DOCUMENT RESUME

ED 403 325 TM 026 071

AUTHOR Crocker, Linda; Zieky, Michael
TITLE Joint Conference on Standard Setting for Large-Scale Assessments (Washington, D.C., October 5-7, 1994). Executive Summary, Volume I.
INSTITUTION Aspen Systems Corp., Rockville, MD.
SPONS AGENCY National Assessment Governing Board, Washington, DC.; National Center for Education Statistics (ED), Washington, DC.
PUB DATE Oct 95
NOTE 34p.; For Volume II, the Proceedings, see TM 026 072.
AVAILABLE FROM U.S. Government Printing Office, Superintendent of Documents, Mail Stop SSOP, Washington, DC 20402-9328.
PUB TYPE Reports - Descriptive (141) -- Collected Works - Conference Proceedings (021)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Achievement; Bias; Educational History; Educational Policy; *Educational Research; Educational Testing; *Research Methodology; Standards; *Testing Programs; Test Use; *Validity
IDENTIFIERS *Large Scale Assessment; *Standard Setting

ABSTRACT

The National Assessment Governing Board and the National Center for Education Statistics sponsored a Joint Conference on Standard Setting for Large-Scale Assessments to provide a forum for technical and policy issues relevant to setting standards at local, state, and national levels. This executive summary conveys the essence of the conference by combining its information into major themes rather than merely providing abstracts of the papers. Renowned educators were invited to present papers on issues within historical, theoretical, methodological, application, or policy perspectives. The six major themes identified in their presentations were: (1) multiple meanings and uses of standards; (2) methods of setting standards; (3) new directions and technical issues in setting standards; (4) fairness and validity in setting standards; (5) problems and controversies; and (6) areas of agreement in setting and using standards. The conference did not result in professional consensus on how standards ought to be set for large-scale assessments, but it did bring together many of the people most active in the field of standard setting, and it did promote an understanding of the multifacted issues involved in standard setting. Abstracts of 19 papers are attached. (Contains four references.) (SLD)

THE NATION'S REPORT CARD

Joint Conference on Standard Setting for Large-Scale Assessments

Executive Summary
Volume I

Prepared by Aspen Systems under contract with the National Assessment Governing Board and the National Center for Education Statistics

What is The Nation's Report Card?

THE NATION'S REPORT CARD, the National Assessment of Educational Progress (NAEP), is the only nationally representative and continuing assessment of what America's students know and can do in various subject areas. Since 1969, assessments have been conducted periodically in reading, mathematics, science, writing, history/geography, and other fields. By making objective information on student performance available to policymakers at the national, state, and local levels, NAEP is an integral part of our nation's evaluation of the condition and progress of education. Only information related to academic achievement is collected under this program. NAEP guarantees the privacy of individual students and their families.

NAEP is a congressionally mandated project of the National Center for Educational Statistics, the U.S. Department of Education. The Commissioner of Education Statistics is responsible, by law, for carrying out the NAEP project through competitive awards to qualified organizations. NAEP reports directly to the Commissioner, who is also responsible for providing continual reviews, including validation studies and solicitation of public comment, on NAEP's conduct and usefulness.

The National Assessment Governing Board (NAGB) is established under section 412 of the National Education Statistics Act of 1994 (Title IV of the Improving America's Schools Act of 1994, Pub. L. 103-382). The Board is established to formulate policy guidelines for the National Assessment of Educational Progress. The Board is responsible for selecting subject areas to be assessed, developing assessment objectives, identifying appropriate achievement goals for each grade and subject tested, and establishing standards and procedures for interstate and national comparisons.

The National Assessment Governing Board

PROCEEDINGS OF THE JOINT CONFERENCE ON STANDARD SETTING FOR LARGE-SCALE ASSESSMENTS OF THE NATIONAL ASSESSMENT GOVERNING BOARD (NAGB) AND THE NATIONAL CENTER FOR EDUCATION STATISTICS (NCES)

Hyatt Regency-Capitol Hill
400 New Jersey Avenue, NW
Washington, DC

October 5-7, 1994

Perspectives on Standard Setting for Large-Scale Assessments

• **HISTORICAL**	Describes what the U.S. and other countries learned in 40 years of education standard setting.
• **THEORETICAL**	Examines approaches to conceptualizing standard setting.
• **METHODOLOGICAL**	Explores alternative models for setting standards in education.
• **APPLICATIONS**	Considers technical qualities that yield appropriate interpretations of data.
• **POLICY**	Examines the policy impact of standards in education.

October 1995

Prepared by Aspen Systems Corporation under contract with the National Assessment Governing Board and the National Center for Education Statistics

National Assessment Governing Board
Roy Truby
Executive Director

National Center for Education Statistics
Emerson J. Elliott
Commissioner

Conference Planning Committee

Mary Lyn Bourque
National Assessment Governing Board

Peggy Carr
National Center for Education Statistics

Sharif Shakrani
National Center for Education Statistics

Daniel Taylor
National Assessment Governing Board

For ordering information on these proceedings, write to:

U.S. Government Printing Office
Superintendent of Documents
Mail Stop SSOP
Washington, DC 20402-9328

The contents of this publication resulted from a conference sponsored jointly by the National Assessment Governing Board and the National Center for Education Statistics of the United States Department of Education. The opinions, interpretations, and conclusions of the authors are their own and do not necessarily represent those of the sponsors.

Foreword

The National Assessment Governing Board (NAGB) and the National Center for Education Statistics (NCES) sponsored this Joint Conference on Standard Setting for Large-Scale Assessments in October 1994. The primary purpose of the conference was to provide a forum to address technical and policy issues relevant to setting standards for large-scale educational assessments at the national, state, and local levels.

Nationally and internationally renowned educators were invited to present papers on specific issues within each of the five perspectives:

1. *Historical*: what the United States and other countries have learned in 40 years of educational standard setting;

2. *Theoretical*: approaches to conceptualizing standard setting;

3. *Methodological*: alternative models for setting performance standards in education;

4. *Application*: relevant technical qualities that yield appropriate interpretations of standard setting data (generalizability, validity, fairness, and clarity of communication); and

5. *Policy*: the policy impact of standard setting at the national, state, and local levels, including the use of standards in constructing large-scale assessments.

The presenters participated in small group sessions that focused on each topic. Interactions and discussions between presenters and participants from throughout the nation added to the quality of information gathered and reported in these Proceedings.

In examining various issues surrounding setting performance standards on educational assessments, the Joint Conference achieved the following goals:

- Established some theoretical and empirical foundations for conceptualizing and designing performance standards in large-scale assessments;

- Identified different areas of concern regarding performance standards for large-scale assessments;

- Provided NAGB and NCES with guidance in examining alternative methodologies for developing standards; and

- Informed NAGB and NCES about relevant issues related to setting student achievement levels for the various subjects included in the National Assessment of Educational Progress (NAEP).

Volume I of the conference Proceedings includes an executive summary of the conference and synopses of the conference papers. Volume II comprises the papers prepared for the conference and summaries of the plenary sessions and small breakout sessions. The conference Proceedings provide a rich and valuable source of information for standard setting and should be of great interest to educators and policymakers.

Acknowledgements

Planning and execution of the Joint Conference on Standard Setting for Large-Scale Assessments involved the participation of numerous staff from the National Assessment Governing Board (NAGB), the National Center for Education Statistics (NCES), Educational Testing Service (ETS), American College Testing (ACT), and Aspen Systems Corporation, the logistical services contractor.

The Planning Committee comprising Mary Lyn Bourque and Daniel Taylor from NAGB and Peggy Carr and Sharif Shakrani from NCES provided key guidance for all phases of conference organization and execution. Arnold Goldstein of NCES worked closely with the Planning Committee in organizing the panels, working with the authors, and coordinating the publication process. Patricia Dabbs and Lawrence Ogle of NCES, and Mary Ann Wilmer of NAGB provided invaluable support in the organization of the conference. Summaries of the break-out sessions were prepared by Ruth Palmer. Lilly Gardner, Munira Mwalimu, Darice Stephenson, and Juanita Taylor from Aspen Systems Corporation provided logistical support for the conference and supported publication of the conference Proceedings.

First, we would like to thank the authors and discussants whose work appears in the two volumes of these Proceedings. Their informative presentations and learned papers provide a collection of some of the best current thinking about setting standards for large-scale performance assessments.

Special thanks also go to James Popham of IOX Assessment Associates who not only moderated a session and summarized the findings from the break-out sessions, but provided insight into standard setting based on his long experience in the area of assessment. We would also like to thank the conference moderators: Joseph Conaty, Office of Educational Research and Improvement; Michael Feuer, National Academy of Sciences; Sylvia Johnson, Howard University; and Michael Nettles, University of Michigan.

The facilitators and recorders of the small-group sessions contributed to a large part of the success of the conference. These individuals are: Mary Crovo, Lawrence Feinberg, Ray Fields, Stephen Swearingen, and Daniel Taylor of the NAGB staff; Susan Ahmed, Robert Clemons, Salvatore Corrallo, Arnold Goldstein, Steven Gorman, Jeanne Griffith, Kristen Keough, Andrew Kolstad, Mary Naifeh, Lawrence Ogle, Alex Sedlacek, Sharif Shakrani, Carolyn Warren, Sheida White, and Shi-Chang Wu of NCES staff; Jules Goodison, Eugene Johnson, Steven Lazer, and Paul Williams of ETS; and Luz Bay, Susan Loomis, and Mark Reckase of ACT.

Executive Summary

Joint Conference on Standard Setting for Large-Scale Assessments

Linda Crocker / Michael Zieky

INTRODUCTION

Purpose

Our purpose is to provide a summary of the most important information derived from the Joint Conference on Standard Setting for Large-Scale Assessments held in Washington, DC, October 5 - 7, 1994.[1] We do not intend to provide abstracts of the contents of each paper in this summary. Our intent, rather, is to convey the essence of the conference by combining the information into major themes.

Overview

At the conference, papers were presented in each of four "perspectives" on setting standards: theoretical, methodological, applications, and policy. In addition, a paper providing a historical perspective was distributed to attendees. The six major themes we present in this summary cut across the various perspectives. These themes are:

1. Multiple meanings and uses of standards
2. Methods of setting standards
3. New directions and technical issues in setting standards
4. Fairness and validity in setting standards
5. Problems and controversies
6. Areas of agreement in setting and using standards

MULTIPLE MEANINGS AND USES OF STANDARDS

To policymakers, educators, and psychometricians, the term "standards" has multiple and sometimes sharply different meanings. Authors of the papers presented at the conference adopted one or more meanings in framing their remarks. Brennan noted that standards could be considered either as (a) goals declared desirable by an agency or authority or (b) the outcomes of a standard-setting process. A simplistic distinction could be made between defining standards qualitatively or quantitatively. Determining the author's definition of standards is critical to understanding each paper.

Single Definitions

Some authors implicitly or explicitly adopted a single perspective on the meaning of standards and offered their ideas exclusively within that context. But Smith and Aldrich concentrated on how the

[1] Unless otherwise noted, all references in this summary are to papers contained in Volume II, Proceedings of Joint Conference on Standard Setting for Large-Scale Assessments.

content of national curriculum standards could affect instructional practice without formally defining their meaning of the word "standards." Collins, on the other hand, cited a definition used by the federal government's Office of Management and Budget which is equally applicable to both educational outcomes and manufactured products, but which was unique among the conference's authors in terms of its breadth.

Performance or Content Standards

Other authors reflected on multiple definitions for standards before adopting a single definition for primary use in their papers. Linn identified three types of standards widely used in education: content standards, performance standards, and opportunity-to-learn standards. In this vein, Shepard characterized content standards as curricular frameworks that specify what should be taught at each grade level, while performance standards designate what students must do to demonstrate various levels of proficiency with respect to specific content. Linn, Shepard, and many other authors concentrated primarily on performance standards once their definitions were clearly established. Messick, however, divided his attention almost equally between content standards and performance standards as he discussed validation strategies. Similarly, Phillips, in discussing legal defensibility of standards, seemed to consider both performance and content standards. Finally, Brennan further delineated several types of outcomes of standard setting as performance standards, cut scores on the score scale of an assessment instrument, or exemplars of performance in the form of specific test items or booklets.

Standards as Cut Scores

The perspective adopted by the majority of authors was the view of standards as cut scores (the numeric outcomes of a standard-setting process), benchmarks on a scale, threshold values between contiguous categories, or numeric values that operationalize "how good is good enough" (Livingston, p. 39). The distinction between performance standards and cut scores was elaborated by van der Linden.

> Contrary to a popular point of view, . . . performance standards are verbal descriptions of achievements that form an important step in the process of specification that leads to the domain of test items represented in the test and selects the cut scores. However, once the domain has been realized and the cut scores selected, performance standards lose their operational meaning. From that point on, conversely, the domain of test items and the cut scores define the empirical meaning of performance standards. (p. 98)

This traditional psychometric interpretation of standards as cut scores seemed to undergird the comments of Berk, Bond, Brennan, Hambleton and Slater, Huynh, Jaeger, Kane, Livingston, Mehrens, Shepard, van der Linden, and Zieky.

Multiple Uses

Even when a particular meaning of standards has been established, the intended uses of those standards may still be in question. Messick stressed that different types of standards would have different uses. For example, content standards should be used as "blueprints for teaching and testing" (pp. 294-296) while performance standards, as levels of accomplishment in a specific form, should be

"challenges or hurdles" (pp. 296-299). In particular, Linn addressed the multipurpose use of performance standards (exemplification, exhortation, accountability, and certification).

Many authors discussed procedures of standard setting from the perspective of one primary purpose, but the purpose was not necessarily the same across all papers. For example, Phillips, Jaeger, Moss, and Bond all offered examples that represent use of standards for certification of the ability level of individual examinees. However, Livingston reminded readers that sometimes the purpose of standards may be to report information on groups instead of making decisions about individual examinees. Hambleton and Slater, who offered extensive recommendations for standards-based reporting of National Assessment of Educational Progress (NAEP) results, illustrate use of standards in describing group performance. Thus when interpreting an author's comments, readers will find it useful to ascertain (or infer) the author's perspective on both the meaning of standards and the purpose or purposes served by those standards.

METHODS OF STANDARD SETTING

When assessments are used for certification of individuals or for determining the proportions of students judged to be in various classifications, the need for explicit standards is inescapable. A major decision in such an enterprise involves choice of a method for standard setting. Readers interested in acquiring a broad overview of methods of standard setting will especially want to consult the paper by Mehrens. Categorizing and describing various methods of standard setting was a common topic among authors. One widely used scheme was the *examinee-centered* vs. *test/item-centered* method of classification (e.g., Brennan, Huynh, Kane, Livingston, and Mehrens). Both types of standard-setting studies involve the use of expert judges. As noted by Livingston, item-centered methods "involve judgments about the question tasks or problems on the test," while examinee-centered methods "involve judgments about actual test takers" (p. 40). The choice of method dictates substantially different activities, and the differences in results obtained from different methods are nontrivial.

Examinee-Centered Methods

These methods are typically characterized by the borderline-group methods and the contrasting-groups methods. Kane and Zieky are among the authors who described these methods. With the *borderline-group method*, judges identify a group of examinees whose level of achievement is at or near the threshold of minimum acceptable performance. The median (or in some cases the mean) score of this group on the assessment is computed and used as the minimum pass score. With the *contrasting-groups method*, judges identify two groups of examinees, one consisting of persons considered to be masters and the other consisting of persons considered to be nonmasters of the content of interest. The assessment is administered to both groups and the score level that best separates them is determined. Typically, the score that will result in fewest misclassifications is set as the passing score.

Test/Item-Centered Methods

These procedures for standard setting are typically characterized by the Angoff or Nedelsky methods. As summarized by Kane in the Angoff procedure, the judges are asked to imagine an examinee whose ability lies at the threshold of minimally acceptable performance. For each item, the judges then estimate a minimum pass level (MPL) for a group of such minimally qualified examinees on that item (i.e., they estimate what percentage of the group would answer the item correctly). The average MPL over judges is defined as the item MPL, and the sum of the item MPLs is the passing score. The

Nedelsky method is similar but is designed exclusively for multiple-choice items. As described by Mehrens, judges are required to

> look at each item and identify the incorrect options that a minimally competent individual would know were wrong. Then for each judge, the probability of a minimally competent student getting an item correct would be the reciprocal of the remaining number of responses. . . . The expected score on the test for a minimally competent student would be the sum of obtained reciprocals across all items. (p. 227)

These expected scores are then averaged over judges to create the minimum pass score.

Other methods

Although the four procedures described above were those most commonly mentioned by authors, a number of additional standard-setting procedures were mentioned or reviewed in individual papers. For example, Mehrens and Zieky both described methods proposed by Ebel, decision-theoretic methods, and an iterative judgment method proposed by Jaeger. Mehrens used the label *compromise models* to describe standard setting, employing both normative and absolute standards. The three models (Hofstee, Beuk, and DeGruijter) he reviewed were similar in requiring judges to set a passing score (often in conjunction with an appropriate passing rate), but then using that score in conjunction with empirical knowledge of how examinees actually perform on the test. Mehrens also described several recently developed procedures for standard setting that have been used in preliminary work by the NAEP, the National Board for Professional Teaching Standards (NBPT), the National Assessment Governing Board (NAGB), and various state departments of education. Zieky included descriptions of standard-setting procedures used in several different countries.

Comparisons of Standard-Setting Methods

With such an abundance of standard-setting procedures available, a question that logically arises is, "Does it matter which procedure is used?" Quite simply, the answer to this question is, "YES." According to Berk, "Probably the only point of agreement among standard-setting gurus is that there is hardly any agreement between results of any two standard-setting methods, even when applied to the same test under seemingly identical conditions" (p. 162). Likewise, in his "meta-review" of standard-setting research prior to 1992, Mehrens reported that 23 studies had been conducted in which cut scores for the same assessment were established by different procedures. A number of these studies had involved the Angoff method, leading to his report that Angoff cut scores generally are set between Nedelsky and Ebel cut scores, that intrajudge consistency has been reported higher for the Angoff method than for the Nedelsky method, and that the Angoff method typically has lower variance across judges. Findings such as these and the simplicity of implementing the Angoff procedure contributed to the widespread reliance on the Angoff procedure or one of its variations for establishing cut scores in large-scale, high-stakes assessment programs noted by Berk and Kane.

Interestingly, a major impetus for the conference ensued from recent challenges to the widespread acceptance of the Angoff procedure and to item/test-centered standard setting in general. Several writers offered a brief history of this recent debate about the choice of a standard-setting method, and, taken together, their accounts provide a valuable context for understanding this set of papers. Generally, these authors traced efforts by NAGB to establish standards for Basic, Proficient, and Advanced levels of achievement using a version of the Angoff method. Berk noted, and supported with

seven citations, that "the test-centered methods used by . . . NAEP to set achievement levels for the 1990 and 1992 assessments were the objects of considerable controversy and, in fact, a heavy barrage of criticism" (p. 162).

The criticisms are well-illustrated in Shepard's paper, in which she described the key findings of the National Academy of Education's (NAE) evaluation of the 1992 NAEP achievement levels. As Shepard recounted, the NAEP originally was charged by Congress to evaluate the effect of extending NAEP to the state level, but

> Because of the salience of standards in educational reform, . . . the National Center for Education Statistics (NCES) asked the NAE Panel to expand its work and conduct an evaluation of the 1992 achievement levels for reading and mathematics. (p.143) The Panel's studies were extensive, including more than a dozen separate field studies and reanalysis of existing data. (p. 145)

Shepard pointed out problems that the NAEP perceived with the pool of NAEP mathematics items and their fit with the standards, with the selection of exemplar items to illustrate the standards, and with the descriptions of the achievement levels.

Comparisons of actual student performance on items to estimated p-values obtained from an Angoff-type procedure led to the conclusion that "judges systematically underestimate the easiness of easy items but overestimate the easiness of hard items" (p. 151). In addition to raising questions about the usefulness of the Angoff procedure, according to Berk, the NAEP recommended that other item-judgment methods should be discontinued in favor of the contrasting-groups approach. These recommendations influenced both the substance and flavor of a number of the papers. For example, Berk and Mehrens mentioned specific recommendations of the NAEP as they reviewed different standard-setting methods. Other authors framed their remarks more generally against the broader canvas of examinee-centered vs. item-centered methods of standard setting for large-scale assessments, but drew specific implications for NAEP assessments as they explored technical issues in standard setting (e.g., Huynh).

NEW DIRECTIONS AND TECHNICAL ISSUES IN STANDARD SETTING

This conference afforded the opportunity for experts in the field to consider new problems in standard setting and proposed solutions as well as issues that might affect use of currently available techniques. The recent surge of interest in performance assessment provided impetus for descriptions of two newly proposed standard-setting techniques. The questions surrounding standard setting for NAEP and the possible use of examinee-centered methods in that assessment spawned exploration of a number of technical and statistical issues. Finally, a growing need for developing recognized criteria or guidelines for the conduct of standard-setting studies was recognized, and evaluative criteria were proposed.

New Procedures

Jaeger and Moss offered descriptions of new procedures for standard setting that are particularly appropriate for performance assessments. Jaeger's empirical test of a policy-capturing study using judgment decision-making (JDM) tasks for judges to evaluate hypothetical profiles of examinee performance on seven performance exercises of the certification assessment for the NBPT illustrates a method that combines aspects of both examinee-centered and item-centered judgments. This paper

illustrates application of linear and nonlinear regression models to test whether judges employ compensatory or conjunctive decision-making strategies for setting cut scores on complex performance assessments.

In contrast to Jaeger's psychometric approach, Moss described a standard-setting technique in which judges review a candidate's entire performance across all exercises, determine the candidate's status, and develop a rationale for that determination. In Moss' technique, performance exemplars would serve as standards for different levels of accomplishment in lieu of numeric cut scores. Moss' integrative approach to standard setting was also proposed for application in the context of the certification process of the NBPT standards.

Technical Issues

Various authors noted a sparsity of work in recent years on examinee-centered standard-setting methods and focused their attention on technical issues worthy of attention if such methods were to be used for large-scale assessments. Technical advances in psychometric theory, such as binary and graded Item Response Theory (IRT) models, utility and decision-theoretic models, and generalizability theory were specifically applied to problems in standard setting.

Huynh demonstrated that the minimal examinee sample size needed for standard setting as sample homogeneity and differences between group means vary. Huynh also addressed the ideal item difficulty level for depicting performance of subgroups at specified proficiency levels, and the use of item category boundary estimates or location estimates when calibrations from a graded IRT model are used to create the scale for different proficiency levels. Kane illustrated how the difficulty level of the criterion used to identify contrasting groups affects the threshold score or standard. Kane also discussed the impact, in terms of false positives and false negatives, as the cut score deviates from the mean for the total group, and he described conditions under which false positive or false negative classification errors will increase.

Livingston considered the effect of statistical bias on standards established with item judgment and with borderline-group methods due to effects of regression to the mean and suggested procedures for moderating these effects. The use of continuous utility functions to assign qualifications (proficiency levels) to achievement distributions were proposed by van der Linden. He also described setting "targets" for achievement distributions of groups with respect to these proficiency levels. Brennan applied generalizability theory to conceptualize the problem of how cut scores would vary if the standard-setting process were replicated with different items, judges, and occasions.

"Standards" for Standard Setting

Several authors specifically addressed the growing need for establishing formal criteria, guidelines, or "standards" for practitioners to consider in selection of a method for standard setting and subsequent conduct of the study. For example, van der Linden discussed the following "Standards for Standard Setting" (pp. 107-111): explicitness, efficiency, unbiasedness, consistency, feasibility, and robustness. Berk offered a list of 10 recommendations for standard setting for item-judgment methods, heavily stressing selection and training of judges, as well as a list of "Top 10 Steps to Standard Setting Salvation" (pp. 170-171) which focused on how the judgment process should be structured and carried

out to arrive at useful results. Mehrens also indicated that beyond the selection of a particular standard-setting model, other decisions vital to the integrity of the process include

> how the judges should be chosen, how many should be involved, how they should be trained, whether they should meet as a group, how their tasks should be monitored, how separate judge's decisions should be aggregated, what should be documented, and what information should be taken to the policy board that makes the final decision. (p. 247)

He then offered specific guidelines for these practices.

Brennan's list of nine standards for conducting and reporting results of standard-setting studies included appropriate level of aggregation for reporting, estimation of standard errors, intra- and inter-rater reliabilities, reporting of anomalous outcomes, and cautions to users about possible misuses. Messick's criteria for validity of performance standards (structural, generalizability, external and consequential aspects of construct validity) are also applicable to the issue of choice of an appropriate standard-setting method.

FAIRNESS AND VALIDITY IN SETTING STANDARDS

With respect to fairness and validity, authors agreed that it is impossible to set fair standards on unfair assessments, and that it is impossible to set valid standards on invalid assessments. Even if the assessments are fair and valid, however, the standards themselves must also be shown to be valid and fair. Concerns were expressed about the fairness of applying standards when examinees lacked equal opportunities to learn the material being tested. The authors discussed ways to help assure fairness in the setting of standards by following due process constraints and by involving people representing the perspectives of all relevant groups. Several authors indicated that the fairness and validity of standards depend on the consequences of using the standards.

Fairness Issues

The fairness of standards depends first of all on the fairness of assessments. As Bond noted, "it is not possible to set interpretable standards of proficiency on an assessment that is itself 'unfair' or biased against specific groups" (p. 313). Bond also pointed out that proficiency classifications such as "basic" or "advanced" should mean the same thing across subgroups to avoid harmful misinterpretations of assessment results. Livingston acknowledged that issues of fairness in using standards may transcend psychometric issues that apply in setting standards. The results of a standard-setting study provide one source of information to policymakers who may find it "important to avoid a large imbalance between different groups of students in the awarding of some educational benefit. . . . The issue is one of educational or social policy, not of psychometrics" (p. 40).

Opportunity to Learn

Bond used forceful language to warn that "national standards of educational achievement . . . are antithetical to considerations of fairness and equity if there is substantial inequality in educational opportunity among the population of affected students" (p. 319). Bond acknowledged that educational reform cannot be expected to end "poverty, racism, broken homes, and despair" and concluded that

a more realistic goal "should be to remove any *official* barriers to educational opportunity for all persons and to encourage universal acceptance of the fundamental premise that all children can learn" (p. 319).

Diversity of Perspectives

Collins stressed that standards should be set "with participants drawn from different key interest groups (so that no single interest dominates the standards development process)" (p. 207). Bond similarly maintained that balance across interest groups in setting standards is important for fairness, and he elaborated: "Diversity of *perspective* should take precedence over ethnic, gender, or 'cultural' diversity, per se, although it is unlikely that the former can be completely assured without some attention to the latter" (p. 316). Messick noted, as did Collins and Bond, that "some means of accommodating diverse viewpoints needs to be considered to make consensus meaningful under conditions of pluralism" (p. 300). Berk stated, "The internal validity of the process hinges on the qualifications of the judges and the procedure used to solicit their judgments" (p. 175).

Process Requirements

Collins discussed ways to achieve fairness in the process of setting standards. She generalized from her experience with standards in industrial settings at the National Institute of Standards and Technology to standards of performance in education. Collins summarized due process requirements for setting standards and isolated five "key principles" for equity: adequate notice of proposed actions; ample provision of opportunities for participation; adequate records of all discussions and decisions; timely distribution of minutes and ballot results; and careful attention to minority opinions (p. 207). Collins cited research showing that standard setting in education generally does not meet the requirements established for standard setting in industry and business.

Legal Issues

Phillips noted that fairness was an important factor in the legal defensibility of standards and stated that fairness in the use of standards requires adequate prior notice to students and school personnel of "the specific knowledge and skills for which students will be held accountable and general guidelines on what constitutes acceptable performance" (p. 382). She warned that the use of new standards will bring renewed attention to differential pass rates. "Although differential performance by itself is not sufficient to invalidate an assessment program, defending such a program against a legal challenge based on alleged discrimination can be costly, time-consuming, and detrimental to public relations" (p. 384). On the same topic, Aldrich cautioned that "fears of litigation about fairness of higher standards may create barriers to the widespread use of high standards" (p. 358).

Dimensionality

Hambleton and Slater wrote that "the validity of the criterion-referenced interpretations depends on the extent to which a unidimensional reporting scale fits the data to which it is applied" (p. 329). Bond also linked standards-based interpretations to unidimensionality: "The setting of a performance standard on a given test implies that a more or less unitary construct is being measured" (p. 316). Messick described concerns that certain assessments may measure multiple dimensions, some of which may be construct irrelevant. "For performance standards to be valid, the increasing achievement levels characterized by such terms as 'basic,' 'proficient,' and 'advanced' . . . should reflect increases in complexity of the construct specified in the content standards and not increasing sources of

construct-irrelevant difficulty" (p. 298). A common example of the type of construct-irrelevant difficulty Messick warned against is the use of convoluted language in framing an arithmetic problem.

Validity of Accommodations

Phillips noted a potential conflict between fairness and validity. She warned that accommodations made to increase the fairness of an assessment for people with disabilities may lower the validity of the assessment. For example, a test designed to measure reading comprehension measures a different construct when it is read aloud to examinees. "Drawing the line between a valid and an invalid accommodation requires consideration of the assessment specifications and the uses for the resulting scores" (p. 391).

Construct Validity

Fairness and validity are closely intertwined. Just as Bond began by pointing out that the fairness of the standard is linked to the fairness of the assessment, Messick began by writing that "the validity of these standards cannot be separated from the validity of the assessment itself" (p. 291). Messick elaborated on the requirements of construct validation efforts by raising the issue of the generalizability of standards-based score interpretations across methods of assessment. He indicated that to interpret performance, "in terms of generic constructs requires evidence of generalizability across measurement methods" (p. 296). In addition, he urged that "attention should be paid not just to convergent evidence of consistency across methods, but also to discriminant evidence of the distinctness of constructs within method" (p. 297). Messick warned that the validity of performance standards "is vulnerable to threats of both construct underrepresentation and construct-irrelevant variance" (p. 297). In the former case, nonmasters may meet the standard, and in the latter case, masters may fail to meet the standard.

Validation Evidence

Messick insisted that the construct validity of the assessment and the construct validity of the performance standard "must be evaluated in the same *evidential* terms" (p. 291). Messick stressed that because of the judgmental nature of standards, "their validity depends on the reasonableness of the standard-setting process and of its outcome and consequences, not the least of which are passing rates and classification errors" (p. 300). Kane noted that using performance standards adds validation burdens beyond those required for use of the scores in the absence of standards and wrote, "it is necessary to choose a standard-setting method judiciously and to generate evidential support for the assumptions embedded in the standards-based interpretation" (p. 120). If examinee-centered methods of setting standards are used, Kane argued, "the criterion used in the examinee-centered standard-setting study should be consistent with the test-score interpretation and, therefore, that the test scores should be validated in terms of this criterion" (p. 135). Berk urged consideration of consequential validity in terms of "measures of the political, economic, social, and/or educational outcomes of decisions about examinees" (p. 176). Ultimately, "Decision validity . . . is the acid test of the worth of a standard-setting method " (p. 176). Berk acknowledged, however, the difficulty of obtaining useful criterion data for professional licensing and certification examinations.

Validity and Values

With respect to evaluating the validity of the outcomes and consequences of the use of standards, Zieky maintained "that there is no objective proof of the validity of a standard," because people may "disagree as to whether a particular outcome is appropriate or not" (p. 32). Outcomes considered valid by individuals who hold one set of values may be considered invalid by those with opposing values.

PROBLEMS AND CONTROVERSIES

It became clear at the conference that standard setters continue to disagree about many aspects of their work. No method of setting standards is universally accepted. The Angoff method, which has been the most widely used means of setting standards, was characterized as "fundamentally flawed" by some authors and defended by others. Not all authors agreed that the use of standards would be beneficial, even if the standards had been appropriately set. Authors elaborated on the difficulties of setting standards, noted the legal vulnerabilities of standards, discussed problems in interpreting the results of using standards, and failed to reach consensus on a number of controversial issues.

Lack of Consensus

One of the most pervasive themes of the conference was the constant airing of the unsolved problems and unresolved controversies associated with setting and using standards. As Bond succinctly stated, "The essential problem stems from the simple fact that there is no way to set performance standards that enjoys consensus among measurement specialists" (p. 312). Berk elaborated on the lack of agreement among practitioners: "What optimal combination of ingredients produces an effective standard-setting procedure? . . . We don't know! The problem is that the measurement community has never reached consensus on a set of criteria that can operationally define the 'effectiveness' of any single standard-setting method" (p. 164).

Controversy Concerning the Angoff Method

As noted above, an important source of controversy was Shepard's conclusion that the Angoff method, the most commonly used procedure for setting standards, is "fundamentally flawed" and presents an extremely complex cognitive task that judges are not capable of doing[2] (p. 151). The effect on many standard setters of the characterization of the Angoff method as fundamentally flawed was captured by Berk. "When I first read this . . . it blew me off my beach chair. The Angoff method . . . has been one of the most trusted names in standard setting. And now, it seems as though it's fighting for its life. . . ." (p. 162).

[2] Zieky (1994, p. 30) questioned the complexity of the cognitive judgments actually made by Angoff judges. He believes, without proof, that the judges are not actually performing the difficult task of estimating the probability that a member of some hypothetical group of examinees would answer an item correctly. Further, he believes that the judges are, rather, engaged in the much simpler task of expressing their own values concerning how well examinees would have to perform on an item before the judges are willing to say the examinees are competent.

A strongly contrasting view of the Angoff method was offered by Mehrens when he cited a number of studies and concluded,

> The review of the literature suggests the general acceptance of the Angoff method as the preferred model, and this is my recommendation. The recommendation is based on the general reasonableness of the standard set, the ease of use, and the psychometric properties of the standard. (p. 231)

Kane also defended Angoff and other test-centered methods, at least for objective, analytically scored tests. He pointed out that the Angoff method has been used "on a host of licensure and certification tests, as well as on numerous state testing programs, without major complaints from the judges involved" (p. 124). Kane offered rebuttals to a number of the studies upon which Shepard based her conclusion and stated,

> The evidence developed in the five studies of the technical properties of the 1992 NAEP standard setting do not seem to justify the conclusion (Shepard et al., 1993, p. 77) based largely on these studies, "that the Angoff procedure is fundamentally flawed because it depends on cognitive judgments that are virtually impossible to make." (p. 129)

No consensus was formed, but it was not the purpose of this conference to reach closure on the appropriateness of the various approaches to setting standards of performance for the NAEP.

Disagreements on Uses of Standards

The description of disagreements was an ongoing motif at the conference. Aldrich cited authors who disagreed about whether or not it was beneficial to students to set standards at all, regardless of how it is done. Aldrich indicated that those opposed to standards characterize them as a "harmful fantasy" that would take attention away from matters of equity, while those who favor standards see them as a strong motivating force that will help all students. Aldrich herself favored standards as "critical navigational aids," as long as the necessary supporting mechanisms were available (p. 358).

Smith also described controversies about the establishment of standards. He reported that "local people . . . are not prepared to surrender what they believe to be the right of local authority to define the nature of their schools. . . . It is unwelcome news to them that the state will tell them what is to be done . . ." (p. 405). He pointed out that many unresolved issues are likely to lead to future controversies and "setting performance standards will create a great many risks for states, educators, and local policymakers" (p. 406).

Difficulties of Setting Standards

Even if agreement can be reached on whether or not to set standards, there will remain great difficulties in defending any particular standards that are set. Jaeger identified setting standards of performance as examples of "judgment or decision-making tasks (JDM)" (p. 57), and noted that "Responses to JDM tasks, including standard-setting tasks, are . . . responses to problem statements that are replete with uncertainties and less-than-complete information" (p. 58). According to Linn, standard setting looks easy, but is actually hard to do.

> Although both the step of converting a description of a performance standard into a defined range of scores on an assessment and the step of describing what students who meet a given standard of performance actually know and can do may appear straightforward, satisfactory accomplishment of these two steps has proven to be extraordinarily difficult. (pp. 369-370)

The difficulty of setting standards was also clearly articulated by Brennan. "Standard setting is a difficult activity, involving many a priori decisions and many assumptions" (p. 285).

Arbitrary Nature of Standards

According to van der Linden, standards are often perceived as arbitrary.

> The feelings of arbitrariness . . . stem from the fact that although cut scores have an "all or none" character, their exact location can never be defended sufficiently. Examinees with achievements just below a cut score differ only slightly from those with achievements immediately above this score. However, the personal consequences of this small difference may be tremendous, and it should be no surprise that these examinees can be seen as the victims of arbitrariness in the standard-setting procedure. (p. 100)

Legal Vulnerability of Standards

That "feeling of arbitrariness," along with the many unsolved problems and disagreements among recognized experts in the field, intensify many problems in the legal defensibility of standards. Phillips enumerated the conditions "most likely to trigger a legal challenge to a high-stakes, large-scale assessment and its associated standards" (p. 380), including,

> adverse impact on historically disadvantaged groups; use of processes perceived to be unfair, arbitrary or capricious; suggestion that specific attitudes or values are being assessed; failure to provide all accommodations requested by the disabled; and assessing knowledge or skills that examinees have not had the opportunity to learn. (p. 380)

(Note that Glass (1978) forcefully and articulately stated that all standards were arbitrary. Thus, it appears that all standards are vulnerable to legal attack.)

In addition, she warned that performance assessments carry additional vulnerabilities, including lack of experience with the methodology, problems in scaling and equating, lack of generalizability, and the use of fallible human judgments in scoring. High stakes assessments and standards may be attacked under the Fourteenth Amendment to the U. S. Constitution unless adequate prior notice of the new testing requirement has been given, and unless both procedural and substantive "due process" has been

followed. That is, both the procedures used to administer the test and the test itself must follow professional standards and be fair[3] to all examinees.

Phillips further cautioned that under First Amendment freedom of speech and free exercise of religion challenges, "parents with definite religious or political views may object to any goal or standard that appears to require the student to espouse a specific belief or point of view" (p. 387). Phillips also alerted standard setters to the need to be very specific about performance requirements and to explicate their assumptions. For example, standard setters "cannot assume that the courts will accept an unstated assumption that the assessment goals intended the measurement to be in English. . . ." (p. 392).

Problems in Reporting Results

Even if standards can survive legal challenge, there remain difficulties in reporting the results. Hambleton and Slater interviewed "policymakers, educators, and people in the media" and found that "many interviewees had problems reading and interpreting the information they were shown" (p. 336) in reports of assessment results in terms of achievement levels. The authors discovered, for example, that only about 10% of the people they interviewed were able to interpret correctly the percentages presented in a table of NAEP results. Even though readers who spent more time on the reports could improve their comprehension, the majority "noted that they did not have the time needed to scrutinize these reports until they could understand them" (p. 340).

Inevitable Nature of Controversy

Are the controversies aired at the conference likely to be resolved over time? Controversial issues have plagued standard-setting activities for years and are not likely to disappear. Zieky, attempting to give a historical perspective, described an "Age of Disillusionment" that began soon after standard setting became a matter of widespread professional attention. He posited that standard setting would necessarily remain controversial because "a standard is *not* a statement of some psychometrically derived truth. A standard is, rather, the embodiment of a set of value judgments. As long as different people hold different values, standard setting will remain a matter of controversy" (p. 29).

AREAS OF AGREEMENT IN SETTING AND USING STANDARDS

Even though controversies and disagreements abounded at the conference, there were some areas of general agreement. Authors agreed that setting standards was a difficult, judgmental task and that the procedures used were likely to disagree with one another. There was clear agreement that the judges employed in the process must be well trained and knowledgeable, represent diverse perspectives, and that their work should be well documented.

[3] Because there are many, sometimes contradictory, definitions of "fair" in the context of testing, it is almost always possible to find a published definition under which a test is either unfair to individuals or unfair to members of historically disadvantaged groups.

Standards as Difficult and Judgmental

Mehrens listed several areas of agreement among the experts in standard setting. The first two points of agreement are probably universal among standard setters. "Although the literature on standard setting is inconclusive on many points, there does seem to be agreement that (a) setting defensible standards is difficult," and that "(b) standard setting is a judgmental process" (p. 224). The difficulty of setting standards was commented on by so many authors that it was selected as a component of one of the main themes for this summary (see p. ES-10). Acknowledgment of the judgmental nature of standards is so widespread that Zieky called it "probably the single greatest area of agreement in the history of setting standards". (p. 28).

Lack of True Standards

Authors were in general agreement that standards are constructed rather than discovered and that there are no "true" standards. Mehrens quoted Jaeger (1989, p. 492) that "a right answer does not exist, except perhaps in the minds of those providing judgments" (p. 224). Shepard pointed out that "a number of reviewers" agreed that "true standards do not exist in nature like parameters waiting to be estimated by a reliable statistical procedure" (p. 158). Similarly, van der Linden commented, "some policymakers or educators seem to believe that true standards do exist independently of methods and judges. . . . This view is not correct" (p. 108). Brennan found it to be "particularly important that users be disavowed of any belief that standard-setting outcomes are anyone's 'truth'" (p. 285).

Imperfection of Standard-Setting Procedures

Authors concurred that standard-setting procedures are merely imperfect mechanisms for the collection of information. According to Shepard, "regardless of how statistical they seem, standard-setting procedures are merely formal processes intended to help judges approach their conceptual task in as systematic a fashion as possible" (p. 158). In a similar vein, van der Linden commented on the limitations of standard-setting procedures by noting, "a standard-setting method is nothing but an instrument to elicit responses from subjects from which an estimate of a quantity is inferred" (p. 112). Jaeger pointed out that standard-setting methods "have been ad hoc constructions . . . that are totally devoid of theoretical grounding" (p. 59).

Differences Across Methods

As noted earlier, there was unanimity that the particular method used to set a standard would affect the results. Shepard reported, "the most pervasive finding [of research on setting standards] is that different standard-setting methods produce different results" (p. 156). Mehrens, van der Linden, and Berk, among others, made the same observation. Unfortunately, participants did not agree on what to do about the problems caused by the fact that different standard-setting methods would give different results.

Need for Documentation

Authors believed that all aspects of the standard-setting process should be well documented. Several authors cited relevant entries from the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council in Measurement in Education, 1985) which indicate areas of broad professional agreement. Standard 6.9

specifically requires that when a cut score is used "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" (p. 43). Brennan attempted to encapsulate areas of agreement by suggesting additional standards for standard setting. His first standard also dealt with the need for documentation. "The characteristics of the judges, the items, and the standard-setting process should be described at an appropriate level of detail for the various types of users" (p. 283). Brennan insisted that documentation "should be sufficiently detailed so that an independent investigator could replicate the process" (p. 283). He also insisted that standard setters should document "unanticipated aspects of the standard-setting process, or anomalous outcomes that might affect interpretations" (p. 285). In addition, van der Linden included standards for standard setting in his paper. His first standard, "explicitness," is relevant to the need for documentation as it requires that, "all steps in a standard-setting experiment be based on explicit definitions and procedures" (p. 107).

Possibility of Misunderstanding or Misuse

There was also agreement that, in certain situations, standards could be in error, be misunderstood, be misused, and have potentially harmful consequences. Hambleton and Slater gave examples of difficulties in understanding standards-based score reports. Brennan warned that people who use the results "should be cautioned about any reasonably anticipated misuses of the outcomes of a standard-setting process," and that "the outcomes can be misunderstood" (p. 285). Zieky warned that people who establish a standard should be "ready to modify that standard if experience demonstrates that it is resulting in inappropriate consequences that outweigh the appropriate ones" (p. 32).

Need for Support for Use of Standards

Authors concurred that merely establishing standards of performance will not have the desired effect on educational practices. The development of the essential support structures will be difficult and time-consuming. Linn exemplified the concerns that were expressed:

> For performance standards to have the desired impact on day-to-day classroom activities, they must be internalized by teachers and students. . . . Because the vision of both content standards and performance standards that are being advanced is a radical departure from the vast majority of current daily classroom practices, the transformation cannot be expected to take place overnight. (p. 370)

Aldrich wrote that it would be necessary to change "teacher beliefs and school culture" (p. 358). Smith asserted that it would be necessary to have "a phase-in plan for standards and their assessment. It must address explaining standards-based reform to educators and the public" (p. 404). Smith also indicated that curriculum materials and tools should be available to teachers before standards are assessed.

Judges

Authors agreed that the judges involved in standard setting must represent diverse perspectives, be well trained, and have the required areas of knowledge. Bond expressed as a "cardinal rule" that "those persons who set the standard should be thoroughly knowledgeable of the content domain that is assessed, the population of examinees who take the assessment, and the uses to which the results will

be put" (p. 316). Messick stated that informed judgments "require knowledge of the subject-matter domain as well as of the students' levels of developing expertise" (p. 300). Both Mehrens (p. 247) and Berk (p. 169) used the same words to note that judges must be "qualified and credible." Mehrens stated that the judges "must be thoroughly trained to do the job" (p. 248). The agreement among authors about the need for training was made clear by Berk who used colorful language to make the same point: "Train these judges till it hurts" (p. 170). The authors' level of agreement concerning the need for diversity among judges is clearly documented (see p. ES-8) and need not be repeated here.

DISCUSSIONS

After the presentation of each set of four related papers at the conference, discussion sessions were held with the authors. Popham was given the task of orally summarizing the contents of the 16 discussion sessions at the conclusion of the conference. According to Popham, discussants generally agreed that the process of setting standards should be as open and as pluralistic as possible, that it was important to define the purposes of setting standards and to evaluate the consequences in the light of those purposes, that policymakers rather than methodologists should be involved in resolving differences, that equity concerns required involving all concerned constituencies, that more research on examinee-centered methods of setting standards was needed, and that improvements in communicating results to the general public were required.

CONCLUSION

The conference did not result in professional consensus on how standards ought to be set for large-scale assessments. The conference did, however, bring together many of the people most active in the field of standard setting under close-to-ideal conditions to state their views, air their differences, and seek solutions to common problems. Participants certainly gained an understanding of the multifaceted issues involved in setting standards and an awareness of the varied points of view that are held about many of the issues. We hope that this summary will encourage readers to study the papers presented at the conference and that the papers will, in turn, encourage expanded research efforts on the problems that remain in setting standards for large-scale assessments.

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.

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

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

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 485-514). New York: American Council on Education/Macmillan.

HISTORICAL PERSPECTIVE

A Historical Perspective on Setting Standards

Michael J. Zieky

Executive Director, Educational Testing Service, New Jersey

ABSTRACT

Passing scores were used on tests for centuries with little or no attention paid to how those standards had been set. That long "Age of Innocence" ended with the rapid growth of the criterion-referenced testing movement and the institution of minimum competency/basic skills testing by states and school districts. In response to the needs engendered by those types of tests, there was an outpouring of publications describing systematic methods of setting standards in a brief but intense "Age of Awakening." That age was, however, quickly overlapped by an "Age of Disillusionment" as researchers discovered that the various methods disagreed with one another, and it became abundantly clear that all the methods, no matter how rigorous they might appear, depended at some point on subjective human judgment. The outpouring of methodologies in the Age of Awakening is tapering off, and the shocked outrage of the Age of Disillusionment is gradually cooling as the "Age of Realistic Acceptance" has begun. Even though techniques for setting standards on conventional tests have evolved to the point at which there are several procedures that are in general use, no method of setting standards is universally accepted.

THEORETICAL PERSPECTIVE

Standards for Reporting the Educational Achievement of Groups

Samuel A. Livingston

Senior Measurement Statistician, Educational Testing Service, New Jersey

ABSTRACT

Standard-setting studies translate the judges' conceptual standards into an implied operational standard for making decisions or reporting group percentages. The operational standards implied by item-judgment studies and by borderline-group studies tend to produce a biased estimate of the percentage of students who meet the judges' conceptual standard. The operational standard that minimizes the number of decision errors tends to produce an estimate that is biased in the opposite direction. However, these two standards can be combined to produce an operational standard that, under certain plausible assumptions, produces an unbiased estimate.

On the Cognitive Construction of Standard-Setting Judgments: The Case of Configural Scoring

Richard M. Jaeger

Center for Educational Research and Evaluation
University of North Carolina, Greensboro

ABSTRACT

Setting performance standards can be regarded as an example of a class of problems identified in the psychological literature as judgment or decision-making (JDM) tasks. Such tasks are characterized by uncertainty of information or outcome or by an outcome dependent on personal preferences. Solutions to JDM tasks can be made more accurate by employing a variety of strategies that include increasing the clarity and completeness of problem statements, managed social interaction, and the use of decision aids. The relevant psychological literature is discussed, and an example of the use of such strategies in the context of a complex, multidimensional standard-setting problem is provided.

Some Technical Aspects of Standard Setting

Huynh Huynh[4]

Professor, College of Education, University of South Carolina

ABSTRACT

This paper deals with some technical aspects in standard setting. First, statistical accuracy and minimum sample size are considered for cut scores based on the contrasting-groups procedure. Next, a number of psychometric procedures are presented for mapping test score levels of performance assessment items, along with implications for standard-setting methods based on item judgment. This is followed by a decision-theoretic approach to adjusting cut scores for situations where adverse consequences of incorrect decisions need to be factored. Finally, remarks are made on practical issues such as local item dependence, number of score levels, and stability of standards in yearly equatings.

[4]Many issues discussed in this paper originated from the author's work with the South Carolina Basic Skills Assessment Program and the Maryland School Performance Assessment Program. Gratitude is extended to Vana Meredith-Dabney, Steve Ferrara, Lynn Mappus, Paul Sandifer, and Joe Saunders for their many beneficial discussions over the years.

A Conceptual Analysis of Standard Setting in Large-Scale Assessments

Wim. J. van der Linden

Professor of Educational Measurement and Data Analysis,
Faculty of Educational Science and Technology,
University of Twente, Enschede, The Netherlands

ABSTRACT

This paper consists of three parts. The first part analyzes the use of cut scores in large-scale assessments and discusses three different functions of cut scores: (a) They define the qualifications used in assessments; (b) they simplify the reporting of achievement distributions; and (c) they allow for the setting of targets for such distributions. The second part of the paper gives a decision-theoretic alternative to the use of cut scores and shows how each of the three functions identified in the first part can be approached in a way that may reduce some of the feelings of arbitrariness that often accompany standard-setting processes. The third part of the paper formulates criteria for standard-setting methods that can be used to evaluate their results.

It has often been stated that setting standards in large-scale assessments is a process with arbitrary results. The purpose of this paper is to precisely identify the elements of arbitrariness in the standard-setting process, to present an alternative approach to the use of cut scores that may reduce some of the feelings of arbitrariness involved in standard setting, and to provide criteria to distinguish better standards from worse. The main philosophy in this paper is that standard setting will always involve a subjective choice, but that some choices are consistent with empirical data and meet important criteria of rationality whereas others do not.

DEVELOPING STANDARDS FOR THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

Examinee-Centered vs. Task-Centered Standard Setting

Michael Kane

Professor, Department of Kinesiology, University of Wisconsin at Madison

ABSTRACT

Alternate approaches to standard setting cannot be compared in terms of accuracy because there is no way of unambiguously determining where the standard and the corresponding threshold score should be. This paper examines three general criteria for evaluating examinee-centered and task-centered standard-setting methods in different contexts: consistency with the model of achievement underlying test interpretation, the level of demand placed on judges, and technical adequacy. As one might expect, neither method comes out a clear winner in all cases; each has strengths and weaknesses that make it relatively more promising for some standard-setting tasks and in some contexts.

Implications for Standard Setting of the National Academy of Education Evaluation of the National Assessment of Educational Progress Achievement Levels

Lorrie A. Shepard

Professor of Education, University of Colorado at Boulder

ABSTRACT

A summary is provided for key findings from the National Academy of Education (NAE) Panel's evaluation of the National Assessment of Educational Progress (NAEP) achievement levels. In the 1992 assessments, items used to exemplify performance failed on both substantive and statistical criteria. External validity studies and internal reanalyses demonstrated inadequacies in the item judgment method used to translate judges' intended standards onto the score scale. Therefore, NAEP achievement level results did not accurately report what students can or cannot do. In the future, assessment tasks should be developed directly to represent intended performance levels; and more holistic standard-setting approaches should be used to allow judges to focus on substantive expectations and avoid false assumptions about item intercorrelations and item difficulties.

METHODOLOGICAL PERSPECTIVE

Standard Setting--The Next Generation

Ronald A. Berk[5]

Professor, Johns Hopkins University School of Nursing

ABSTRACT

Specific directions for future standard-setting practices are proffered by answering three key questions: (1) Where have we been? (2) What have we learned? and (3) So what's a practitioner to do? The answers are presented in the form of three "Top 10 Lists." Technical issues related to reliability and validity evidence for the standard-setting process and a standard-setting awards finale complete the presentation.

[5]The author gratefully acknowledges the assistance of Ellen Spies in the preparation of this paper and of all the materials for the presentation upon which the paper was based.

An Interpretive Approach to Setting and Applying Standards

Pamela A. Moss[6]

Assistant Professor, University of Michigan, School of Education
Educational Studies Program

ABSTRACT

Interpretive research traditions suggest alternatives to assessment practices developed within the psychometric tradition. They also provide perspectives from which mainstream assessment practices can be critiqued. In this paper, I compare psychometric and interpretive approaches to setting and applying standards in the context of high-stakes individual assessment. Perhaps the most striking contrast between the two approaches involves the ways in which standards are typically applied to individual cases. Within the psychometric approach, single performances are evaluated independently, the scores algorithmically combined, and the composite score compared to a pre-established cut score. Within the interpretive approach, readers consider all the available evidence on an individual, evaluating each part in light of the whole, and engage in a consensus-seeking discussion to reach a decision about whether standards have been met. In comparing these approaches, I consider the quality of information provided, the potential consequences to various stakeholders, and the implications for developing and maintaining an assessment system under each approach.

The Consensus Process in Standards Development

Belinda L. Collins

Director, Office of Standards Services,
National Institute of Standards and Technology

ABSTRACT

The consensus process in the development of standards is considered to include balance, openness, agreement, and due process for all affected parties. The American National Standards Institute (ANSI), all major standards developers in the United States, the National Institute of Standards and Technology (NIST), and the International Organization for Standardization (ISO) have each defined procedures for developing consensus. While they differ slightly from one another, these procedures typically concentrate on elements such as committee membership--including balance among interest groups without domination by any single entity--announcement of and openness during meetings, voting procedures, review processes, communications, and appeals. Although these procedures are generally applied to the development of product and process standards to ensure quality, safety, health, and/or environmental integrity, they may provide a useful model for evaluating the effectiveness of procedures

[6]I am grateful to Gail Baxter for helpful comments on an earlier draft of this paper. The paper was written while the author was working under a fellowship from the National Academy of Education/Spencer Postdoctoral Fellowship Program.

and standards developed in the field of education. The general consensus process is described here, with particular focus on relevance for developing standards for education.

Methodological Issues in Standard Setting for Educational Exams[7]

William A. Mehrens

Professor of Education, Michigan State University

ABSTRACT

This presentation focuses on some fairly broad but interrelated concerns to keep in mind while holding discussions regarding standard-setting methodologies. A brief review of the literature, including some recent approaches to standard setting, is presented. Points of agreement and disagreement among experts are noted. The presentation concludes with some issues that must be considered that are separable from, but related to, choosing a method and some thoughts on preparing policymakers.

APPLICATIONS PERSPECTIVE

Standard Setting From the Perspective of Generalizability Theory[8]

Robert L. Brennan

Professor of Educational Measurement, Director of Iowa Testing Programs

ABSTRACT

In education currently, the term "standards" has numerous meanings. In this paper, the word standards is used to refer to standards for standard-setting procedures. Otherwise, "standards" refers either to standard-setting processes or to outcomes of such processes. A substantial part of the audience for this paper includes persons with considerable experience in setting or evaluating National Assessment of Educational Progress (NAEP) achievement levels, which are certainly standards. While it is clearly intended that this paper address issues of generalizability in establishing achievement levels, the focus on standards considered here is broader than just achievement levels.

This paper is organized into four principal parts. The first part provides a conceptual framework for considering standard setting, broadly conceived, in the context of generalizability theory. The second

[7]A minor portion of this manuscript was adapted from *Standard Setting for Medical Board Exams* prepared by the author for the American Board of Emergency Physicians.

[8]The author gratefully acknowledges many helpful discussions with Michael T. Kane that influenced several parts of this paper.

part considers the role of various facets (including judges, items, occasions, and methods) in characterizing the dependability of the outcomes of a standard-setting process. The third part treats some collateral issues, including different perspectives on dimensionality and the role of score scales in standard setting. Finally, the fourth part suggests some particular standards for standard setting. This paper deals largely with reliability-related issues, but it also considers aspects of validity.

Standards-Based Score Interpretation: Establishing Valid Grounds for Valid Inferences

Samuel Messick[9]

Vice President for Research, Educational Testing Service, New Jersey

ABSTRACT

The construct validity of content standards is addressed in terms of their representative coverage of a construct domain and their alignment with the students' cognitive level of developing expertise in the subject matter. The construct validity of performance standards is addressed in terms of the extent to which they reflect increasing levels of construct complexity as opposed to construct-irrelevant difficulty. Also critical is the extent to which performance standards characterize the knowledge and skills operative at each level both to accredit specific accomplishment and to serve as goals for further learning. All of this depends on construct-valid assessment attuned to the content standards and the development of dependable scoring rubrics and measurement scales for representing the performance standards.

Ensuring Fairness in the Setting of Performance Standards

Lloyd Bond

Professor, School of Education, University of North Carolina at Greensboro

ABSTRACT

Standards of performance on educational measures should distinguish the same levels of knowledge, proficiency, and ability for all test takers. That is, a performance standard that purports to identify "advanced" proficiency on some measure should result in classifications that are equally accurate across all groups of test takers. In like manner, standards that purport to identify those in need of remediation should not result in misclassification rates that vary as a function of ethnicity, gender, or other characteristics that are irrelevant to the construct being assessed.

[9]I gratefully acknowledge helpful comments on the manuscript provided by Ann Jungeblut, Robert Mislevy, and Michael Zieky.

This presentation discusses (a) methods for investigating the extent to which performance standards may result in differential rates of misclassification of subgroup members in the population of examinees, and (b) procedures for setting initial performance standards that minimize the likelihood of such misclassifications.

Using Performance Standards to Report National and State Assessment Data: Are the Reports Understandable and How Can They Be Improved?

Ronald K. Hambleton, Professor of Education, and Sharon Slater, Graduate Student
University of Massachusetts at Amherst

ABSTRACT

Considerable evidence suggests that policymakers, educators, the media, and the public do not understand national and state test results. The problems appear to be twofold: the scales on which scores are reported are confusing, and the report forms themselves are often too complex for the intended audiences.

This paper addresses two topics. The first is test score reporting scales and how to make them more meaningful for policymakers, educators, and the media. Of special importance is the use of performance standards in score reporting. The second topic is the actual report forms that communicate results to policymakers, educators, and the public. Using some results from a 1994 interview study with 60 participants using the Executive Summary from the 1992 National Assessment of Educational Progress mathematics assessment, the paper highlights problems in score reporting and suggests guidelines for improved score reporting.

POLICY PERSPECTIVE

The Impact of Standards on How Teachers Behave in the Classroom: Promises and Perils

"For every complex issue there is a simple answer, and it is wrong." (McMillan, 1994, p. 466)

Phyllis W. Aldrich

Coordinator of the Gifted and Talented Program, Director of the National Javits Language Arts Research Project, WSWHE Board of Cooperative Educational Services

ABSTRACT

The delineation of standards at a national level could have a powerful and positive effect on what actually happens in the classroom, if teachers are directly involved at their schools in discussing interpretation and potential applications to their own specific students and curriculum. Expectations for student achievement and standards are always present but rarely articulated and not readily

comparable. Careful description of student performance levels, clearly described with many examples, could provide valuable "calibration" or "anchors." This would inform individual teachers on what might be possible for students to achieve at a specific grade level in a specific area of study. The linkage between nationally derived student performance levels and adaptation to local practices could be a useful tool in the campaign to improve student learning.

The Likely Impact of Performance Standards as a Function of Uses: From Rhetoric to Sanctions

Robert L. Linn

Professor of Education and Codirector of the National Center for Research on Evaluation, Standards, and Student Testing, University of Colorado at Boulder

ABSTRACT

Performance standards have many potential uses, ranging from relatively benign reporting mechanisms to the determination of rewards and sanctions. The impact of performance standards will depend upon the specific uses to which they are put. A range of potential uses of performance standards will be identified. Plausible unintended effects, as well as the intended impact of each use, will be discussed using examples from existing assessment programs where possible.

Legal Defensibility of Standards: Issues & Policy Perspectives

S. E. Phillips

Professor, Michigan State University

ABSTRACT

Legal challenges to standards may be expected when standards create adverse impact on historically disadvantaged groups; use processes perceived to be unfair, arbitrary, or capricious; imply that specific attitudes or values will be assessed; fail to consider accommodations for the disabled; or appear to assess knowledge/skills that examinees have not had the opportunity to learn. This paper will discuss the legal framework surrounding these issues, provide guidance on developing policies and procedures that are legally defensible, and consider the types of empirical data that may be relevant in the event of litigation.

Issues in Standards Setting: Making a Decentralized System Responsive to Centralized Policy Making

H. Wesley Smith

Superintendent, Newberg Public Schools, Oregon

ABSTRACT

Standards setting is an effort to make a decentralized public school system responsive to national and state policy decisions. This presentation will explore the hopes, fears, and issues emerging at the local level in the context of state and national standards. In most states, public schools have developed under local control. The result is a tradition-bound, provincial system that is, ostensibly, beyond the influence of central authority. Consequently, as new cultural, social, and economic dynamics demand greater school flexibility and responsiveness to the larger world, the public school system has proven remarkably resistant to change. Policymakers are concerned about how they will find the leadership and resources to overhaul the present system to respond to achievement standards. The professional development initiative required to reach the goal is monumental.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☐ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☒ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").