16

Special Topics and Next Steps

As we reflect on our description of the several standard-setting methods detailed in preceding chapters, we recognize that many topics we would have wanted to include did not seem to fit nicely into an existing chapter. A number of loose ends remain to be tied up; a variety of important issues warrant special treatment; there are many aspects of standard-setting practice that merit additional attention by researchers, and new methods are likely to be introduced to address emerging and more complex assessment contexts. For example, in Chapter 14, we described vertically-moderated standard setting (VMSS), a process or set of processes by which sets of performance standards are adjusted in order to provide a meaningful set of standards across several grades. We noted that this procedure is relatively new in the field of standard setting, in that VMSS is not a method itself for setting individual performance standards, but is in reality a method for *adjusting* a system of individual performance standards using one of the several methods appropriate for doing so.

The concept of VMSS is a perfect exemplar of what we mean by a loose end, a problem in need of additional research. It highlights what we hope will be at least one of the next steps in research on standard setting—the development and investigation of methods for adjusting cut scores. Of course, policymakers and others have made adjustments to individual cut scores for some time; however, none of the adjustment methods (save those for confronting the relative seriousness of false positive and false negative classifications) has any particularly scientific—or even procedural—grounding to provide strong support for its use.

Naturally, we hope that readers have benefited from the practical orientation of this book so far. In the remainder of this concluding chapter, we will try to maintain a focus on the pragmatic aspects of standard setting, while introducing a few topics for which clear answers are not yet available and for which synthesis of field-based experience may prove to be as useful as sustained academic investigation. We include in this list of applied issues methods of adjusting cut scores and the issue of how to incorporate uncertainty inherent in the measurement and judgmental processes into the final result, the problem of how/when to round values that result from a standard setting procedure, the use of multiple methods of setting performance standards, and concern about improving the quality of participant training to perform the standard-setting task.

The reader who is engaged in the art and practice of standard setting likely would be able to suggest other issues that could be addressed here. Although this chapter concludes this book, we hope that it does not close the conversation; we look forward to the reactions, insights, and suggestions of readers with whom we join to advance the state of the art in the challenging and high-profile endeavor of standard setting.

Rounding

To begin our attention to some special topics in standard setting, we start with what might appear to be one of the simplest issues: the problem of rounding. Few standard-setting methods yield an exact cut score. Most commonly used procedures produce an average percentage correct, a mean theta estimate, or some other measure that is then converted into a raw score or scale score cutoff. Both the object of the conversion (e.g., the percent correct or theta value) and the result of the conversion (e.g., the raw- or scale-score cutoff) are almost always fractional values.

For example, it is plausible that an Angoff procedure would yield a cut score of 27.4 on a 40-item test. A Bookmark procedure might yield an average theta of 1.19 (itself a rounded value), which corresponds to a raw cut score somewhere between 27 and 28 (of course the actual interpolated value of 27.36 would likely be used, but again that value has been rounded). In these two cases, what are the values that should be recommended as cut scores? If half-point scores are possible on the test, the decision might depend on the particular rule adopted, and cut scores of 27, 27.5, or 28 may be "correct" for either situation. If half-point scores are not possible, both of these examples would still yield cut scores of either 27 or 28. If, by rule or custom, we round to the nearest score, both yield cut

scores of 27. If we take a more conservative approach that holds that 27 does not meet the cut score of 27.4, then the first obtainable score that does meet the cut is 28.

In typical mathematical applications, rounding rules are clear and well-known. However, in standard setting, because the grist of the standard-setting data analytic mill is more a matter of combining judgments than quantities, it makes some sense to ask the question "How much difference does the choice of a rounding approach make?"

In our experience, the answer is "Quite possibly, a very substantial difference." On a typical statewide student achievement test involving 75,000 students, there are likely to be 1,000 or more students at each raw score point in the vicinity of the cut score. If only to the 1,000 students who just missed the cut (and their parents!) because it was rounded upward or to the 1,000 students who just made it because it was rounded down, the difference is highly consequential. To the six new thoracic surgeons who just made the cut (pun intended)—and to their prospective patients—it can literally make the difference between life and death.

Although the "life and death" phrasing may be somewhat over the top, we hope the point that rounding is not a trivial matter is clear. The issue of when and how rounding should occur should not be relegated to a post hoc matter of last-minute cleanup. Directions for rounding and accompanying rationales should be specified in advance of standard setting. As with consideration of how and when to incorporate information related to uncertainty, rounding rules should be seen as an integral part of the standard-setting process and should be spelled out, debated, and finalized in advance. Of course, as with other critical aspects of standard setting, these rules should be open to discussion later in the process. In addition, such discussion should be clearly focused and documented so that the final results can be properly interpreted.

Methods of Adjusting Cut Scores

It is sometimes the case that an entity responsible for setting performance standards is dissatisfied with the cut scores recommended to it by a standard-setting panel. The entity may wish to modify the performance standards for a variety of reasons. For one, the agency may have additional information to bring to bear, perhaps information that was available but not provided to the standard-setting participants because of concerns about participants' ability to make use of the data in an already complex standard-setting context. Or the information may have become available subsequent to the standard-setting

meeting. A testing context may have changed between a standard-setting meeting and the time the operational test is given, such as a change in time limits, the mode of administration (e.g., paper-based vs. computer-based), or any number of other factors that may have played a part in participants' original cut score recommendations. Some deviation from intended standard-setting processes may have arisen during the standard-setting meeting, casting doubt on the appropriateness of the results (e.g., a facilitator may have become ill and had to leave, an opinionated participant may have inappropriately dominated group discussions, an entire panel might have deviated from the standards-referenced intention of a chosen standard-setting method and applied norm-referenced perspectives in making their judgments, and so on). Finally, the responsible entity may choose to adjust a panel's recommended standards on purely policy, political, or economic grounds.

Of course, measurement specialists have long understood that the need to adjust performance standards would likely arise, and every standard-setting method mentioned in this book has an accompanying adjustment method. For example, calculation of a standard deviation of estimates of cut scores made by individual judges is sometimes used to adjust recommendations stemming from the Angoff (1971), Ebel (1972), and Nedelsky (1954) methods. Similarly, the standard deviation of the distribution of scores of the Borderline Group is sometimes used to adjust final recommendations emanating from use of that method.

More recently introduced standard-setting methods also sometimes include the calculation of an error estimate. The Bookmark procedure derives cut scores by averaging individual theta estimates of examinees at the cut score; those averages have associated standard deviations, also expressed in theta units, which can be used for cut score adjustment. Although not unique to the Body of Work method, it has become somewhat common for adjustments to those results to consider two sources of adjustment information, one based on the variability of participants' judgments and one based on the standard errors of estimate of the logistic regression coefficients. The analytic judgment method and similar procedures likewise include a mechanism for calculating standard error.

A notion underlying most standard-setting procedures is that the cut score is an estimate, not in the sense of a population parameter, but a statistic that is subject to random fluctuation and that would differ to some extent in replications of the procedure under similar conditions, with a different (though equivalent) group of participants, and so on. The cut score is a statistic, derived by taking an average (usually a mean or a median) over participants. Like any statistic, cut scores can thus be thought of as having

associated standard errors of the mean (SE), derived typically by the well-known equation

$$SE = S_x/\sqrt{n} \qquad \text{(Equation 16-1)}$$

where S_x is the standard deviation of the observations of variable x, and n is the number of observations (e.g., examinees for methods such as Borderline Group; participants for others).

To illustrate the use of this standard error, let us suppose that 16 participants using a modified Angoff method recommended a cut score of 32 out of 50 points. The individual estimates of the 16 participants ranged from 29.0 to 35.0, with a standard deviation of 4.0 points. Applying Equation 16-1, we would obtain an SE of 1.0. A board or agency responsible for actually setting the performance standard on the test might consider the final recommended cut score of 32 points and the associated standard error of 1.0 to reach the conclusion that if the standard-setting activity were replicated, the same procedure would result in a recommended cut score between 31 and 33 about two-thirds of the time.

A somewhat related psychometric concept, the *standard error of measurement* (SEM), is also sometimes used as a basis for adjusting cut scores. Whereas the SE focuses attention on variability in the participants' judgments, an agency may also want to take into consideration the reliability of test scores when making a final decision about a cut score. As is also widely known, no test yields perfectly reliable data, and the degree of unreliability can be quantified in classical test theory terms as the SEM as shown in Equation 16-2:

$$SEM = S_x \sqrt{1 - r_{xx'}} \qquad \text{(Equation 16-2)}$$

where S_x is the standard deviation of examinees' observed scores on the test, and $r_{xx'}$ is an estimate of the reliability of the test scores. This is, of course, the simplest expression of the SEM and can be thought of as an average degree of uncertainty across the range of observed scores. For tests constructed and scored based on an item response theory (IRT) approach, an estimate of measurement error at each scale point (literally, the standard error of the estimate of theta) is easily calculated using the inverse of the square root of the information at that point, as shown in Equation 16-3:

$$SE(\theta) = 1/\sqrt{I(\theta)} \qquad \text{(Equation 16-3)}$$

where $I(\theta)$ is the amount of information provided by the test at a given value of ability (i.e., θ) and is obtained by taking the sum of the information yielded

by each item in a test at the given value of theta. Item information can be calculated using equations provided in introductory IRT textbooks (see, e.g., Hambleton & Swaminathan, 1985) or obtained from output of modern IRT software programs (e.g., Bilog, WINSTEPS). These IRT-based errors of estimation can be thought of more generally as conditional standard errors of measurement (CSEMs). Although somewhat less easily obtained, CSEMs can also be obtained using classical test theory methods. The reader interested in a more detailed exploration of this topic is directed to an in-depth treatment of score reliability and decision consistency produced by researchers at ACT (see Colton et al., 1997).

Regardless of the method chosen for considering this type of information, it is clear from the *Standards for Educational and Psychological Testing* that such information is important data that should be reported when cut scores are used (see AERA/APA/NCME, 1999, Standard 2.14). The reason for this requirement is easily illustrated. An examinee's observed score on a test is an estimate of the examinee's true score or latent ability; that estimate has an associated interval that is defined by the standard error of measurement and that is directly related to the reliability of the test. For tests with equal variances, the test that yields more reliable scores will have a smaller interval for a given score point.

For example, consider two 50-item mathematics tests—Form A and Form B—intended to be equivalent and measuring the same construct. On both tests, a cut score of 32 is established, and both tests have a raw score standard deviation of 6.0 points. Form A has a reliability coefficient of .91, while Form B has a reliability coefficient of .84. Now consider two individuals who obtain scores of 31, one on Form A and one on Form B. According to Equation 16-2, the SEM for Test A is 1.8 points, while the SEM for Form B is 2.4 points. For any selected confidence level, the interval for Form B will be wider than for Form A. Consequently, the probability that a student earning 31 points on Test B might have a true score of 32 or higher is considerably greater than that for a student earning 31 points on Test A. Thus it seems important to at least consider whether these two outcomes should be treated the same, or whether test reliability should be taken into account when an agency makes a final decision about a cut score.

Deciding How to Incorporate Uncertainty

Given that there are many ways in which uncertainty is inherent in the standard-setting process and that methods exist for quantifying those levels of uncertainty, the question now is "What do we do with this information?"

One option would be to simply report the information along with the cut score to those responsible for setting the performance standards—though perhaps also to those who are consumers of score information. Or should some use be made of the information in terms of making an adjustment to the cut score? Given the requirements of the *Standards* and perhaps the requirements of ethical science, the first alternative would seem to be mandatory, although it would not preclude the second. Knowledge of the variability of the estimates of the cut score (or viewed from another perspective, the level of agreement or disagreement among the participants) would seem to be crucial to the adopting, or policy-making, body. Similarly, measurement specialists are obligated to report not only reliability coefficients but also various standard errors of measurement. Informed state boards of education, certifying agencies, licensing boards, and other authorities would likely view cut scores with small standard errors (of mean and of measurement) differently from those with larger standard errors.

Let us for a moment explore further the second alternative and use uncertainty information to make adjustments. What sorts of adjustments should be considered? State superintendents and boards of education have been known to lower all cut scores by a fraction of an SEM or even a whole SEM for high-stakes tests, reasoning that in such cases, the student should be given the "benefit of the doubt"—a phrase that seems benevolent, but to some extent masks an implicit policy that favors false positive classification errors over false negative ones.

Of course, by and large the decision to adopt or adjust a cut score is itself essentially a policy decision. An interesting—though perhaps not unique—example from a real statewide student testing program highlights this point. In that state, performance standards were adopted on two different components of the same program (e.g., reading and mathematics), with the primary difference between the two adoptions being that the meetings for setting the cut scores occurred a year apart, with a change in state superintendent also occurring in the intervening year. The first superintendent adjusted the recommendation of the standard-setting panel by lowering the cut score by one-half of an SEM; the adjusted standard was presented to the state board of education, which adopted it. The following year, the new superintendent, fully aware of the practice of the former superintendent, made no adjustment to the panel's recommendation; the unadjusted cut score was presented to the same board of education, which adopted it. For many years afterward, the percentages of students passing the two tests remained remarkably different.

Which superintendent was right? Were they both right? Both acted legally and responsibly, within the bounds of their oaths of office. But the

effects of the decisions were quite different. Moreover, it is relevant to consider whether the decisions would have been equally appropriate if the context had been a medical licensure program, or one pertaining to the certification of nuclear power facility operators, rather than a statewide educational achievement testing issue. Perhaps in these situations it would have been more appropriate to use information about cut score variability or test reliability to adjust the cut score upward, rather than downward. Or to adopt a standard-setting panel's recommendations without any adjustment. How will we know?

One procedurally sound method for considering a response to that question has roots in Nedelsky's (1954) standard-setting method, which included an adjustment factor in the formulation for the cut score, or to use Nedelsky's terminology, the minimum passing level (MPL) used to distinguish between two groups (the F and D groups in Nedelsky's formulation, hence the "FD" subscripts in the following equation):

$$MPL = M_{FD} + kS_{FD} \qquad \text{(Equation 16-4)}$$

where M_{FD} is the mean of participants' summed cut scores (i.e., such that each participant's cut score is the sum of his or her item probabilities), S_{FD} is the standard deviation of that distribution, and k is a variable, undefined in Nedelsky's formulation, but clearly intended by Nedelsky as a value that could take on a range of values depending on how large an adjustment in the cut score, upward or downward, was considered.

Although such a conceptualization would not be well-received in the context of today's standards-referenced measurement methods, Nedelsky suggested that k could be a number that would fix the number or percentage of passing examinees at some desirable level.

Along the same lines, Emrick (1971) introduced the notion of *ratio of regret* (RR) into the calculation of a cut score C_x, as shown in Equation 16-5:

$$C_x = \frac{\log[\beta/(1-\alpha)] + [1/n^*(\log RR)]}{\log[\alpha\beta/(1-\alpha)(1-\beta)]} \qquad \text{(Equation 16-5)}$$

where β is the probability of a false negative (Type 2) error, α is the probability of a false positive (Type 1) error, n is the number of items, and RR is the ratio of regret calculated in such a way that the log of RR would be negative if the cut score needed to be lowered or positive if it needed to be raised.

The raising or lowering of a cut would depend on which type of error was considered to be of greater consequence, more serious, or more to be

avoided. RR would be negative if erroneously failing an examinee were worse, or positive if erroneously passing an examinee were worse. It should be noted that Emrick's formulation was applied at the item level and summed over very short diagnostic tests. Further, it is perhaps most accurate to note that Equation 16-5 is not really an adjustment, but yields a value of the cut score itself.

Given the fact that Emrick's original formulation focused on very short tests and the fact that most modern tests are not only much longer but much more complex in their composition, application of Equation 16-5, in its entirety, may be impractical. Furthermore, we now have a much more effective arsenal of procedures for calculating cut scores than were available in 1971. We include this historical information on Nedelsky's and Emrick's procedures largely as a starting point for cut score adjustments based in decision theory. As we move forward, we will leave most of the mathematical notation behind, but we will rely on the concept of RR in the discussion that follows.

Let us return to the practice alluded to earlier in which the state superintendent adjusted a panel's recommended cut score downward by one-half of an SEM. What was the rationale for the choice of one-half an SEM? It was simply the custom of that superintendent, who had routinely made the same adjustment when presented with panel recommendations on several previous occasions. Again, this was within the purview of his authority as the state's chief state school officer. But the decision was an opaque one. It could have been made much clearer, and it could have involved more stakeholders in its derivation.

Broader dissemination of a cut score adjustment alone—that is, without a compelling rationale—can be troublesome. In Pennsylvania, for example, the selection of an adjustment factor for a set of cut scores for statewide assessments (increasing them by one-fourth of a standard error of the mean) generated controversy and negative publicity that might have been avoided had the rationale for the decision been made more explicit and the choice of the specific adjustment been more open and made before, rather than after, the standard-setting procedure was conducted (Helfman, 2002). And, although one-fourth of an SEM might appear on the surface to be a trivial adjustment, as with all statistics there is the companion issue of practical significance. In the Pennsylvania example, the relatively minor technical adjustment resulted in the classification of 8,000 students as failing who would have passed if no adjustment had been made.

These and other similar scenarios that have been seen in diverse standard-setting contexts highlight the fact that while substantial progress has been made in the methods used to derive cut scores during standard-setting

meetings, considerably less progress has been made in methods and procedures for applying cut score adjustments. We offer a modest proposal to open a dialogue related to the concept of ratio of regret in order to quantify the positions of one or more decision makers or stakeholders in every standard-setting activity.

The concept of ratio of regret is necessarily situation specific. Our first suggested step would be to consider relevant research on the benefits presumed to accrue from making particular classification decisions. For example, if the body of early elementary grades retention research suggests that little is to be gained by holding back marginally proficient (or even clearly skill-deficient) third graders, that information would at least need to be considered if one potential outcome of adopting a performance standard on a third-grade achievement test was retention in grade. If, alternatively, the same test were to be used to identify struggling third graders for additional help during the first half of the fourth grade, and if the proposed program of remediation had proven effective, then the ratio might be reversed.

Beyond initial research to establish a baseline for adjusting cut scores, we propose that individuals who are likely to be involved at the final stage of standard setting (i.e., those who act in official capacities to accept, reject, or modify cut scores recommended by a standard-setting panel) be formally identified early in the process and polled regarding their personal ratios of regret. The process might actually be carried out in much the same way that standard setting is conducted, that is, consisting of steps in which a panel of "experts" is identified, presented with information, polled for their recommendations (in one or more rounds, with or without feedback) and in which summary value is calculated. At minimum, our recommendation for a next step in this area will be to begin formal research and development into much more rigorous and systematic methods and procedures for adjusting cut scores.

Generalizability of Standards

Several approaches to examining the generalizability of recommended performance standards have been suggested. Unfortunately, many applications of standard-setting methods cannot be studied for dependability because they involve only one measurement occasion. That is, the result of the procedure is a single value (usually a mean) based on a single unit of observation (i.e., a single panel of participants).

However, in some standard-setting applications, it is feasible and desirable to conduct the procedure by splitting a larger group of participants into

two randomly equivalent panels. For example, a group of 20 participants would receive common orientation, training, and practice using a standard-setting method. The group of 20 would then be divided into subgroups of 10 members each, which would produce ratings, engage in discussions, and so on independently. Although the data from such a design would likely still be used in the aggregate (i.e., based on all 20 participants), a study design in which independent subgroups were formed would afford the opportunity to estimate a standard error of the resulting performance standards.

One simple method for estimating this quantity has been documented by Brennan (2002). Brennan's approach involves calculation of the standard error of a mean when there are only two observations. Using the means of the two independent groups as the observations, the standard error is calculated as

$$\sigma'_x = |x_1 - x_2|/2 \qquad \text{(Equation 16-6)}$$

where x_1 and x_2 are the cut scores recommended by each of two independent panels and σ'_x is the estimate of the standard error of the performance standard.

More sophisticated approaches to estimating the dependability of recommended cut scores can be implemented. One particularly powerful approach relies on a generalizability theory approach (see Brennan, 1983; Shavelson & Webb, 1991). However, these methods are most appropriate when there are more than two measurement (i.e., rating) occasions—a rare configuration in applied standard-setting practice.

Decision Consistency and Decision Accuracy

Consider an examinee taking a test as part of the process required to earn a high school diploma or obtain certification in a professional field. Obviously, the score on which the graduation/certification decision will be made is less than perfectly reliable. Moreover, the performance standard (i.e., cut score) that the examinee must attain has been set by a committee who did not completely agree on where the cut score should be set but, more than likely, agreed to set it (or them) at some average of all the individual cut scores recommended by the group of participants. Thus an examinee faces a test with a very fixed cut score overlying a matrix of possible cut scores, given the variability of both test score (as measured by a standard error of measurement of the test) and the cut score (as measured by the standard error of the mean of participants' recommendations).

Table 16-1 Hypothetical Classification Frequencies (Proportions) and Calculation of Decision Consistency Estimate

	Classification on Second Administration		
Classification on First Administration	*Fail*	*Pass*	*Total*
Fail	28 (.14)	14 (.07)	42 (.21)
Pass	16 (.08)	142 (.71)	158 (.79)
Total	44 (.22)	156 (.78)	200 (1.00)

NOTES: Agreement coefficient, $p_0 = (28 + 142) / 200 = .85$

Proportion of chance agreement, p_c $= \Sigma\,(p_{k\bullet})(p_{\bullet k})$
$= (.21)(.22) + (.14)(.71) \cong .15$

Decision consistency, κ $= (p_o - p_c)/(1 - p_c)$
$= (.85 - .15)/(1 - .15)$
$= .70/.85$
$\cong .82$

So far, we have examined some of the adjustments that might be made on the basis of one or the other of the two sets of measures affecting our examinee. In this section, we explore some recent advances in decision consistency and then consider both sources of instability simultaneously.

Early work in decision consistency approached the problem of estimating consistency using a strategy parallel to estimating reliability using the test-retest method. Assuming that an examination was administered twice to a sample of examinees (without differences in motivation, effort, knowledge, etc.) and that examinees were classified into performance categories (e.g., pass and fail) on both occasions, the proportion of consistent decisions, symbolized p_o, can be directly calculated. Table 16-1 shows a 2×2 matrix of hypothetical results when a test form was administered to 100 examinees and the same cut score was applied creating passing and failing classifications.

Taking the total number of consistent classifications—that is, the number of examinees classified as passing on both administrations ($n = 142$) plus the number of examinees classified as failing on both occasions ($n = 28$)—and dividing that sum by the total number of classifications (200) yields an estimate of decision consistency. This estimate is sometimes referred to as the agreement coefficient and symbolized as p_o. Based on the

hypothetical data shown in the table, the value of p_o in this case is equal to .85.

In practice, it is often desirable to adjust the value of p_o based on the likelihood that some of the consistency in decision making can be attributed to chance. The proportion of consistent classifications attributable to chance is symbolized by p_c, and the calculation of that value is also shown in the table. Based on the data in the table, the value of p_c is approximately .15. Finally, the value of p_c is then used to adjust p_o; this yields the kappa coefficient (κ) proposed by Cohen (1960). The calculation of κ is also shown in Table 16-1 and, based on the hypothetical data in the table, yields a decisions consistency coefficient, corrected for chance consistency of classifications, of approximately .82.

Of course, one practical limitation inherent in obtaining all of the coefficients just mentioned is that they require the same group of examinees to take the same examination on two occasions. Subsequent research by Subkoviak (1976, 1988) and others has provided the tools for estimating the likelihood that an examinee classified as passing (or failing) on one administration of an examination will be classified similarly on a second administration.

For example, Subkoviak (1988) has provided a straightforward and computationally simple method of estimating an agreement coefficient (p_0) and a kappa (κ) coefficient based on a single administration of a test using a reliability estimate for the total test scores and the absolute value of Z, computed from the following formula:

$$Z = (C_x - M - 0.5)/(S_x) \qquad \text{(Equation 16-7)}$$

where C_x is the cut score for the test, M is the observed test mean, and S_x is the standard deviation of observed scores on the test. Absolute values of the statistic, Z, are then used to obtain the estimates of the agreement coefficient and kappa from look-up tables provided in Subkoviak's (1988) publication and reproduced in Tables 16-2 and 16-3, respectively. To illustrate use of the tables, suppose a test of 100 items was administered to a sample of examinees, that the sample mean and standard deviation were 85.5 and 8.0, respectively, that a cut score of 74 was used to make pass/fail decisions, and that the total score reliability was .70. In this case, the calculated value of Z would be $[(74 - 85.5 - 0.5)/8.0)] = -1.50$. Using Table 16-2, the agreement coefficient, p_0, is found by locating the intersection of the row containing the absolute value of Z (1.50) and the column containing the reliability estimate of .70. The single-administration estimate of p_0 in this case is .92, indicating that a high proportion of consistent decisions would be expected if the

Table 16-2 Approximate Values of the Agreement Coefficient (p_0) for Various Values of Reliability

Approximate Values of the Agreement Coefficient (p_0)									
	Total Test Reliability Estimate ($r_{xx'}$)								
| Z |	.10	.20	.30	.40	.50	.60	.70	.80	.90
.00	.53	.56	.60	.63	.67	.70	.75	.80	.86
.10	.53	.57	.61	.63	.67	.71	.75	.80	.86
.20	.54	.57	.61	.64	.67	.71	.75	.80	.86
.30	.56	.59	.62	.65	.68	.72	.76	.80	.86
.40	.58	.60	.63	.66	.69	.73	.77	.81	.87
.50	.60	.62	.65	.68	.71	.74	.78	.82	.87
.60	.62	.65	.67	.70	.73	.76	.79	.83	.88
.70	.65	.67	.70	.72	.75	.77	.80	.84	.89
.80	.68	.70	.72	.74	.77	.79	.82	.85	.90
.90	.71	.73	.75	.77	.79	.81	.84	.87	.90
1.00	.75	.76	.77	.77	.81	.83	.85	.88	.91
1.10	.78	.79	.80	.81	.83	.85	.87	.89	.92
1.20	.80	.81	.82	.84	.85	.86	.88	.90	.93
1.30	.83	.84	.85	.86	.87	.88	.90	.91	.94
1.40	.86	.86	.87	.88	.89	.90	.91	.93	.95
1.50	.88	.88	.89	.90	.90	.91	.92	.94	.95
1.60	.90	.90	.91	.91	.92	.93	.93	.95	.96
1.70	.92	.92	.92	.93	.93	.94	.95	.95	.97
1.80	.93	.93	.94	.94	.94	.95	.95	.96	.97
1.90	.95	.95	.95	.95	.95	.96	.96	.97	.98
2.00	.96	.96	.96	.96	.96	.97	.97	.97	.98

SOURCE: Subkoviak (1988).

Table 16-3 Approximate Values of Kappa (ê) for Various Values of Reliability

Approximate Values of Kappa (κ)									
	Total Test Reliability Estimate (r_{xx}')								
| Z |	.10	.20	.30	.40	.50	.60	.70	.80	.90
.00	.06	.13	.19	.26	.33	.41	.49	.59	.71
.10	.06	.13	.19	.26	.33	.41	.49	.59	.71
.20	.06	.13	.19	.26	.33	.41	.49	.59	.71
.30	.06	.12	.19	.26	.33	.40	.49	.59	.71
.40	.06	.12	.19	.25	.32	.40	.48	.58	.71
.50	.06	.12	.18	.25	.32	.40	.48	.58	.70
.60	.06	.12	.18	.24	.31	.39	.47	.57	.70
.70	.05	.11	.17	.24	.31	.38	.47	.57	.70
.80	.05	.11	.17	.23	.30	.37	.46	.56	.69
.90	.05	.10	.16	.22	.29	.36	.45	.55	.68
1.00	.05	.10	.15	.21	.28	.35	.44	.54	.68
1.10	.04	.09	.14	.20	.27	.34	.43	.53	.67
1.20	.04	.08	.14	.19	.26	.33	.42	.52	.66
1.30	.04	.08	.13	.18	.25	.32	.41	.51	.65
1.40	.03	.07	.12	.17	.23	.31	.39	.50	.64
1.50	.03	.07	.11	.16	.22	.29	.38	.49	.63
1.60	.03	.06	.10	.15	.21	.28	.37	.47	.62
1.70	.02	.05	.09	.14	.20	.27	.35	.46	.61
1.80	.02	.05	.08	.16	.18	.25	.34	.45	.60
1.90	.02	.04	.08	.12	.17	.24	.32	.43	.59
2.00	.02	.04	.07	.11	.16	.22	.31	.42	.58

SOURCE: Subkoviak (1988).

examination procedure were repeated. Using Table 16-3, the corrected decision consistency coefficient agreement coefficient, κ, is found by locating the intersection of the same values of Z and $r_{xx'}$. The table reveals a single-administration estimate of κ for this situation of .38, indicating the test procedure is adding only modestly to consistency in decision making. Two reasons for such a result are (1) the cut score is located somewhat far away from the area of greatest density of the observed score distribution, and (2) the reliability estimate for the test scores is modest.

One of the recent advances in decision consistency permits estimation of that quantity for tests that result in more than dichotomous pass/fail classifications. For example, Livingston and Lewis (1995) have proposed a four-step method for calculating the decision consistency of a single measure, using the minimum and maximum obtainable scores on the test, the reliability of the test, the length of the test, and the cut scores. They suggest four steps for estimating decision consistency:

1. Estimate the effective test length (n).
2. Estimate the distribution of the proportional true scores (Tp).
3. Estimate the conditional distribution of classifications on another form of the test, for test takers at each true-score level.
4. Estimate the joint distribution of classifications based on true scores and scores on another form of the test. Transform the category boundaries linearly.

Completing these four steps requires the construction of two parallel half-length tests from the original test in such a way that the content, means, standard deviations, and reliabilities of the two half-length tests are identical or nearly so. The remainder of the procedure involves essentially the calculation not just of scores but of categorical classifications on each half-length test.

To illustrate this procedure, suppose that the *Proficient* or passing cut score on an 80-point test were set at 50. Two half-length tests of 40 points each would be constructed to conform proportionally to the same blueprint as the 80-point test (only with half as many items). Now, to set cut scores for the two half-length tests, we would note the percentile ranks for the cut scores on the full-length test. As noted earlier, the Proficient/Pass cut score was 50 out of 80. Even with 40-point half-length tests, we would not necessarily have cut scores of 25. Instead, we would note the percentile rank of the score of 50 on the full-length test and match that rank to the corresponding raw score on the half-length test. Thus, if a raw score of 50 (out of 80) represented a percentile rank of 63, then the cut scores for the two

half-length tests would be the raw scores closest to the 63rd percentile for that half-length test.

The next step in the procedure involves the establishment of a 2×2 contingency table (or, more generally, an $n \times n$ contingency table, with n representing the number of categories into which examinees can be classified) and the calculation of the agreement statistics. Livingston and Lewis (1995) used straight agreement rate (i.e., the sum of the diagonal entries representing exact agreement between the two half-length tests with regard to the category placement of each examinee). However, other agreement indices, such as kappa (Cohen, 1960), could also be used.

A Demonstration of Computing Decision Consistency and Decision Accuracy on Complex Tests

Software has been developed by Brennan (2004a) to simplify the generation of decision consistency and decision accuracy estimates for tests with multiple cut scores (e.g., *Basic, Proficient, Advanced*) and tests that do not consist exclusively of equally weighted, dichotomously scored items (i.e., tests that comprise a mix of select-response and constructed-response items, or tests comprised exclusively of constructed-response, performance tasks, or other polytomously scored formats). The software, titled *BB-CLASS,* is available for download at http://www.education.uiowa.edu/CASMA/DecisionConsistencyPrograms.htm. The zipped file package, bb-class.zip, contains the executable program, a user's manual, sample data sets, and output. The software provides results for the Livingston and Lewis (1995) procedure described previously and is based on either a two- or four-parameter beta binomial model. In addition, *BB-CLASS* provides results for a method proposed by Hanson and Brennan (1990) although that method was designed for tests consisting exclusively of equally weighted, dichotomously scored items.

Running *BB-CLASS* to obtain decision consistency and decision accuracy estimates based on the Livingston and Lewis procedure requires the user to supply only a reliability estimate for the test under consideration, the cut scores to be applied (expressed in terms of both raw and percentage correct scores), and to select the number of parameters to be used (two for a two-parameter beta true score distribution or four for a four-parameter beta true score distribution). (Another program called *IRT-CLASS,* available at the same site indicated in the preceding paragraph, allows the user to input scores in an IRT [i.e., theta or ability] metric.)

Input data for *BB-CLASS* can consist of a list of raw scores or a frequency distribution of raw scores (although the program can also provide

Table 16-4 Sample Input Control File for Estimating Decision Consistency and Decision Accuracy

```
                              11111111112
Column No.          12345678901234567890
Line 1              LL 0.9 4
Line 2              "LL DATA" F 1 2
Line 3              3 140.0 160.0
```

SOURCE: Adapted from Brennan (2004b).

results using only the first four moments of the raw score distribution as input). Although additional options can be added, Table 16-4 shows the basic input control cards required for running *BB-CLASS* based on the sample data set provided with the zipped software package.

The file shown in Table 16-4 consists of three lines of program control information. A *BB-CLASS* control file must be a text-only file (i.e., saved in DOS-TEXT or ASCII format as a .txt file). Note that the information shown in the table regarding "Column No." and the line labels "Line 1," "Line 2," and "Line 3" are *not* included in the control file; these are included in Table 16-4 for reference only.

The first line of the program control file consists of three pieces of information. The characters LL appear in columns 1 and 2 of Line 1; these characters indicate that the Livingston and Lewis (1995) method has been selected. (The characters HB would be substituted if the Hanson and Brennan method were desired.) A space (or tab) follows, then the reliability estimate for the test is entered. In this case, the reliability estimate of 0.9 appears in columns 4–6, followed by another space. Finally, the number of parameters of the desired beta distribution is entered in column 8 (a four-parameter option is shown in Table 16-4).

The second line of the program control file supplies the source of the data file that *BB-CLASS* will use. The data file must be located in the same directory as the control and program files, and it must be enclosed in double quotation marks. In Table 16-4, the date file name "LL DATA" appears in columns 1–9. (The data file used here consists of scores and associated frequencies for 1,000 examinees and is the same data file supplied with the zipped *BB-CLASS* package.) In column 11 of Line 2, the character F indicates that input data are in the form of frequencies (this would be changed to R if the input were raw data). The final two values appearing on Line 2 of the control file specify the location in the data file of the scores and their

associated frequencies. In this case, column 1 of the tab-delimited data file contains the scores; the frequencies associated with each score are found in column 2 of the data file. All entries in a command line are separated by a single space.

The final line of the program control file provides input regarding the number of categories and values for the cut scores used. In this case, the number of classification categories (in this case, 3) appears in column 1. The next two values give the cut scores in raw score units. Columns 4–8 indicate the first cut score (in this case, 140.0); columns 10–14 indicate the second cut score (i.e., 160.0).

Selected output from running *BB-CLASS* using the data set provided in the zipped package and the controls described in the preceding paragraphs are provided in Table 16-5. The table consists of two panels. Information on decision accuracy is provided in the upper portion of the table (16-5a). This table compares classification decisions actually made based on observed scores and classifications that would be made based on estimated true scores. Among the information of interest in this panel are the values shown in bold type at the bottom of the table, including overall probability of correct classification (0.83988) and false positive and false negative classification rates (0.07695 and 0.08318, respectively).

Information on decision consistency is provided in the lower portion of the table (16-5.b). This table compares classification decisions actually made based on expected and observed scores. Among the information of greatest interest in this panel are the values shown in bold type at the bottom of the table, including overall percentage of consistent classification (0.77634), the value of the kappa statistic (consistent classifications corrected for chance agreement; in this case, 0.64685), and the overall probability of inconsistent classification (0.22366).

Other Decision Consistency Procedures

Another straightforward method of calculating decision consistency has been suggested by Brennan and Wan (2004). The method applies to single test administrations of complex tests and utilizes a bootstrap technique. Their approach begins with an examinee's item responses to the full-length test and then randomly from that examinee's response vector a large number of times, calculating a percentage correct score that is compared to the observed percentage correct score. If a sample-based classification agrees with the original decision (e.g., Pass-Pass or Fail-Fail), then the two scores agree; otherwise they do not. Over a large number of sample comparisons, an agreement index is calculated for that examinee. This process is then

Table 16-5 Sample Output from *BB-CLASS*

16-5.a Accuracy Relative to Actual Observed Scores

	Observed Category Classification			
True Category Classification	*1 (Lowest Category)*	*2 (Middle Category)*	*3 (Highest Category)*	*Marginal Values*
1 (Lowest category)	0.17639	0.02541	0.00002	0.20182
2 (Middle category)	0.03759	0.24202	0.05152	0.33113
3 (Highest category)	0.00002	0.04557	0.42146	0.46705
Marginal values	0.21400	0.31300	0.47300	1.00000

Overall probability of correct classification = **0.83988**
False positive rate = **0.07695** False negative rate = **0.08318**

16-5.b Consistency Using Expected (Row) vs. Actual (Column) Observed Scores

	Actual Category Classification			
Expected Category Classification	*1 (Lowest Category)*	*2 (Middle Category)*	*3 (Highest Category)*	*Marginal Values*
1 (Lowest category)	0.16806	0.04193	0.00068	0.21068
2 (Middle category)	0.04527	0.20712	0.07116	0.32355
3 (Highest category)	0.00066	0.06395	0.40116	0.46577
Marginal values	0.21400	0.31300	0.47300	1.00000

Overall percentage of consistent classifications (p_c) = **0.77634**
Percentage of consistent classifications attributable to chance agreement (p_{chance}) = **0.36667**
Estimated kappa ($\hat{e}$) = **0.64685**
Probability of misclassification = **0.22366**

SOURCE: Adapted from Brennan (2004b).

repeated over all examinees so that an overall agreement index may be calculated.

The Brennan and Wan (2004) procedure is made all the more attractive by the availability of computer programs to carry out the bootstrapping procedure. Although the original samples described by those authors were much smaller than most large-scale assessments (129 cases vs. more than 100,000 examinees for many statewide assessments), the programs seem well suited for much larger populations. Their methodology is also adaptable to multiple cut scores; although Brennan and Wan refer to Pass-Fail decisions, it would be just as easy to consider each cut score as a dichotomous decision point and repeat the procedure at each cut score. The primary advantage of the bootstrapping approach over previous approaches is the fact that this approach does not require the construction of new tests. The advantage is not so much the time saved (though that is considerable) as the fact that the requirement to create parallel half-length tests introduces an unknown estimation bias into the process, similar to the bias introduced when estimating reliability using a split-halves approach. Such tests will hardly ever be truly parallel, and the decision consistency estimate will always be dependent on the particular way in which the half-tests were created.

Lee (2005), also working with Brennan and his colleagues at the Center for Advanced Studies in Measurement and Assessment (CASMA) at the University of Iowa, has developed procedures for calculating decision consistency for a compound, multinomial model, along with a computer program (MULT-CLASS). This and other work being performed at CASMA offers considerable promise for the future.

Summary and Future Directions

In conclusion, although sound procedures exist for calculating decision consistency for tests from single administrations, it is clear that these procedures focus on only one aspect of the measurement problem, namely, the reliability of the test, or more specifically, the reliability of the classification of examinees with respect to a fixed cut score. They do not address the stability of the cut score itself or what to do with the information yielded by decision consistency estimates. We can consider these issues with reference to the hypothetical examinee we described previously in this section. That examinee is still facing a fixed cut score that is based on a test and a standard-setting process that leave some room for ambiguity. Livingston and Lewis (1995), Brennan and Wan (2004), and Lee (2005) have suggested how we might at least estimate the variability of one aspect of the classification

decision. Nedelsky (1954), Emrick (1971), and others have suggested how we might estimate the variability of the other dimension. Is there any way to combine estimates of both types of variability and do something with the information?

Let us consider a very simple though highly plausible example. Assume that, in the region of the cut score (50), a certain test has a standard error of measurement (SEM) of 2 raw score points. An examinee with an observed raw score of 49 would obtain a score between 47 and 51 about 68% of the time, if tested repeatedly without fatigue or learning. This score interval includes the cut score 50. Moreover, the cut score was set by a committee on the basis of a final round of standard setting that yielded a mean of 50 and a standard deviation of 5. With a committee of 25 individuals, the resulting standard error of the mean (SE) would be 1 point.

Now let us examine the situation in a slightly different light. Let us start with a cut score of 50. Our same examinee earns the same 49 points, but now we have to interpret the result slightly differently. We have the same 68% confidence interval for the examinee's score, but we also have a 68% confidence interval for the cut score itself. It might be reasonably argued that the cut score should be lowered to 49 (or raised to 51) to reflect the committee's lack of unanimity (examinee passes) or that the examinee's true score could easily be 51 (examinee passes at 49, 50, or 51). A matrix illustrating the scenarios just described is shown in Table 16-6.

Practically speaking, then, how do we use the information such as that presented in Table 16-6? We know how to calculate the stability of one aspect of our decisions. We have not focused on exactly what we should do with those calculations once we have them. Clearly, this is one of the pressing pragmatic issues that has lacked much attention in the applied

Table 16-6 Example of Decision Matrix

	Cut Score		
Examinee Score	*49 (–1 SE)*	*50 (Observed)*	*51 (+1 SE)*
47 (–1 SEM)	Fail	Fail	Fail
49 (Observed)	Pass	Fail	Fail
51 (+1 SEM)	Pass	Pass	Pass

NOTES: Cut score = 50; Examinee score = 49 (SE for cut score = 1 point; SEM for raw score = 2 points)

psychometric literature and that represents a "next step" for research and development in standard setting.

Using Multiple Methods of Standard Setting

A somewhat intuitively appealing idea proposed every now and then is that standard setting should include multiple methods. On the surface, the idea might seem like the perfect solution to the potential problem of the cut score resulting from implementing a single method being unsatisfactory. Training participants in multiple methods and requiring them to apply each method to the same data (i.e., test form or group of examinees) will likely result in two, three, or more possible "answers" to the standard-setting question. This smorgasbord of standards can then be forwarded to the appropriate entity with authority to actually set the standards, and that body then has the luxury of a diversity of choices for the final decision.

We believe that the surface appeal of such an idea stops, well, at the surface. For one thing, the cost of conducting even a single standard-setting procedure is substantial. Subject matter experts must be persuaded to contribute a large amount of time to the endeavor, which can extent to four or five days when the procedure includes several rounds of judgments of individual test items. Logistical arrangements—for such things as transportation, lodging, meeting space, materials, and so on—are also costly. Given the fact that an entity is likely to have finite and limited resources to expend on the standard-setting effort (we think that high-quality test *development* is important too!), it does not seem sensible to spread those limited resources too thinly at the point of standard setting.

Beyond consideration of resources, however, is the fact that a standard-setting procedure should be selected because it presents a strong match with the format of the assessment, the purposes of testing, the skills of the participants, and other factors. Thus, in a given context, it is likely that a single standard-setting method is better aligned with those factors than would be other methods, and the use of the best aligned approach would be preferred.

Finally, we are aware of only a few contexts in which multiple standard-setting methods were used. We are not aware of even a single documented instance in which a systematic, replicable process has been documented for synthesizing the results of the multiple procedures. For example, one high-profile use of multiple standard-setting methods has been described in the context of setting performance standards for a statewide student achievement testing program in Kentucky. A very costly design was followed in

which three methods—Bookmark, Contrasting Groups, and Jaeger-Mills (2001)—were all used to arrive at different possibilities for a system of cut scores on the assessments (see CTB/McGraw-Hill, 2001). However, although the description of each of the individual procedures and their results was adequate, details concerning precisely how a synthesis panel used the discrepant results to arrive at final recommendations are essentially absent from the documentation. The available documentation fails to describe the procedure beyond reporting that the three methods "offered guidance to [synthesis participants] in their efforts to weight particular results and to consider on which information to rely most heavily" (CTB/McGraw-Hill, 2001, p. 23).

Since the time of the work in Kentucky, little if any progress has been made in research and development of methods for combining the results of multiple standard-setting procedures. No methodology currently exists for satisfactorily addressing the challenge that arises when multiple standard-setting procedures result in different answers to the standard-setting question.

It is easy in this case to conclude that research may be needed to clarify the issue. The careful reader will notice that the conclusion just stated was that "research *may* be needed." To be less coy about our position, we will state directly that, for many of the previously cited reasons, we believe that the use of multiple methods is ill-advised currently and in the near future. Our optimism at the possibility that such research will be fruitful is slight, however. The prospect of using multiple methods reminds us of an aphorism attributed to Lee Segal and often referred to as "Segal's Law." We plead ignorance of any biographical detail related to Mr. Segal, but not ignorance about how best to think about the result of implementing multiple standard-setting methods. According to Segal, "A man with a watch knows what time it is. A man with two watches is never sure." Because there is no equivalent of an atomic clock in the field of standard setting, our recommendation is simply for practitioners to invest in a single watch of greatest quality given available resources.

Improving Participant Training

In the 1990s, a standard-setting dustup occurred when a group was opposed to what they perceived to be unrealistically high performance levels set by the National Assessment Governing Board (NAGB) for the National Assessment of Educational Progress (NAEP). The group attacked the method used to set those standards in a widely cited report. The report claimed that the method

used (the Angoff method) was "fundamentally flawed" (Shepard et al., 1993, p. xxiv) and that it presented participants with a "nearly impossible cognitive task" (p. xxiv). It urged that the NAEP performance standards be rejected.

The opinions offered by Shepard et al.—rooted perhaps more in political than scientific grounds—were broadly and conclusively rejected, a rejection with which we also concur. An uncharacteristically frank rebuttal to the Shepard et al. (1993) report was coauthored by an unprecedented collection of psychometricians—11 in all (Hambleton et al., 2000). In the rebuttal, the work of Shepard et al. was evaluated as being marked by a "lack of logic" (p. 8), failing to incorporate research published in scientific journals (p. 7), and "weak[ness] with respect to other aspects of the scientific approach" (p. 7). The report was dismissed as "one-sided, incomplete and inaccurate" and "a disservice to NAGB, educational policy makers, educators, and the public" (p. 13).

Although the initial report itself may have been roundly refuted, it may have had the unintended (or intended) consequence of prompting greater attention to the cognitive processes engaged in by participants in standard-setting procedures. Whether or not it was the NAEP achievement-levels conflagration that has resulted in greater research on the factors considered by standard setters when they make the judgments required by particular methods, we enthusiastically support this endeavor. Without question, we need to know much more about how participants make their judgments, what kinds of information they consider, and how they weight different kinds of information. Much good work is just beginning to be done in this area, and the preliminary results suggest that their cognitions are complex, sometimes idiosyncratic, and clearly warrant further research.

For example, in one recent study, the researchers concluded that participants differed in their understanding of the purpose of the standard setting and the performance categories that had been adopted, they used feedback inconsistently across modes of student assessment, and their understanding of the rating task may be related to the time available for the standard-setting task and their work rate (Skorupski & Hambleton, 2005). In another recent article, participants using an Angoff-based approach and generating low ratings were found to be using a more norm-referenced perspective to judge item performance than participants who generated high or moderate ratings and who tended to apply a more criterion-referenced perspective (Ferdous & Plake, 2005).

One particularly vexing issue requiring practical answers is the question of *when* to provide standard-setting participants with information about the consequences of their judgments in terms of the percentages of examinees that will likely be assigned to various performance categories based on

the proposed cut scores. For example, in procedures involving three rounds of ratings, impact information might be presented to participants after just one round of judgments, as late as the end of Round 3, or at each stage. It is our experience that there is a tendency to provide normative information quite early in the process and more regularly, while impact information is usually presented later (and sometimes not at all). It is also our experience that impact information tends to have a greater influence on participants' judgments the earlier it is provided. However, current research has not provided firm guidance regarding how participants process impact information or regarding interaction effects when various kinds of information (e.g., impact and normative) are provided concurrently. In conclusion, and broadening this line of inquiry, we suggest a next step would be to devote as much research attention over the next decade to studying the larger participant decision-making process as has been devoted to developing standard-setting procedures themselves during the past decade. In our opinion, we now know a great deal about how to set standards but relatively little about what people are thinking while they are doing it.

And, extending this research agenda beyond those who participate in a standard-setting meeting, we recognize that we know virtually nothing about how those who actually *set* standards process the information they are given, namely, superintendents, chief executives of licensing and certification agencies, and other policymakers. While we believe that stating *a priori* a position about how standards should be adjusted is a desirable goal and, as we have indicated, that decision theory provides an effective set of tools for doing so, we first need to find out more about the way this elite group makes decisions. Finally, once the processes of both participants and decision makers are better understood, it is our hope that the technology of instructional design can be brought to bear in order to provide more effective training to both groups so that they are able to complete their important task with fidelity to the method and to the purposes for setting standards in the first place.