the intended domain. Before pilot testing, items should undergo a preliminary bias–sensitivity review where representative stakeholders evaluate items and suggest revising or eliminating any items that have the potential to disadvantage any test-takers. The *Code of Fair Testing Practices* notes that test developers should "avoid potentially offensive content or language when developing test questions and related materials" (Joint Committee on Testing Practices, 2004, p. 4). The *Standards* requires that "To the extent possible, test content should be chosen to ensure that intended test scores are equally valid for members of different groups of test takers" and "The test review process should include empirical analyses and, when appropriate, the use of expert judges to review items and response formats" (AERA, APA, & NCME, 1999, p. 44).

Pilot and field testing is an essential part of the measurement process and can help mitigate concerns related to test fairness. Because items are often selected for operational use based on their qualities in item tryouts, it is important that examinee samples used in this process are as large and representative as possible. Technically, IRT does not require that the pilot or field test groups be representative samples as long as they are sufficiently large and include the full range of performance in the intended population. However, given the potential for differential item functioning to occur—a sure threat to test fairness—it is desirable that pilot and field test samples are as representative as possible. Otherwise, items that appear to function well in a pilot or field test may have less validity evidence to

support their operational use in the intended population. As indicated in the *Standards*:

> When item tryouts or field tests are conducted, the procedures used to select the sample(s) of test takers for item tryouts and the characteristics of the sample(s) should be documented. When appropriate, the sample(s) should be as representative as possible of the population(s) for which the test is intended. (AERA, APA, & NCME, 1999, p. 44)

Likewise, the *Code of Fair Testing Practices* indicates that test developers should "obtain and provide evidence on the performance of test-takers of diverse subgroups, making significant efforts to obtain sample sizes that are adequate for subgroup analyses [and] evaluate the evidence to ensure that differences in performance are related to the skills being assessed" (Joint Committee on Testing Practices, 2004, p. 4).

It is also important that the testing conditions for a pilot or field test should be as close as possible to those of operational test administrations. If pilot or field tests are conducted as stand-alone procedures, steps should be taken to investigate and document any conditions that may affect examinee behavior and subsequent performance. For example, if items for a high school graduation test are pilot tested as a stand-alone exercise that has no consequences, low motivation is likely to affect the students' performance. This could compromise the accuracy of test results and test fairness because the item statistics generated from the pilot and field tests are typically used to select the items for the operational test that will be used for the high-stakes decisions.

To address the concern about the accuracy of item statistics from pilot or field tests, it is usually preferable to use embedded field testing procedures (i.e., where the trial items are interspersed with operational items and examinees have no knowledge of which items count toward their score). This way, testing conditions are similar to the operational administration conditions and are therefore less likely to adversely impact the results of the pilot or field tests.

There are several key purposes of item tryouts. First, pilot or field testing data can be analyzed to select the items with the best qualities that are most likely to represent the test content and minimize the potential for unfairness. Second, to maximize the precision of measurement of a test to which a cut score will be applied, it is desirable to select items that are highly discriminating in the range where a decision is made (i.e., in the area of any cut score). Third, differential item functioning (DIF) analyses can be performed to determine whether there are differences in performance on individual items where focal and reference group abilities are equivalent (Camilli, 2006). Items that are flagged for displaying statistically significant DIF routinely undergo additional review to determine

whether they should be included on an operational test, or they may simply be eliminated due to any potential for unfairness.

Finally, item tryouts permit the evaluation of scoring guides or rubrics for performance tasks or constructed-response items. Analyses are performed to ensure that each score category is functioning as intended, that the boundaries between score categories are clear, that the rubric or scoring guide can be interpreted as intended and applied consistently by any raters, and to permit adjustments to the scoring procedures when unanticipated examinee responses shed light on gaps in the category descriptions.

Test Form Development

After the development and review of test items and tasks, test forms are created. At this juncture, ethical issues also must be addressed. If multiple forms will be developed for a single test administration, or if new forms are developed across test administrations, the forms must be developed according to the same content and statistical specifications, including targets for difficulty and reliability. Failure to develop equivalent forms can reduce confidence that equating (described later in this chapter) will correct for variations in difficulty, or that examinees' scores on different forms can be interpreted in the same way.

An additional method for promoting consistency in form development procedures and match to test specifications is found in *alignment analyses*. Alignment analyses include both judgmental review and quantitative indices of the degree to which a test matches the content standards it was intended to assess. According to Porter (2006), there are two general ways that a test may be imperfectly aligned with its content standards: (a) Some areas specified in the content standards are not measured by a test; or (b) some areas assessed on a test are not part of the content standards. The latter condition—that is, when a test includes material not specified for coverage—often accounts for examinees' informal evaluations of a test as "unfair."

Various quantitative procedures have been developed to gauge and help address concerns about alignment. Among the most commonly used are the Survey of Enacted Curriculum (Porter & Smithson, 2001) and the Webb alignment method (Webb, 1997, 2002). Each of these methods results in an index ranging from 0.0 (no alignment) to 1.0 (perfect alignment). The method proposed by Webb is the most commonly used method for gauging alignment between the content standards and assessments used by states as part of federally mandated annual student testing. The method provides quantitative summaries of various aspects of alignment, including *categorical concurrence* (i.e., the extent to which a test contains an adequate number of items measuring each content standard); *depth of*

knowledge (i.e., the extent to which the items or task in a test are as cognitively demanding as suggested by the content standards); *range of knowledge correspondence* (i.e., the extent to which at least half of the subobjectives for a content standard are covered by the test); and *balance of representation* (i.e., the extent to which the objectives for a content standard included in a test are addressed in an even manner). Overall, consistency in form development over time and attention to alignment help promote fairness to the extent that examinees' test results are not advantaged or disadvantaged because of the particular test form they were administered, nor because the domain covered by the test was an uneven or unrepresentative sample of the content standards to which scores on the test are referenced. This principle is reflected in the *Standards for Educational and Psychological Testing*, which states that "test developers should document the extent to which the content domain of a test represents the defined domain and test specifications" (AERA, APA, & NCME, 1999, p. 45).

Standard Setting

Whereas the term *content standards* refers to the collections of statements regarding what examinees are expected to know and be able to do, *performance standards* refers to the levels of performance required of examinees on a test designed to assess the content standards. Although subtle and sometimes important distinctions can be made (see Cizek, 2006; Kane 1994), the term *cut score* is often used interchangeably with *performance standard*. Further, it is important to note that, although cut scores are typically derived as result of the procedures described in this section, it would be inaccurate to say that the panels of participants who engage in such procedures "set" the performance standards. Rather, such panels almost always serve as advisory to the entity with the legal or other authority to determine the cut scores that will be applied to examinees' test performances.

According to the *Standards for Educational and Psychological Testing*:

> A critical step in the development and use of some tests is to establish one or more cut points dividing the score range to partition the distribution of scores into categories. . . . [C]ut scores embody the rules according to which tests are used or interpreted. Thus, in some situations, the validity of test interpretations may hinge on the cut scores. (AERA, APA, & NCME, 1999, p. 53)

In the context of licensure and certification testing, the *Standards* notes that "the validity of the inferences drawn from the test depends on whether the standard for passing makes a valid distinction between adequate and inadequate performance" (p. 157).

The performance standards for a test are used to define various categories of performance, ranging from a simple dichotomy (e.g., *pass–fail*) used for many licensure or certification examinations, to more elaborate classifications such as *basic*, *proficient*, and *advanced* used in many student achievement testing programs. Performance standards may be expressed in a raw score metric (e.g., number correct), an IRT metric (e.g., theta value), or another metric (e.g., transformed or scaled scores).

There are five steps common to all standard-setting procedures: (1) choice of standard setting method, (2) selecting and training qualified participants, (3) providing feedback to participants, (4) calculating the cut score(s), and (5) gathering validity evidence. Each of these steps involves ethical concerns. The following portions of this chapter address steps 1, 2, 3, and 5.

Although numerous methods exist, the chosen standard setting method should be related to the purpose, format, and other characteristics of the test to which it will be applied. Detailed descriptions of the possible methods are presented elsewhere (see Cizek, 2001; Cizek & Bunch, 2007) and are beyond the scope of this chapter. Whatever method is selected, there are two primary goals—transparency and reproducibility—with the former a necessary condition for the latter. The first key goal—transparency—requires that the process for gathering judgments about cut scores should be carefully and explicitly documented. The *Standards* indicates that "when a validation rests in part on the opinions or decisions of expert judges, observers or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described" (AERA, APA, & NCME, 1999, p. 19). In addition, transparency helps to ensure that the chosen standard setting method is well aligned to the purpose of the examination. The goal of reproducibility also requires careful following and documentation of accepted procedures, but it also requires an adequate number of participants so that the standard error of participants' judgments about the cut scores is minimized. Fundamental fairness requires that any cut scores should be stable and not a statistical anomaly; if the standard-setting procedure were repeated under similar conditions, it is important to have confidence that similar cut scores would result. The *Standards* (AERA, APA, & NCME, 1999) provides some guidance on representation, selection, and training of participants (called "judges" in the *Standards*). For example, regarding the number of participants that should be used, the *Standards* indicates that "a sufficiently large and representative group of judges should be involved to provide reasonable assurance that results would not vary greatly if the process were replicated" (AERA, APA, & NCME, 1999, p. 54). In practice, logistical and economic factors must be considered when determining the sample size for standard-setting studies, but in general as large a group as feasible should be used to enhance reproducibility.

When selecting and training participants for a standard-setting activity, the ethical concerns center on the qualifications and representativeness of those who will participate in the process, and on how well prepared the participants were to engage in the standard-setting task. Potential participants must be knowledgeable regarding the content to be tested and the characteristics of the examinee population. Here, the *Standards* (AERA, APA, & NCME, 1999) requires that "the qualifications of any judges involved in standard setting and the process by which they are selected" (p. 54) should be fully described and included as part of the documentation of the standard setting process and that the standard setting process "should be designed so that judges can bring their knowledge and experience to bear in a reasonable way" (p. 60).

Whereas decisions about representation on standard setting panels are ultimately a policy matter for the entity responsible for the testing program, some ethical guidelines apply. The concern about appropriate qualifications can be illustrated in the contexts of achievement and credentialing tests. For the high school mathematics test used as a part of diploma-granting decisions, qualified participants would need to know about the mathematics curriculum and content covered by the test, the characteristics of the high school students who must pass the examination, and the mathematical knowledge and skill required in the variety of contexts that the students will encounter after graduation. Thus, it would be appropriate to include high school mathematics teachers on the standard-setting panel, but also representatives of higher education, business, the military, parents, and other community members. In contrast, in standard setting for licensure or certification, the primary purpose is often public protection. According to the *Standards*, "the level of performance required for passing a credentialing test should be dependent on the knowledge and skills necessary for acceptable performance in the occupation or profession" (AERA, APA, & NCME, 1999, p. 162). Thus, it would be most appropriate to include entry-level practitioners or those who already hold the credential that is the focus of the examination, as well as public representatives whose interests are ultimately served.

According to the *Standards*:

> Care must be taken to assure that judges understand what they are to do. The process must be such that well-qualified judges can apply their knowledge and experience to reach meaningful and relevant judgments that accurately reflect their understandings and intentions. (AERA, APA, & NCME, 1999, p. 54)

Thus, the training of participants in the selected standard-setting procedure is also a critical step; the method used should be one that allows

participants to make the judgments described, and the training should adequately prepare them to do so. One of the mechanisms for accomplishing this is to provide participants with various kinds of information to help them make their judgments.

There are three basic kinds of information. *Normative data* permit panelists to compare their judgments with those of other participants, and it is perhaps the most common type of feedback used in standard-setting studies. Normative data consist of information such as a distribution of item ratings or overall cut scores, minimum and maximum ratings, and mean or median ratings in the form of frequency distributions, bar graphs, or other data summaries that are easy for the participants to interpret and use. *Reality data* are provided to assist participants in generating realistic judgments. Reality information typically consists of item difficulty indices (i.e., *p* values) or theta-scale values for individual items. Reality information can be computed based on complete samples of test-takers or may be computed based on subsamples, such as examinees around a given point in the total score distribution. Panelists use this information to help them gauge the extent to which their judgments relate to the performance of examinees or test items. Finally, *impact data* (also called *consequence data*) are provided to aid panelists in understanding the consequences of their judgments. Typically, impact data consist of information about the number or percentage of examinees who would pass a test, or who would fall into a given performance level if a recommended cut score was implemented. The three types of information are typically provided across "rounds" of judgments, so that participants have opportunities to revise their judgments in response to the information provided.

The ethical implications of providing these data are clear. First, normative data are provided so that participants can gauge the extent to which their judgments concur with other qualified participants—and to make revisions as they deem appropriate. Reality data are provided so that participants' judgments are grounded in the performance of the examinees who are subject to the test, and not merely dependent on the leniency, stringency, or perspectives of the participants. Impact data are provided so that those who make cut score judgments are aware of the effect on test-takers. For example, it would not seem reasonable to deny graduation to all high school seniors (if a cut score was set too high) or to certify all examinees who take a test in brain surgery (if a cut score was set too low).

Gathering validity evidence to support the cut score recommendations is another ethical responsibility of those who engage in standard setting. Hambleton (2001) and Pitoniak (2003) have outlined several sources of potential validity evidence. These include *procedural* fidelity and appropriateness, as well as *internal* evidence (e.g., the degree to which participants

provide ratings that are consistent with empirical item difficulties, the degree to which ratings change across rounds, participants' evaluations of the process) and *external* evidence (e.g., the relationship between decisions made using the test and other relevant criteria such as grades, supervisors' ratings of job performance, performance on tests measuring similar constructs, etc.).

Evaluation of standard setting must attend to the omnipresent reality that classifications made based on the cut scores may be incorrect. To the extent that sample size, alignment, representativeness or qualifications of the participants, or other factors are compromised, the cut scores resulting from a standard-setting procedure might unfairly classify as "failing" some examinees who truly possess the knowledge, skills, or abilities deemed necessary to be classified into a certain performance category. Conversely, some examinees who do not possess the knowledge, skills, or abilities deemed necessary may be mistakenly classified as "passing." These classification errors are often referred to as *false-negative* and *false-positive* errors, respectively. Although it is true that nearly all test-based classification decisions will result in some number of classification errors, an ethical obligation of those who oversee, design, and conduct standard-setting procedures is to minimize such errors.

Finally, a specific unethical action is highlighted in the *Standards* related to setting cut scores for licensure or certification examinations. According to the *Standards*:

> The level of performance required for passing a credentialing test should be dependent on the knowledge and skills necessary for acceptable performance in the occupation or profession and should not be adjusted to regulate the number or proportion of persons passing the test. (AERA, APA, & NCME, 1999, p. 162)

Validation

Among all criteria by which tests are evaluated, validity is universally endorsed as the most important. For example, the *Standards* asserts that validity is "the most fundamental consideration in developing and evaluating tests" (AERA, APA, & NCME, 1999, p. 9). A necessary (although insufficient) precondition for the ethical use of any test is the collection and evaluation of adequate validity evidence.

Refining Messick's (1989) definition, in this chapter we define *validity* as the degree to which scores on an appropriately administered instrument reflect variation in the characteristic it was developed to measure and support the intended score inferences. By extension, we define *validation* as the ongoing process of gathering relevant evidence for generating

an evaluative summary of the degree of fidelity between scores yielded by an instrument and inferences about standing on the characteristic it was designed to measure. That is, validation efforts amass and synthesize evidence for the purpose of articulating the degree of confidence that intended inferences are warranted.

Validation centers on a concern about the quality of the data yielded by an instrument. That concern is heightened whenever a test is one part of procedures for making important decisions in countless situations in which the information yielded by a test has meaningful consequences for persons or systems. Because it would be unethical to use information that is inaccurate, misleading, biased, or irrelevant—that is, lacking validity—to make such decisions, validity is rightfully deemed to be the most important characteristic of test scores. The topic of validity is treated in substantial depth elsewhere in this Handbook (see Carrig & Hoyle, Chapter 5, this volume); thus, we will only briefly summarize six broadly endorsed tenets of validity here.

First among the accepted tenets is that validity pertains to the inferences that are made from test scores. Because latent traits and abilities cannot be directly observed, these characteristics must be studied indirectly via the instruments developed to measure them, and inference is required whenever it is desired to use the observed measurements as an indication of standing on the unobservable characteristic. Because validity applies to the inferences to be made from test scores, it follows that a clear statement of the intended inferences is necessary to design and conduct validation efforts.

Second, validity is not a characteristic of instruments but rather of the data generated by those instruments. Grounded in the position first articulated by Cronbach (1971), the current *Standards* notes that "it is the interpretations of test scores that are evaluated, not the test itself" (1999, p. 9).

Third, the notion of discrete kinds of validity (i.e., content, criterion, construct) has been supplanted by the realization that, ultimately, all evidence that might be brought to bear in support of an intended inference is evidence bearing on the responsiveness of the instrument to variation in the construct measured by the instrument. This conceptualization of validity is referred to as the *unified view of validity*, and validity is now generally regarded as a singular phenomenon. In describing the unified view, Messick has indicated that "What is singular in the unified theory is the kind of validity: All validity is of one kind, namely, construct validity" (1998, p. 37).

Fourth, judgments about validity are not absolute. As Zumbo has stated, "Validity statements are not dichotomous (valid/invalid) but rather are described on a continuum" (2007, p. 50). There are two reasons why this must be so. First, in a thorough validation effort, the evidence is routinely mixed in terms of how directly it bears on the intended inference,

its weight, and its degree of support for the intended inference. Second, because validation efforts cannot be considered "completed" at a specific juncture, evidence amassed at any given time must necessarily be considered tentative and a matter of degree.

Fifth, validation is an ongoing enterprise. Just as it is incorrect to say that *a test* is valid, so it is incorrect to say that the validity case for an intended inference is closed. Many factors necessitate a continuing review of the empirical and theoretical information that undergirds the inferences made from test scores. For example, replications of the original validation efforts, new applications of the instrument, new sources of validity evidence, new information from within and beyond the discipline about the construct of interest, and theoretical evolution of the construct itself all represent new information that can alter original judgments about the strength of the validity case.

The final tenet of modern validity theory is that the process of validation necessarily involves the exercise of judgment and the application of values. Because searching validation efforts tend to yield equivocal evidence, the available evidence must be synthesized, weighed, and evaluated. Kane (2001) has observed that "validity is an integrated, or unified, evaluation of the [score] interpretation" (p. 329). Validation efforts result in tentative conclusions about the degree to which evidence supports confidence that a test, administered to its intended population under prescribed conditions, yields accurate inferences about the construct it is intended to measure. Lacking such evidence, or when the evidence fails to support a desired level of confidence, the use of test data to make important decisions about examinees would be unethical.

The *Standards* lists and describes five sources of validity evidence. They include (a) evidence based on test content, (b) evidence based on response processes, (c) evidence based on internal structure, (d) evidence based on relations to other variables, and (e) evidence based on consequences of testing. Common threats to the validity of test score inferences include *construct underrepresentation*, in which "a test fails to capture important aspects of the construct," and *construct irrelevant variance*, in which "test scores are affected by processes that are extraneous to its intended construct" (AERA, APA, & NCME, 1999, p. 10). An example of the former would be a licensure test of automobile driving ability that involved only a written, knowledge component; an example of the latter would include a test of English composition ability for which scores were influenced by examinees' handwriting. The ethical aspects of validity in these examples are clear: It would be inappropriate to license drivers who lacked driving skill; it would be unfair if examinees of equal writing ability were assigned different scores based on a characteristic (handwriting legibility) that the test was not intended to measure.

Test Administration and Scoring

Ethical concerns are also present when tests are administered and scored. The following subsections of this chapter describe aspects of test administration and scoring where quantitative procedures can be invoked to advance the goal of ethical testing practice. These aspects include (a) test registration and test preparation; (b) test administration conditions, accommodations, and security; and (c) scoring procedures.

Registration and Examinee Test Preparation

When preparing to administer a test, it may be necessary first to determine whether examinees meet the eligibility guidelines for taking the test. Such guidelines may include, among other things, academic preparation requirements, age or residency requirements, and completion of required internships and supervised or independent practice. Where such qualifying criteria exist, it is an ethical obligation of the entity responsible for the testing program to ensure that only eligible candidates are permitted to take the examination.

Once it has been determined that an examinee has met the eligibility requirements, the entity should provide candidates with information about ethical and unethical test preparation activities and should follow rigorous procedures to preclude examinees from having improper prior access to test materials. According to the *Standards*, "test users have the responsibility of protecting the security of test materials at all times" (AERA, APA, & NCME, 1999, p. 64). Likewise, the *Code of Fair Testing Practices* (Joint Committee on Testing Practices, 2004) requires that test developers "establish and implement procedures to ensure the security of testing materials during all phases of test development, administration, scoring, and reporting" (p. 6) and that test users should "protect the security of test materials, including respecting copyrights and eliminating opportunities for test takers to obtain scores by fraudulent means" (p. 7).

Second, where there is reason to believe that examinees may be inappropriately advantaged or disadvantaged by aspects of the test administration (e.g., computer-based mode of administration, test format), the responsible entity should take steps to address such concerns. For example, if specialized software or an unfamiliar computer interface will be used for testing, examinees should be provided with opportunities to practice with the software or interface, ideally well in advance of the day of testing. Whereas it might be reasonable to assume that test-takers are familiar with multiple-choice item formats, some tests may contain formats that would be less familiar to examinees, such as gridded-response formats, drag-and-drop completion items, or other novel response formats. In such

cases, and also to provide examinees with an opportunity to gauge the content coverage, level of difficulty, and other test-related factors, it is desirable to provide examinees who register for an examination with a practice test form that they can complete after registration but before taking the operational test.

All the major ethical standards support these recommendations. For example, they are in line with the relevant guidelines in the *Standards*, which note, among other things, that "Instructions should … be given in the use of any equipment likely to be unfamiliar to test takers [and] opportunity to practice responding should be given when equipment is involved" (AERA, APA, & NCME, 1999, p. 63). Similarly, according to the *Rights and Responsibilities of Test Takers* (Joint Committee on Testing Practices, 1998), testing professionals should "make test takers aware of any materials that are available to assist them in test preparation" (p. 8) and should "provide test takers with information about the use of computers, calculators, or other equipment, if any, used in the testing and give them an opportunity to practice using such equipment" (p. 10). Finally, the *Code of Fair Testing Practices* indicates that test users should "provide test takers with an opportunity to become familiar with test question formats and any materials or equipment that may be used during testing" (Joint Committee on Testing Practices, 2004, p. 7).

Test Administration Conditions

A primary ethical obligation in testing is to ensure that test scores accurately reflect the true knowledge, skill, or ability of the test-taker. One way to help accomplish this goal is to establish testing conditions that do not advantage or disadvantage any test-takers. This means, among other things, ensuring that the test setting is conducive to examinees providing their best performance and configured to deter the potential for unethical behavior. Accomplishing these goals requires more than simply providing adequate lighting and seating; test security must be maintained throughout the testing process, including during test administration. According to the *Standards*, "the testing environment should furnish reasonable comfort and minimal distractions" (p. 63) and "reasonable efforts should be made to assure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent means" (AERA, APA, & NCME, 1999, p. 64). The *Rights and Responsibilities of Test Takers* also indicates that testing specialists should "take reasonable actions to safeguard against fraudulent actions (e.g., cheating) that could place honest test takers at a disadvantage" (Joint Committee on Testing Practices, 1998, p. 11).

Beyond ensuring test security, it is an ethical responsibility of testing specialists to ensure that neither the testing conditions nor surface features of the test itself interfere with accurate measurement. The *Standards* notes this

goal in the first of 12 standards related to testing individuals with disabilities, noting that "test developers, test administrators, and test users should take steps to ensure that the test score inferences reflect the intended construct rather than any disabilities and their associated characteristics extraneous to the intent of the measurement" (AERA, APA, & NCME, 1999, p. 106).

To accomplish this, some test-takers with special physical or other needs may require adjustments to the testing conditions to demonstrate their knowledge or skills. In general, there are two broad categories of such adjustments. One category, testing *modifications*, involves an alteration in the testing conditions that also alters the construct intended to be measured by the test and reduces confidence in the validity of interpretations of the examinee's test score. The other category, testing *accommodations*, also involves altered testing conditions, but in such a way as that the construct of interest and intended score inferences are unchanged. For example, allowing an examinee to wear glasses or contact lenses would be an accommodation for a reading comprehension test because the construct of interest is reading comprehension and the use of corrective lenses is unrelated to the measurement of that construct. However, the same adjustment in testing conditions would be considered a modification if the examinee were taking a vision test. In that case, the adjustment *is* related to the characteristic being assessed and would adversely affect the accuracy of conclusions about the examinee's vision. A complete classification system for testing accommodations has been developed by Thurlow and Thompson (2004) and is shown in Table 8.1 with examples of each type of accommodation. Overall, testing specialists must carefully evaluate any alterations in testing conditions to ensure fairness; that is, to ensure that accommodations are obtained by those who need them (and not by those who do not), and to ensure that any alterations do not affect the validity of test scores.

Like test preparation, all the major ethical guidelines for testing address test conditions. According to the *Standards*:

> If the test developer indicates that the conditions of administration are permitted to vary form on test taker or group to another, permissible variation in conditions for administration should be identified and a rationale for permitting the different conditions should be documented. (AERA, APA, & NCME, 1999, p. 47)

The *Rights and Responsibilities of Test Takers* requires that test-takers, if they have a disability, should be advised that "they have the right to request and receive accommodations or modifications in accordance with the provisions of the Americans with Disabilities Act and other relevant legislation" (Joint Committee on Testing Practices, 1998, p. 10). And, according to the *Code of Fair Testing Practices* (Joint Committee on Testing Practices, 2004), test developers should "make appropriately modified forms of tests

TABLE 8.1

Categories of Accommodations

Accommodation Type	Example
Setting	Accessible furniture; individual or small group administration
Timing	Extra time; frequent breaks during testing
Scheduling	Multiple testing sessions; different test days or times
Presentation	Audio, Braille, large-print, or other language version of a test
Response	Scribe to record student's answers; oral, pointing to indicate responses
Other	Highlighters, dictionaries, "reading rulers," or other aids

Source: G. Walz (Ed.), *Measuring Up: Assessment Issues for Teachers, Counselors, and Administrators,* Pro-Ed, 2004.

or administration procedures available for test takers with disabilities who need special accommodations" (p. 4); test users should "provide and document appropriate procedures for test takers with disabilities who need special accommodations or those with diverse linguistic backgrounds" (p. 6).

Scoring Procedures

After administration of a test, examinees' responses must be evaluated. A key fairness issue in evaluating the responses centers on the objectivity and reproducibility of the scoring. When responses are entered directly by examinees via computer or onto a form for optical scoring, the degree of objectivity and reproducibility is typically greater than if the responses involve performances, constructed responses to open-ended test items, or other response types that require human scoring. Of course, objectivity and reproducibility are issues even when scoring is automated. For example, subjective judgments must be made regarding the sensitivity settings on optical scanning equipment; judgments must be made when configuring algorithms for automated essay scoring; and so on. Thus, although it is possible to increase objectivity with these methods, it is not possible to eliminate all subjectivity in scoring.

Fairness in scoring has two aspects alluded to previously in this chapter. A necessary but insufficient condition for fair scoring is that it is consistent. That is, examinees who give the same responses should receive the same scores. Second, the scoring should be valid. That is, variation in scores assigned to responses should reflect variation in the characteristic that the instrument is intended to measure and, to the extent possible, no

other, unintended characteristics. The *Standards* (AERA, APA, & NCME, 1999) provides at least three specific recommendations related to scoring items and tasks:

- "The process of selecting, training, and qualifying scorers should be documented by the test developer. The training materials, such as the scoring rubrics and examples of test takers' responses that illustrate the levels on the score scale, and the procedures for training scorers should result in a degree of agreement among scorers that allows for the scores to be interpreted as originally intended by the test developer. Scorer reliability and potential drift over time in raters' scoring standards should be evaluated and reported." (p. 48)
- "The criteria used for scoring test takers' performance on extended-response items should be documented. This documentation is especially important for performance assessments, such as scorable portfolios and essays, where the criteria for scoring may not be obvious to the user." (p. 46)
- "Procedures for scoring and, if relevant, scoring criteria should be presented by the test developer in sufficient detail and clarity to maximize the accuracy of scoring." (p. 47)

Several quantitative and qualitative procedures can be implemented to facilitate the goals of reproducibility and accuracy. First, raters should be thoroughly trained in the analytical features of the response (i.e., performance, task, essay, etc.) that they will be scoring. Effective training focuses on ensuring that raters attend to the features of responses that are the intended object of measurement. For example, if scoring handwritten essays, training would focus on ensuring that raters evaluate the predetermined aspects of the essay specified in the directions to examinees (e.g., content, word choice, organization, style) and not that aspects deemed to be irrelevant (e.g., handwriting, spelling, neatness).

The process of *rangefinding* is used to help operationalize the boundaries of scoring categories. For example, suppose that a constructed-response mathematics problem appeared on the high school graduation test, and the directions required examinees to solve the problem and explain their solution. A scale might be used that assigned 0 points for missing or completely incorrect response, 1 point for an attempted but incorrect solution, 2 points for a partially correct solution lacking an explanation, 3 points for a correct response, but with missing or inadequate explanation, and 4 points for a correct solution with a complete, accurate explanation. In this scenario, the borderline between score points 3 and 4 is critical and hinges on judgments about the adequacy

of the explanation provided. Extensive work would be required to identify the range of responses that should be considered "inadequate" (and therefore receive a score of 3) versus responses that should be considered sufficiently "adequate" (and assigned a score of 4). Such judgments—admittedly subjective—would need to be made in advance of rater training, and training would need to be included in the training to ensure that similar responses were judged consistently.

Validation samples are used to gauge the efficacy of rater training. Validation samples are actual or model responses that exemplify the scale points in a scoring rubric or essential elements a response must contain to be assigned a certain score. After training, scorers evaluate the validation samples, and targets are established that specify the agreement rate a scorer must attain to qualify to rate operational responses. Raters who do not meet the qualification targets receive additional training or are disqualified from rating operational responses.

In the scoring of operational responses, raters continue to be monitored for their accuracy and consistency in evaluating examinees' responses. To gauge consistency, multiple raters may be assigned to rate the same responses independently, and rater agreement or rater reliability indices may be calculated (see von Eye & Mun, 2005). To monitor accuracy, typical procedures include the insertion (blind to the raters) of validation samples to assess the extent to which raters' scores agree with the (known) scores of the validation responses and assessment of the extent to which raters' exhibit errors of leniency, stringency, central tendency, or drift. These procedures reflect best practices in psychometrics and align with ethical standards of the profession, such as those found in the *Code of Fair Testing Practices* (Joint Committee on Testing Practices, 2004), which indicate that test developers should "provide procedures, materials and guidelines for scoring the tests, and for monitoring the accuracy of the scoring process. If scoring the test is the responsibility of the test developer, [test developers should] provide adequate training for scorers" (p. 6).

Score Reporting and Use of Test Results

The third phase of testing where ethical concerns arise occurs when test scores are reported and test results are used. The following subsections of this chapter describe three aims of score reporting and use where quantitative methods are applied toward ensuring that ethical considerations are attended to when scores are calculated, reported, and used. The three aims include (a) promoting score comparability, (b) protecting confidentiality, and (c) ensuring score integrity.

Score Comparability

A fundamental ethical issue in testing relates to the process of assigning scores to examinees who are administered different test forms (i.e., versions of a test that do not contain identical sets of test questions). Because of security concerns, most high-stakes testing programs use multiple forms. However, it is an issue of fairness that, when examinees are administered different forms, scores across the differing forms should be equivalent in meaning. As Angoff has noted regarding two test forms, X and Y, scores yielded by the forms are comparable if it is "a matter of indifference to [examinees] of every given ability level θ whether they are to take test X or test Y" (1980, p. 195). This means, for example, that it would of great ethical concern if two students of equal ability were differentially likely to pass a high school graduation test solely because they were administered different forms of the test. In general, examinees should not be penalized (or rewarded) for receiving one test form that may be slightly harder (or easier) than another.

Although equivalent forms are intended to be similar in difficulty, it is nearly impossible in practice to construct multiple test forms with exactly the same level of difficulty. *Equating* is a process used to adjust for slight differences in difficulty between test forms so that scores from all forms can be placed on the same scale and used interchangeably. For example, an equating analysis may result in performance of 24 of 30 questions on Form X being determined as equivalent to obtaining 26 of 30 questions correct on Form Y. Although examinees might perceive Form X as being slightly harder and an appearance of "unfairness," this variation in difficulty would not pose an ethical concern if scores on the two forms were properly equated.

There are several different data collection designs used for equating, detailed by Kolen and Brennan (2004) and briefly described here. The first type of equating design is the *random groups* design, in which examinees are randomly assigned to different forms, and each examinee group is assumed to be equivalent in ability. This design could pose ethical concerns if examinee groups are not equivalent, so it is necessary that the assumption of randomly equivalent groups be evaluated. The second design is the *single group with counterbalancing* design, where each examinee takes both (or all) forms of an examination, and the order of the forms is counterbalanced to control for order effects. The third design is the *common-item nonequivalent groups* design, where groups of examinees (that are not necessarily equivalent) take different forms that contain a subset of the same items; these common or "anchor" items are used to place all items on the same scale. The extent to which this design results in valid and ethical score interpretations depends in large part on the characteristics of the common items, which should be evaluated to determine the

extent to which they can be considered a representative subsample of the full test.

Technical details on equating methods are beyond the scope of this chapter, but interested readers should consult Kolen and Brennan (2004) for a thorough discussion of the appropriate methods for each equating design. Some of the most common equating methods include mean and linear equating, equipercentile methods, and IRT methods. In the simplest equating procedure, *mean equating*, scores are converted by adding a constant to account for differences in the mean scores on each form. In addition to the mean, the standard deviation is also taken into account during the transformation process when using linear equating. Equipercentile equating involves transforming the score scales by setting percentile ranks of scores on different forms to be equal. IRT methods achieve score transformation by using item parameters of established items to calibrate new items and estimate examinee ability. Whatever equating method is used, it is important to recognize that scores and subsequent decision making may be affected by both the equating process itself and the resulting error associated with the equating process. The *Standards* (AERA, APA, & NCME, 1999) addresses score comparability, including the importance of providing evidence that scores on different test forms are interchangeable and assuring that relevant assumptions for equating procedures have been satisfied. For example, the *Standards* requires that:

> "A clear rationale and supporting evidence should be provided for any claim that scores earned on different forms of a test may be used interchangeably" (p. 57).
>
> "When claims of form-to-form score equivalence are based on equating procedures, detailed technical information should be provided on the method by which equating functions or other linkages were established and on the accuracy of the equating functions" (p. 57).
>
> "In equating studies that rely on the statistical equivalence of examinee groups receiving different forms, methods of assuring such equivalence should be described in detail" (p. 58).

Confidentiality

Because test data represent information about individuals that can be used in both beneficial and harmful ways, it is an ethical responsibility of those who report test results to ensure that scores are used appropriately. Many testing situations may require that test results be released confidentially and only with the permission of the examinee to those he or she authorizes.

According to the *Standards*, "Test results identified by the names of individual test takers, or by other personally identifying information, should be released only to persons with a legitimate, professional interest in the test taker or who are covered by the informed consent of the test taker." (AERA, APA, & NCME, 1999, p. 87). Additional ethical obligations apply even if the test scores are released appropriately to such persons: "Professionals and others who have access to test materials and test results should ensure the confidentiality of test results and testing materials." (p. 132). The *Code of Fair Testing Practices* indicates that both test developers and test users should "develop and implement procedures for ensuring the confidentiality of scores" (Joint Committee on Testing Practices, 2004, pp. 7, 8).

Confidentiality concerns apply not only to test scores but also to other aspects of testing. For example, according to the *Rights and Responsibilities of Test Takers*, those responsible for test information should also "keep confidential any requests for testing accommodations and the documentation supporting the request" (Joint Committee on Testing Practices, 1998, p. 19). Federal regulations, such as the *Family Educational Rights and Privacy Act* (FERPA, 1974) also apply to test results, which are considered "educational records" under FERPA. According to the FERPA law, except for narrow exclusions, educational records cannot be disclosed within or outside the educational institution to those who do not have a legitimate educational interest in the information. When the test scores or other records are those of a minor, a parent's or guardian's written consent for disclosure must be obtained; if the records are those of an adult, the adult's consent must be obtained.

Finally, even group reporting of test results can lead to inadvertent breaches of confidentiality. The problem of what has been called *deductive disclosure* is of increasing concern in many testing situations. Deductive disclosure occurs when an individual's identity or confidential test information can be deduced using other known characteristics of the individual. For example, suppose that a high school released individual performance results—with students' names removed—from a mathematics examination used as part of granting diplomas. In even moderately large high schools, if the individual test performance data were accompanied by collateral information about each student (e.g., sex, race/ethnicity, class level, number of test retake opportunities, middle school attended), it may be possible to determine the identity of an individual test-taker and his or her test performance.

Ensuring Score Integrity

One of the most important aspects of the test-reporting process is communicating information about the test scores to examinees and other interested parties. A primary ethical consideration in score reporting is the importance of providing information about confidence in test scores

(and/or resulting decisions that are made). All test scores contain a certain amount of uncertainty, and error may be a result of sampling, measurement, and other sources.

Both the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 1998) and the *Standards* (AERA, APA, & NCME, 1999) stress the importance of conveying information about error and precision of test results to the intended audiences. For example, the *Standards* indicates that "The standard error of measurement, both overall and conditional (if relevant), should be reported . . . in units of each derived score recommended for use in score interpretation" (p. 31); they also require that those involved in scoring tests "should document the procedures that were followed to assure accuracy of scoring [and] any systematic source of scoring errors should be corrected" (p. 64). The *Code* recommends that test developers should "provide evidence that the technical quality, including reliability and validity, of the test meets its intended purpose" (p. 4); "provide [test-takers with] information to support recommended interpretations of the results" (p. 6); and "advise test users of the benefits and limitation of test results and their interpretation" (p. 6).

Information about error and precision should include appropriate sources of error and estimates of their magnitude. In addition, the information should be most relevant to the intended uses of the test. For example, technical documentation on a high school graduation test should not be limited to an estimate of reliability or overall standard error of measurement but should also include information about decision consistency and the standard error of measurement at the cut score. The magnitude of error near a performance standard (e.g., a cut point that separates pass and fail categories) would be of primary interest. Speaking directly to the issues of accuracy and precision, respectively, the *Standards* requires that "when a test or combination of measures is used to make categorical decisions, estimates should be provided of the percentage of examinees who would be classified in the same way on two applications of the procedure, using the same form or alternate forms of the instrument" and that "standard errors of measurement should be reported in the vicinity of each cut score" (p. 35).

Another ethical aspect of score integrity is the importance of providing appropriate interpretive aids and avenues for appeal. According to the *Rights and Responsibilities of Test Takers*, test-takers have "the right to receive a written or oral explanation of [their] test results within a reasonable amount of time after testing and in commonly understood terms" (Joint Committee on Testing Practices, 1998, p. 6). Both the *Standards* (AERA, APA, & NCME, 1999) and the *Code of Fair Testing Practices* (Joint Committee on Testing Practices, 2004) stress the importance of communicating how test scores should (and should not) be interpreted, appropriate uses, and the error inherent in the scores. It is important that this information is

conveyed in simple language to all interested parties. According to the *Standards*:

> In educational testing programs and licensing and certification applications, test takers are entitled to fair consideration and reasonable process, as appropriate to the particular circumstances, in resolving disputes about testing. Test takers are entitled to be informed of any available means of recourse. (AERA, APA, & NCME, 1999, p. 89)

The *Code* requires that "test developers or test users should inform test takers about the nature of the test, test taker rights and responsibilities, the appropriate use of scores, and procedures for resolving challenges to scores" and that test-takers should be provided with information about their "rights to obtain copies of tests and completed answer sheets, to retake tests, to have tests rescored, or to have scores declared invalid" (p. 10).

If subscores are reported, it is important that psychometric analyses be conducted at the subtest level. In many cases, subtests may not contain a sufficient number of items to be diagnostically useful, and the test may not have adequate validity evidence for this level of inference. For example, a high school graduation test in mathematics is likely to comprise several subtopics, including geometry and algebra. If scores for the algebra items are to be reported separately, it is important to have psychometric support for such a practice; that is, to demonstrate that the algebra items form a cohesive group and are sufficiently reliable. Comparisons between subscores should take into account the reliability of difference scores. If student performance were compared across different subtests without taking the reliability of difference scores into account, judgments of apparent differences might be entirely due to error in the scores. According to the *Standards*, when such scores are used, "any educational decision based on this comparison should take into account the extent of overlap between the two constructs and the reliability or standard error of the difference score" (p. 147).

A final aspect of score integrity that has ethical implications is the use of test scores for secondary purposes. The *Standards* (AERA, APA, & NCME, 1999) is clear that tests require evidence in support of each intended purpose, and appropriate evidence for one purpose may not support (and may even detract from) evidence needed for a different purpose. Evidence supporting a high school graduation test likely centers on analyses relating student performance on the test to skills needed outside of school. Whether it is ethical to use the results from high school graduation tests for other purposes, such as making inferences about teachers or schools, depends on the extent to which evidence has been collected for those purposes as well. No test is equally justifiable for all purposes, and the

intended inferences must be taken into account when using the scores. According to the *Standards*, "If validity for some common or likely interpretation has not been investigated, or if the interpretation is inconsistent with available evidence, that fact should be made clear and potential users should be cautioned against making unsupported interpretations" (p. 18).

Conclusion

The context of testing—especially high-stakes testing—often comprises decision-making processes that result in classifications that can be consequential for people, groups, organizations, or systems. Personnel hiring and promotion decisions, psychological diagnoses, licensure and credentialing decisions, and educational admission, placement, retention, promotion, and graduation decisions are only a few examples of contexts in which test scores are used, most often in conjunction with other relevant information, to provide or withhold credentials, treatments, opportunities, and so on. All these situations are fraught with junctures at which insufficient psychometric safeguards could adversely affect those affected by test scores. It is only somewhat of an exaggeration to label the ethical concerns as life-or-death matters. In fact, the psychometric technology of standard setting was an important aspect of a U.S. Supreme Court case (*Atkins v. Virginia,* 2002) in which a convicted murderer, Daryl Atkins, had been sentenced to death. The sentence was overturned by the Supreme Court because Atkins' measured IQ of 59, derived from administration of the *Wechsler Adult Intelligence Scale*, fell below a cut score of 60. The execution of mentally retarded individuals was considered by the Court to be "cruel and unusual" and hence prohibited by the 8th Amendment (cited in Cizek & Bunch, 2007, p. 6).

Different, less dramatic, circumstances involving tests occur for individuals in many aspects of their lives, but the ethical concerns are the same. The science and practice of psychometrics has evolved and developed methods that are responsive to these concerns toward the goals of enhancing the accuracy of the information yielded by social science instruments and promoting the appropriate use of test results. The armamentarium of the assessment specialist currently comprises many quantitative tools for facilitating these goals. However, research and development efforts must continue to improve current methods and develop new ones that will equip those who develop and use tests with the tools to improve outcomes for the clients, students, organizations, and others who are the ultimate beneficiaries of high-quality test information.

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.

American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, DC: Author.

Atkins v. Virginia. (2002). 536 U.S. 304.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221–256). New York: Praeger.

Cizek, G. J. (Ed.). (2001). *Setting performance standards: Concepts, methods, and perspectives.* Mahwah, NJ: Erlbaum.

Cizek, G. J. (2006). Standard setting. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 225–258). Mahwah, NJ: Erlbaum.

Cizek, G. J., & Bunch, M. (2007). *Standard setting: A practitioner's guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

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

Crocker, L., & Algina, J. (1986). *An introduction to classical and modern test theory*. Orlando, FL: Holt, Rinehart and Winston.

Family Educational Rights and Privacy Act. (1974). 20 U.S.C.1232.

Gould, S. J. (1996). *The mismeasure of man.* New York: Norton.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89–116). Mahwah, NJ: Erlbaum.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Joint Committee on Testing Practices. (1998). *Rights and responsibilities of test takers: Guidelines and expectations.* Washington, DC: Author. Retrieved from http://www.apa.org/science/ttrr.html

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education.* Washington, DC: American Psychological Association, Joint Committee on Testing Practices.

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

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38,* 319–342.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

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

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research, 45,* 35–44.

Miyazaki, I. (1976). *China's examination hell: The civil service examinations of Imperial China.* New York: Weatherhill.

National Council on Measurement in Education. (1995). *Code of professional responsibilities in educational measurement.* Washington, DC: Author.

Pitoniak, M. J. (2003). *Standard setting methods for complex licensure examinations.* Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Porter, A. C. (2006). Curriculum assessment. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 141–159). Mahwah, NJ: Lawrence Erlbaum.

Porter, A. C., & Smithson, J. L. (2001). *Defining, developing, and using curriculum indicators* (CPRE Research Report Series No. RR-048). Philadelphia: University of Pennsylvania Graduate School of Education, Consortium for Policy Research in Education.

Raymond, M., & Neustel, S. (2006). Determining the content of credentialing examinations. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 181–224). Mahwah, NJ: Erlbaum.

Raymond, M. R. (1996). Establishing weights for test plans for licensure and certification examinations. *Applied Measurement in Education, 9,* 237–256.

Thurlow, M. L., & Thompson, S. J. (2004). Inclusion of students with disabilities in state and district assessments. In G. Walz (Ed.), *Measuring up: Assessment issues for teachers, counselors, and administrators* (pp. 161–176). Austin, TX: Pro-Ed.

von Eye, A., & Mun, E. Y. (2005). *Analyzing rater agreement*. Mahwah, NJ: Erlbaum.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6. Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.

Webb, N. L. (2002). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states.* Washington, DC: Council of Chief State School Officers.

Webb, N. L. (2006). Identifying content for student achievement tests. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 155–180). Mahwah, NJ: Erlbaum.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice, 16,* 33–45.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics* (pp. 45–79). Amsterdam: Elsevier Science.