DOCUMENT RESUME

ED 384 614 TM 023 200

AUTHOR Cizek, Gregory J.
TITLE Standard Setting as Psychometric Due Process: Going a Little Further Down an Uncertain Road.
PUB DATE Apr 95
NOTE 30p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 19-21, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Criteria; Decision Making; *Due Process; *Educational Assessment; *Evaluative Thinking; Knowledge Level; *Psychometrics; Standards; Synthesis
IDENTIFIERS *Standard Setting; Standards for Educational and Psychological Tests

ABSTRACT

The concept of due process provides an analogy for the process of standard setting that emphasizes many of the procedural and substantive elements of the process over technical and statistical concerns. Surely such concerns can and should continue to be addressed. However, a sound rationale for standard setting does not rest on this foundation. Standard setting on educational assessments will continue to be a fundamental concern because it inescapably involves the collection and synthesis of human judgment. This paper uses the due process analogy to develop suggestions for improving the synthesis of judgment, including: (1) refining clarity of purpose prior to setting standards; (2) pursuing new knowledge related to methods for selecting and training standard setting participants; (3) reevaluating participant consensus as a criterion for successful standard setting; (4) reevaluating the desirability of various "adjustments" used in standard setting; and (5) collecting and expanding professional guidelines for standard setting. An appendix identifies the standard setting references in the "Standards for Educational and Psychological Tests." (Contains 27 references.) (Author/SLD)

Standard Setting As Psychometric Due Process:

Going a Little Further Down an Uncertain Road

BEST COPY AVAILABLE

April 1995

Gregory J. Cizek
Assistant Professor of Educational Research and Measurement
350 Snyder Hall
University of Toledo
Toledo, OH 43606-3390

Draft prepared for the 1995 annual meeting of the National Council on Measurement in Education, San Francisco, CA; please do not cite or quote without author's permission.

ABSTRACT

The concept of due process provides an analogy for the process of standard setting which emphasizes many of the procedural and substantive elements of the process over technical and statistical concerns. Surely such concerns can and should continue to be addressed. However, a sound rationale for standard setting does not rest on this foundation.

Standard setting on educational assessments will continue to be a fundamental concern because it inescapably involves the collection and synthesis of human judgment. This paper uses the due process analogy to develop suggestions for improving the synthesis of judgment, including: refining clarity of purpose prior to setting standards; pursuing new knowledge related to methods for selecting and training standard setting participants; reevaluating participant consensus as a criterion for successful standard setting; reevaluating the desirability of various "adjustments" used in standard setting; and collecting and expanding professional guidelines for standard setting.

Standard Setting As Psychometric Due Process:

Going a Little Further Down an Uncertain Road

In 1993, I wrote an article for the Journal of Educational Measurement, entitled "Reconsidering Standards and Criteria" (Cizek, 1993). Although I was pretty excited about reopening a discussion regarding the fundamentals of standard setting, the article was received with overwhelming apathy. Thus, I was somewhat surprised to learn that a proposal to extend those thoughts was accepted for this conference. I suspect that, perhaps, my mother was one of the reviewers.

In a nutshell, the article sought to reopen a debate about standards and criteria that had begun, formally, in another issue of JEM, back in 1978, and to make explicit and examine some of the assumptions inherent in setting performance standards on educational assessments. The article I wrote was the result of several years of my confusion. Perhaps you could sense that if you read it. My confusion was caused, I think, by at least three factors. First, at the time I wrote the article, I had only the briefest educational exposure to standard setting in my graduate coursework. I consider that whatever increase I might now claim in knowledge about standard setting is similarly modest. Perhaps some of my confusion can be traced to my own limited attention to an arguably narrow topic in applied educational measurement.

A second factor contributing to my confusion I attribute to the fact that, prior to writing the article, I spent about five years actually engaged in various standard setting

activities. I value this training, too. In fact, I tried to keep some track of the numerous nettlesome inconsistencies that kept cropping up along the way. As I attempted to reconcile all of these discordant details of standard setting practice, I simply became more confused about what exactly I was doing.

I don't think I've come much further as of today. In fact, I can tell you about a particularly distressing, short, informal conversation I had recently with an unnamed professor of measurement and statistics. He asked me, bluntly, "Do you really think we can set standards?" I recall answering with an externally-confident, "Well of course we can set standards; we have to; we do it all the time; what do you mean can we set standards?... blah, blah, blah."

Standard Setting As Psychometric Due Process

All of which is to say that I have only marginally increased my level of confidence in the state of the art. And, I am still fairly comfortable with the major conclusion of my earlier paper on standard setting. In that paper, I reviewed what measurement was supposed to be about and I leaned on some analogies taken from the field of law.

First, I thought that standard setting had come in for some particularly harsh criticism. I recall one measurement specialist who referred to standard setting as simply "a partially rigged plebiscite" (Madaus, 1986, p. 13). I tried to bring standard setting completely into the psychometric fold, finding it to simply be a special case of all other measurement. Ghiselli (1964) had defined measurement as:

"going through a prescribed set of operations according to a specified set of rules utilizing specified procedures, instruments, or devices which result in specified descriptions of individuals" (p. 21).

This sounded like standard setting to me.

And, the definition of measurement seemed, to me, to be describing what in the legal profession is called "due process," and which has a fairly firm foundation in constitutional law. Specifically, in describing the powers of the federal government, Article V of the constitution provides that no person shall "be deprived of life, liberty, or property without due process of law."

This due process clause of the constitution has been interpreted over the years to require two aspects of due process: substantive and procedural. Substantive due process requires what is called "fundamental fairness." The procedural aspect requires that individuals, faced with a governmental action that would deprive them of their life, liberty, or property, must be provided with adequate notice, must be afforded an opportunity to be heard, must be provided with a fairly conducted hearing.

My alternative for conceptualizing standard setting combined a definition of measurement with a traditional, legal, notion of fairness. I concluded that defensible standard setting can be viewed as the proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states or degrees of performance. I tried to distance this conceptualization from another

tradition, that of parameter estimation. Instead, I confessed that the alternative conceptualization of standard setting rests on a foundation--like its function--of the ability of human beings to rationally derive, consistently apply, and explicitly describe procedures by which inherently judgmental decisions must be made. Recent courtroom activities have caused me to think that the analogy is even better than I had initially supposed: These proceedings seem less oriented toward getting at an objective truth, and more oriented toward ensuring that all nits get picked in a procedural sense.

In the balance of this presentation, I will try to strain the analogy of psychometric due process a little further, by exploring some of the many other uncertainties I still have about standard setting, and I will try to suggest some possible answers, directions, responses, or dodges for the questions that linger.

Need for Standard-Setting Standards

If there is a discernable void in the literature on standard-setting, it is the absence of assembled guidelines available to those who are charged with establishing standards (Zieky, 1994). Although some researchers have provided discrete suggestions (see, Jaeger, 1990), overall, there is little collected wisdom concerning how to coordinate, design and evaluate standard setting processes and how to evaluate the results of those processes. In short, how can one tell when a standard setting design is "good" or when the results of a standard setting procedure are "satisfactory"? Surely, this is a troubling state of affairs for those who must establish standards and who are concerned that the standards be established in a psychometrically sound and defensible manner.

A logical source to begin looking for guidance concerning evaluating standard-setting would be the Standards for Educational and Psychological Testing (AERA/APA/NCME, 1985). A compilation of the relevant standards is presented in an appendix to this paper. A review of these standards reveals a treatment of standard setting in the Standards that is disjointed, inconsistent, and lacks the specificity afforded to other topics addressed in that document. The Standards themselves acknowledge that many developing areas in testing were not satisfactorily addressed in the 1985 version of the document, noting that: "These standards are concerned with a field that is evolving. Therefore, there is a continuing need for monitoring and revising this document as knowledge develops" (AERA/APA/NCME, 1985, p. 2).

It is hoped that the currently ongoing revision of the Standards (Testing Standards to be Revised, 1993) will provide an appropriate forum for instituting a necessary coordination and revision of guidelines on standard setting. Such a revision could go a long way toward providing a more integrated, interpretive, useful, and generalizable document.

From the standpoint of due process, a revision of the Standards would be welcomed. There are, certainly, numerous ways to conduct a "fair hearing;" guidelines that suggest appropriate frameworks for conducting standard setting would provide a sound basis for those who must implement standard setting to design defensible procedures. There is also, I believe, a need to assemble and explicate rules of evidence for standard setting. For example, what are characteristics of "successful" standard setting? What would make some results acceptable and others unacceptable? How can competing technical and political concerns be weighed? How can the process of standard setting be maintained as procedures

that are accessible and comprehensible to those effected by the results?

Clarity in Standard Setting Purposes

Two aspects of standard setting related to the reason for setting standards in the first place deserve to receive more attention than they currently do. First, I have long believed, but long hesitated to express the following observation: I believe that a lot of needless standard setting occurs. If testing in general should not occur without a compelling purpose or reason for doing so, then the implementation of a performance standard should be accompanied by sufficient justification. Kane (1994a) has summarized the issue of purpose:

> "Before embarking on any standard setting method, however, it is important to consider the fundamental issue of whether it is necessary or useful to employ a passing score.... Assuming that it is necessary or useful to employ a passing score, it is important to be clear about what we want to achieve in making pass/fail decisions, so that our goals can guide our choices at various stages in the standards-setting process" (p. 427).

It is instructive to consider the proliferation of the various licensure and certification testing programs, whether in education or the professions. Many of these offer little clarity of purpose beyond the admittedly noble but rarely substantiated claim of "protecting the public."

Two suggestions are provided here. The first suggestion would be for entities that set

standards to be substantially more specific about the potential for public harm that is ameliorated, and the benefit accrued to the public by the application of a standard. Also, entities responsible for standard setting should also be encouraged to provide some evidence that the application of the standard is actually, demonstrably related to that benefit. This is necessarily a matter of discovering and constructing an argument for validity (see Kane, 1992) based on an evaluative synthesis of available evidence. My observation is that, often times the validity evidence could be mined, but only the surface of the ground has been scratched. A rigorous, well supported argument would, frequently, provide a firm foundation for standard setting and a clarity of purpose that would suffuse the entire credentialling endeavor.

On the other hand, I am persuaded that, in many instances, a rigorous amassing and evaluation of the evidence would reveal that the costs of standard setting far outweigh the real, potential benefits.

A second suggestion is that entities responsible for standard setting carefully, honestly explicate the rationale for setting standards in the first place. It is the stated intention of many licensure and certification programs to award credentials based on demonstrated acquisition of knowledge, skill, competence, or whatever. Nonetheless, a primary raison d'etre of many entities is to enhance the prestige of a vocation, to promote a sense of professionalism for those practicing a vocation, to secure economic benefits for members of a profession, or to maintain the value a credential by, among other things, limiting its acquisition. In education, these concerns are attested to daily in debates about setting standards for high school graduation that maintain "the value of the diploma."

It is my opinion that, in some cases standard setting methods are prostituted to provide a ply of psychometric propriety to a patently political pursuit. Because these are political questions, their answers should derive primarily from political, not statistical, processes. It is difficult for me to see how any traditional standard setting procedures could be defended as the right tool for these jobs.

A second aspect of clarity of purpose is the specification of what the result of a standard setting procedure is intended to reflect. Recent controversies involving setting achievement levels for the National Assessment of Educational Progress (NAEP) has clarified this aspect in my mind. In a review of a National Academy of Education (NAE) report on the NAEP levels setting process, it occurred to me that there were different ways to conceptualize the purpose of standard setting. Even though a performance standard might only be attained by a few examinees, I considered this to be appropriate if the standard was intended to be a distillation of opinion about aspirations, as opposed to a standard based upon judgments about expectations.

Although I had only begun to think about these differences in purpose, Robert Linn (1994) had already carefully explored the same issue. Linn developed a framework for thinking about the various purposes of standards which he derived from the common uses of standards. The four uses consist of: 1) exhortation, 2) exemplification, 3) accountability for educators, and 4) certification of student achievement. Linn notes that these classifications are not necessarily mutually exclusive.

Overall, I believe that much work remains to be done related to clarity in purposes standard setting. The clarity issues can, I think, be classified as "endogenous" or

"exogenous." Endogenous clarity relates to the purpose of standard setting relative to internal needs within the credentialling entity, profession, or system that the standard setting process is intended to serve. It involves the construction of arguments and the collection of evidence that bear on the rationale for even having a test and a passing standard. Endogenous clarity is enhanced by the traditional validation process.

Exogenous clarity refers to the impact and uses of standard setting; exogenous clarity begins with specifying the external purpose of the standard setting endeavor and provides a point of reference--for example, a point of reference that represents a future-oriented aspiration of performance, or a present-time expectation of accomplishment. Exogenous clarity may also involve the collection of evidence to support the reasonableness of the expectation, aspiration, or exhortation.

Selection and Training of Standard Setting Participants

Two areas in standard setting which desperately require attention are the selection and training of participants in the standard setting process. Good advice on these subjects has been provided elsewhere (see Jaeger, 1991; Reid, 1991). I have two concerns about what might be missing from the advice and research evidence that has accumulated so far.

First, continuing the legal analogy, a fundamental guarantee that we often take for granted is the right to a trial by a jury of one's peers. I would not contend that standard setting panels for a statewide 4th grade student competency test be comprised of 9 and 10-year-olds. This would probably result in the first known case of having to adjust a passing score up by several standard errors.

However, I am suggesting that the opposite situation is one that may be occurring with apparent regularity. On many testing programs I am aware of, a standard is "recommended" by a group of content experts who have followed an accepted methodology for setting passing scores. We are careful to call the resulting passing score a recommendation, because it is often adjusted by the entity that is actually responsible for setting the standard. In all of these cases I can think of, the direction of the adjustment is downward. I suspect that the reason for this has something to do with characteristics of the standard setting participants.

In the area I am most familiar with--professional licensure and certification testing--standard setting panels are rarely if ever constituted in a way that reflects how the literature suggests: by drawing a random sample of all qualified judges. The bias I see in the sample of judges is that the persons selected often comprise senior scholars or practitioners in a field who possess a (very) high degree of expertise that is uncommon in the field they represent. Their judgments about the probable performance of minimally competent aspiring entrants into a field and their conceptualizations of adequate knowledge and performance are necessarily a manifestation of their own levels of experience and professional competence. Although some research exists on the interaction between judge competence and expected performance levels, the devil is surely in the details in this area. Additional guidance would be desirable regarding how the technical gaol of empaneling a random sample of qualified persons can be more closely approximated in practice.

A second area in desperate need of attention is the training provided to standard setting participants. In the legal arena, it can take months to select a jury and to educate

them about the task they will be asked to perform. Indeed, in other areas such as athletics, music, or drama, the vast majority time is spent in training for but a fleeting moment of performance. In contrast, the training provided to standard setting participants is often minimal compared to the task they will be asked to perform.

Cone and Foster (1991) writing about the importance of measurement coursework for graduate students in psychology, observed that students "learn complex, sophisticated statistical procedures to test data obtained in elegant, internally and externally valid designs, but they are rarely exposed to the training needed to evaluate whether the data they obtain so cleverly and analyze so complexly are any good in the first place" (p. 653).

This observation applies to standard setting (which is measurement) also. In the NAE report referred to earlier, Shepard and her colleagues (1993) wondered whether critical conceptualizations (e.g., minimally competent) often relied upon in standard setting methods might be too difficult or impossible for standard setting participants to acquire and to adhere to once acquired. In essence, this concern seems to be about the credibility of the testimony provided by participants in the standard setting process.

In order to enhance the credibility of the testimony, I think that two strategies promise some success. First, the easiest thing to propose is that more time be devoted in standard setting to the training of participants. Time spent acquiring consistent conceptualizations of key constructs and becoming proficient in applying a methodology is often shorted in favor of the time required for rating items, reading essays, or "recalibrating" participants. An interesting line of research would be one that actually investigated the return on the investment of additional training.

Second, the content of not merely the amount of training is a critical issue for future study. It will be important to find ways to enhance the training of participants and to ensure that they comprehend and consistently apply the training they receive. I am optimistic about the potential for collaboration between measurement personnel and specialists in instructional design and about the possibility of designing efficient instruction that targets areas of participants' greatest difficulties in applying standard setting methodologies.

In summary, the issue of training is probably one of the least well investigated, yet one that holds great promise for improving the state of the art. In addition to the positive benefit of more reproducible standard setting results, careful attention to the selection and preparation of participants in standard setting exercises is also likely to diminish the need for post hoc "adjustments" to recommended standards. It seems preferable to avoid confronting the issue of how to adjust standards by rigorously implementing procedural safeguards which have the potential to obviate the need for adjustments. This principle has been expounded as a caveat about the power of statistics by Light, Singer, and Willett who note that: "You can't fix by analysis what you bungled by design" (1990, p. viii).

Consensus

Much of the literature on standard setting describes procedures for promoting and measuring interjudge consensus. And, for much of the time I have spent considering standard setting, I have shared the widespread concern about achieving consensus within the group of standard setting participants. However, I have come to believe that consensus is not a necessary or reasonable criterion for evaluating a standard setting procedure.

The following seem to me to be practical questions about which standard setters could use additional guidance: Why is consensus desirable? When does everyone have to agree? Under what circumstances is a hung jury acceptable? If consensus is desirable, what are legitimate, non-Machiavellian procedures for attaining it?

In order to provide a starting point for these discussions, I will assert that what is called interjudge consensus is not a feasible, rational, necessary, or possibly even always a desirable criterion for assessing the outcome of standard setting. However, I will also assert that there are some aspects of standard setting for which consensus is critical, and that consensus can be interpreted in different ways.

First, I begin with a definition of consensus as some level of agreement among participants in a standard setting procedure. There are two areas in which consensus is ordinarily sought. The first area in which consensus is sought is in the establishment of common frames of reference for the participants. These frames include: the content area, the examinee population, the purpose of the test, the purpose of the cutting score, descriptive frameworks used to develop items or tasks, and scoring rubrics. In each of these areas consensus is necessary. A second area in which consensus is sought is in participants' judgments about item ratings, task difficulty levels, etc.

From my experience, I believe that more attention is given to evaluating the degree of consensus about the latter than the former. I also believe that the opposite case should prevail. It seems absolutely critical that each participant in a standard setting study be in complete agreement about the frames mentioned above. Consensus may be promoted in this area by greater attention to training the participants, which I have already mentioned. Jurors

may have differing levels of doubt about their decisions, nonetheless, all of them must be contemplating the same defendant, the same crime, the same evidence, etc.

Accordingly, consensus of judgment about item ratings, task difficulty and so on is not necessarily a reasonable criterion for evaluating the success of standard setting. Although there are certainly situations in which consensus reflects the authentic attainment of a common understanding of the frames of reference (described above), of critical conceptualizations (e.g, "minimal competence"), and of the standard setting methodology. However, these factors are confounded with other factors which can promote an undesirable movement toward consensus, such as the social influences described by Fitzpatrick (1989). Or, as another example, consider that standard setting panels are often specifically constituted to reflect a diversity of perspectives, interests, or constituencies. In such cases, consensus might mean that group process effected a homogenization of viewpoint that would not reflect a desirable synthesis of the perspectives sought. Indeed, failure to achieve consensus could be taken as evidence for effective representation.

Validity of Standard Setting

One of the most elegant descriptions of what a passing score is has been provided by Kane (1994a) in which cutting scores are subsumed under matters of score interpretation, which in turn is subsumed by under the broader heading of validation:

> "It is useful to draw a distinction between the passing score, defined as a point on the score scale, and the performance standard, defined as the minimally

adequate level of performance for some purpose. Validation then consists of a demonstration that the proposed passing score can be interpreted as representing an appropriate performance standard. The performance standard is the conceptual version of the desired level of competence, and the passing score is the operational version" (p. 426, emphasis in original).

In another place, Kane (1994b) has suggested that standard setting methods can be conceived of as representing holistic models "which assume that achievement or skill is highly integrated" and analytic models "which assume that achievement can be assessed using relatively small parts or samples of performance" (p. 4-5)

From one perspective, Kane's work provides additional motivation to actually do what has regularly been called for in educational testing generally--gather and interpret evidence that bears on the validity of inferences based upon test scores (see, for example, Ebel, 1961). From another perspective, however, Kane's work also suggests that validation of passing scores is a substantially more messy endeavor than is usually undertaken.

For example, many of the frequently used standard setting methods seem to implicitly assume an analytic model, despite the fact that such a model may not be the best representation of the underlying cognitive processes for the contexts in which standard setting is utilized. The findings resulting from much of the current work reexamining the links between testing and cognitive psychology (see, e.g., Mislevy, 1993; Mislevy, Yamamoto, & Anacker, 1991; Nichols, 1994; Shepard, 1991) suggests that the integrated, complex characterization--what Kane calls holistic--may be a better match with many testing situations.

For standard setting practice, when there are complex characteristics or abilities assessed, it is unlikely that there will always be a neat one-to-one correspondence between the versions of the passing score identified by Kane; that is, between the conceptual version of the desired level of competence and the passing score as the operational version. This means that validation, if it requires a demonstration that the proposed passing score can be interpreted as representing an appropriate performance standard, is likely to be an even more difficult undertaking.

Currently, a few researchers are studying standard setting procedures that address complex performances (see, for example, the multi-stage, dominant profile method suggested by Putnam, Pence, & Jaeger, 1994). Research along these lines might, in the future, lead to the elimination of the very term "passing score" as the use of a single point along a score scale to define passing and failing status is replaced by "passing profiles" which describe combinations of cognitive states deemed acceptable by knowledgeable observers.

This kind of standard setting practice is reminiscent of a U.S. Supreme Court case, in which the court faced the difficult task of defining pornography. After extensive, serious deliberation, a satisfactory definition could not be derived. In his famous comment revealing the frustration, Justice Potter Stewart remarked:

> "I have reached the conclusion...that criminal laws in this area are constitutionally limited to hard-core pornography. I shall not today attempt to further define the kinds of material I understand to be embraced within that shorthand description; and perhaps I could never succeed in intelligibly doing so. But I know it when I see it" (Jacobellis v. Ohio, 1964, at 197).

By analogy, the success of procedures such as the multi-stage dominant profile method may rest on the ability of cognitive psychologists and psychometricians to develop easily-implemented processes which help tease apart the multiple, complex judgments that contribute to shorthand descriptions such as "minimally competent."

Procedural vs. Substantive Due Process

One of the two aspects of due process is called 'procedural due process.' This aspect requires that individuals must be provided with adequate notice, must be afforded an opportunity to be heard, must be provided with a fairly conducted hearing. In standard setting, the opportunity to be heard is provided via the chosen instrument, and the fairness of the hearing is linked to the validity of the inferences made based on applications of the passing score.

As described earlier, a due process conceptualization of standard setting, procedural due process refers to the derivation of defensible standards through the proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states or degrees of performance. By extension, the rigorous adherence to the procedures delineated in an accepted standard setting methodology can provide a strong source of evidence for the validity of score interpretations. This type of validity evidence gained by fidelity to carefully-followed procedures has been called procedural validity (Kane, 1994a).

Due process also includes a second aspect called substantive due process. In testing, substantive due process would require that the use of a passing score must address a valid

objective and must be fundamentally fair (i.e., not arbitrary or capricious). At minimum, this aspect would argue for increased conceptual accessibility to the standard setting process on the part of those affected by the results of the process.

Recently, I have begun to wonder about the fundamental fairness of the various sorts of "adjustments" made to passing scores at various stages of the standard setting process. The adjustments usually take one of three forms: 1) adjustments to participants; 2) adjustments to the data provided by participants; and 3) adjustments to the final standard (passing score). An examination of the adjustment procedures from a due process perspective brings the phrase "jury tampering" to mind. The following sections describe each of the kinds of adjustments, examine the rationales for each type, and provide an overall recommendation for their use in standard setting practice.

Adjustments to Participants

One way of adjusting the data provided by participants in the standard setting process is to adjust the participants themselves. The most extreme form of adjustment discussed by various researchers is disqualifying a standard setting participant. Geisinger (1991) provides an overview and examples of adverse influences on participants' judgments. He suggests that when "ratings made by individual judges appear inappropriate...these data should be eliminated" (p. 20). Further, Geisinger recommends that, when problems exist within entire panels of participants, "there may be nothing to do but convene another panel" (p. 20). Although Geisinger provides examples of when such action may be warranted, he does not broach the delicate issue of whether or how to inform panelists that they will be eliminated

from the process.

Adjustments to Judgments

The "elimination" of participants carries a fairly ominous connotation; another class of adjustment procedures seems considerably less harsh. These types of adjustments are sometimes referred to as "correcting for rater effects" and have been described in detail elsewhere (see, for example, Houston, Raymond, & Svec, 1991; Raymond & Viswesvaran, 1993). The rater effects in need of correction are usually limited to "leniency" and "stringency", although it is possible that statistical corrections could also be extended to correcting for participants' internal inconsistencies.

Although such corrections have the intuitive appeal of reducing variability in participants' judgments, it is an open question whether reducing variability is more defensible if attained through improved training, by more thorough group consensus building, or by statistical methods. As Jaeger has observed, statistical adjustments aimed at reducing variation may be "antithetical to the more fundamental goal of seeking the informed judgments of one or more samples of judges who represent the population or populations of persons who have a legitimate stake in the outcome" (1988, p. 29).

Adjustments to Passing Scores

Probably the most common adjustment technique involves adjusting the overall passing score that results from combining (e.g., averaging) all of the participants' individual standards. Often, the adjustment consists of raising or lowering the passing score by a

fraction (or multiple) of the standard error of measurement for the test; in the colloquial terminology associated with the practice, it is often referred to as "giving the examinee the benefit of the doubt." When multiple standard-setting methods are applied, some way of reconciling differences between the results of the methods is necessary. Also, when passing scores are established for two or more components of an assessment system (e.g., for a state-level Math and Reading competency test) there is logical need for consistency in the rationales offered for making any adjustments.

In a passionate attempt to compel policy makers to face the effects of such adjustments, Mehrens (1986) argued that the relative effects of incorrect decisions be considered and that the values underlying those considerations be made explicit. He recounts the cases of adjusting the passing scores on teacher licensure tests by three standard errors and the concomitant dramatic effects on false positive licensures and false negative decisions.

Also, the culmination of a standard-setting procedure must also address the criterion of reasonableness or feasibility (van der Linden, 1994). If adjustments to the passing score are to be made, evidence should be presented that supports the reasonableness (i.e., validity) of such an adjustment in terms of other indicators of performance, cost-benefit analyses, or other relevant considerations. Perhaps the optimal approach to this problem is to avoid after-the-fact adjustments by providing standard-setting participants with the relevant ancillary information a priori.

All three of the kinds of adjustments described above are, in essence, adjustments to standards based on post-hoc judgments about the ratings, the standard setting participants, or the passing score. It may be tempting to justify these adjustments on the basis that all

standard setting involves the synthesis of judgment, and the adjustments merely reflect the incorporation of additional information or judgment; this position is tenuous at best. Geisinger has suggested a more cautious approach:

> "We should explicitly decide whether or not to modify our passing scores on the basis of established techniques using factors such as those listed above. We must be clear as to the rationale for such adjustments..." (1991, p. 21, emphasis added).

However, as can be seen from the reason the adjustments seem necessary (e.g., participants did not understand the task, domination of the consensus process by an individual participant, judgments that are unreasonable, and so on) the only truly satisfactory solution to the question of whether or not to adjust standards is to avoid the problem in the first place. All of the motivations for adjusting standards essentially reflect inadequacies in the standard setting procedures implemented (e.g., sampling, participant training, consensus building, group monitoring, provision of normative data, etc.).

In summary, it seems always preferable to avoid confronting the issue of whether and how to adjust standards by rigorously implementing procedural safeguards to potentially obviate the need for adjustments. If it is decided that adjustments to participants, individual judgments, or passing scores are necessary, a detailed explication of the rationale, method, and effect of the adjustment is clearly warranted in a report on the standard setting procedure.

Conclusion

The concept of due process provides an analogy for the process of standard setting which admittedly emphasizes many of the procedural and substantive elements of the process over technical and statistical concerns. Surely such concerns can and should continue to be addressed. However, a sound rationale for standard setting does not rest on this foundation.

It is difficult to imagine a future in which standard setting on educational assessments does not continue to be a fundamental concern. Perhaps the major reason that standard setting will always be a vexing issue is that it inescapably involves the collection and synthesis of human judgment. The suggestions described in this article for improving the synthesis of judgment include: refining clarity of purpose prior to setting standards; pursuing new knowledge related to methods for selecting and training standard setting participants; reevaluating participant consensus as a criterion for successful standard setting; reevaluating the desirability of various "adjustments" used in standard setting; and collecting and expanding professional guidelines for standard setting.

Appendix - Standards on Standard Setting

The Standards for Educational and Psychological Testing (AERA/APA/NCME, 1985) contain several mentions of relevant standard-setting principles. Six individual references to standard setting are listed, with five of the six designated as "Primary" (see Standards 1.24, 5.11, 6.9, 8.6, and 10.9) and one guideline described as "Secondary" (Standard 2.10).

Primary Standards

1) Standard 1.24: "If specific cut score are recommended for decision making (for example, in differential diagnosis), the user's guide should caution that the rates of misclassification will vary depending on the percentage of individuals tested who actually belong in each category."

2) Standard 5.11: "Organizations offering automated test interpretation should make available information on the rationale of the test and a summary of the evidence supporting the interpretations given. This information should include the validity of the cuts scores or configural rules used and a description of the sample from which they were derived."

3) Standard 6.9: "When a specific cut score is used to select, classify, or certify test takers, the method and rationale for setting that cut score, including any technical analyses, should be presented in a manual or report. When cut scores are based primarily on professional judgment, the qualifications of the judges also should be documented."

4) Standard 8.6: "Results from certification tests should be reported promptly to all appropriate parties, including students, parents, and teachers. The report should contain a description of the test, what is measured, the conclusions and decisions that are based on the test results, the obtained score, information on how to interpret the reported score, and any cut score used for classification."

5) Standard 10.9: "A clear explanation should be given of any technical basis for any cut score used to make personnel decisions. Cut scores should not be set solely on the basis of recommendations made in a test manual."

Secondary Standard

1) Standard 2.10: "Standard errors of measurement should be reported at critical score levels. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported for score levels at or near the cut score."

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Cizek, G. J. (1993). Reconsidering standards and criteria. Journal of Educational Measurement, 30(2), 93-106.

Cone, J. D., & Foster, S. L. (1991). Training in measurement: Always the bridesmaid. American Psychologist, 46(6), 653-654.

Ebel, R. L. (1961). Must all tests be valid? American Psychologist, 16(10), 640-647.

Fitzpatrick, A. R. (1989). Social influences in standard-setting: The effects of social interaction on group judgments. Review of Educational Research, 59(2), 315-328.

Geisinger, K. F. (1991). Using standard setting data to establish cutoff scores. Educational Measurement: Issues and Practice, 10(2), 17-22.

Jacobellis v. Ohio, 378 U. S. 184 (1964).

Houston, W. M., Raymond, M. R., & Svec, J. (1991). Adjustments for rater effects in performance assessment. Applied Psychological Measurement, 15(3), 409-421.

Jaeger, R. M. (1988). Use and effect of caution indices in detecting aberrant patterns of standard-setting judgments. Applied Measurement in Education, 1(1), 17-31.

Jaeger, R. M. (1991). Selection of judges for standard-setting. Educational Measurement: Issues and Practice, 10(2), 3-6, 10, 14.

Kane, M. T. (1992). An argument-based approach to validity. Psychological Bulletin, 112(3), 527-535.

Kane, M. (1994a). Validating the performance standards associated with passing scores. Review of Educational Research, 64(3) 425-461.

Kane, M. (1994b). Examinee-centered vs. task-centered standard setting. Paper presented at the Joint Conference on Standard Setting for Large-Scale Assessments, Washington, DC.

Light, R. J., Singer, J. D., & Willett, J. B. (1990). By design: Planning research on higher education. Cambridge, MA: Harvard University Press.

Linn, R. L. (1994, October). The likely impact of performance standards as a function of uses: From rhetoric to sanctions. Paper presented at the Joint Conference on Standard Setting for Large-Scale Assessments, Washington, DC.

Madaus, G. F. (1986). Measurement specialists: Testing the faith--A reply to Mehrens. Educational Measurement: Issues and Practice, 5(4), 11-14.

Mehrens, W. A. (1986). Measurement specialists: Motive to achieve or motive to avoid failure? Educational Measurement: Issues and Practice, 5(4), 5-10.

Mislevy, R. J. (1993, April). Test theory reconceived. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Mislevy, R. J., Yamamoto, K., & Anacker, S. (1991, April). Toward a test theory for assessing student understanding (RR-91-32-ONR). Princeton, NJ: Educational Testing Service.

Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessments. Review of Educational Research, 64(4), 575-603.

Putnam, S. E., Pence, P., & Jaeger, R. M. (1994, April). A multi-stage dominant profile method for setting standards on complex performance assessments. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Raymond, M. R. & Viswesvaran, C. (1993). Least squares models to correct for rater effects in performance assessment. Journal of Educational Measurement, 30(3), 253-268.

Reid, J. B. (1991). Training judges to generate standard-setting data. Educational Measurement: Issues and Practice, 10(2), 11-14.

Shepard, L. (1991). Psychometricians' beliefs about learning. Educational Researcher, 20, 2-9.

Shepard, L., Glaser, R., Linn, R., and Bohrnstedt, G. (1993). Setting performance standards for student achievement. Stanford, CA: National Academy of Education.

Testing Standards to be Revised. (1993, September/October). Psychological Science Agenda, 6(5), pp. 1,4.

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