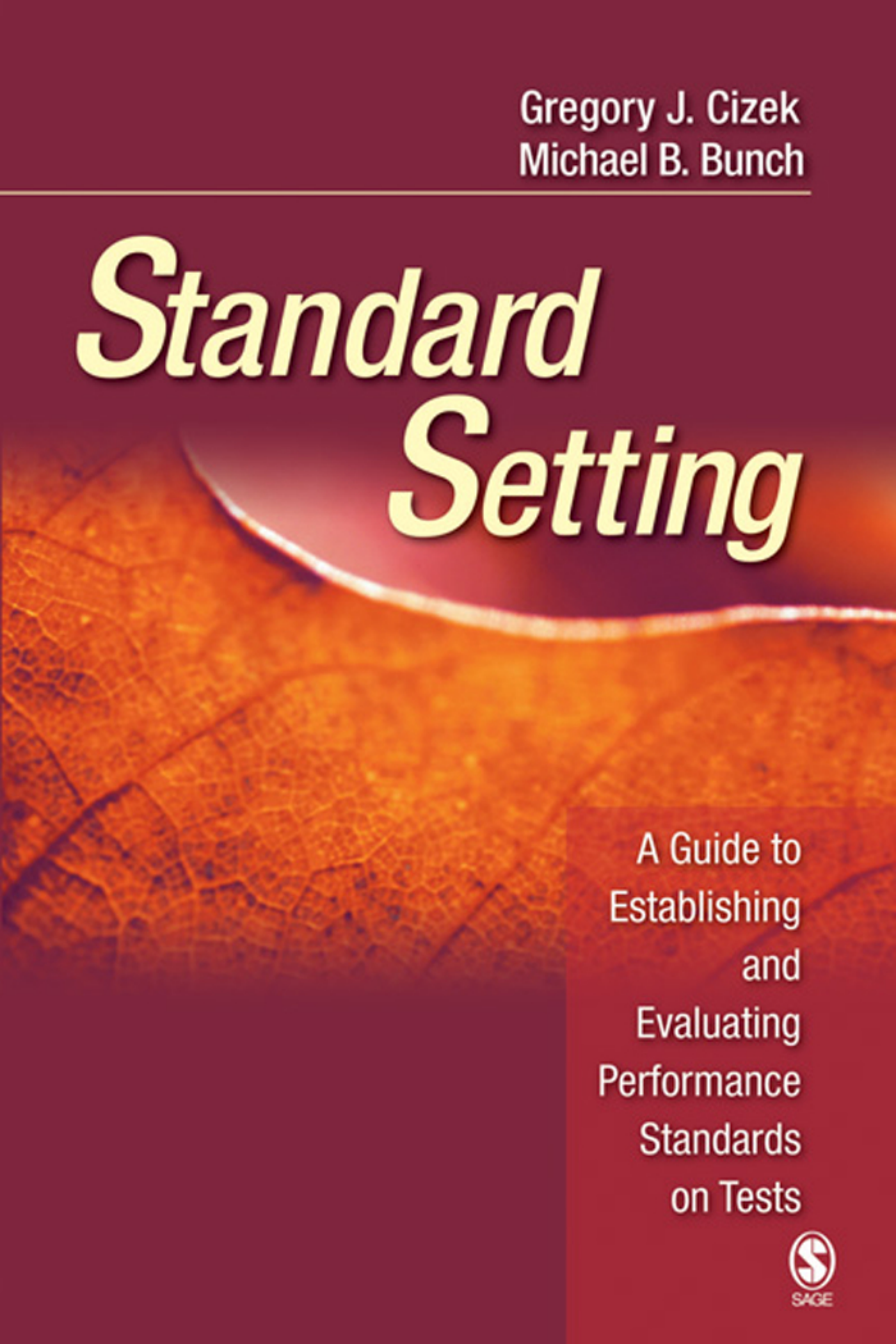
Gregory J. Cizek
Michael B. Bunch

Standard Setting

A Guide to Establishing and Evaluating Performance Standards on Tests



Standard Setting

Standard Setting

A Guide to Establishing and Evaluating Performance Standards on Tests

Gregory J. Cizek
University of North Carolina at Chapel Hill

Michael B. Bunch
Measurement Incorporated

SAGE Publications
Thousand Oaks ▪ London ▪ New Delhi

For information:

Sage Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

Sage Publications Ltd.
1 Oliver's Yard
55 City Road
London EC1Y 1SP
United Kingdom

Sage Publications India Pvt. Ltd.
B-42, Panchsheel Enclave
Post Box 4109
New Delhi 110 017 India

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Cizek, Gregory J.
Standard setting: A guide to establishing and evaluating
performance standards for tests / Gregory J. Cizek, Michael B. Bunch.
p. cm.
Includes bibliographical references and index.
ISBN 1-4129-1682-8 or 9-781-4129-1682-0 (cloth)
ISBN 1-4129-1683-6 or 9-781-4129-1683-7 (pbk.)
1. Achievement tests—United States—Evaluation. 2. Examinations—Scoring—Statistics. 3. Educational tests and measurements—United States.
I. Bunch, Michael B. II. Title.
LB3060.3.C58 2007
371.26′2—dc22

2006017002

This book is printed on acid-free paper.

06 07 08 09 10 11 10 9 8 7 6 5 4 3 2 1

Acquisitions Editor:	Lisa Cuevas Shaw
Editorial Assistant:	Karen Greene
Production Editor:	Denise Santoyo
Copy Editor:	Mary Tederstrom
Typesetter:	C&M Digitals (P) Ltd.
Cover Designer:	Candice Harman

Contents

Preface

In what now seems like a distant and much simpler time, the extremely popular 71-page booklet on setting cut scores was written by Livingston and Zieky (1982). That little volume, called *Passing Scores,* summarized the most popular methods of that time for setting performance standards.

As readers of this preface almost surely know, much has changed since that time. **Cut scores**—or what are now more commonly called **performance standards**—have become more necessary and more consequential. The needs range from a single cut score separating passing from failing performance on a licensure or certification examination, to three or more cut scores creating ordered categories of performance on a single high school graduation test, to a system of cut scores spanning several grades and subjects across the elementary and secondary grades. Numerous choices exist for how to derive those performance standards.

This book is written for those who must make those choices and for those who conduct, oversee, or are responsible for standard-setting procedures. Thus this book—like that written many years ago now by Livingston and Zieky—focuses more on the practicalities and "how-to" of standard setting than on the theory or academic concerns that unarguably underlie the practice. Compared to the psychometric state of affairs when *Passing Scores* was written, however, the number and complexity of standard-setting options available today has expanded considerably, as has the range of issues that must be considered.

Our goals in this book are to provide the reader with a brief but solid foundation of the basic theory and logic underlying standard setting, and then to dive directly into a nuts-and-bolts survey of the current inventory of standard-setting methods. On the one hand, we hope to provide enough specific, practical information based on our own knowledge and experiences in standard setting so that a reader of this book could conduct the procedure of his or her choice. On the other hand, we recognize that the specific steps we outline for any of the methods surveyed here must often

be changed, adapted, deleted, or supplemented in any specific standard-setting application. Indeed, because in nearly all instances readers who conduct a standard setting will need to modify the basic step-by-step procedures of a given method, we have included at the end of each chapter a section titled "Alternative Procedures and Limitations" in which we provide some suggestions—though certainly not exhaustive—for ways in which a basic procedure has been or could be successfully adapted to the demands of differing contexts.

This volume is organized into three sections. In the first section, we address what can be thought of as the fundamentals of standard setting. Chapter 1 situates the practice of standard setting as a necessary, important activity, with benefits for individuals, organizations, and society. The many contexts in which standard-setting methods are utilized are described, along with various political realities that have influenced the reach and consequences of the activity. Chapter 2 presents a definition of **standard setting**, differentiates between **content standards** and **performance standards**, addresses the relationship between **item scoring criteria** and total-test performance standards, and introduces some of the key policy and practical decisions that must be considered before beginning a standard-setting procedure. Chapter 3 surveys the common elements—such as **performance level labels** (PLLs), **performance level descriptions** (PLDs), and so on, that must be addressed in any standard-setting procedure, regardless of the specific method selected. This chapter also provides some important considerations that guide the choice of a specific standard-setting method—it provides information on selecting and training standard-setting participants, presents a framework for evaluating the implementation of a standard-setting procedure, and provides relevant guidelines and professional standards that help shape sound and defensible standard-setting practice.

The second section of the book provides step-by-step activities for actually conducting a standard-setting procedure. The collection of chapters in this section covers each of 13 methods that have at least some degree of current use. Some of the methods described may be unfamiliar to some readers, in part perhaps because certain standard-setting methods are more widely used in some areas (e.g., licensure and certification contexts) than in others (e.g., elementary and secondary school achievement testing). Readers may also notice that the chapters in Section II differ rather markedly in their lengths. For the most part, this is attributable to the relative simplicity of some methods (e.g., Nedelsky) compared to the relative complexity of those introduced more recently (e.g., Bookmark).

The final section of the book describes challenges and likely future directions in standard-setting practice. Chapter 13 provides a framework

to aid those responsible for conducting standard-setting procedures in accomplishing one of the most practical but difficult aspects of standard setting: scheduling and integrating all of the requisite activities. Chapter 14 provides an introduction to an emerging technology necessitated by requirements in elementary and secondary education that a system of tests span several grades and subject areas. That technology, sometimes referred to as **vertically-moderated standard setting**, helps promote a coherent system of performance standards across diverse levels and content standards. Chapter 15 provides practical information on another problem unique to K–12 education contexts: the need to set performance standards on tests that are specially developed for students with severe learning disabilities, including requirements for testing format, content, language **accommodations**, or other special needs. Such tests, sometimes referred to as **alternate assessments**, present special standard-setting problems. The final chapter, Chapter 16, provides a brief look at some special topics in standard setting and discusses some of the likely standard-setting challenges on the horizon.

Also included in the book are some special features that we hope readers will find helpful. For example, as readers may have already noticed, some terms are set in **bold** type. These terms are included and defined in a glossary at the end of the book. At various junctures in the book, we have also included links to data sets, Excel programs, and other electronic resources that readers can download and use to practice the methods or use as models for developing their own customized resources and tools.

We must note that our work on this book has been aided significantly by many people. We are both grateful to the National Council on Measurement in Education (NCME) for permission to adapt some material we have published previously as installments in the Instructional Topics in Educational Measurement (ITEMs) series of the NCME journal, *Educational Measurement: Issues and Practice* (e.g., Cizek, 1996a, b). We note too that the proportion our contribution to the standard-setting literature represents is small compared to the abundance of excellent work contributed by others; thus we also humbly acknowledge that our summary of the advances in standard setting represented in this volume would not have been possible without the diligent work of many dedicated scholars and practitioners in the field of applied psychometrics. We are also indebted to Sage Publications, which has a long and successful history of publishing important works in the social sciences. We are particularly grateful to Lisa Cuevas Shaw, Rachel Livsey, and Phyllis Cappello of Sage for their advice and continuing enthusiasm for this project, and to Mary Tederstrom for her helpful editorial support.

Individually, the authors would like to acknowledge others who have shown great support for our work. First, although it is not customary to thank one's co-author, I (MBB) would like to do just that. I was honored when Greg asked me to collaborate on this book. I had conducted standard setting for many years but had never written anything on the subject except reports for clients. My work brings me into contact with Greg and other academicians serving on **technical advisory committees** (**TACs**). I believe it was a standard-setting plan I had submitted to one of these TACs that may have induced Greg to approach me about more broadly disseminating some of the practical work we had done on standard setting. Along those lines, I am also grateful to other TAC members for asking questions that forced me to think more precisely with regard to methodology, and to the clients for providing numerous opportunities to conduct standard setting in a variety of settings with a vast array of methods. I particularly wish to thank staffs of the Arkansas, Georgia, New Jersey, Ohio, and Virginia state departments of education and of the Council of Chief State School Officers (CCSSO) for opportunities to conduct standard setting for them and for their cooperation during the planning and implementation of those standard-setting activities.

In addition, I would also like to express appreciation to my boss, Hank Scherich, for the opportunity to do this work, and to my colleagues, past and present, at Measurement Incorporated who have helped me plan and carry out standard setting over the last 20-plus years. Many of the ideas presented in this book came directly from them or benefited from their critiques. Elliot Inman deserves special credit in this regard. Colleagues in the field also deserve special recognition, particularly Robert Lissitz at the University of Maryland (UMD) for inviting me to teach a module on standard setting at UMD, forcing me to begin to work out many ideas that eventually found their way into this book, and Thomas Hirsch of Assessment and Evaluation Services for asking just the right questions to get me to devise clear explanations of crucial aspects of complex procedures.

Finally, my wife, Kathryn, has been especially supportive through the writing and rewriting of my portions of this book. Contractors have strange schedules, and we're not home very much. Adding a book to our schedules is like adding another contract, except that the nights away are actually nights at home (or during vacations or holidays) but out of sight for hours on end. I appreciate Kathryn's listening to bits and pieces of my prose and her comments on the parts that didn't make sense the first time around. But mostly I appreciate her certainty that I could actually do this.

I (GJC) have been fortunate to collaborate with Mike Bunch not only on this project but also on many other measurement matters. Located as I am

in an academic setting, I value every opportunity to collaborate with those who must apply standard-setting methods in diverse applied contexts. This chance to work with Mike has been of tremendous professional benefit to me; he is clearly one of the most thoughtful, analytical, and rigorous practitioners one could encounter. More important, I have appreciated Mike's insights, diligence, integrity, and friendship.

In addition to Mike's substantial contributions, I acknowledge the many informal conversations about standard setting with outstanding scholars in the field of psychometrics; those conversations have helped clarify my own thinking and influenced the content of this book in ways that surely strengthened it. I happily acknowledge the insights of Professor Ronald Hambleton of the University of Massachusetts, Professor Gregory Camilli of Rutgers University, Professor William Mehrens (emeritus) of Michigan State University, Professor Barbara Plake (emeritus) of the University of Nebraska, Dr. Jon Twing of Pearson Educational Measurement, Dr. Steve Ferrara of the American Institutes of Research, Dr. E. Roger Trent of the Ohio Department of Education (emeritus), Dr. Matthew Schulz of ACT, Inc., Dr. Elliot Inman of SAS, Inc., and Heather Koons, doctoral candidate at the University of North Carolina at Chapel Hill. In addition, our original plans for the manuscript benefited from the input of Professor Steve Sireci of the University of Massachusetts. All of the sage advice of these scholars notwithstanding, any remaining errors in conceptualization or fact should be attributed solely to the authors.

I also appreciate the support for this work provided by the School of Education at the University of North Carolina at Chapel Hill (UNC) and the encouragement of Dean Thomas James to complete this work. I am particularly grateful for the support provided by the Kenan research leave program at UNC.

Finally, I am so grateful for the continuing support of my wife, Rita and our children, A. J., Caroline, David, and Stephen, who I join in thanking God for showering his abundance on the American educational system and in pleading his continuing favor.

MBB/GJC
Durham and Chapel Hill, North Carolina

SECTION I

Fundamentals of Standard Setting

In general, this is a book about standard setting for people who want to understand the theory and practice of setting performance standards; it is for those who will conduct standard setting; it is for those who must receive, review, and critique reports about standard setting; it is for those who must oversee or make decisions about standard setting. In short, this book is about the practical aspects of standard setting.

In this introductory grouping of chapters, we begin with an orientation to the fundamental questions about standard setting—what it is, what it is not, why it is necessary, what it entails, and what it accomplishes. Of course, not everyone with expertise in setting performance standards agrees on the answers to questions such as these, and we think that it is important for readers to understand the diversity of perspectives that exists.

This section reflects the varied perspectives of the two authors: one of us engages in research, teaches graduate students, and advises agencies about the conduct of standard setting; one of us contracts with those agencies to plan and actually carry out standard setting. We have attempted to balance the theoretical underpinnings of standard setting and the enduring quest for new knowledge in the area with the practical need for a set of cut scores and documentation that will allow an agency (i.e., a licensing or certification board, a state education agency, or the federal government) to make

well-informed, defensible decisions about human beings that will serve all stakeholders—including examinees themselves—well.

We have attempted to address the needs of standard setters in as many contexts as possible—licensure, certification, and educational assessment in elementary and secondary schooling contexts. We recognize a need for well-conceived and executed standard setting in all of these situations, regardless of the number of individuals involved or the size of the budget for such activities.

The chapters in this first section of the book provide the foundational understandings and practical information necessary to begin the course of setting performance standards. Chapter 1 addresses the enduring need for standard setting. Testing has been—and remains—an important source of information for decision making in nearly limitless aspects of modern life. At some point in all our lives, someone looks at a test score and makes a decision about us. In some instances—as we will illustrate in Chapter 1—the stakes are literally life or death. Even in less dire circumstances, the need to justify the propriety of the decisions made about people should be clear. Far from being simply arbitrary, decisions made on the basis of well-researched and effectively implemented standard-setting procedures can have great benefit.

Chapter 2 answers the question "What is standard setting?" In this chapter we focus on what standard setting is—and what it is not—in an effort to disentangle standard setting from a host of related concepts and activities that are easily and often confused. We address some of the policy issues directly related to standard setting as well as a variety of very practical matters that must be addressed in every situation in which standard-setting methods are used to establish cut scores for a test to determine who passes and who fails, who is certified and who is not, and so on.

Although there exists a wide variety of methods for setting performance standards, each of the methods has something in common with the others. Chapter 3 surveys the elements common to all standard-setting activities. Regardless of the procedure chosen, the standard setter will always need to involve people and judgments. How those people render their judgments and how those judgments are processed vary with the procedure chosen, but the involvement of people who must exercise judgment is a constant. They have to be trained, in either the content, or the process, or both. They typically participate in one or more rounds of making judgments, with feedback of some sort between rounds. One common element that may not have even been mentioned a few years ago is the performance level description or definition (PLD). Recalling Angoff's (1971) admonition to keep in mind a "minimally acceptable person," as a prelude to standard setting, we

note that Angoff's admonition is now replaced with a fairly detailed description of such a person and perhaps other classes of persons as well. In years to come, there may be other common elements that permeate all standard-setting procedures, which are only now emerging as elements of specialized procedures. The field has evolved rapidly in the last 50 years, and it will be exciting to see what is yet to come. For now, we turn to a close examination of why the science and practice of standard setting is important—and why it is crucial that is it done right.

1

Contemporary Standard Setting

An Enduring Need

In the next chapter, we set forth a more extensive definition of standard setting. However, it seems appropriate to introduce an abridged definition of this a concept so fundamental to this book from the outset. In brief, *standard setting* refers to the process of establishing one or more cut scores on examinations. The cut scores divide the distribution of examinees' test performances into two or more categories.

For example, in the context of licensure and certification testing programs, it is often the case that only a single cut score is required, the application of which results in the creation of two performance categories, such as *Pass/Fail, Award/Deny* a license, or *Grant/Withhold* a credential. In other contexts, multiple cut scores may be required, the application of which results in the creation of more than two performance categories. For example, in elementary and secondary education, the now familiar *No Child Left Behind Act* (NCLB, 2001) and the *Individuals with Disabilities Education Act* (IDEA, 1997) have mandated that achievement for all students in specified school grade levels and subjects be reported using the performance categories *Basic, Proficient,* and *Advanced.* To classify test performances into these three categories, two cut scores are needed—one to define the border between *Basic* and *Proficient,* and another to define the border between *Proficient* and *Advanced.*

Clearly, the activity of creating the boundaries of the performance categories is one of the most important tasks in the test development, administration, and reporting process. That importance is due, in large measure, to

the consequences that can be associated with the resulting classifications. The important consequences are not limited to tests used for credentialing, for example, nurses, financial planners, chiropractors, attorneys, otolaryngologists, real estate brokers, or cosmetologists, or for gauging the achievement of U.S. students or their schools in areas such as reading and mathematics. As one researcher has pointed out, establishing cut scores can literally be a matter of life and death:

> The choice of cut score was important in the deliberations of the U.S. Supreme Court in the case of *Atkins v. Virginia* (2002). The petitioner, Daryl Renard Atkins, had been sentenced to death for a murder he committed. This sentence was overturned on June 20, 2002, by the Supreme Court because Atkins was deemed to be retarded and the execution of mentally retarded individuals was "cruel and unusual" and hence prohibited by the Eighth Amendment to the U.S. Constitution. Atkins had a measured IQ of 59 on the Wechsler Adult Intelligence Scale (WAIS-III). In a 1986 case in Georgia, Jerome Bowden was convicted of murder and sentenced for execution. The same defense was invoked, and his execution was stayed while he was tested. He was found to have an IQ of 65 and was executed the following day. If the court based its decision solely on the IQ test results, we would be forced to conclude that the cut score that determined the decision of life or death must lie between 59 and 65. (Wainer, 2006, p. 63)

Regardless of whether the specific consequences associated with a cut score are great or small in a given context, in our opinion standard setting is often (mistakenly) considered later than it should be in the test development and administration cycle. Standard setting is best considered early enough to align with the identified purpose of the test; to align with the selected test item or task formats; when there is ample opportunity to identify relevant sources of evidence bearing on the validity of the categorical assignments; when that evidence can be systematically gathered and analyzed; and when the standards can meaningfully influence instruction, examinee preparation, and broad understanding of the criteria or levels of performance they represent. Indeed, the initial planning for any test that will be used to sort or classify individuals into performance categories should include a detailed discussion of the nature of those categories. A clear understanding of the performance categories will then influence all phases of test development, reporting, score interpretation, and validation efforts.

The Need to Make Decisions

As the reader probably knows, there are many arenas in which such classifications must be made. Perhaps the classifications are required by law, as in state testing for the granting of a driver's license. Or the desire to impose

classifications may arise from within a profession, where credentialing examinations and the resulting classifications can serve as a means of assisting the public in choosing from among more and less qualified practitioners. In American education, postsecondary institutions often establish their own cutoffs for a variety of tests such as the Test of English as a Foreign Language (TOEFL), Graduate Record Examinations (GRE), the SAT, and the ACT. These cutoffs and the resulting decisions regarding admissions and scholarships help students, their families, and institutions make the best use of limited resources and align with the American tradition of rewarding merit.

As has been argued elsewhere (see Mehrens & Cizek, 2001), there is simply no way to escape making such classifications. If, for example, some students are graduated from high school and others are not, categorical decisions have been made, regardless of whether or not a formal high school graduation test is part of the decision-making process. Coaches of major league baseball teams make decisions about who will be the starting first baseman, though the "test" that is used is considerably less formal and more performance based.

Classifications like these are unavoidable in many situations. For example, high school music teachers make decisions about who will be first chair for the flutes, colleges make decisions about admission to their undergraduate and graduate programs, university faculties make decisions to tenure their colleagues (or not), state governments are charged with regulating who will be permitted to practice medicine, and many professions recognize advanced levels of practice via credentialing examinations and the accordant cut scores that categorize test takers into those who will receive such recognition and those who will not.

At the present time, it is not conceivable that every baseball player can be named a starter; that every flautist can be first chair; that colleges can admit all students who apply; that universities will establish policies to retain all instructors regardless of quality; or that consumers of health care will reject the protection afforded by state action restricting the practice of medicine, nursing, pharmacy, and other areas to those who are able to demonstrate some acceptable level of knowledge and skill. Thus, for the foreseeable future, each of the preceding examples of classifications is unavoidable. In their textbook on measurement and evaluation, Mehrens and Lehmann (1991) summarized the need to make decisions and the relevance of measurement to that task:

> Decision making is a daily task. Many people make hundreds of decisions daily; and to make wise decisions, one needs information. The role of measurement is to provide decision makers with accurate and relevant information. . . . The most basic principle of this text is that *measurement and evaluation are essential to sound educational decision making*." (p. 3, emphasis in original)

The Benefits of Standard Setting

The psychometric aspects of standard setting help ensure that any decisions or classifications are based on high-quality data and that the data are combined in a systematic, reproducible, objective, and defensible manner. From a broader perspective, it can be asserted that if categorical decisions must be made, they will be fairer, wiser, more open, more valid, more efficient, and more defensible when they utilize established, systematic processes that result in cut scores that are based on nonarbitrary, explicit criteria.

Besides being fairer, valid, and so on from a psychometric perspective, standard setting yields other benefits. For one thing, it is a well-accepted learning principle that the presence of explicit performance criteria increases attention to successful performance on the criteria. When the specific standards to be met are made explicit, more time, money, and effort are likely to be spent by those who seek a license, certification, or classification (e.g., students or other test takers) and by those whose role it is to assist them in doing so (e.g., teachers, residency program directors, test-prep course instructors). It seems almost certain that increased attention to mastery of the explicit criteria would, in most cases, result in increased competence for those affected, and increased competence seems like a good thing.

In addition, the explicit standards are likely to result in increased understanding by and trust on the part of the public. As has been argued elsewhere (Cizek, 2001a; Mehrens & Cizek, 2001), the impetus for tests of student competence (as well as tests for teacher licensure and likely many other tests) is due in large part to the lack of public trust in the soundness of criteria in place prior to such tests. In the 1970s, complaints of some business and industry leaders that "we are getting high school graduates who have a diploma, but can't read or write!" provided the impulse for student competency testing. In 1978, Popham wrote that "minimum competence testing programs . . . have been installed in so many states as a way of halting what is perceived as a continuing devaluation of the high school diploma" (p. 297). As Burton observed, "The criterion-referenced testing movement [in education] can be seen as an attempt to transfer responsibility for some important educational decisions from individual teachers to a more uniform, more scientific, technology" (1978, p. 263).

Although they may not be as readily recognized as the concrete benefit of keeping unsafe motorists from behind the wheel of an automobile, the intangible benefits of standard setting such as the bolstering of public trust and confidence in educational, medical, professional, and vocational areas should not be underestimated. For example, Lerner (1979) described the important role that standard setting can play in her portrayal of shared

societal concepts about standards as "mortar, holding the multicolored mosaics of civilizations together" (p. 15).

Nor should the difficult task of standard setting itself be underestimated. Lerner also recognized the inherent difficulty in deriving the cutoffs that define the categories, observing that "the cutoff score problem is very similar to one that judges and lawyers deal with all the time; the question of where and how to draw the line" (1979, p. 28).

General Approaches to Standard Setting

In subsequent chapters of this book, we describe many specific methods for setting standards. Measurement specialists have attempted to devise classification schemes for organizing the variety of methods into logical groupings. By and large, we think the groupings suggested to date are either inaccurate or not helpful. For example, a well-known grouping scheme was suggested by Jaeger (1989) in his chapter in *Educational Measurement, Third Edition.* According to Jaeger, standard-setting methods could be classified into two categories—those that are *examinee-centered* and those that are *test-centered.* Jaeger classified a method as test-centered if standard-setting participants primarily ground their cut score recommendations in careful scrutiny about test content or test items; he classified a method as examinee-centered if participants' judgments are primarily about the test takers themselves.

Kane (1994a) has suggested an alternative way of classifying standard-setting methods. He suggests that methods can be classified as either *holistic models* "which assume that achievement or skill is highly integrated" or *analytic models* "which assume that achievement can be assessed using relatively small parts or samples of performance (pp. 4–5).

A third way of classifying methods might lump the various approaches into those that are more *norm-referenced* and those that are more *criterion-referenced.* **Norm-referenced** methods would include those in which standards are established with respect to relative standing or performance of examinees. For example, a consistent standard might be established that passed (or credentialed, or licensed, etc.), say, 75% of test takers who were administered a test given on an annual basis. Such an approach might be defensible if relevant characteristics of the test taker population were stable from year to year, if test forms were constructed to consistent specifications, and if there was evidence that the standard adequately served the purpose of the testing program. Norm-referenced standard setting was very common until roughly the late 1950s and continues to be used and justifiable in some situations today. In fact, as late as 1976, Andrew and Hecht

reported that "at present, the most widely used procedures for selecting . . . pass-fail levels involves norm-referenced considerations in which the examination standard is set as a function of the performance of examinees in relation to one another" (p. 45).

In contrast to norm-referenced standard setting (and more commonly used currently) is what can be called **criterion-referenced** standard setting. Criterion-referenced methods are also sometimes referred to as "absolute" methods in contrast to the "relative" nature of norm-referenced methods. An early work by Nedelsky (1954) pioneered the concept of absolute standards, as opposed to the then-prevalent use of relative standards. Nedelsky classified a standard as absolute "if it can be stated in terms of the *knowledge and skills* a student must possess in order to pass the course" (p. 3, emphasis in original). Explaining the distinction between criterion-referenced and norm-referenced approaches, Nedelsky argued that a passing score should be

> based on the instructor's judgment of what constitutes an adequate achievement on the part of a student and not on the performance by the student relative to his class or to any other particular group of students. In that sense the standard to be used for determining the passing score is absolute. (p. 3)

In K–12 educational contexts, a variation of criterion-referenced standard setting—called **standards-referenced**—has gained popularity. Whether labeled criterion-referenced or standards-referenced, the methods have in common that they seek to express a standard not in terms of what an examinee's relative performance should be with respect to a relevant comparison group, but in terms of the specific knowledge, skills, objectives, content, or proportion of some domain to be mastered.

While two-dimensional categorization schemes such as examinee-centered versus test-centered, holistic versus analytical, or norm-referenced versus criterion-referenced have some surface appeal, the demands and nature of standard setting in practice compel us to conclude that no simple distinctions between methods can be made and that well-conceived and implemented standard setting must recognize that any procedure requires participants to rely on both dimensions to effectively carry out their task.

For example, as we consider the test-centered/examinee-centered distinction, we think it is obvious that any standard-setting procedure necessarily requires participants to bring to bear information about both test content and test takers. It would not be possible for a standard-setting participant to make a judgment about the difficulty of an item or task without relying on his or her knowledge or expectations of the abilities of examinees in the

target population. Conversely, it would not be possible for a participant to express judgments about examinees without explicit consideration of the items or tasks presented to the examinees. Along the same lines, it does not seem possible that a standard-setting participant could express a judgment about an "absolute" level of performance without incorporating his or her knowledge or opinion about the general levels of skill or ability in the examinee group; that is, criterion-referenced judgments cannot be made without at least implicit consideration of normative information.

In conclusion, while we retain and use some of the terminology (such as "test-centered" or "standards-referenced") that has come to be associated with various ways of grouping standard-setting methods, we believe we must also alert practitioners to the limitations of those labels and to the reality that setting cut scores requires a more sophisticated understanding of the nature of the task than more simplistic classifications can promote.

Current Contexts for Standard Setting

The variety and demands of current contexts for standard setting further support our contention that the activity must be viewed as sophisticated and challenging. In the licensure arena, those responsible for conducting standard-setting activities and who have the authority to establish cut scores must take into account competing, and sometimes conflicting, interests. The setting of cut scores in these instances must delicately balance the commitment and ambitions of examinees in pursuit of a vocation of their choosing with the duty of the licensing authority to protect the public from harm attributable to less-than-adequately prepared practitioners. Credentialing boards must balance the potentially competing aims of promoting increased professional knowledge and skill, providing recognition for advanced levels of practice, and enhancing the status of member professionals and the value of the credential.

The context of educational **achievement testing** is equally challenging—or even more so. Previously mentioned legislation such as NCLB and IDEA has mandated assessment of all students. In the past, standards could be (and were) established with a "typical" student population in mind; standard-setting activities could take advantage of relatively easy-to-implement methods based largely on testing formats and conditions that were homogeneous and traditional. Presently, however, assessments must be developed and standards established for contexts, content, student populations, formats, and conditions that are strikingly diverse and for which existing standard-setting techniques must be dramatically reconfigured or new techniques developed altogether.

For example, NCLB and IDEA mandates that all students be assessed have been interpreted to mean, well, that *all* students be assessed. Tests—and, importantly, comparable standards of performance—are required for students who are best measured in languages other than English. Tests and comparable standards are required for students who are severely cognitively, emotionally, or physically impaired. Content specifications, tests, and performance standards are required at grade levels and in content areas where, in many states, no such content specifications, tests, or standards had existed previously. And because some knowledge and skills are best measured using performance tasks or **constructed-response** formats, standard-setting methods appropriate to those contexts are necessary. Adding to the complexity is the fact that, because the mandated assessments and standards must apply to a broad range of grade levels, some mechanism for promoting a logical coherence of performance standards across that range is required. The latest standard-setting technology (sometimes called **vertically-moderated standard setting**; see Chapter 14) has been recently and hurriedly introduced to meet this demand.

Finally, regardless of the arena—licensure, certification, education, or elsewhere—all standard setting is unavoidably situated in a complex constellation of political, social, economic, and historical contexts. Those who set standards must be cognizant of these contexts that may play as great a role (or greater) in determining the outcome of a given standard-setting procedure than the particular participants involved or the specific method utilized. As we hope we have convincingly argued, however, the complexity of contexts notwithstanding, standard setting is a necessary and beneficial activity that must be conducted with attention to the multitude of technical and nontechnical factors that present themselves and that ultimately bear on the validity and usefulness of the results.

2

What Is Standard Setting?

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

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

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

Kinds of Standards

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

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

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

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

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

Definitions of Standard Setting

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

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

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

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

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

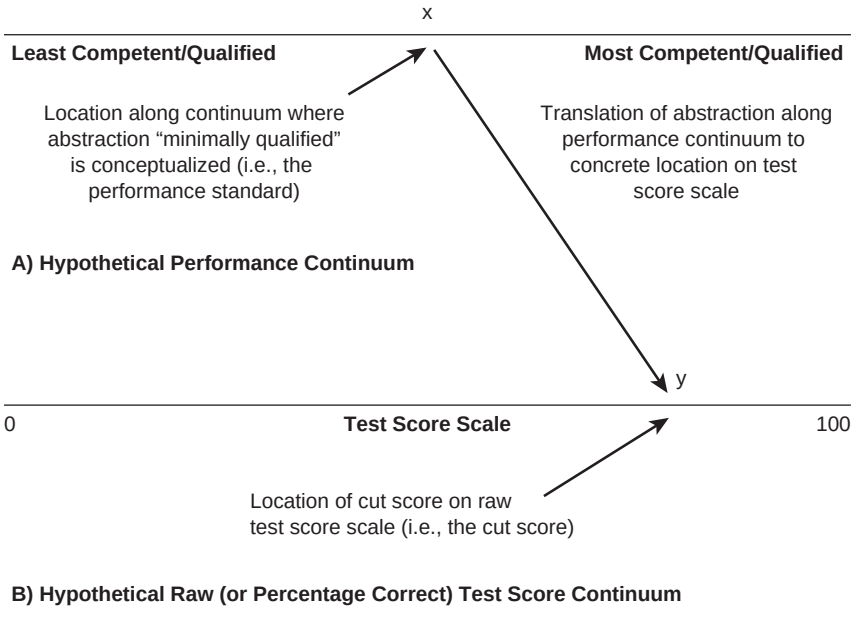


Figure 2-1 Relationship Between Performance Standard and Cut Score

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Policy Issues and Standard Setting

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

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

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

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

Scoring Models

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

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

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

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

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

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

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

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

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

Research on Standard Setting

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

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

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

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

Rounding

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

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

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

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

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

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

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

Classification Errors

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

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

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

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

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

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

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

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

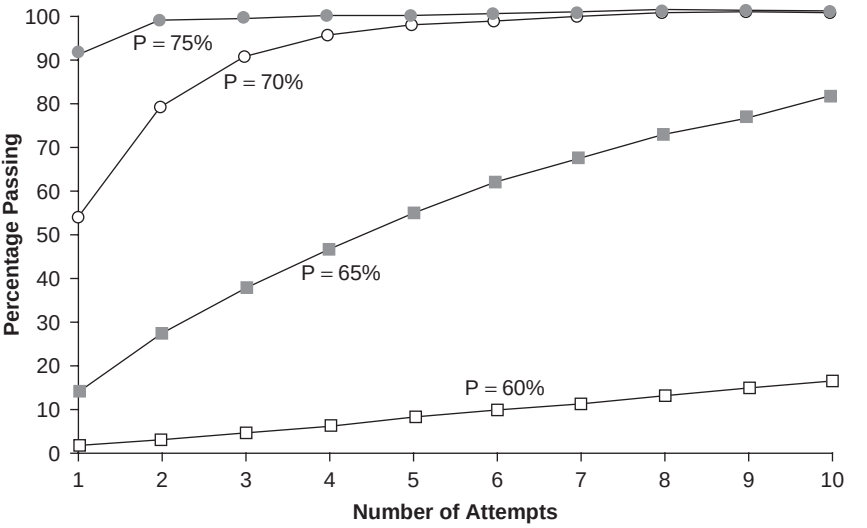


Figure 2-2 Percentage of Examinees at Four Levels of Competence Expected to Reach a 70% Passing Standard as a Function of the Number of Attempts Permitted

SOURCE: Millman (1989).

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

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

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

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

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

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

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

Item Scoring Criteria and Total-Test Performance Standards

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

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

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

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

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

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

Table 2-2 Performance Level Descriptions for a Reading Test

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

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

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

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

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

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

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

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

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

Conclusions

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

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

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

3

Common Elements in Setting Performance Standards

Regardless of the standard-setting method selected for use in a specific context, there is a good chance that the chosen method will have many things in common with any of the alternatives that were considered. Across specific methods, the general series of steps to be followed in a standard-setting procedures is often quite similar. Table 3-1, adapted from Hambleton (1998), shows a generic listing of common steps required for standard setting.

Of course, much more is required of a defensible standard-setting process than choosing and implementing a specific method, and any general listing of steps masks the reality that the key to successful standard setting lies in attention to decisions about many consequential details. For example, those responsible for standard setting must attend to identification and training of appropriately qualified participants, effective orientation and facilitation of the standard-setting meeting, monitoring and providing feedback to participants, and well-conceived data collection to support whatever validity claims are made.

In this chapter, we identify several commonalities that cut across standard-setting methods. For example, we consider performance level labels, performance level descriptions, selection and training of standard-setting participants, and the kinds of information about their judgments provided to participants during the standard-setting process. We also present information on criteria to be addressed when choosing a standard-setting method in the

Table 3-1 Generic Steps in Setting Performance Standards

Step	*Description*
1	Select a large and representative panel.
2	Choose a standard-setting method; prepare training materials and standard-setting meeting agenda.
3	Prepare descriptions of the performance categories (i.e., PLDs).
4	Train participants to use the standard-setting method.
5	Compile item ratings or other judgments from participants and produce descriptive/summary information or other feedback for participants.
6	Facilitate discussion among participants of initial descriptive/summary information.
7	Provide an opportunity for participants to generate another round of ratings; compile information and facilitate discussion as in Steps 5 and 6.
8	Provide for a final opportunity for participants to review information; arrive at final recommended performance standards.
9	Conduct an evaluation of the standard-setting process, including gathering participants' confidence in the process and resulting performance standard(s).
10	Assemble documentation of the standard-setting process and other evidence, as appropriate, bearing on the validity of resulting performance standards.

SOURCE: Adapted from Hambleton (1998).

first place and criteria that can be used to evaluate the processes and outcomes of a standard-setting activity. We begin this chapter, however, with what is likely the most important common consideration: purpose.

Purpose

According to the *Standards for Educational and Psychological Testing,* "The proper use of tests can result in wiser decisions about individuals and programs than would be the case without their use and also can provide a route

to broader and more equitable access to education and employment" (AERA/APA/NCME, 1999, p. 1). It follows that clarity regarding the purpose of a test is the most important aspect of any test development, administration, and reporting system. The purpose of a test drives what the test will look like, the conditions of administration, the kinds of information reported and to which audiences, what sources of validity evidence must be gathered, and so on. Thus a critical touchstone necessarily preceding any standard-setting activity is the development and articulation of a specific testing purpose. Standard setting must align with that stated purpose.

For example, if the purpose of a medical licensure test were to protect the public from harm that might arise from practitioners who were unable to demonstrate possession of knowledge and skills deemed essential for safe practice, then a standard-setting procedure on, say, an anatomy test would be necessary to facilitate identification of those examinees who possessed at least that presumably high level of knowledge or skill. Conversely, a somewhat lower standard might be established on the same anatomy test if the test were to be given to beginning anatomy students and used as part of a process of identifying which students will be permitted to continue in a medical training program. In the end, cut scores are the mechanism that results in category formation on tests, and the purpose for making those categorical decisions must be clearly stated.

Because of this, we contend that the first question that must be asked before setting performance standards is a fundamental one: "What is the purpose of setting these standards?" The issue of purpose underlies all standard setting; indeed, it is the foundational issue that provides the starting point for validation of any test score inference. Kane (1994b) has summarized the centrality of purpose in this way:

> 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)

We concur. And we suspect that explicit attention documenting the purpose of setting standards may be overlooked in many situations, and even attention to the task of standard setting may often (mistakenly, we believe) be incorporated very late in the test development and administration process. Table 3-2 provides a general overview of the key steps or activities that comprise a typical test development cycle. Standard setting is listed as Step 12 in the table because that is the point at which a standard-setting procedure

Table 3-2 Typical Steps in Test Development and Reporting Process

Step Number	*Description*
1	Identify/clarify purpose of test.
2	Implement process to delineate domain to be tested (e.g., curriculum review, job analysis).
3	Develop specific objectives, content standards.
4	Decide on item and test specifications, formats, length, costs.
5	Develop items, tasks, scoring guides.
6	Review items/tasks (editorial, appropriateness, alignment with content, sensitivity).
7	Field test items (i.e., item tryout).
8	Review field-test item/task performance.
9	Create item bank/pool.
10	Assemble test form(s) according to specifications.
11	Administer operational test forms (perhaps with additional pretest items).
12	Establish performance standards, cut score(s).
13	Evaluate preliminary item/task and examinee performance; verify scoring keys.
14	Score test; apply standards.
15	Report scores to various audiences.
16	Evaluate test; document test cycle.
17	Update item pool, revise development procedures, etc.; repeat Steps 5–10, 12–16.

SOURCE: Adapted from Cizek (2006).

would typically be conducted (although standards must sometimes be established without examinee data from an operational test administration; such a situation would place standard setting as Step 11). However, we also note that consideration of the need for performance standards and of the specific method to be used should occur much earlier than Step 12, arguably as early as Step 4 when deciding upon the format of test items and tasks, or even as early as Step 1 when the purpose of the test is first articulated.

Because contexts vary, the ideal time for standard setting to occur is not a simple question to address, and there are advantages to locating it earlier or later in the test development process. On the one hand, the placement of standard-setting activities early in the test development process and before the test has been administered reflects the due process concern for fundamental fairness. Namely, it seems somewhat inappropriate to require examinees to submit to an examination requirement without being able to communicate to them in advance about the level of performance that will be required in order to pass or achieve some performance level. On the other hand, establishing cut scores after an operational test has been administered is likely to yield more dependable results. Because many standard-setting procedures involve the provision of examinee performance data to participants during the course of various rounds of judgments, actual performance data from an operational test administration is necessary. Examinee data based on results from field testing or item tryouts are notorious for their instability. An unknown degree of bias in the information is introduced because such data are often collected during a "no-stakes" test administration context or other situation in which motivation levels of examinees are not likely equivalent to those of examinees who will take the test under standard, operational conditions.

In summary, we return to the fundamental concern about purpose. Primacy of purpose requires that the role, timing, and method of standard setting are best considered early enough to align with the identified purpose of the test, the selected test item or task formats, and when there is ample opportunity to identify relevant sources of evidence bearing on the validity of the categorical assignments and to gather and analyze that evidence.

This first common element to be considered when setting standards—purpose—may not be particularly difficult to address, although it is often overlooked or given insufficient attention. Those responsible for setting performance standards can often refer to legislative mandates, the public interest, selection constraints, and so on to provide a sound rationale for using a test to aid in decision making in the first place. For example, a publication of the American Board of Medical Specialties describes an intent to "provide assurance to the public that the diplomate has successfully completed . . . an evaluation to assess knowledge, experience, and skills requisite to the provision of high quality medical care" (Langsley, 1987, p. 11). Although descriptions like the preceding are important, the rationale could be expanded to include, for example, descriptions of what the public is being protected from; what knowledge, skills, and experience are assessed; what constitutes high-quality medical care; and so on.

Addressing the purpose of a test as it relates to the establishment and interpretation of performance standards most often focuses on description

of a **construct**, as opposed to a more directly observable characteristic. In one measurement textbook, the authors describe constructs as hypothetical concepts that are "products of the informed scientific imagination of social scientists" (Crocker & Algina, 1986, p. 4). In credentialing contexts, constructs might include such characteristics as *preparation for independent practice, professional knowledge,* or *minimal competence.* In education, constructs would include aspects of human behavior such as *mathematics computation skill, reading comprehension, creativity, life skills,* and so on. In practice, specific constructs are often further delimited and contextualized (e.g., minimum competence for ninth graders in writing three-page narrative essays in response to a prompt). Our experience tells us that constructs that play a key role in test development and standard setting are also often left disconcertingly vague or ill-defined at the outset of the testing and standard-setting processes.

Perhaps this tendency toward ambiguity or failure to make explicit a focused purpose is related to the fact that in only rare cases are the demands placed on test results limited and focused themselves. For example, consider the context of licensure and certification testing. A primary purpose often stated by credentialing entities is to award credentials based on demonstrated acquisition of knowledge or skill. However, the licensing or certification entity may also have as a purpose the maintenance or enhancement of the status of the profession and the members of an association. On the one hand, then, standard setting might serve the purpose of heightening the value of the credential by, among other things, limiting its acquisition (although consideration of supply and demand factors when setting cut scores on licensure tests is rejected by many testing specialists; see, e.g., Shimberg, Esser, & Kruger, 1973). On the other hand, professional associations are not immune from market forces themselves, and survival of an organization may depend on increasing membership in the organization, providing an incentive to make the credential more widely available.

The same tensions can be seen in other settings where performance standards are established. For example, in educational achievement testing, it is common to encounter rhetoric regarding the need for high standards on high school graduation tests that maintain "the value of the diploma." Competing against that purpose may be pressures flowing from federal mandates (e.g., the adequate yearly progress requirements of the *No Child Left Behind Act* [NCLB, 2001]) that K–12 educational systems demonstrate continuous improvement in graduation rates, achievement gaps, and getting all students to a *Proficient* level of performance within a fixed time period. The state-level, standards-referenced, every-pupil testing mandated by NCLB must be configured in such a way as to report at least three categories

of performance, which requires at least two cut scores to distinguish between those categories. However, the resulting performance standards are almost never exclusively used for NCLB purposes, but expand to include demands for diagnostic information, state accountability or "report card" uses, instructional utility for teachers, local system information for planning and curriculum adjustment, and others.

In conclusion, the first commonality cutting across virtually all standard-setting methods is the need for clarity of purpose. Prior to beginning any standard-setting procedure, it is important for those responsible for the testing program to clearly and fully describe the construct(s) or characteristic(s) assessed by the test and to articulate the relationship of the proposed cut score(s) to the construct(s) of interest and the purpose that the test and performance standards are expected to serve. We are not naive to the fact that multiple, sometimes conflicting, purposes for testing programs and performance standards can and do exist and that diversity of purpose can work against the focus and level of clarity desirable for standard setting. Nonetheless, we also believe that it is only if the purpose of testing and standard setting is clearly stated that the selection of a specific standard-setting method can be justified and the necessary validity evidence supporting the application of cut scores and the interpretation of performance classifications be gathered.

Choosing a Standard-Setting Method

According to the *Standards for Educational and Psychological Testing,* "There can be no single method for determining cut scores for all tests or for all purposes, nor can there be any single set of procedures for establishing their defensibility" (AERA/APA/NCME, 1999, p. 53). Although development in the field of standard setting has generated new methodological alternatives that can be applied to increasingly diverse standard-setting context, and research in the field has helped refine existing methods so that results are more stable, the choice of which method to use in any specific context is not always clear.

It is a well-established research finding that the choice of a standard-setting method influences the resulting cut scores. William Angoff, the namesake of one of the most commonly used standard-setting methods, commented on the variability in results attributable to both within- and between-method variation:

> [Regarding] the problem of setting cut-scores, we have observed that the several judgmental methods of setting cut-scores not only fail to yield results that agree with one another, they even fail to yield the same results on repeated application. (1988, p. 219)

Some research has identified sources of consistent variation. For example, use of the Nedelsky method (1954; see Chapter 4, this volume) has been shown to result in somewhat lower cut scores and, correspondingly, higher pass rates because of an inherent bias in the way item ratings must be expressed using this method. However, no well-established hierarchy exists that ranks methods in terms of the overall stringency or leniency of results produced. In our experience, several other variables (e.g., clarity of purpose, backgrounds and qualifications of participants, adequacy of training, timing and quality of feedback to participants, etc.) have substantially greater influence on the eventual performance standards recommended by a panel than the particular standard-setting method used.

When choosing a standard-setting method, a number of considerations must be taken into account. In the following paragraphs, we identify and provide rationales for including what we believe to be the most important of the considerations.

First, as we have already stated, the methodological choice for establishing performance standards should be related to the purpose of the test. To use a simple example, if the purpose for setting a cut score were to consistently identify a certain number or percentage of test takers for some award, placement, or position, a simple norm-referenced method would be well matched to that purpose. For another example, it is possible that standard setting might serve dual purposes of measuring student achievement in, say, writing, and of teacher professional development. In such a case, the standard-setting method of choice could be narrowed to one that involved classroom teachers who, as part of orientation to the test, received training in the use of a rubric used to score student writing samples and, as part of the standard-setting activity itself, reviewed and judged actual samples of student writing.

Second, when selecting a standard-setting method, the choice of method should be related to the level of complexity of the knowledge, skills, and abilities assessed by the test. For example, a procedure such as Nedelsky's (1954; see Chapter 4, this volume) is appropriate for dichotomously scored, multiple-choice items, where the characteristic assessed is basic knowledge; a procedure such as the Body of Work method (Kingston, Kahl, Sweeney, & Bay, 2001; see Chapter 9, this volume) would be more appropriate where the material on which judgments are to be made consisted of more complex multidimensional work samples. This consideration is reminiscent of Kane's (1994a) classification of standard-setting models as either analytic or holistic.

A third consideration, and one related to the complexity of the constructs and products to be judged, is that of test format. Some methods can

be applied exclusively to a particular item format. Some methods can be adapted for use with different formats. Some methods can be used with tests comprising items and tasks of varying formats but with some constraints. Some methods are indifferent as regards format. Examples of these situations would include, respectively: the Nedelsky (1954) procedure, which can be used only with multiple-choice format items; the Angoff and extended Angoff methods (1971; see Chapter 6, this volume), which can be adapted for use with selected- or constructed-response formats; item-mapping procedures such as the Bookmark method (Mitzel, Lewis, Patz, & Green, 2001; see Chapter 10, this volume), which can be used with tests comprising a mixture of selected- and constructed-response formats, provided that the ratio of selected- to constructed-response items is relatively large (i.e., there are proportionally fewer constructed-response items and score points); and procedures such as the Contrasting Groups (Livingston & Zieky, 1982) and Hofstee (1983) methods for which item or task format is irrelevant as regards the judgmental task.

A fourth consideration is the number of performance categories (i.e., cut scores) that are required. For some tests, only a single cut score is required for the creation of two categories (e.g., Pass/Fail). Any of the methods described subsequently in this volume could be used in that context.

However, other testing programs require that multiple cut scores be derived to create three or more categories. The number of categories that are required can limit the options for standard-setting methods and, more importantly, can often place severe demands on the ability of the test to support the number of distinctions demanded and the ability of any standard-setting technique to derive reliable and valid results. For example, in the educational achievement testing mandated by the NCLB (2001) legislation, the three performance categories of *Basic, Proficient*, and *Advanced* must be created for all covered tests. For tests at upper grade levels (e.g., high school) and in certain subject areas where test takers are relatively developmentally advanced and test lengths can be relatively longer, setting two cut points on, say, a 50-item mathematics tests may not pose a substantial challenge. However, for the necessarily shorter tests in the primary grades (e.g., third grade) or for other subject areas, setting two cut points on, say, a 25-item mathematics test or a 1-item writing test may be daunting. Some individual states have also gone beyond the number of categories required by NCLB and created standard-setting situations that clearly tax the abilities of both participants and methods to accomplish the task. As one example of this, legislators in Ohio have required that all state-level student **achievement tests** support five levels of performance (*Below Basic, Basic, Proficient, Accelerated*, and *Advanced*) regardless of age, grade level, or subject area.

A fifth consideration—and, admittedly, an extremely practical one—involves the extent of resources available for standard setting. In our experience, entities responsible for testing programs have multiple demands on the organization and its members in terms of money, time commitments, scheduling, and so on. A very intensive standard-setting procedure, conducted over several days and involving multiple rounds of judgments, may be highly desirable in some contexts, but not in others, such as when the more intensive procedure diverts resources from other testing activities such as bias and sensitivity reviews, item development, field testing, and so on, when the most qualified participants cannot be recruited to participate in the more intensive procedure or when the more intensive procedure would strain the alacrity of participants to the point of introducing construct-irrelevant aspects into the judgmental process.

Finally, it is our experience that a question often arises as to the desirability of using multiple methods of setting standards and somehow integrating the results. On balance, we believe that the use and analysis of multiple methods is a policy decision, largely related to the position an entity has related to research on standard setting as an aim in itself (see Chapter 2). Because we know of no research base that would guide those who implement multiple methods in synthesizing the results or in choosing one result over another, we strongly caution against such a course. And, from the practical perspective just presented, we assert that an entity's limited resources are in nearly all cases better utilized implementing a single procedure very well than in implementing two (or more) methods less well.

Performance Level Labels

A third crosscutting aspect of most standard-setting endeavors is the creation and use of **performance level labels** (PLLs). PLLs refer to the (usually) single-word terms used to label performance categories. For much standard setting in the recent past, the concept of PLLs was not as salient as it is in contemporary contexts because, historically, usually only two performance levels were required; the labels Pass and Fail were adequately descriptive of the performance categories to which examinees were assigned. Although nothing more elaborate than Pass and Fail may be required for some testing contexts (e.g., licensure and certification), increasingly, because multiple cut scores are needed to create more than two categories, multiple PLLs are also required. The now-familiar PLLs *Basic, Proficient*, and *Advanced* associated with the National Assessment of Educational Progress (NAEP) and the NCLB (2001) Act are examples of such labels. A wide variety of categorical labeling systems exists,

however. A few examples of PLLs taken from large-scale student achievement testing programs used across Grades K–12 are shown in Table 3-3.

It may be obvious that PLLs have little technical underpinning, although there are measurement concerns associated with the choice of a set of labels. From a semantic perspective, PLLs such as *Basic, Proficient*, and *Advanced* could be exchanged for other creative, one-word labeling schemes such as Deficient, Magnificent, and Omniscient depending on the extent to which PLLs are a primary score-reporting mechanism, on the informational needs of the consumers of test information (e.g., examinees, the public), on the aims of those responsible for communicating about test performance, and on the risks associated with misinterpretations based on the labels.

Their lack of technical underpinning notwithstanding, it is clear that PLLs carry rhetorical value as related to the purpose of the standard setting and, much as the label affixed to the test itself, must be selected with care. As Cronbach (1971) has indicated in reference to the choice of a title for a test, the matter is "of great importance. . . . If this phrase is a poor summary of the sophisticated professional interpretation, it promotes possibly serious misconceptions" (pp. 460–461). In a similar manner, care

Table 3-3 Sample Performance Level Labels (PLLs) From K–12 Achievement Testing Programs

Labels	*Source*
Basic, Proficient, Advanced	National Assessment of Educational Progress (NAEP)
Starting Out, Progressing, Nearing Proficiency, Proficient, Advanced	TerraNova, Second Edition (CTB/McGraw-Hill)
Limited, Basic, Proficient, Accelerated, Advanced	State of Ohio Achievement Tests
Far Below Basic, Below Basic, Basic, Proficient, Advanced	State of California, California Standards Tests
Did Not Meet Standard, Met Standard, Commended Performance	State of Texas, Texas Assessment of Knowledge and Skills
Prefunctional, Beginner, Immediate, Advanced, Fully English Proficient	English Language Development Assessment (Council of Chief State School Officers)
Partially Proficient, Proficient, Advanced Proficient	State of New Jersey, High School Proficiency Assessment

(and restraint) should be exercised in the choice of PLLs, as they have the potential to convey a great deal in a succinct manner vis-à-vis the meaning of classifications that result from the application of cut scores. From a psychometric perspective, PLLs should be thoughtfully chosen to relate to the purpose of the assessment, to the construct assessed, and to the intended, supportable inferences arising from the classifications.

Performance Level Descriptions

A fourth commonality found in many standard-setting approaches is actually an extension of the concept of PLLs. **Performance level descriptions** (PLDs) are (usually) several sentences or paragraphs that provide fuller, more precise explication of what the one-word PLLs attempt to convey and to more completely describe what performance within a particular category connotes about a test taker so classified. PLDs vary in their level of specificity, but they have in common the characteristic of being verbal elaborations of the knowledge, skills, or attributes of test takers within a performance level.

According to Lewis and Green, "all commonly used standard setting methods utilize PLDs to some degree" (1997, p. 1). Other researchers have asserted that PLDs are essential to the validity and defensibility of the standard-setting process and the validity of the resulting performance standards (see, e.g., Hambleton, 2001). However, little research has been done to provide guidance on how to develop or evaluate PLDs. Lewis and Green have described the development of PLDs for Angoff and Bookmark procedures, and they offer suggestions for further developmental efforts. Mills and Jaeger (1988) have also described a procedure for developing PLDs; they conducted a study comparing more general and more specific content standards–based PLDs and reported that participants found greater utility in the more specific versions.

From a procedural perspective, it is highly desirable for PLDs to be developed in advance of standard setting by a separate committee for approval by the appropriate policy-making body. In some cases, elaborations of the PLDs may be developed by participants in a standard-setting procedure as a first step (prior to making any item or task judgments) toward operationalizing and internalizing the levels intended by the policy body. Sample PLDs, in this case those used for the NAEP Grade 4 reading assessment, are shown in Table 3-4 (see also Table 2-2 for another example).

Once developed by the appropriate policy body and (possibly) elaborated prior to participants engaging in a standard-setting procedure, PLDs can then be used by participants during standard setting as a critical referent for their judgments. The same PLDs used in standard setting can also be used after a test is given, when scores and performance classifications are reported to aid in interpretation of what the classifications mean.

Table 3-4 NAEP Performance Level Descriptions for Grade 4 Reading Tests

Performance Level Label	*Performance Level Description*
Advanced	Fourth-grade students performing at the *Advanced* level should be able to generalize about topics in the reading selection and demonstrate an awareness of how authors compose and use literary devices. When reading text appropriate to fourth grade, they should be able to judge texts critically and, in general, give thorough answers that indicate careful thought. For example, when reading literary text, *Advanced*-level students should be able to make generalizations about the point of the story and extend its meaning by integrating personal experiences and other readings with ideas suggested by the text. They should be able to identify literary devices such as figurative language. When reading informational text, *Advanced*-level fourth graders should be able to explain the author's intent by using supporting material from the text. They should be able to make critical judgments of the form and content of the text and explain their judgments clearly.
Proficient	Fourth-grade students performing at the *Proficient* level should be able to demonstrate an overall understanding of the text, providing inferential as well as literal information. When reading text appropriate to fourth grade, they should be able to extend the ideas in the text by making inferences, drawing conclusions, and making connections to their own experiences. The connections between the text and what the student infers should be clear. For example, when reading literary text, *Proficient*-level fourth graders should be able to summarize the story, draw conclusions about the characters or plot, and recognize relationships such as cause and effect. When reading informational text, *Proficient*-level students should be able to summarize the information and identify the author's intent or purpose. They should be able to draw reasonable conclusions from the text, recognize relationships such as cause and effect or similarities and differences, and identify the meaning of the selection's key concepts.
Basic	Fourth-grade students performing at the *Basic* level should demonstrate an understanding of the overall meaning of what they read. When reading text appropriate for fourth graders, they should be able to make relatively obvious connections between the text and their own experiences and extend the ideas in the text by making simple inferences. For example, when reading literary text, they should be able to tell what the story is generally about—providing details to support their understanding—and be able to connect aspects of the stories to their own experiences. When reading informational text, *Basic*-level fourth graders should be able to tell what the selection is generally about or identify the purpose for reading it, provide details to support their understanding, and connect ideas from the text to their background knowledge and experiences.

Key Conceptualizations

A fifth commonality cutting across all standard-setting methods is the necessity for participants in a procedure to form conceptualizations that will guide their judgments. These conceptualizations become key referents that participants revisit frequently during the standard-setting process and that are also helpful in terms of interpreting the meaning of resulting cut scores.

Examples of these key conceptualizations are many and varied. The attributes of hypothetical examinees in or between the PLLs of Basic, Proficient, and Advanced and the schemas participants bring to bear when applying the chosen **response probability (RP) criterion** when an item-mapping approach is used are examples of key conceptualizations that are foundational to standard setting. When the Angoff method (see Chapter 6, this volume) is used, participants are directed to review each item in a test and to estimate "the probability that the 'minimally acceptable person' would answer each item correctly" (Angoff, 1971, p. 515). The ability of participants to conceptualize the hypothetical person—a construct—is critical to the success of that method. Likewise, in the Borderline Group method (Livingston & Zieky, 1982; see Chapter 8, this volume), participants must be able to retain a focus on whatever description has been developed of a "borderline" examinee, and in the Contrasting Groups method (Livingston & Zieky, 1982; see Chapter 8, this volume) participants must frame and apply the conceptualizations of "master" and "non-master" or similar abstraction. An umbrella term that subsumes a number of more specific ones (such as "minimally competent examinee," "master," "borderline," etc.) used in standard setting is *target examinee*.

These key conceptualizations are, at minimum, difficult to articulate and to apply consistently. For example, research on implementation of the Angoff method for the NAEP revealed that the necessary conceptualizations can be difficult for standard-setting participants to acquire and to maintain once acquired (see Shepard, Glaser, Linn, & Bohrnstedt, 1993). However, more recent research has demonstrated that when participants are provided with specific descriptions or are given adequate time to form clear understandings of the target examinee, these key conceptualizations can be applied consistently. Research by Giraud, Impara, and Plake (2005) found that

> the definition and panelists' discussion of the characteristics of the target examinee influences judges' perceptions. . . . This finding suggests that *a priori* definitions of performance that describe the target examinee in certain ways and more or less exactly can substantially influence judges' operational notion of target competence. (p. 230)

In summary, because of the difficulty and importance of the task, whatever key conceptualizations are necessary warrant specific and extended attention in the planning of a standard-setting process and in the training and evaluation of the process.

Selecting and Training Standard-Setting Participants

As we have mentioned previously, the particular group of persons selected to participate in a standard-setting procedure, and the effectiveness of the training the participants receive, are two common aspects of all standard setting that can affect the eventual standards that are recommended as much as, or more than, the specific standard-setting method used. Participants in the standard-setting process are critical to the success of the endeavor and are a source of variability of standard-setting results.

Recruitment and selection of participants is the first step. In the most general sense, the most appropriate group of participants to empanel would be a sufficiently large and representative sample of the population of possible participants that meet whatever criteria the responsible entity has established as defining "qualified" for the standard-setting task. Of course, there are no correct operationalizations of "qualified," and reasonable attempts to define that construct can be highly variable. One commonality is that any operationalization should be related to the purpose of setting standards in the first place. For example, on the one hand, one medical specialty board establishing a single cut score for a recertification test may wish members of a standard-setting panel to be the most experienced physicians in the field and those with long-standing memberships in a professional association. On the other hand, the same board establishing a cut score for entry into the specialty may wish to empanel newly credentialed practitioners in the specialty who have completed their professional training within the last five years.

Had the choices in the preceding example been reversed, very different standards may have resulted. It is not uncommon, for example, for standard-setting panels comprising the most experienced and sophisticated practitioners to set very high and demanding performance levels, and for less experienced panels to recommend standards more in line with their more limited exposure to the breadth and depth of what is encountered over years of practice in a field. Depending on the field, this situation may be reversed as well. In our experience, sometimes very experienced practitioners—who are also temporally well beyond their own initial credential—may set comparatively lower standards on an entry-level examination covering academic training than would be set by those who have just completed their academic preparation.

While it is often recommended that participants have special expertise in the area for which standards will be set, in practice this can mean that standard-setting panels consist of participants whose perspectives are not representative of all practitioners in a field, all teachers at a grade level, and so on. Such a bias might be desirable if the purpose of standard setting is to exhort those covered by an examination program to high levels of achievement, though less so if the purpose of standard setting is to certify competence of students for awarding a high school diploma.

As these examples illustrate, the specification of the pool of potential participants in a standard-setting procedure is first a policy matter. As we have recommended vis-à-vis other policy matters in standard setting, we believe that the entity responsible for setting standards should explicitly discuss and adopt a position on the issue in advance of standard setting. And whatever position is articulated should strongly relate to the purpose of the examination and the classifications that are to be made based on the cut score(s).

The *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999) provides guidance on representation, selection, and training of participants (called "judges" in the *Standards*). Regarding the number and characteristics of participants, the *Standards* indicates that "a sufficiently large and representative group of judges should be involved to provide reasonable assurance that results would not vary greatly if the process were repeated" (p. 54). Further, the *Standards* recommends that "the qualifications of any judges involved in standard setting and the process by which they are selected" (p. 54) should be fully described and included as part of the documentation of the standard-setting process.

The *Standards* also addresses the training of participants:

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

In our experience—and as reflected by the comparatively thinner research base on training standard-setting participants—the topic of training is one of the least well developed. Some excellent, chapter-length treatments of the topic have recently been published (see Raymond & Reid, 2001), but much work remains to be done in the area. In the following paragraphs, we outline some of the key components of participant training that, whether supported by available research or not, seem prevalent and currently accepted professionally as best practice.

After participants are selected, perhaps the first exposure they will have to the standard-setting process will be informational materials prepared by

the entity responsible for the testing program. These materials are often delivered to participants in advance of the actual standard setting, allowing sufficient time for participants to read the materials, familiarize themselves with a schedule or agenda, begin understanding the purpose of the standard setting, and communicate with the appropriate persons regarding any questions or concerns. The advance materials may include copies of any content standards, job analyses, test specifications, sample test items or tasks, historical context on the standard setting, statement of purpose for setting standards, and logistical materials such as information related to any honorarium, travel arrangements, and so on. Noticeably absent from the advance materials is specific information on the particular standard-setting method to be used; it is considered important that all participants have the same exposure to this information under uniform conditions with opportunity for immediate clarification so that accurate and consistent explanation and comprehension of that information are facilitated.

After reviewing advance materials, the next aspect of training participants encounter will likely be at the actual standard-setting meeting. Research has demonstrated the importance of activities planned to orient participants to the purpose of standard setting and to train them in the method they will use. According to Skorupski and Hambleton (2005), who studied the initial conceptualizations of participants in a standard-setting meeting,

> It is clear that panelists arrived at the meeting with very different ideas about the performance level descriptors and with very different perspectives on why they were there, the tasks to be accomplished, and the importance of the process. . . . The goal is to ensure that all participants are "on the same page," when it comes to purpose, task, and understanding of training. (p. 248)

A first key aspect of that meeting—and an important part of the training for participants—is usually an oral welcome and introduction to the purpose of standard setting. It is desirable that this introduction to purpose be delivered by a person with a high degree of credibility, status, and recognition within the profession or area covered by the assessment on which standards are to be set. For example, the introduction to purpose might be delivered by the chairperson, president, or executive director of a professional association, an officer of a board of directors, a state-level superintendent or director of assessment, and so on. There are critical outcomes of orientation to standard setting that must be accomplished. First, it is important that whatever orientation is given in the introduction be clearly laid out in advance to ensure that an accurate message is communicated regarding the purpose of standard setting. Second, it is essential that what the task participants will be asked to perform is clearly circumscribed (e.g., that

participants are also informed about what the purpose and tasks are *not,* if appropriate).

Following this introduction, it is common for the next aspect of training to consist of an opportunity for participants to self-administer and self-score a form of the test on which they will set standards, or in some other way (e.g., review a sample of items, tasks, etc.) become more familiar with the content and demands placed on examinees. Obviously, in some situations it may not be practical for standard-setting participants to actually complete a test form, particularly if the test is lengthy, if it comprises constructed-response format items that would be difficult to score, or if the test consists of performance tasks, demonstrations, oral components, and so on. Nonetheless, it is widely considered to be an effective and important aspect of training for participants to have experience with the assessment on which they will be making judgments. From the perspective of amassing validity evidence and documenting the qualifications (broadly speaking) of participants to engage in the standard-setting task, it would seem inappropriate for them *not* to have experience with the assessment.

If, as suggested previously, PLDs are developed in advance of the standard-setting meeting, the next aspect of training participants is to review and discuss them to ensure that participants have a strong grasp of that essential aspect of the process. Following this, training would then focus on providing clear instruction in the specific standard-setting method to be used. (Critical aspects of each method to be emphasized in training are highlighted in Section II within the chapters covering individual methods.)

In addition to a somewhat didactic introduction to the particular method, it is common to provide participants with limited opportunities to practice rating items, making judgments, placing bookmarks, reviewing work samples, or whatever else the standard-setting task will involve. Usually, a specially selected subset of items or tasks is assembled for this purpose, with the intention of broadly representing the kinds of items, formats, work samples, and so forth that participants will encounter during the actual standard-setting process, and with the intention of stimulating further discussion and refined understandings of the performance categories, concepts such as "borderline" or "Proficient" and so on. During this aspect of training, particularly challenging items, tasks, or rating situations can be brought to the attention of the entire group of participants so that common conceptualizations are facilitated.

As the final component of training, at least one juncture in the training portion of standard setting, participants are usually administered a survey, self-assessment, evaluation, or data collection instrument. The purpose of the data collection is to gather information on several aspects of

the training that can support the validity of the standard-setting process. For example, an instrument might question participants about

- their understanding of the purpose of the standard setting,
- their grasp of key conceptualizations,
- their understanding of the PLLs and PLDs,
- their comprehension regarding what kinds of judgments they are to make and how those judgments are to be made,
- their understanding of what they are to do when they are uncertain about a judgment,
- their understanding of the specifics of the standard-setting method, and
- their evaluation and recommendations regarding the effectiveness of all aspects of training (i.e., advance materials, orientation, logistics, self-administered test, instruction method, opportunities for discussion and clarification, etc.).

This list of potential survey topics is only a brief summary, intended to give an overview of this component of the training process. Additional specific information, along with sample survey items and other information, is presented later in this chapter, under the heading "Evaluating Standard Setting."

Providing Feedback to Participants

Another commonality cutting across nearly all standard-setting procedures is the provision of feedback to participants following the expression of their judgments about items, tasks, examinees, and so on, depending on the particular standard-setting method used. The feedback is often provided individually and confidentially to each participant; it is also common for feedback summaries (e.g., frequency distributions, means, etc.) to be provided with the identities of individual participants removed to assure their anonymity.

Many standard-setting approaches comprise "rounds" or iterations of judgments. At each round, participants may be provided various kinds of information, including a summary of their own judgments, a summary of their internal consistency, an indication of how their judgments compare to the judgments of other participants, an indication of variability in participants' ratings, and the likely impact of the individual and/or group judgment on the examinee population.

The kinds of feedback information provided to participants can be grouped into three categories according to the type of information communicated. The categories include *normative information,* which provides participants with data on how their ratings compare with other participants' judgments; *reality information,* which provides participants with data on

Table 3-5 Examples of Three Types of Feedback

Normative Information
1. cut scores by participant
2. distribution of bookmarks by participant
3. matrix of holistic rating by participant

Reality Information
1. item *p* values
2. item *b* values (or other scale score values)
3. response probability values (e.g., RP50 or RP67)
4. Reckase charts
5. correspondence of participants' ratings with external criterion variable (e.g., program director judgments, supervisor ratings, other test data)

Impact Information
1. raw (or scaled) score distributions
2. percentages of examinees in each performance category

the actual performance of examinees on items, tasks, and so on; and *impact information* (sometimes referred to as *consequence information*), which provides participants with data on the effect of their judgments on the examinee group. The following paragraphs elaborate on each type of feedback. Table 3-5 provides a listing of examples of each type of feedback, and more elaborate illustrations of each type of feedback listed in the table are available at www.sagepub.com/cizek/feedback.

We have one caveat with regard to Table 3-5, which gives examples of each of the three kinds of feedback. Because the way in which feedback is communicated to participants varies depending on the specific standard-setting method used, the number of participants, and decisions such as whether a group mean or median will serve as the eventual recommended standard, the examples shown in the table should be used as a guide only; the particular form of feedback information will need to be adapted to the user's specific context. For method-specific examples, readers may wish to consult specific sources such as Reckase (2001).

Normative Feedback

The first type of feedback, *normative information,* is data that permit each participant to perceive how his or her ratings compare to the other participants' judgments. Normative information is usually provided in the form of frequency distributions that show, for example, each participant's rating or judgment in an array with all other participants' judgments. Such

information can help participants see the extent to which they are judging more harshly or leniently relative to other participants, whether they are "outliers" compared to other participants, and how discrepant their judgments are from some measure of central tendency of the group's ratings.

In providing normative information, participants are usually assigned a unique identification code, so that results can be displayed and anonymity ensured, although individual participants can use their codes to locate their own ratings and summary information. Typically, summary statistics for the distribution of ratings, such as the mean, median, and standard deviation, are also provided. When normative information is provided to participants, participants should be assured that the purpose of collecting their individual judgments is to tap individual expertise, diversity of perspective, and informed opinions. It is important that training in how to use normative information include assurances that the purpose of the information is solely to provide participants with information about their judgments with respect to each other, possible outlier status, variability, and so on. The purpose of providing this feedback is *not* to suggest that they align their individual judgments with a group mean or alter their judgments based solely on relative stringency or leniency compared to other participants.

Reality Feedback

The second type of feedback, *reality information,* is data that help participants perceive the accuracy of their judgments. Reality information usually consists of item difficulty information, giving participants a rough idea of how their judgments about items or tasks compare to actual examinee performance. Reality information is usually provided to participants in a manner consistent with the test model in use and with deference to the ability of participants to comprehend and use the information. Typical difficulty (i.e., reality) information may be expressed as p values, b values, mean task performance, and so on.

For example, in some standard-setting methods (e.g., Angoff, 1971), the participants' task is to estimate the probability that a hypothetical, minimally competent or borderline examinee will answer an item correctly. Of course, because the borderline examinee group exists only in concept until the cut score is set, it is impossible to provide reality information based on the hypothetical group. Instead, assuming that operational test data are available, item difficulty information is commonly based on total examinee group performance. Participants can use the total group difficulty data as a rough gauge as to the accuracy of their ratings. For example, suppose that the total population of examinees covered by an examination was generally

regarded to be highly capable and homogenous, with very few examinees of borderline knowledge or skill. In this case, if total group item-difficulty information were provided to participants, participants would need to consider that their ratings for performance of a borderline group are likely to be close to, but perhaps slightly lower than, the total group *p* values. That is to say, the reality information would be statistically biased (i.e., total group *p* values would likely be systematically higher than borderline group *p* values), although precise estimates of exactly how much higher could not be determined and provided to participants.

Alternatively, reality information could be calculated using an initial estimate of the location of the cut score (from the first round of judgments), and *p* values or other indices can be calculated based on the performance of examinees located within, say, one standard deviation of the tentative cut score. In this case, participants could use the reality information in a fairly straightforward manner. In either case, it would be important to instruct participants regarding how to use the reality information (including, e.g., the potential that guessing on multiple-choice tests may contribute to borderline group performance on items and should be taken into account when making use of reality information) as part of the training and practice aspects of the standard-setting meeting.

Impact Feedback

The third type of feedback, *impact information,* is data that help participants understand the consequences of their judgments on the population of examinees that will be subject to application of the cut score(s) recommended by the standard-setting panel. Impact information might consist of overall passing or failure rates, the percentages of examinees likely to be classified in each of the performance level categories, the differential impact on examinees from relevant demographic groups, and so on. Accurate feedback information of this type is only possible if operational test data for a currently tested group (or a similar reference group) are available. Training in the use of impact information should highlight that such information should not be used in a normative sense (i.e., to achieve a preconceived passing rate) but as an aid to participants in understanding the consequences of the judgments and cut score recommendations. As such, impact information forms a key piece of evidence in evaluating standard setting, to the extent that it can be demonstrated that participants knew the consequences of their recommendations, had opportunity to discuss and revise their recommendations based on sound information, and expressed confidence in the final standards and associated impact.

Professional Guidelines for Standard Setting

In addition to fundamental clarity about the purpose for setting standards, a critical step in planning for a standard-setting activity is to become familiar with professionally recommended quality control steps that should be built into the process. A number of sources provide a compilation of guidelines related to the conduct of standard setting. Four comprehensive sources are Cizek (1996b), Hambleton (1998, 2001), and Kane (2001).

However, the single most authoritative source of guidance related to standard setting is the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999). This document, now in its sixth edition since 1954, represents the joint effort of the three sponsoring professional associations whose missions include the advancement of sound testing practice; for this reason, the *Standards for Educational and Psychological Testing* is frequently referred to as the *Joint Standards*. The three professional associations that sponsor the *Standards* include the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME).

As Linn (2006) has described, the *Standards* provides guidance on a wide range of test development and administration related activities. However, it gives specific attention to the process of standard setting. According to the *Standards,*

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

In addition to general standards related to aspects of test development, administration, and technical characteristics of tests (e.g., **reliability** and **validity**), the *Standards* also contains a number of specific statements pertaining to standard setting. For example, in the special context of licensure and certification testing, the *Standards* notes that "the validity of the inferences drawn from the test depends on whether the standard for passing makes a valid distinction between adequate and inadequate performance" (AERA/APA/NCME, 1999, p. 157).

Table 3-6 provides an abridged collection of these statements (numbered sequentially, with each individual statement also called a "Standard"). Two caveats are in order related to Table 3-6. First, although Table 3-6 lists each Standard related to setting cut scores, it is abridged in the sense that, in the

Table 3-6 AERA/APA/NCME Standards Related to Setting Cut Scores

Standard Number	*Standard*
1.7	When a validation rests in part on the opinions or decisions of expert judges, observers or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described. The qualifications and experience of the judges should be presented. The description of procedures should include any training and instructions provided, should indicate whether participants reached their decisions independently, and should report the level of agreement reached. If participants interacted with one another or exchanged information, the procedures through which they may have influenced one another should be set forth.
2.14	Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score.
2.15	When a test or combination of measures is used to make categorical decisions, estimates should be provided of the percentage of examinees who would be classified in the same way on two applications of the procedure, using the same or alternate forms of the instrument.
4.19	When proposed interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be clearly documented.
4.20	When feasible, cut scores defining categories with distinct substantive interpretations should be established on the basis of sound empirical data concerning the relation of test performance to relevant criteria.
4.21	When cut scores defining pass-fail or proficiency categories are based on direct judgments about the adequacy of item or test performances or performance levels, the judgmental process should be designed so that judges can bring their knowledge and experience to bear in a reasonable way.
6.5	When relevant for test interpretation, test documents ordinarily should include item level information, cut scores . . .
14.17	The level of performance required for passing a credentialing test should be dependent on the knowledge and skills necessary for acceptable performance in the occupation or profession and should not be adjusted to regulate the number or proportion of persons passing the test.

SOURCE: Adapted from AERA/APA/NCME (1999).

Standards document, each individual Standard is followed by more detailed, elaborated commentary designed to assist the user in applying the Standard appropriately. Second, the list of Standards shown in Table 3-6 represents a general set of principles. The list should not be viewed as the entirety of essential information regarding standard setting but rather as the fundamental guidelines that testing professionals consider to be important aspects of sound practice. Any entity responsible for planning, conducting, or implementing standard setting should be familiar with the full *Standards* document and should review other sources as appropriate to the specific standard-setting task at hand.

Evaluating Standard Setting

It is an important aspect of validation that any standard-setting process gather evidence bearing on the manner in which the particular standard-setting method was designed and implemented, and the extent to which participants in the process were able to understand, apply, and have confidence in the eventual performance standards they recommend. Evaluation of standard setting is a multifaceted endeavor with many potential sources of evaluation information (see Hambleton & Pitoniak, 2006). A complete listing of possible evaluation elements, based on the work of Pitoniak (2003), is provided in Table 3-7.

Evaluation of standard setting can be thought of as beginning with a critical appraisal of the degree of alignment between the standard-setting method selected and the purpose and design of the test, the goals of the standard-setting agency, and the characteristics of the standard setters. This match should be evaluated by an independent body acting on behalf of the entity that authorizes the standard setting and is responsible ultimately for the choice of cut scores.

Evaluation continues with a close examination of the application of the standard-setting procedure: To what extent did it adhere faithfully to the published principles of the procedure? Did it deviate in unexpected, undocumented ways? If there are deviations, are they reasonable adaptations, specified and approved in advance, and consistent with the overall goals of the activity?

The preceding questions reflect an "external" focus of the evaluation. Other evaluation activities can be thought of as more "internal" to the process. For example, a measure of the degree to which standard-setting participants seem to achieve consensus or converge toward a common standard from one round of judgments to the next can indicate that the selected method is working as intended. Trained facilitators can assess the extent to

Table 3-7 Standard-Setting Evaluation Elements

Evaluation Element	*Description*
Procedural	
Explicitness	The degree to which the standard-setting purposes and processes were clearly and explicitly articulated a priori
Practicability	The ease of implementation of the procedures and data analysis; the degree to which procedures are credible and interpretable to relevant audiences
Implementation	The degree to which the following procedures were reasonable and systematically and rigorously conducted: selection and training of participants, definition of the performance standard, and data collection
Feedback	The extent to which participants have confidence in the process and in resulting cut score(s)
Documentation	The extent to which features of the study are reviewed and documented for evaluation and communication purposes
Internal	
Consistency Within Method	The precision of the estimate of the cut score(s)
Intraparticipant Consistency	The degree to which a participant is able to provide ratings that are consistent with the empirical item difficulties, and the degree to which ratings change across rounds
Interparticipant Consistency	The consistency of item ratings and cut scores across participants
Decision Consistency	The extent to which repeated application of the identified cut scores(s) would yield consistent classifications of examinees
Other Measures	The consistency of cut scores across item types, content areas, and cognitive processes
External	
Comparisons to Other Standard-Setting Methods	The agreement of cut scores across replications using other standard-setting methods
Comparisons to Other Sources of Information	The relationship between decisions made using the test to other relevant criteria (e.g., grades, performance on tests measuring similar constructs, etc.)
Reasonableness of Cut Scores	The extent to which cut score recommendations are feasible or realistic (including pass/fail rates and differential impact on relevant subgroups)

SOURCE: Adapted from Pitoniak (2003).

which deliberations or discussions are freely engaged in by all participants or are driven by one or more influential participants.

In-progress evaluations of the process of standard setting also serve as an important internal check on the validity and success of the process. Minimally, two evaluations should be conducted during the course of a standard-setting meeting, and both usually consist of mainly forced-choice survey questions with a few open-ended items. The first evaluation occurs after initial orientation of participants to the process, training in the method, and (when appropriate) administration to participants of an actual test form. This first evaluation serves as a check on the extent to which participants have been adequately trained, understand key conceptualizations and the task before them, and have confidence that they will be able to apply the selected method. The second evaluation is conducted at the conclusion of the standard-setting meeting and mainly serves the purpose of gathering information on participants' level of confidence in and agreement with the final to-be-recommended standard. A sample survey, which includes both kinds of questions and which users should modify to their particular context, is shown in Table 3-8. An adaptable electronic version of the form can be found at www.sagepub.com/cizek/evaluationform.

Much of the preceding information on evaluation has focused on process-related aspects of standard setting. Of course, the product or result (i.e., the actual cut scores) of standard setting is arguably even more important. Two commonly employed evaluation criteria related to results include *reasonableness* and *replicability*.

The reasonableness aspect can perhaps initially be assessed in the first "product" of standard setting, which in most instances are PLLs and PLDs or a written description of referent examinees or groups. The utility and comprehensibility of these descriptions are essential. For a given field, subject, or grade level they should accurately reflect the content standards or credentialing objectives. They should be reasonably consistent with statements developed by others with similar goals.

The aspect of reasonableness can be assessed by the degree to which cut scores derived from the standard-setting process classify examinees into groups in a manner consistent with other information about the examinees. For example, suppose that a state's eighth-grade reading test and the NAEP were based on common content standards (or similar content standards that had roughly equal instructional emphasis). In such as case, a standard-setting procedure for the state test resulting in 72% of the state's eighth graders being classified as Proficient, while NAEP results for the same grade showed that only 39% were proficient, would cause concern that one or the other set of standards was inappropriate.

Table 3-8 Sample Evaluation Form for Standard-Setting Participants

Directions: Please indicate your level of agreement with each of the following statements and add any additional comments you have on the process at the bottom of this page. Thank you.

Item	*Statement*	*Strongly Disagree*	*Disagree*	*Agree*	*Strongly Agree*
1	The orientation provided me with a clear understanding of the purpose of the meeting.				
2	The workshop leaders clearly explained the task.				
3	The training and practice exercises helped me understand how to perform the task.				
4	Taking the test helped me to understand the assessment.				
5	The performance level descriptions (referent examinee descriptions) were clear and useful.				
6	The large and small group discussions aided my understanding of the process.				
7	There was adequate time provided for discussions.				
8	There was an equal opportunity for everyone in my group to contribute his/her ideas and opinions.				
9	I was able to follow the instructions and complete the rating sheets accurately.				
10	The discussions after the first round of ratings were helpful to me.				
11	The discussions after the second round of ratings were helpful to me.				
12	The information showing the distribution of examinee scores was helpful to me.				
13	I am confident about the defensibility and appropriateness of the final recommended cut scores.				
14	The facilities and food service helped create a productive and efficient working environment.				

15 Comments:_____

SOURCE: Adapted from Cizek, Bunch, & Koons (2004).

Local information can also provide criteria by which to judge reasonableness. Do students who typically do well in class and on assignments mostly meet the top standard set for the test, while students who struggle fall into the lower categories? In licensure and certification contexts, past experience with the proportions of candidates who have been deemed competent and the experiences of those who oversee preservice internships or residencies or have other interactions with candidates for the credential can be brought to bear to assess reasonableness.

Replicability is another aspect in the evaluation of standard setting. For example, in some contexts where substantial resources are available, it is possible to conduct independent applications of a standard-setting process to assess the degree to which independent replications yield similar results. Evaluation might also involve comparisons between results obtained using one method and an independent application of one or more different methods. Interpretation of the results of these comparisons, however, is far from clear. For example, Jaeger (1989) has noted that different methods will yield different results, and there is no way to determine that one method or another produced the wrong results. Zieky (2001) noted that there is still no consensus as to which standard-setting method is most defensible in a given situation. Again, differences in results from two different procedures would not be an indication that one was right and the other wrong; even if two methods did produce the same or similar cut scores, we could only be sure of precision, not accuracy.

Conclusions and a Foreword

There exists an extensive and growing list of methodological options for setting standards. However, many of the methods have similar components; this chapter has surveyed many aspects that are common to a range of standard-setting methods. Prior to actually implementing a standard-setting procedure, these aspects must be attended to, as they can have as great an influence on the defensibility of the procedure and the resulting cut scores as whatever specific standard-setting method is used and how it is carried out.

The ultimate goal of all the standard-setting methods that will be described in the following chapters is to aid participants in bringing their judgments to bear in ways that are reproducible, informed by relevant sources of evidence, and fundamentally fair to those affected by the process. The first three steps in accomplishing that goal are (1) to ensure that a standard-setting method is selected that aligns with key features of the purpose and format of the test on which standards will be set, (2) to identify and select

an adequate panel of participants, and (3) to train them regarding any key conceptualizations that must be applied, the kinds of judgments they are to make, and how to incorporate feedback related to those judgments. Toward that goal also, those responsible for designing and implementing a standard-setting procedure should ground those activities in relevant professional guidelines for standard setting and should engage in systematic evaluation of the procedure at key junctures in order to gather evidence regarding the validity of the procedure and results.

In the following chapters, many methods for actually implementing specific standard-setting methods are described. The concepts covered in this chapter, however, are *not* repeated in each of the following chapters. For the sake of avoiding redundancy, specific references are not made to selecting and training participants, developing key conceptualizations, evaluating the standard-setting process, and so on for each method. Readers should be alert, however, that the need for attention to these important commonalities remains. Instead, the focus in the following chapters will be on the unique aspects of implementing each method described, the specific contexts in which they are best suited (or ill suited), and any unique challenges to implementing the method.

SECTION II

Standard-Setting Methods

The chapters in this section introduce several of the most frequently used standard-setting methods. The one common element of all these methods is that they involve, to one degree or another, human beings expressing informed judgments based on the best evidence available to them, and these judgments are summarized in some systematic way, typically with the aid of a mathematical model, to produce one or more cut scores.

Each method combines art and science. Although it can be shown that different methods will yield different results (cut scores and percentages of examinees falling into each performance level), it has not been—and may never be—demonstrated that a particular set of results based on a particular method is always superior to another set of results based on another method. It *can,* however, be demonstrated that some methods are better suited to certain types of tests or circumstances, but even here there are few hard-and-fast rules that a particular method must or must not be used with a particular type of test or in a particular circumstance.

As we mentioned previously, there was a period in educational and psychological assessment when norm-referenced testing and evaluation methods dominated. Percentile ranks were used to determine a variety of outcomes, including selection, eligibility, admissions, guidance, certification, and so on. Decisions about academic placement, hiring, promotion, raises, retention, and honors and recognition were routinely based on an examinee's rank vis-à-vis all other examinees. As we have acknowledged, test-based decisions will, to some degree, always have a norm-referenced

aspect as long as academic admissions, scholarships, promotions, career advancement, and a host of other decisions exist within a zero-sum, competitive environment. However, in contemporary large-scale assessment, there is an increased emphasis on absolute accomplishment, and criterion-referenced tests (CRTs) and standards-referenced tests (SRTs) have gained prominence. Accordingly, the need for a proper way to express their outcomes and appropriate score inferences has intensified.

Three independent currents led to an increased scrutiny of previously existing methods of setting performance standards and to a proliferation of new methods. The first was the advent of standards-based education. The second was the rise of multiple performance levels. The third was the increasing use of constructed-response (CR) items in large-scale assessments. All three of these currents are closely associated with the National Assessment of Educational Progress (NAEP).

NAEP, first authorized by Congress in 1969, is also known as the *Nation's Report Card*. From its inception, NAEP was designed to show how well students in the United States were performing in a wide range of academic subjects. For each subject, a national panel of content experts devised a set of content standards—topics on which students should be able to demonstrate some degree of mastery. As the content standards were promulgated, other panels developed test items and others created descriptions of various performance levels with specific reference to these content standards. Ways to express student performance on these items relative to the content standards and performance-level descriptions were dependent on the standard-setting methodology of the time. It soon became evident that existing methods were not entirely adequate.

Loomis and Bourque (2001) provide an excellent review of the emergence of new methods for setting performance standards for NAEP. The authors, associated with NAEP for many years, lend both authenticity and a sense of the struggle that NAEP and contractor staff (American College Testing, or ACT) encountered on the frontiers of large-scale educational assessment very much in the public eye. Over the course of several years, the NAEP/ACT team pilot tested numerous existing and emerging standard-setting methodologies. Many innovations in standard setting, many of the recently introduced methods, and much of the extensive documentation and research on standard setting are a direct result of these pioneering efforts.

The chapters in this section describe both long-standing and more recently introduced methods of setting performance standards, with each chapter devoted to a specific method (or highly related methods). Most of the methods described are designed for establishing so-called absolute or criterion-referenced cut scores, or cut scores on standards-referenced tests. The section

also includes so-called **compromise methods** as well that attempt to strike a balance between criterion-referenced and norm-referenced concerns.

For each method, we provide a brief historical overview. However, the bulk of each chapter focuses on providing a detailed description of the psychometric foundations and applications of the method, circumstances for which the method is particularly well suited, and specific procedures for implementing the method and analyzing the data resulting from use of the method in order to obtain whatever cut scores are needed by the user.

Finally, the reader may have noticed in Chapter 3 a link to a Web site containing examples of various kinds of standard-setting feedback. In each chapter of Section II, we continue this practice of including links to additional information, forms, software, training materials, data sets, and other aids to foster understanding of the concepts covered, as well as to provide practitioners with useful templates that can be easily adapted to different contexts and needs.

4

The Nedelsky Method

Though not nearly as commonly implemented as many other methods, and used primarily in credentialing as opposed to educational contexts, a method proposed by Nedelsky (1954) for setting cut scores remains in use today. The longevity of the method is perhaps due to the facts that it is intuitive to and easily performed by participants, that it is comparatively time efficient, and that the kind of item format to which it can be applied—multiple choice—remains a format used in many testing programs. Another advantage to the method is that it can be implemented with or without examinee performance data on the items or test form on which standards are based. That is, a Nedelsky cut score can be derived before or after operational administration of the test requiring performance standards to be set.

Though used primarily in the credentialing arena, it was in an educational context that the Nedelsky (1954) method was developed. It was, at the time it was proposed, perhaps the first widely disseminated criterion-referenced method for setting cut scores. At the time of its introduction, norm-referenced methods were the most commonly used approach for setting standards, assigning students' grades in courses, and so on. By norm-referenced standard setting and evaluation here, we mean that passing/failing or the awarding of grades such as A, B, C, D, and F are determined not in an absolute sense to reflect an individual's actual achievement but in a comparative sense, reflecting the relative standing or achievement of the individual within the appropriate comparison group. Nedelsky's method is included here because it remains in use today, but it is particularly noteworthy because it was one of

the first to shift the focus from relative performance to what Nedelsky termed "absolute" levels of performance. In many situations today, norm-referenced procedures would be judged as fundamentally unfair and would raise serious validity concerns.

Procedures for the Nedelsky Method

To use the Nedelsky (1954) method, members of a standard-setting panel assign probabilities to multiple-choice test items based on the judged likelihood that a specific, though abstract, group of examinees should be able to rule out incorrect options. The examinees used as a reference group are hypothetical examinees on the borderline between inadequate and adequate levels of performance, or what is sometimes referred to as the borderline between mastery and nonmastery of some domain of knowledge, skill, or ability.

Nedelsky, a university professor of physical science, was concerned with assigning grades in his course based on students' performance on a final examination. He proposed considering the characteristics and performance of a hypothetical borderline examinee that he referred to as the "F-D student," where "F" and "D" refer to clearly failing and just passing grades in a course, respectively. The F-D student lies right on the cusp of the failing and passing categories. Accordingly, the first portion of a standard-setting meeting in which the Nedelsky method is used consists of participants' discussion, description, and clarification of the borderline examinee.

After this important conceptualization has been formalized, participants using the Nedelsky (1954) method inspect, individually, each item in a multiple-choice examination, with particular attention to the options (response choices) for each item. According to Nedelsky, on an individual item,

> Responses [i.e., options] which the lowest D-student should be able to reject as incorrect, and which therefore should be attractive to [failing students] are called F-responses. . . . Students who possess just enough knowledge to reject F-responses and must choose among the remaining responses at random are called F-D students. (p. 5)

Participants then review each item in a test form and, for each item, identify the options that they believe a hypothetical minimally competent examinee would rule out as incorrect. The reciprocal of the remaining number of options becomes each item's "Nedelsky value." That value is interpreted as the probability that the borderline student will answer the item correctly. For example, suppose that participants judged that, for a certain five-option item, borderline examinees would be expected to rule out two of

the options as incorrect, leaving them to choose from the remaining three options. The Nedelsky rating for this item would be 1/3 = .33. Repeating the judgment process for each item would give a number of Nedelsky values equal to the number of items in the test (n_i). The sum of the n_i values can be directly used as a raw score cut score. For example, a 50-item test consisting entirely of items with Nedelsky ratings of .33 would yield a recommended passing score of 16.5 (i.e., $50 \times .33 = 16.5$). This example also illustrates what we first broached in Chapter 2; that is, a standard-setting procedure may not result in a recommended passing score that is a whole number, and it will often be necessary to implement a policy decision regarding how the fractional value should be handled (e.g., rounded, truncated, etc.).

Of course, it is not likely that all items in a test would be judged to have a Nedelsky rating of .33. Table 4-1 provides a more realistic illustration of the kind of data that would result from using the Nedelsky method and how the data would be analyzed to arrive at a cut score. The table presents the hypothetical ratings of 15 items by 6 participants. In the example, participants were not required to reach consensus about the Nedelsky values for each item; rather, the mean of their judgments is used as the final rating for each item. The sum of the 15 ratings, 6.89, or approximately 7 items correct, would be the recommended passing score on the 15-item test. This is the number of items that the borderline student would be expected to answer correctly. The same data and an Excel spreadsheet for calculating the cut score based on the hypothetical data shown in Table 4-1 can be found at www.sagepub.com/cizek/nedelsky.

Alternative Procedures and Limitations

There are at least three ways of arriving at the Nedelsky rating for each item. One way—the one illustrated in Table 4-1, is to permit all participants to provide judgments for all items independently, then average the independent ratings (see the column labeled "Item Means"). One modification of this approach would be to have all participants provide their own ratings, but to encourage group discussion of items (perhaps even progressing as a group, item by item) before collecting the ratings. Strictly speaking, these ratings would not be independent, but would perhaps be less variable, and participants' accuracy would likely benefit from the interactions. Another variation would be to configure the procedure as either independent or encouraging group discussion, then gather two rounds of ratings for each item. Between rounds, participants could receive normative information (see Chapter 2) so as to see how their ratings compare with other participants.

Table 4-1 Hypothetical Data for Nedelsky Standard-Setting Method

Rater ID Number							
	1	*2*	*3*	*4*	*5*	*6*	
Item	Nedelsky Values						Item Means
1	.33	.50	.50	.33	.33	.33	.39
2	.50	1.00	.50	.25	1.00	1.00	.71
3	.25	.33	.25	.25	.25	.33	.28
4	1.00	1.00	.50	1.00	.50	1.00	.83
5	.33	.33	.33	.33	.25	.33	.32
6	.25	.33	.25	.25	.25	.33	.28
7	.25	.20	.25	.33	.20	.20	.24
8	1.00	.33	1.00	.50	1.00	.50	.72
9	.20	.33	.25	.20	.33	.25	.26
10	.50	1.00	1.00	.50	.50	1.00	.75
11	.50	.50	.50	1.00	.50	.50	.58
12	.50	.33	.33	.50	.33	.33	.39
13	.20	.20	.20	.20	.20	.20	.20
14	.25	.20	.33	.25	.33	.25	.27
15	1.00	.50	.50	.50	1.00	.50	.67
					SUM = 6.89		

As another alternative to the basic procedure, Nedelsky originally proposed an adjustment of the simple sum of the ratings that, in essence, would take into account the relative costs of incorrect pass/fail decisions. Suppose, for example, that each participant in a Nedelsky procedure arrived at his or her own summed rating across the items in a 15-item test. We can represent each participant's sum as Σx_j, where x is the sum of the participant's Nedelsky ratings and the subscript j runs from 1 to n_j, where n_j is the total number of participants. Assuming that participants were not required to reach consensus on their Nedelsky values for each item, participants' sums would vary.

The mean of those sums, *M*, obtained by $\Sigma x_j/n_j$ would be one option for the recommended cut score on the total test. The adjustment to that cut score proposed by Nedelsky takes into account the nature of the responses that participants identified as those the borderline (i.e., F-D) student would recognize as incorrect. According to Nedelsky, there are two kinds of such responses. The first kind is very clearly or egregiously incorrect; the second kind is less obviously incorrect. For a test comprising the first kind of incorrect answers, Nedelsky suggested adjusting the cut score so that very few of the borderline or F-D students would pass; for a test comprising primarily the second type of incorrect answers, a less severe adjustment would be made.

The actual adjustment took into account the kinds of items in the test by calculating a factor that Nedelsky labeled σ_{FD}. When participants' item ratings were such that they judged 0, 1, 2, 3, or 4 options to be recognizable by F-D students with approximately equal occurrence, σ_{FD} would take on a value of $.41\sqrt{n_i}$, where n_i is the total number of items in the test. Where extreme ratings (i.e., those identifying none or all incorrect options as recognizable by the F-D student) were comparatively rare, Nedelsky suggested using a value for σ_{FD} of $.50\sqrt{n_i}$. In addition, Nedelsky proposed the use of a constant, *k*, that when multiplied by σ_{FD} gave the final adjustment to the mean cut score (*M*). Nedelsky suggested using values of *k* such as –1, 0, 1, and 2. These values would result in failing 26%, 50%, 84%, and 98% of F-D students, respectively, for a test where $.41\sqrt{n_i} < \sigma_{FD} < .50\sqrt{n_i}$. A higher value of *k* would be preferred when the cost of a false positive was judged to be high, or when the test under study contained a large proportion of clearly or egregiously incorrect answer choices. A lower value of *k* would be preferred when the cost of a false negative was judged to be high, or when the test under study contained a comparatively small proportion of clearly or egregiously incorrect answer choices.

Although the modification proposed by Nedelsky is rarely if ever applied when the method is used, we agree with the concept of incorporating consideration of the relative costs of classification errors. Other approaches to calculating cut scores that explicitly take into account the relative costs of errors have been proposed, such as the "ratio of regret" (see, e.g., Bunch, 1978; Emrick, 1971). Regardless of whether a statistical or judgmental approach to the issue is taken, we recommend that entities responsible for testing programs consider adopting a policy regarding the concern of classification errors a priori.

In addition to illustrating the Nedelsky method, Table 4-1 also illustrates some of the limitations of the method that have been noted in the literature. As mentioned previously, the method can only be used with multiple-choice format items. However, not all multiple-choice items in a test to which the

method is applied would need to have the same number of options; the same procedure of taking the reciprocal of the number of remaining options is used even if the items in the test vary in the number of response choices.

Another limitation of the Nedelsky method is that it essentially permits participants only a very limited number of probabilities that they can assign to an item. For example, for a five-option item, the only Nedelsky values that could result from an individual participant's ratings would be .20, .25, .33, .50, and 1.00 (see Berk, 1984). In addition, another well-known limitation of the method is that there are not equal intervals between those possibilities. Because raters tend not to assign probabilities of 1.00 (i.e., to judge that a borderline examinee could rule out all incorrect response options), this tends to create a downward bias in item ratings (i.e., a rating of .50 is assigned to an item instead of 1.00) with the overall result being a somewhat lower passing score than the participants may have intended to recommend, and somewhat lower passing scores compared to other methods (Shepard, 1980).

A third limitation of the Nedelsky method is that, ostensibly, it cannot be used in situations where more than one cut score is required on the same test (e.g., one cut score to separate *Basic* and *Proficient* performance levels, another to distinguish between *Proficient* and *Advanced*). We are not aware of any situation in which this modification of the basic Nedelsky procedure has been tried. Such a modification seems difficult—though not impossible—to conduct with success, as it parallels such modifications that have been attempted successfully with other methods. If such a modification were attempted with the Nedelsky method, participants would need to provide two ratings for each item.

For example, if a *Basic/Proficient* and a *Proficient/Advanced* cut score were needed, a first rating session would require participants to identify those options that could be eliminated by an examinee on the *Basic/Proficient* borderline. A second rating session would require participants to identify options that could be eliminated by an examinee on the *Proficient/Advanced* borderline. Presumably, during the second rating session, participants would begin examining items with information already in hand regarding which options they had identified during the first session, which would make identification of the second (higher) cut score relatively easier and more efficiently accomplished. However, to be successful, this modification would also seem to place a greater burden on test construction. For example, items would need to be written with attention to creating options of varying appeal to examinees of differing ability levels. The modification would not likely be useful if items had fewer than five options. And it does not appear that the modification could be used in any context requiring more than two cut scores.

5

The Ebel Method

Robert Ebel introduced the standard-setting method that now bears his name in 1972. Like the method named after William Angoff (1971), the introduction of a new standard-setting method was not the primary focus of Ebel's (1972) work, but was a comparatively minor topic in *Essentials of Educational Measurement*, a book he had written primarily as an introduction to testing for upper-level undergraduates and graduate students taking a first course in applied measurement and evaluation. Ebel's method is currently used mainly in medical and health-related contexts and perhaps most commonly as a method for setting the passing level on classroom achievement tests (see, e.g., Downing, Lieska, & Raible, 2003). It is generally applied to tests comprising multiple-choice format items. One advantage of the Ebel method is that, although it does not appear to have been applied to other formats, there does not seem to be any reason why it could not be utilized whenever item scoring is dichotomous, and it may even be possible to extend the method to tests comprising a mix of polytomously and dichotomously scored items.

Procedures for the Ebel Method

Like other so-called test-based approaches, the method proposed by Ebel (1972) requires participants to make judgments about individual test items. However, unlike some other methods, Ebel's method requires participants to make two judgments for each item: one judgment being an estimate of the

difficulty of each item, the other judgment regarding the relevance of the item. Participants' judgments are not expressed as probabilities but as category placements. For example, participants are typically required to classify items into one of three difficulty categories (Easy, Medium, and Hard) and into one of four relevance categories (Essential, Important, Acceptable, and Questionable). These judgments effectively cross-categorize all items into one of 12 cells (e.g., items that are judged to be of Medium difficulty and Acceptable relevance, items that were judged to be Hard and Important, and so on). It is important to note at this point that participants' judgments about the characteristics of Difficulty and Relevance are *not* made with respect to a hypothetical examinee but with respect to the purpose of the test and the population of examinees. Thus participants are simply required to judge whether they believe individual items are Hard, Medium, or Easy for examinees as a group, and whether the individual items on the test form used for standard setting are Essential, Important, and so on as regards the pass/fail or other classification decision to be made.

Specific reference to a target examinee group such as "borderline" or "minimally qualified" examinees is incorporated into the next step in the Ebel method. After making two judgments about individual items, it is at this point that participants make another judgment regarding the proportion of each of the 12 cells that should be answered correctly. Specifically, participants are asked to judge, for each Difficulty-by-Relevance cell, the percentage of items in that cell that hypothetical minimally qualified or borderline examinees should answer correctly.

Table 5-1 illustrates the use the Ebel procedure and data analysis for that method. The first step in obtaining a cut score is to gather and summarize item judgments regarding Relevance and Difficulty. Table 5-1 presents the hypothetical item classifications for 100 items by a panel of 5 participants, yielding a total of 500 item judgments. Judgments for each of the 12 cells are shown in the column labeled "Number of Items Judged to Be in Category (A)." With five participants, each item would be judged by each of the participants, resulting in 500 total judgments (100 items times 5 judges). For the data shown in the table, 94 of the Relevance judgments classified an item as "Essential," 259 as "Important," 125 as "Acceptable," and 22 as "Questionable. Of the total number of judgments regarding item difficulty, the table shows that, for example, items were classified as "Easy" by the participants a total of 228 times ($94 + 106 + 24 + 4 = 228$). The total number of judgments for any dimension (i.e., Relevance or Difficulty) is somewhat difficult to interpret precisely, however, because the judgments could have been obtained in a variety of ways. For example, the total of 228 "Easy" item classifications could have been obtained in many ways (e.g., participants

Table 5-1 Illustration of Ebel Standard-Setting Method

Relevance Category	*Difficulty Category*	*Number of Items Judged to Be in Category (A)*	*Judged Percentage Correct (B)*	*Product (A × B)*
Essential	Easy	94	100%	9,400
	Medium	0	100%	0
	Hard	0	100%	0
	Subtotal	**94**		
Important	Easy	106	90%	9,540
	Medium	153	70%	10,710
	Hard	0	50%	0
	Subtotal	**259**		
Acceptable	Easy	24	80%	1,920
	Medium	49	60%	2,940
	Hard	52	40%	2,080
	Subtotal	**125**		
Questionable	Easy	4	70%	280
	Medium	11	50%	550
	Hard	7	30%	210
	Subtotal	**22**		
TOTALS		**500**		**37,630**
		Passing percentage (C_x) = 37,630/500 = 75.46%		

SOURCE: Adapted from Ebel (1972).

differing in their judgments about which items were "Easy" across a large number of items, or participants frequently in agreement about a comparatively smaller proportion of items being "Easy").

The second judgmental task for participants using the Ebel method is to assign values to each of the 12 relevance-by-difficulty cells that represent the participants' judgment regarding what percentage of items in a cell the

target (e.g., borderline, minimally competent, etc.) examinee should answer correctly in order to be considered passing. Often, this activity is done as a consensus activity by the group of participants, involving reference to the characteristics of the hypothetical borderline examinee, group discussion, and so on. Table 5-1 also shows the result of group consensus regarding the cell percentages in the column labeled "Judged Percentage Correct (B)." For example, participants agreed that borderline examinees should be expected to answer all of the "Essential" items correctly, regardless of whether the item was judged to be Easy, Medium, or Hard in terms of difficulty. Participants in this example assigned different percentages to items judged to be Important, however; the table shows that the panel indicated that examinees should answer correctly 90% of items judged to be Easy, 70% of items judged to be of Medium difficulty, and only 50% of items judged to be Hard.

To obtain a cut score using the Ebel method, the number of times items are judged to be in a category (A) is multiplied by the percentage of those items participants have deemed must be answered correctly (B). The product of A and B is shown in the final column of Table 5-1. The individual row products for each cell are added together, and the sum appears at the bottom of the final column. Dividing the sum of the cell products (37,630) by the total number of judgments (500) yields a recommended passing percentage correct. In the hypothetical case illustrated in Table 5-1, the passing percentage correct would be 75.26% or approximately 75–76 of the 100 test items correct. The same data and an Excel spreadsheet for calculating the cut score based on the hypothetical data shown in Table 5-1 can be found at www.sagepub.com/cizek/ebel.

Alternative Procedures and Limitations

One modification of the basic Ebel procedure that seems preferable to implement when item performance data (i.e., *p* values) are available is to delete the step in the procedure where participants are asked to make judgments about the difficulty levels of the individual items in a test. Obviously, if item performance data are available, it makes more sense to rely on the empirical values than to solicit participants' estimates of those values. The difficulty estimation step cannot be eliminated entirely, however, but only substituted with a different judgment about item difficulty. In place of the step as originally conceived by Ebel, participants would still need to decide on the borderlines of the Easy, Medium, and Hard categories. By reviewing individual items with performance data, participants might decide, for

example, that items with *p* values from 0.00 to 0.49 be classified as Hard, those with *p* values between .50 and .79 be categorized as Medium, and those with *p* values of .80 or greater be considered Easy.

A final alternative to the basic Ebel method would be to eliminate the need for group consensus regarding the "Judged Percentage Correct" for each of the 12 Difficulty-by-Relevance cells. Group discussion of the range of values considered and the rationales for various values would still be desirable. However, instead of requiring a consensus about such a value that may be difficult to attain, the percentages in column B of the table could be designated by policy, or the mean of individual participants' judgments about the "Judged Percentage Correct" for each of the 12 difficulty-by-relevance cells could be used.

As noted previously, one advantage of the Ebel method is that it can be used with item formats other than multiple choice. It is clear that the method can be used whenever a test consists of dichotomously scored items. It is possible, however, that the method might also be used with items scored according to a scoring rubric (e.g., constructed-response items that are polytomously scored). In such a case, the polytomous items would be classified in the same way as regards their Difficulty (i.e., into one of the four Difficulty categories); however, the "Judged Percentage Correct" applied to items in a category would be applied and interpreted in two different ways. For the dichotomously scored items, the interpretation would be the same as in the basic Ebel method (i.e., the percentage of items judged to be in a particular cell that borderline examinees should answer correctly). However, for polytomously scored items, the cell percentage would indicate the percentage of rubric score points that a borderline examinee would be expected to obtain. In practice, because of variability in scoring rubrics, variability in the distributions of score points for individual items, and varying difficulty of obtaining the same score point across different items, it would seem appropriate to calculate the product (A × B) separately for dichotomously scored items and for each polytomously scored item, then sum those products.

Although adaptable and easily implemented, the Ebel method has some limitations. Berk (1984) has suggested that it may be challenging for standard-setting participants to keep the two dimensions of difficulty and relevance distinct because those dimensions may, in some situations, be highly correlated. And, as already noted, the method requires participants to make judgments about item difficulty that may not be necessary (i.e., when empirical item difficulty values are available). Inclusion of a difficulty-estimation step under these circumstances may even present some concerns about the validity of the participants' difficulty estimates to the extent that they differ

markedly from the known values. For these reasons, the item difficulty estimation step of the Ebel method should probably be avoided whenever operational item data can be used.

Another limitation—actually, another validity concern—has to do with judgments about item relevance. As Table 5-1 illustrates, for a total of 22 times, items in the hypothetical 100-item test were judged to be of "Questionable" relevance. Because the "Questionable" item category appears after the "Acceptable" category in the ordinal sequence of category labels used in the example, it is a reasonable conclusion that items judged to be "Questionable" are, in fact, not acceptable as regards their relevance for classifying examinees for licensure, certification, graduation, and so on. As such, the inclusion of items judged to be of questionable relevance appears on its face to weaken the validity evidence supporting defensible interpretation of the total test scores.

6

The Angoff Method and Angoff Variations

Unlike most other standard-setting methods, the procedure attributed to William Angoff (and its many variations) was not the primary focus of the original work in which it was published. The method initially appeared as a very short portion—a sidebar almost—in the 92-page chapter Angoff wrote on scaling, norming, and equating for the measurement reference book *Educational Measurement, Second Edition* (Thorndike, 1971). That original chapter contains what is now referred to as "the Angoff method," although Angoff attributed the method to a colleague at Educational Testing Service, Ledyard Tucker. Thus it might have been more appropriate for the method to be called the "Tucker method," but we will choose not to swim against that tide.

The method as described by Angoff is rarely used exactly as it was proposed. Rather, slight reconfigurations of the basic approach—each variation referred to as "modified Angoff method"—are now considerably more common, although precisely what constitutes a "modified" Angoff method is somewhat unclear. However, in almost all cases in current practice where an Angoff method is said to be used, it is almost certain to be a modified Angoff approach, which we will describe later in this chapter. In addition to the basic Angoff and modified Angoff approaches, so-called extended Angoff procedures have been developed, and recently a procedure called the "Yes/No method" has been introduced that is highly similar to the original Angoff approach.

Despite the lack of clarity in labeling, it is certain that the Angoff method (and all its variations) is the most commonly used method for setting performance standards in contemporary use in licensure and certification contexts. Although it has become somewhat less frequently used in K–12 education settings, recent published surveys involving credentialing examination programs conclude that the Angoff approaches are prevalent (see, e.g., Meara, Hambleton, & Sireci, 2001; Sireci & Biskin, 1992).

Shortly after its introduction, a substantial amount of research was conducted on the Angoff method in the 1980s, and the use of the Angoff method in current comparative studies suggests that it remains a method of choice. It is safe to say that the Angoff methods have become the most thoroughly researched of all standard-setting methods, and they have been included in dozens of comparison studies with other methods. In 1988, Mills and Melican reported that "the Angoff method appears to be the most widely used. The method is not difficult to explain and data collection and analysis are simpler than for other methods in this category" (p. 272). In a 1981 study comparing the Angoff, Ebel, and Nedelsky methods, Colton and Hecht reported that "the Angoff technique and the Angoff consensus techniques are superior to the others" (p. 15). In his 1986 review of standard-setting methods, Berk concluded that "the Angoff method appears to offer the best balance between technical adequacy and practicability" (p. 147).

As originally proposed, the Angoff method is well suited for tests comprising multiple-choice format items. In licensure and certification testing contexts and in medical and health-related professions, where tests often consist exclusively of multiple-choice items, the Angoff method remains the most widely used. In K–12 education contexts, where tests increasingly comprise a mix of multiple-choice and constructed-response format items, Angoff and modified Angoff approaches are being used less frequently than when those tests comprised multiple-choice items exclusively. However, a more recent variation of the basic Angoff approach—called the "Extended Angoff method"—has been developed and is often used in mixed-format assessments.

Procedures for the Angoff Method

Like other methods that require participants to make judgments about test items, Angoff (1971) proposed that participants review each operational item to be included in a test form and provide estimates of the proportion of a subpopulation of examinees who would answer the items correctly. The subpopulation of interest is that group of examinees who would be considered

"just barely passing," "minimally competent," or just over the hypothetical borderline between acceptable and unacceptable performance. Angoff's original suggestion was as follows:

> A systematic procedure for deciding on the minimum raw scores for passing and honors might be developed as follows: keeping the hypothetical "minimally acceptable person" in mind, one could go through the test item by item and decide whether such a person could answer correctly each item under consideration. If a score of one is given for each item answered correctly by the hypothetical person and a score of zero is given for each item answered incorrectly by that person, the sum of the item scores will equal the raw score earned by the "minimally acceptable person." (pp. 514–515)

Almost as an aside, Angoff described—in a footnote to the basic description of the method—an alternative to asking participants to simply assign zeros and ones to each item in a test form. The alternative he described has become standard practice for implementations of the method. According to Angoff's footnote,

> A slight variation of this procedure is to *ask each judge to state the probability* that the "minimally acceptable person" would answer each item correctly. In effect, judges would think of a number of minimally acceptable persons, instead of only one such person, and would estimate the proportion of minimally acceptable persons who would answer each item correctly. The sum of these probabilities would then represent the minimally acceptable score. (1971, p. 515, emphasis added)

In this description, Angoff was obviously not referring to the acceptability of an examinee as a *person,* but to the qualifications of the examinee vis-à-vis the purpose of the test. Subsequent to Angoff's introduction of this method, the phrase "minimally competent examinee" has been substituted when the Angoff procedure is used. It should be clear, however, that this idea—that is, the minimally competent or borderline examinee—is a key referent for this standard-setting method. It is common for much of the training time afforded to the actual standard-setting meeting to be devoted to helping participants refine and acquire this essential conceptualization.

Although the conceptualization of the minimally competent examinee is somewhat unique to the Angoff method, the other aspects of implementing the method are common to most other standard-setting approaches. Namely, qualified participants are selected and are oriented to the task; they are grounded in the content standards or essential knowledge, skills, and abilities (KSAs) upon which the test was built; they are (usually) required to take the test themselves; and so on.

The methodological description provided in Angoff's footnote defines what would be called an "unmodified" Angoff approach, and the probabilities referred to (i.e., borderline examinee item difficulty estimates) are now commonly referred to as "Angoff ratings." At present, the nearly ubiquitous variations of the basic approach—that is, what are called modified Angoff approaches—are characterized by the requirement that participants providing the probability estimates generate more than one rating for each item. The multiple estimates occur because the process is modified to allow participants to reconsider their probabilities in iterations or "rounds" of ratings. Usually no more than three rounds are included in a modified Angoff application. What makes the iterative process desirable is that, between each round, participants are provided with one or more types of feedback (i.e., normative, reality, or impact information; see Chapter 3). The provision of such feedback often has the effect of "convergence" by participants on a cut score, as they tend to make item probability estimates that are more accurate, the amount of between-participant variation in ratings is reduced, and/or the participants become aware of (and usually respond to) the consequences of their ratings vis-à-vis the projected passing rates for the examination.

A hypothetical illustration of a modified Angoff data set and analysis is provided in Table 6-1. The table illustrates a scaled-down example of the data that might be collected as part of a basic Angoff standard-setting procedure, along with some key summary statistics. The table shows two iterations or "rounds" of the item judgment process in which 10 participants rated 13 multiple-choice items once (in Round 1) and again (in Round 2). The purpose of the iterative rating process is to permit participants to discuss their opinions and view normative data between rounds of ratings, the purpose of which is to provide participants with feedback on their ratings and to reduce group variability in the estimates (i.e., to promote movement toward a consensus standard of performance).

Reality information may also be presented between rounds in the form of actual item difficulty indices (e.g., p values) based on total group performance if operational test data are available. In theory and if practicable, however, it is preferable if the p values are based on a subset of examinees whose performance locates them in a borderline region. Because the exact location of the border will not be known until the process is completed, it is possible to use participants' first round ratings to identify a preliminary cut score. With this cut score in hand, those conducting the standard-setting meeting can recalculate p values for feedback to participants based only on the performance of examinees scoring within, say $\pm$ 1 standard error of measurement of the preliminary cut.

Finally, impact information would ordinarily also be provided. To the extent that operational test data are available for a representative group of examinees, the impact information would consist of percentages of examinees who would be classified as passing or failing, or into one of two or more performance categories such as *Basic, Proficient*, and *Advanced*.

As is often the case when using the Angoff approach, participants were instructed to imagine a group of 100 minimally competent examinees and to estimate the number out of that 100 who would answer a given item correctly. To make the task easier, participants were given a form on which to record their estimates, and they were asked to provide their estimates in multiples of 10 only (though this is not a requirement of the Angoff method). After rating all 13 items in Round 1, participants were given normative information. The normative information consisted of a table showing the distribution of ratings for each item and the means and standard deviations for each participant. Participant ID numbers were used to protect the anonymity of the participants' ratings; however, each participant could see how his or her ratings compared to those provided by other members of the panel.

In the hypothetical data shown in Table 6-1, the ratings generated by each participant are reproduced as two lines of data. The first line represents the participant's judgments about the difficulty of each item for the borderline examinee in the first round of ratings; the second line is the participant's Round 2 ratings. The means and standard deviations for each participant are shown at the end of each participant's line of ratings. For example, in Round 1, Rater 1 produced ratings of 90, 90, and 100 for Items 1, 2, and 3, respectively. In Round 2, Rater 1 generated ratings of 80, 90, and 90 for the first three items. Across all 13 items, Rater 1 produced a mean rating of 90.8 in the first round and 88.5 in the second round. Rater 1's ratings were somewhat more variable in Round 1 (SD = 10.38) than they were in Round 2 (SD = 8.01). Across all raters in Round 2, Item 13 was judged to be the easiest (M = 79.0) and Item 10 the most difficult (M = 63.0).

A review of the column labeled "Means *(SD)*" indicates that, for example, in Round 2, Rater 1 produced the most stringent ratings (M = 88.5) and Rater 2 the most lenient (M = 66.9). In Round 2, Rater 2 produced ratings were the most variable (SD = 13.78), while Rater 4's ratings were the least variable (SD = 5.99). The four values in the bottom right corner of the table are the Round 1 and Round 2 grand means across raters and items and the standard deviations of the ratings for each round. For example, the mean Round 1 rating (calculated by averaging across either raters or items) was 72.6. The standard deviation of the raters' means in Round 1 was 9.11. The reduced variability of the Round 2 ratings (SD = 6.35) suggests that the

Table 6-1 Hypothetical Data for Angoff Standard-Setting Method

Rater ID Number	*Item Number*													*Means (SD)*
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	
1	90 80	90 90	100 90	100 100	100 90	90 90	90 100	90 90	90 80	60 70	90 90	100 90	90 90	90.8 (10.38) 88.5 (8.01)
2	60 70	80 80	50 60	60 70	70 80	90 90	70 80	60 70	30 40	40 50	40 60	50 60	70 60	59.2 (17.06) 66.9 (13.78)
3	90 90	70 80	80 90	80 70	100 80	60 60	80 70	80 80	80 80	60 60	50 60	90 90	80 70	76.9 (13.77) 75.4 (11.27)
4	70 70	60 70	70 60	80 70	90 80	80 80	80 70	70 70	70 70	60 70	50 70	90 80	90 80	73.9 (12.61) 72.3 (5.99)
5	90 80	60 70	90 90	40 60	80 80	60 60	80 70	70 70	60 70	60 70	90 80	70 70	80 70	71.5 (15.19) 72.3 (8.32)
6	60 70	60 60	80 70	60 70	70 70	70 70	80 70	80 80	60 60	50 50	70 70	80 80	90 90	70.0 (11.55) 70.0 (10.00)
7	90 80	50 60	80 80	60 70	60 60	70 70	70 60	70 80	70 80	60 50	80 80	80 70	70 80	70.0 (10.80) 70.8 (10.38)
8	80 70	50 50	70 80	80 70	40 50	90 90	70 70	70 80	60 70	60 70	70 70	70 80	80 80	68.5 (13.45) 71.5 (11.44)
9	80 90	70 70	60 70	70 70	60 60	80 80	50 60	60 70	60 70	30 60	50 60	60 70	90 80	63.1 (15.48) 70.0 (9.13)
10	80 80	90 70	90 90	40 60	100 100	80 80	100 90	70 80	80 70	90 80	100 80	70 80	80 90	82.3 (16.41) 80.8 (10.38)
Means	79.0 78.0	68.0 70.0	77.0 78.0	67.0 71.0	77.0 75.0	77.0 77.0	77.0 74.0	72.0 77.0	66.0 69.0	57.0 63.0	69.0 72.0	76.0 77.0	82.0 79.0	72.6 (9.11) 73.9 (6.35)

provision of normative information between Rounds 1 and 2 had the desired effect of tempering outlier ratings and of helping participants converge toward greater consensus in their individual item judgments.

Derivation of a recommended passing score using the Angoff method is accomplished by averaging either the rater or item means. Usually the calculations are based on the final round of ratings. Thus, in this example, the use of the ratings generated in Round 2 would yield a recommended passing score of 73.8% correct, or approximately 9.6 of the 13 items on the test. (As we have indicated previously, the decision about how to handle noninteger results is a policy decision, and the result of 9.6 would be rounded up or truncated according to such a policy.) The same data and an Excel spreadsheet for calculating the cut score based on the hypothetical data shown in Table 6-1 can be found at www.sagepub.com/cizek/angoff.

Procedures for Angoff Variations

One advantage of the Angoff method is its flexibility. It can be used in a variety of applications other than tests consisting of exclusively multiple-choice or other **selected-response** formats. Although the method was originally introduced as a procedure for rating dichotomously scored multiple-choice items, it has easily been adapted to **constructed-response** format items such as writing samples or other polytomously scored performance tasks where examinees' responses are evaluated according to a scoring rubric.

The Extended Angoff Method

One adaptation of the Angoff method for constructed-response items is called the Extended Angoff method (Hambleton & Plake, 1995). Instead of providing conventional probability estimates of borderline examinee performance for each multiple-choice item in a test, participants estimate the number of scale points that they believe borderline examinees will obtain on each constructed-response item. These procedures can be combined as well when a test comprises a mix of item formats, with participants providing conventional probability estimates for dichotomously scored items and scale point estimates for polytomously scored items. As will be illustrated shortly, cut scores for the Extended Angoff approach are calculated in the same way as with the original and modified Angoff methods.

Table 6-2 presents hypothetical data for the ratings of 6 participants in two rounds of ratings using the Extended Angoff method. Participants

rated 8 items that were scored on a 0–4 rubric with the total possible points ranging from 0 to 32. The upper and lower entries in each cell of the table represent participants' first and second round ratings, respectively. For example, in Round 1, Rater 1 judged that the minimally competent examinee would obtain 2 out of 4 points on the first item; in Round 2, Rater 1 revised his rating to 3 out of 4 points on Item 1. The means and standard deviations for each rater and item are also presented for each round. Based on her Round 2 judgments, Rater 5's mean rating across all 8 items was 2.88—making her the most lenient of the participants. Rater 5's ratings in Round 2 were also the least variable (*SD* = .354). Rater 6's mean Round 2 rating across all 8 items was 2.25—the most stringent. The rater exhibiting the greatest variability in Round 2 was Rater 6 (*SD* = .707). Item 5 was viewed across participants as being the most challenging, with a mean rating in Round 2 of 1.83. Items 4, 7, and 8 were judged to be the easiest, with each obtaining a mean rating in Round 2 of 3.00.

The standard deviations shown in the lower right corner of Table 6-2 represent the variability in raters' Round 1 and Round 2 mean ratings. As would be expected, the mean ratings were less variable in Round 2 (*SD* = .221) than they were in Round 1 (*SD* = .358). As indicated by the grand means in the lower right corner of Table 6-2, participants' mean ratings increased somewhat between Round 1 and Round 2. At the end of Round 2, participants judged that the minimally competent examinee should earn, on average, 2.69 out of 4.00 possible points per item for an average proportion correct score of approximately .67. The final recommended Extended Angoff passing score for the eight-item constructed-response test could be derived in one of two ways that produce the same result. Multiplying the Round 2 mean rating across items or participants by the number of items in the test ($2.69 \times 8 = 21.52$) yields, with allowance for rounding of the mean rating to two decimal places, the same result as multiplying the average proportion by the total number of points possible ($.67 \times 32 = 21.44$). Thus the recommended cut score, in raw score units, would be approximately 21.5 out of 32, which represents the number of points participants judged that a minimally competent examinee should earn in order to pass the examination. The same data and an Excel spreadsheet for calculating the cut score based on the hypothetical data shown in Table 6-2 can be found at www.sagepub.com/cizek/extendedangoff.

The Yes/No Method

Another one of the many variations of the Angoff method has been suggested by Impara and Plake (1997). The method they introduced, called the

Table 6-2 Hypothetical Data and Example of Extended Angoff Standard-Setting Method

	Rater ID Number						
Item	*1*	*2*	*3*	*4*	*5*	*6*	*Means (SD)*
1	2 3	3 3	2 3	2 3	3 3	1 2	2.17 (.753) 2.83 (4.08)
2	1 2	2 2	1 2	2 2	2 3	1 2	1.50 (.548) 2.17 (.408)
3	2 3	2 3	2 3	2 3	2 3	2 2	2.00 (.000) 2.83 (.408)
4	3 3	3 3	2 3	2 3	3 3	2 3	2.50 (.548) 3.00 (.000)
5	1 2	1 2	2 2	1 2	2 2	1 1	1.33 (.516) 1.83 (.408)
6	2 3	3 3	3 3	2 3	3 3	2 2	2.50 (.548) 2.83 (.408)
7	3 3	2 3	2 3	2 3	3 3	2 3	2.33 (.516) 3.00 (.000)
8	2 3	3 3	3 3	2 3	3 3	2 3	2.50 (.548) 3.00 (.000)
Means (SD)	2.00 (.756) 2.75 (.518)	2.38 (.744) 2.75 (.463)	2.13 (.641) 2.75 (.463)	1.88 (.354) 2.75 (.463)	2.63 (.518) 2.88 (.354)	1.63 (.518) 2.25 (.707)	2.10 (.358) 2.69 (.221)

Yes/No method is nearly identical to the method originally introduced by Angoff in 1971 that called for participants to make zero and one judgments about items. Downing et al. (2003) describe a procedure, citing the previous work of both Impara and Plake (1997) and Subhiyah, Featherman, and Hawley (2002), called the "Direct Borderline" method, which is also essentially the same as the original Angoff approach.

In response to suggestions that the modified Angoff method may be difficult for some participants to apply because of its demands that they make difficulty estimates for a hypothetical group, the Yes/No and Direct Borderline methods were proposed as means of diminishing one challenge inherent in the rating task—namely, they do not require participants to estimate probabilities. (For simplicity hereafter, we will refer to both the Yes/No and the Direct Borderline methods as the Yes/No approach.)

Indeed, one of the most appealing features of the Yes/No method is its reduced demands on participants. In typical implementations of modified Angoff procedures, participants must maintain a concept of a group of hypothetical examinees and must estimate the proportion of that group that will answer each item correctly. Clearly, this is an important—though perhaps difficult—task. The Yes/No method simplifies the judgment task by reducing the probability estimation task to a dichotomous outcome. In a study designed to assess this effect, Impara and Plake (1998) found that the Yes/No method ameliorated some of the difficulty of the probability estimation task. They reported:

> We believe that the yes/no method shows substantial promise. Not only do panelists find this method clearer and easier to use than the more traditional Angoff probability estimation procedures, its results show less sensitivity to performance data and lower within-panelist variability. Further, panelists report that the conceptualization of a typical borderline examinee is easier for them than the task of imagining a group of hypothetical target candidates. Therefore, the performance standard derived from the yes/no method may be more valid than that derived from the traditional Angoff method. (p. 336)

The specific directions to participants about the rating task are straightforward. Whereas Angoff's original proposal was for standard-setting participants to simply judge whether or not a hypothetical minimally competent examinee would answer an item correctly or not, the question addressed by standard-setting participants can also be answered "Yes" or "No" for each item. In the method, as described by Subhiyah et al. for use with classroom examinations, the directions are similar: "[Participants] first considered the borderline student and then reviewed every question on the examination, rating each item as 'yes, the borderline student will pass this item' or 'no, the borderline student will not pass this item'" (2002, p. S85).

In the Yes/No method described by Impara and Plake (1997), participants are directed to

> read each item [in the test] and make a judgment about whether the borderline student you have in mind will be able to answer each question correctly. If you think so, then under Rating 1 on the sheet you have in front of you, write in a Y. If you think the student will not be able to answer correctly, then write in an N. (pp. 364–365)

Full implementation of the Yes/No method comprises the same features as most common standard-setting approaches. After training and discussion of the characteristics of the minimally competent examinee, participants rate a

set of operational items (usually an intact test form) to complete a first round of judgments. Following this, participants would ordinarily be provided with feedback on their Round 1 ratings—typically normative information and/or reality information along with group discussions—then they would generate a second round of yes/no judgments for each item. If not provided previously, at the end of Round 2 participants would receive additional feedback, which would ordinarily include impact information (i.e., the percentages of examinees would be predicted to pass/fail based on their judgments). Regardless of how many rounds of ratings occur, calculation of the final recommended passing score would be based on data obtained in the final round.

Table 6-3 presents hypothetical data for the ratings of 6 participants in two rounds of ratings using the Yes/No method. Participants rated 12 items by responding "Yes" or "No" to the question about minimally competent examinee performance described previously; participants' yes and no responses were coded as 1s and 0s, respectively, as shown in the data set presented in the table. The upper and lower entries in each cell of the table again represent participants' first and second round ratings. For example, in Round 1, Rater 1 judged that the minimally competent examinee would answer Item 1 correctly; in Round 2, Rater 1 maintained that judgment. In only a very few cases did participants change their judgments. For example, in Round 1, Rater 2 indicated that a minimally competent examinee would not need to answer Item 1 correctly; in Round 2 she indicated that a correct response to that item should be required of such an examinee.

Table 6-3 also shows the means and standard deviations for each rater and item for each round. Based on the Round 2 judgments, Raters 5, 6, and 11 were the most lenient, based on their mean ratings across all 12 items of 0.00. Raters 4, 7, 8, and 9 viewed minimally competent performance quite differently; based on their mean ratings across all 12 items of 1.00, they judged that borderline examinees should answer all 12 items correctly. The Round 2 ratings of Rater 1 were least variable (SD = .142); the Round 2 ratings of Rater 3 showed the greatest variability (SD = .151).

In Round 2, Item 3 was viewed across participants as being the most challenging, with a mean rating across participants of 0.50; Item 1 was viewed as the easiest, with a mean rating in Round 2 of 0.67. With regard to variability, Item 1 showed the greatest variability in ratings in Round 2 (SD = .224); Items 4–9 and 11 showed the least variability (SD = .000), as they were each rated as all "yes" or all "no" by participants.

The standard deviations shown in the lower right corner of Table 6-3 represent the variability in raters' Round 1 and Round 2 mean ratings. The mean ratings for the two rounds of judgments were similar in variability,

with the Round 1 ratings (SD = .041) slightly less variable than the Round 2 ratings (SD = .054). As indicated by the grand means in the lower right corner of the Table 6-3, participants' mean ratings increased slightly between Round 1 and Round 2. At the end of Round 2, participants judged that the minimally competent examinee should be required to attain a proportion correct of .58. On this 12-item test, the mean recommended passing raw score would be approximately 7 out of 12 ($.58 \times 12 = 6.96$). The same data and an Excel spreadsheet for calculating the cut score based on the hypothetical data shown in Table 6-3 can be found at www.sagepub.com/cizek/yesno.

Alternative Procedures and Limitations

One advantage of the Modified Angoff, Extended Angoff, and Yes/No methods is that they can be applied to tests comprising mixed formats. For example, referring to Tables 6-2 and 6-3, we might consider the tests illustrated not as separate tests but as two components of a larger test. In this case, the larger test would consist of 20 items: 12 multiple-choice items and 8 constructed-response items. Hypothetically, the cut score on the first component (i.e., the constructed-response items) could have been set using an Extended Angoff approach, while the cut score on the second component (i.e., the multiple-choice items) could have been set using a Yes/No procedure. The results of the two procedures could then be combined to arrive at a cut score on the total test. Assuming that a compensatory model was deemed appropriate, the cut score on the full 20-item test would be approximately 28 of the 44 total possible raw score points [$(2.69 \times 8) + (.58 \times 12) = 28.48$].

Another alternative application of these methods involves the variety of ways in which results from the two components can be combined to arrive at a total test cut score. Although we will not describe the possible variations here, there are more complicated ways of combining the conventional Modified Angoff or Yes/No probabilities for dichotomously scored items with the scale point estimates for the constructed-response items obtained using an Extended Angoff approach. Such variations would be used, for example, when it is judged desirable to weight the contribution of the two components differentially. Rather than describe possible variations here, we refer the interested reader to Hambleton (1998) for additional information.

Finally, the Modified Angoff method has been used to set more than one cut score. For example, in the case where three performance categories (i.e., two cut scores) are required, the Modified Angoff approach can be modified further to require participants to generate two sets of ratings in each round.

Table 6-3 Hypothetical Data and Example of Yes/No Standard-Setting Method

	Rater ID Number						
Item	*1*	*2*	*3*	*4*	*5*	*6*	*Means (SD)*
1	1 1	0 1	0 0	1 0	0 0	1 1	0.50 (.224) 0.50 (.224)
2	0 0	0 0	0 0	0 1	0 0	0 0	0.00 (.000) 0.17 (.168)
3	1 1	1 1	0 0	1 1	1 1	1 1	0.83 (.168) 0.83 (.168)
4	1 1	1 1	1 1	1 1	1 1	1 1	1.00 (.000) 1.00 (.000)
5	0 0	0 0	0 0	0 0	0 0	0 0	0.00 (.000) 0.00 (.000)
6	0 0	0 0	0 0	0 0	0 0	0 0	0.00 (.000) 0.00 (.000)
7	1 1	1 1	1 1	1 1	1 1	1 1	1.00 (.000) 1.00 (.000)
8	1 1	1 1	1 1	1 1	1 1	1 1	1.00 (.000) 1.00 (.000)
9	1 1	1 1	1 1	1 1	1 1	1 1	1.00 (.000) 1.00 (.000)
10	1 1	1 1	1 1	0 0	1 1	1 1	0.83 (.168) 0.83 (.168)
11	0 0	0 0	0 0	0 0	0 0	0 0	0.00 (.000) 0.00 (.000)
12	0 1	0 0	1 1	0 1	0 1	0 0	0.17 (.168) 0.67 (.210)
Means *(SD)*	.58 (.149) .67 (.142)	.50 (.151) .58 (.149)	.50 (.151) .50 (.151)	.50 (.151) .58 (.149)	.50 (.151) .58 (.149)	.58 (.149) .58 (.149)	.53 (.041) .58 (.054)

Supposing that cut scores were needed to identify the boundaries between *Basic/Proficient* and *Proficient/Advanced,* training would be implemented to help participants form two key conceptualizations involving borderline performance for the two boundaries. Participants would then generate two sets of ratings based on these conceptualizations, although in most cases they would not generate two ratings for each item in a single round. Instead, participants would rate all items with one boundary (i.e., one hypothetical examinee) in mind, then they would rate all items again with the second boundary in mind. Ordinarily, if one boundary category was more consequential than the other (e.g., if performance below the *Proficient* level resulted in a student being retained in grade), then the consequential cut score would be the focus of the first round of ratings. At least theoretically, this would be done so that the hypothetical examinee at this boundary would be the only key conceptualization participants would be exposed to at the time they generated these item ratings, reducing the potential for contamination when the second key conceptualization was introduced.

Although the Modified Angoff method has received much attention by researchers, it should be noted that many Angoff variations have not received as much attention. In one study, the Yes/No method (called the Direct Borderline method in the study) was compared to other common standard-setting methods (e.g., Nedelsky, Ebel, and Hofstee) and found to perform reasonably well in a classroom testing situation (Downing et al., 2003). Overall, however, comparative studies involving the Yes/No and Extended Angoff approaches are comparatively fewer.

Another limitation of some Angoff methods is that they have not been commonly used in situations involving polytomously scored items. Although it would be possible to extend methods such as the Yes/No approach to polytomously scored items by having participants respond "Yes" or "No" to whether a borderline examinee would achieve each possible score point, the method to date has only been applied in contexts where the outcome is dichotomous (e.g., with multiple-choice items).

A potential weakness of the Yes/No method lies in the potential for either positive or negative bias in item ratings, depending on the clustering of item difficulty values in the test. The potential for bias arises because the method is based on an implicit judgment of whether the probability of correct response at the cut score is greater than .5. To illustrate, suppose that a test were composed of identical items that all had a probability of correct response at the cut score of .7. An accurate rater would assign ratings of "1" to each item, and the resulting performance standard would be a perfect score—clearly not the intent of the rater or a realistic expectation based on the difficulty of the test.

Finally, although the Angoff family of procedures remains the most widely used and researched of any of the standard-setting approaches, the method has been criticized. For example, a report of the National Academy of Education studied implementation of a modified Angoff approach used to set standards for the National Assessment of Educational Progress (NAEP). The report provided some evidence related to the inability of standard-setting participants to form and maintain the kinds of conceptualizations required to implement item-based procedures, suggesting that abstractions such as minimally competent or borderline candidates may be impossible for participants to acquire and to adhere to once acquired. The report also criticized the Angoff method for not allowing participants to adequately form integrated conceptions of proficiency. The report concluded that the Angoff procedure was "fundamentally flawed" and recommended that "the use of the Angoff method or any other item-judgment method to set achievement levels be discontinued" (Shepard et al., 1993, p. xxiv). To date, these hypotheses have not received much empirical attention, and the critical view of the Angoff method has been refuted by a number of leading psychometricians with expertise in standard setting (see, e.g., Hambleton, Brennan, Brown, Dodd, Forsyth, Mehrens, et al., 2000). Overall, it is likely that item judgment methods—and in particular the Angoff methods—will continue to see widespread use in the future.

7

The Direct Consensus Method

One of the known disadvantages of item-based methods for setting cut scores (i.e., Angoff and Nedelsky) is the amount of time required for participants to review each item, particularly when more than one round of review and ratings is incorporated in the process. As described in the preceding chapter, a criticism of the Modified Angoff method has been that the combination of participants' needing to form and maintain a conceptualization of a minimally competent examinee and use that conceptualization to generate probability estimates imposes a significant cognitive burden on them. Although the validity of that concern is questionable, it has seemed sensible to reduce the time required to complete an item-based standard-setting procedure and to lighten the cognitive burden on participants.

The Direct Consensus method was recently introduced by Sireci, Hambleton, and Pitoniak (2004) to address these concerns. Among the reasons cited by the authors, they sought a method that could be completed quickly to reduce the amount of time that would be demanded of expert participants—often three or more days. The lengthy time required to participate in a standard-setting procedure can limit the pool of potential participants and may be viewed as unreasonable by some participants given that they are often uncompensated for their participation.

Sireci et al. also specifically sought to "improve upon some of the perceived shortcomings of the Angoff method and to give panelists more direct control in recommending where the passing score is set" (2004, p. 21). As indicated previously, the perceived shortcoming in the Angoff method

involves the complicated nature of the rating task. As we illustrate shortly, the Direct Consensus method greatly reduces that burden. To some extent, the Direct Consensus method can be seen as a special case of the Yes/No method described in Chapter 6.

Perhaps the most important and salient feature of the Direct Consensus method is the incorporation of strategies for permitting participants to directly express their opinions—in the number-correct metric—about their preferred location of the cut score. According to the developers of the method,

> The direct control comes from seeing the passing score being recommended by the panel fairly early in the process. When changes are proposed to the passing score, they are direct, and so their impact on a panelist's passing score and the recommended panelists' passing score are seen immediately. With the Angoff method, if a panelist wants to revise his/her passing score, it must be done through item level ratings, and panelists are never certain until the item rating process is completed just exactly how they have changed their passing score. (Sireci et al., 2004, p. 21)

Procedures for the Direct Consensus Method

Like other standard-setting methods, it is assumed that qualified participants are selected and that they are introduced to the purpose of the testing program and are familiar with the content standards or objectives covered by the test. As in the Angoff method, participants in the Direct Consensus method must form and become familiar with a key conceptualization before generating any ratings. In the Direct Consensus method as introduced by Sireci et al. (2004), the key conceptualization involves the hypothetical "just-qualified candidate" (p. 22).

Unlike the Angoff method, however, participants do not proceed item-by-item through a test form. Rather, an intact test form is reorganized into sections, and entire sections are reviewed and judged. It is assumed that the test on which a standard will be set comprises several subtests or content subareas, which contain whatever numbers of items are indicated in the test specifications for the total test. Sections are formed based on the subareas that would ordinarily represent more homogenous content than the total test.

Participants begin with the first subarea, and they are instructed to indicate those items that they believe the just-qualified candidate will answer correctly. According to Sireci et al. (2004), participants are then directed to sum the number of items they have indicated, producing a number correct

passing score for that subarea. Once all participants have completed this procedure on the remaining sections, the first phase of ratings is complete. At this point, each participant has generated an expected number correct score for each subarea. The sum of each participant's subarea scores is taken as that participant's recommended passing score for the total test. These data are summarized for participants and provided at the end of the first phase of the procedure as normative information.

Table 7-1, adapted from Sireci et al. (2004), illustrates hypothetical data and results of implementing the Direct Consensus method. The method was applied to a 46-item test comprising five subareas of varying lengths and labeled A through E. Eight participants marked the number of items they judged that the just-qualified examinee should answer correctly in order to pass. For example, the table shows that Subtest Area A consisted of 9 items; Rater 1 marked 7 items in Subtest Area A that the borderline examinee should answer correctly. Reading across the row of Subtest A, it appears that the participants were fairly uniform in their expectations, with 6 of

Table 7-1 Hypothetical Data and Example of Direct Consensus Standard-Setting Method

	Rater ID Number									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*		
Subtest Area (Number of Items)	*Participants' Recommended Number of Items Within a Subtest Area That the Just-Qualified Examinee Should Answer Correctly*								*Subtest Area Means* (SD)	*Percentage of Number of Items in Subarea*
A (9)	7	7	6	7	7	7	7	6	6.75 (0.46)	75.0
B (8)	6	6	7	6	5	7	7	7	6.38 (0.74)	79.8
C (8)	5	6	7	6	6	6	5	6	5.88 (0.64)	73.5
D (10)	7	8	8	7	7	8	7	7	7.38 (0.52)	73.8
E (11)	6	5	7	7	7	7	8	6	6.63 (0.92)	60.2
									Grand Mean *(SD)*	Percentage
Rater sums	31	32	35	33	32	35	34	32	33.00 (1.51)	71.7

the 8 participants agreeing that the just-qualified examinee should answer 7 items in that area correctly and the other two participants indicating that 6 items correct was acceptable.

Overall, the mean number correct across raters for Subtest Area A was 6.75, as shown in the column labeled "Subtest Area Means *(SD)*." This column contains the mean number correct across raters for each subtest area as well as the standard deviations of the ratings within the areas. The standard deviations indicate that raters' judgments were least variable with regard to Subtest Area A ($SD = 0.46$) and most variable for Subtest Area E ($SD = 0.92$). The far right column shows the mean number correct from the previous column as a percentage of the total number of items in a subtest area.

The bottom row of Table 7-1 shows the total number correct recommended by each participant. The totals were obtained by summing the subtest area judgments for each participant. These sums show that Rater 1's recommended cut score was the lowest (31 correct out of 46), while Raters 3 and 6 would set the highest performance standard (35 correct out of 46). However, it should be noted that although their total score recommendations are the same, it is unlikely that the same 35 items were identified by Raters 3 and 6 as required for the just-qualified examinee to pass.

A recommended standard is obtained by taking the mean of the individual cut scores in the bottom row of the table. In this case, the recommended cut score would be 33 out of 46. The value in parentheses following the grand mean is the standard deviation of the participants' recommended cut scores (i.e., the standard deviation of the raters' sums). In this case, the value reveals that the participants were fairly homogenous in their performance standard recommendations ($SD = 1.51$). The same data and an Excel spreadsheet for calculating the cut score based on the hypothetical data shown in Table 7-1 can be found at www.sagepub.com/cizek/directconsensus.

The results in Table 7-1 would be shown to participants as normative information following the first round of judgments. In many standard-setting methods, participants would ordinarily be assigned an identification number during the provision of normative feedback. However, in the Direct Consensus method, the process often includes open discussion in which participants are asked to support, defend, or elaborate on the rationales for their judgments. Thus, in an actual implementation of the Direct Consensus approach, the ID number shown in the table could be replaced with the actual names of the participants. As with other methods, the normative information then serves as the basis for group discussions where participants are encouraged to discuss their ratings and the reasons for their ratings. The discussions are facilitated sequentially by subtest area so that the

discussions remain focused on the content of that area and on the characteristics of the just-qualified examinee. A second round of ratings may be incorporated during which participants can make use of the normative information and discussion to revise their subtest area ratings. One feature of the Direct Consensus method is that, during this round, participants can clearly see how changes in their subtest area judgments directly affect their individual total test cut score recommendations.

The third and final phase of the Direct Consensus method involves encouraging participants to come to consensus regarding a single recommendation for the cut score. Discussion begins with consideration of the group mean recommendation—in this case, a raw score of 33, which represents 71.7% correct. As with discussions at the previous step, participants are encouraged to indicate their rationales for support of the group mean as a potential final cut score, and participants who believe that the final recommendation should be a higher or lower cut score are encouraged to give reasons for those options.

Ideally, of course, the final discussion phase would foster participants' agreement on a consensus cut score, although some alternative procedure will be required in the event that consensus is not reached. One suggestion is that the group mean of the final individual recommended cut scores be used.

In their description of the Direct Consensus method, Sireci et al. (2004) describe the results of two implementations of the method and the results of two studies comparing Direct Consensus with other standard-setting procedures. The contexts involved a professional licensure examination and a certification test in an information technology field, and the comparison methods were an Angoff method and an approach the authors called the "item cluster method." As would be expected, in both implementations the Direct Consensus method proved to be more time efficient than the comparison method, requiring 20% and 67% less time, respectively.

In the comparison study involving the Direct Consensus and Angoff procedures, the total group of participants was split into two subgroups not only to allow for comparison of the two different standard-setting methods but also to observe the stability of results within the procedures. The cut scores recommended by the Direct Consensus panels were similar to each other (34 and 35 out of 50) and similar to the cut scores recommended by the panels using the Angoff method (34 and 31 out of 50). Interestingly, in only one of the panels using the Direct Consensus method was consensus actually reached by the group; in the other panel, discussion failed to promote consensus, and the group mean (34) was used as the final recommended cut score.

Results for the comparison of the item cluster and Direct Consensus methods were less similar. In that study, different test forms were used for the comparison, and a 7% difference in recommended cut scores was observed. The authors do not report which method yielded a higher cut score, and the use of different forms is a confounding factor in interpreting the results.

Alternative Procedures and Limitations

Because the Direct Consensus method is a very recent introduction, few variations have been suggested and few potential limitations of the method have been explored. One alternative suggested by Sireci et al. (2004) is that reality information be provided to participants at the end of one of the rounds of ratings. According to the authors, "It would be possible to show the average item performance of examinees in each content area" (p. 22). This suggestion to provide item difficulty data (i.e., *p* values) from operational use of the items in the test form on which performance standards were being set seems reasonable. However, implementation of the suggestion would also increase training time. In addition, the difficulty data may be challenging for participants to use appropriately if it is based on the performance of a total group of examinees as opposed to only on the performance of a borderline or just-qualified group.

As noted in a previous portion of this chapter, another variation of the Direct Consensus method involves the second round of item judgments within subtest areas. While the method is already considerably more time efficient than many other standard-setting approaches, it can be conducted by proceeding from the first round of ratings directly to discussion of the final cut score.

One recommendation to improve the process may seem trivial, though it still warrants mentioning. In the implementations of the Direct Consensus method described by Sireci et al. (2004), participants were asked to sum up their individual subtest area counts and submit a total score that represented the participant's recommended cut score for the total test. In many cases, there are not likely to be errors introduced at this point. However, it would seem prudent to introduce quality control checks in two ways. Before the participants' data are provided as normative information to the group, a check should be made to ensure that the correct counts within subtest areas are reported, and the participants' sums should be checked for accuracy. In fact, in our experience it is not unusual for subject matter experts to make errors when recording page numbers from Bookmark standard-setting

procedures or transferring a raw cut score or theta value from a table to a recording form.

Several limitations of the Direct Consensus method should be also considered. First, we note that the method has only been tried with tests that can be divided into distinct content-based subareas. Indeed, the method was developed with such a context in mind. It is clear that, for tests that would be considered unidimensional or homogenous as regards content, the Direct Consensus method is a special case of the Yes/No method described in Chapter 6.

Second, it is perhaps overly optimistic to believe that the method will always—or even usually—produce consensus. The failure of a group of panelists to agree at the end of the session on a single cut score may be quite probable, particularly when the group of participants is larger or more diverse, or when the test contains a larger number of items than in the situations studied by Sireci et al. (2004). The extent to which facilitators should press participants to reach consensus—which skilled facilitators may be able to accomplish quite well—is another of the policy issues that should be decided upon in advance of the standard-setting implementation by the entity responsible for the testing program.

Third, one of the advantages of the Direct Consensus method—namely that it allows participants to exert more direct influence on the final cut score recommendation—may also be somewhat of a cause for concern. Although feedback provided to participants during the process in the form of normative information is presented without participant identification, in the discussion, phase anonymity will almost certainly be reduced as participants provide their rationales for higher or lower values for the recommended cut scores. Consequently, facilitators for the Direct Consensus method will need to be particularly attuned to group interaction dynamics and work diligently to ensure that the views of one or more vocal, influential, or persuasive participants do not dominate the discussion, stifle discussion, or force the appearance of consensus when a true consensus does not exist. In addition to skilled facilitation, one way to assess the potential for this would be to collect and analyze information from an in-process evaluation in which participants anonymously report the extent to which they are comfortable with the procedure, they have had an opportunity to contribute effectively to discussions, they believe the process fairly incorporates their views, and they have confidence in the recommended cut scores.

Finally, in published information on the method to date, it has not been reported that impact information has been incorporated into the process. Following what is fairly common practice with other methods, it would make sense to provide participants with some idea of the consequences of

their final recommended cut scores in the form of impact information. The logical point for introducing this information would be prior to the final discussion phase so that participants would have the benefit of knowing the consequences of their final cut score recommendations when they are considering raising or lowering their cut scores or supporting the cut score represented by the group mean.

8

The Contrasting Groups and Borderline Group Methods

The majority of standard-setting methods in use today require participants, in one way or another, to consider each item or task in a test, or the collection of items and tasks, and to make judgments about how a hypothetical examinee (such as the minimally qualified candidate or a student just at the *Basic/Proficient* borderline) will perform on those items and tasks. Jaeger (1989) characterized these methods as "test centered," and that characterization is accurate to some extent; Jaeger also defined a second class of methods, which he termed "examinee centered."

In this chapter, we focus on two methods that, if Jaeger's classification scheme were used, would be considered to be examinee centered. However, as we have commented previously, any judgmental standard-setting process necessarily requires participants to bring to bear information on both test content and the population on which the performance standards will be set. No method can focus on either test items or examinees to the exclusion of the other; instead, methods differ primarily in the degree to which one or the other focus is explicitly emphasized.

Thus, although it is perhaps more accurate to describe the methods presented in this chapter as having examinees themselves in the foreground and test content as a background, we will retain the customary label of "examinee centered" for the sake of consistency with common usage in the field. Our faint support for the labeling system notwithstanding, we concur completely

with the rationale that some standard-setting experts have articulated to support their preference for examinee-centered methods. Namely, it has been asserted—and to some extent demonstrated—that examinee-centered methods represent a task that participants often find to be familiar and comprehensible. That is, participants—who routinely come to the standard-setting task from their roles as teachers, clinical directors, professors, and so on—are usually well versed in the task of judging whether a specific student or examinee possesses an acceptable level of knowledge, skill, or ability vis-à-vis a set of content standards to merit placement into a specific performance category. Thus "examinee-centered" methods are thought to pose a task that may be more competently engaged in by participants than the comparatively unfamiliar "test-centered" tasks.

Two examinee-centered methods are described in this chapter: the Borderline Group method and the Contrasting Groups method. Both methods require participants to make direct judgments about the status of real—that is, not hypothetical—test takers who are known to them. In essence, each method derives a passing score for a test by combining participants' judgments about examinees with information about the examinees' actual performances on the test. The methods differ in the specific judgments that participants are required to make, in the subgroup of examinees that is the focus of the judgments, and in how the judgments are analyzed to derive cut scores.

Procedures for the Contrasting Groups Method

The Contrasting Groups method was first described by Berk, who referred to it as "an extension of the familiar known-groups validation procedure" (1976, p. 4). The historical context of Berk's observation was a time when the pedagogical technique of mastery learning (Bloom, 1974) was also being advocated and researched. The known-groups procedure was used to obtain a cut score differentiating between instructed and uninstructed students, or students who had attained mastery of some content and those who had not. Implementation of the known-groups validation procedure gathered data from two groups of examinees—one group that was known to have received effective instruction covering the content covered by a test and another group that had not been so instructed. Both groups were administered a test over the content of instruction. Distributions of the total test scores for the two groups could be plotted and examined to find a point on the score scale that maximized the differentiation between those examinees who had received effective instruction (called "true masters") and those who had not (called "true nonmasters"). Although the popularity of the mastery learning

instructional method has faded, the Contrasting, or "known," Groups procedure has remained useful, and the mastery/nonmastery labels have been retained.

As commonly configured, the Contrasting Groups method requires examinees to take a test before any cut score(s) are yet known so that those test scores can be used in the data analysis that will produce a cut score. Participants are impaneled who have personal knowledge of individual, real examinees' levels of knowledge or skill level with respect to the characteristic being assessed. The participants, who are unaware of examinees' actual test scores, make judgments about each examinee as to their mastery/nonmastery status; that is, they judge whether the examinee overall should be categorized as passing or failing. Participants' category judgments about the individuals are used to form distributions of total test scores for each of the two groups. Specifically, one distribution of total scores is formed for the examinees categorized judgmentally as nonmasters and another distribution of total scores is formed for the examinees categorized as masters. The two distributions are then plotted and analyzed to arrive at a cut score that distinguishes group membership.

Figure 8-1 provides a hypothetical illustration of the two distributions. In the figure, the horizontal axis provides information about total test scores, with the raw score scale ranging from 0 to 80. The vertical axis provides information about the proportion of each group obtaining each of the raw score levels. (The axis could also have been configured to represent frequency.) The figure shows that the two distributions are of different size; namely, participants judged that more examinees were truly masters or were considered passing than were judged to be nonmasters or failing.

Of course, as with nearly every aspect of setting performance standards, the reality of practice is never nearly as tidy as the theory. Typical test score distributions are almost never as neat as the distributions shown in Figure 8-1. At least theoretically, one explanation for the jaggedness of the distributions may be sampling error. Depending on a number of factors, the plotted distributions may be rather "toothy" and not conforming to easy analysis. Among these factors are the sample sizes of the group judged to be masters and the group judged to be nonmasters—with smaller samples often yielding less smooth distributions—and the total number of total score points—with less smooth distributions resulting when the number of possible score points is limited and when some score points along the scale are not observed or not attainable. Accordingly, a common recommendation is that a smoothing procedure be implemented prior to locating a cut score (Livingston & Zieky, 1982). Many types of smoothing techniques are possible; however, a complete discussion of smoothing is beyond the scope of this book, and specialized

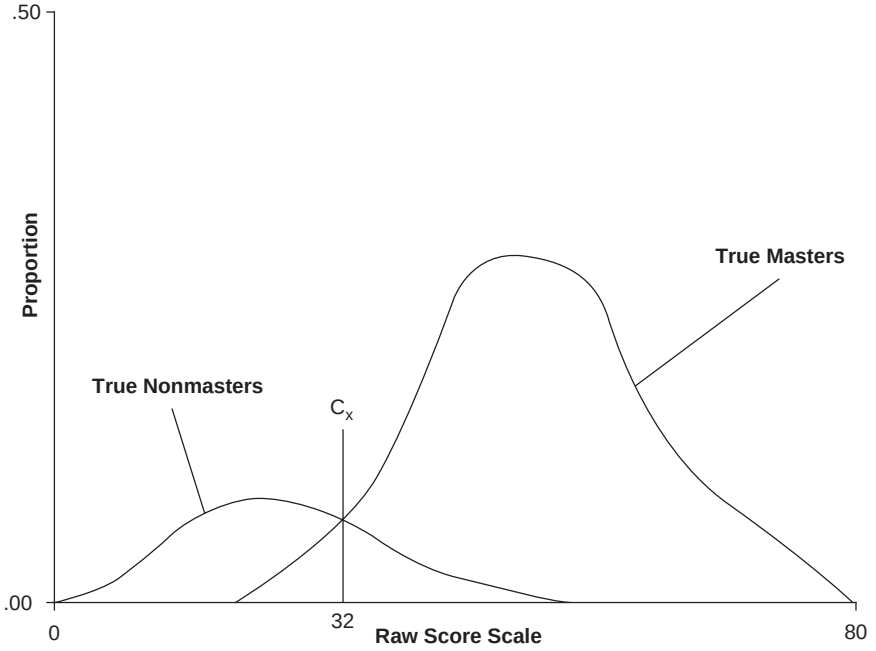


Figure 8-1 Hypothetical Distributions Illustrating Contrasting Groups Method

software is often necessary to accomplish smoothing. Users with access to nearly universally available software programs such as Excel can implement rudimentary smoothing using a command found by choosing the "Tools" drop-down menu in that program, then clicking on "Data Analysis" and choosing "Analysis Tools." It should be noted, however, that this option is not automatically installed with Excel, but must be added as a user option later from the Excel installation disk.

Whether smoothed—as are the distributions shown in Figure 8-1—or unsmoothed, one method for deriving a cut score is to select the point of intersection of the two distributions. For the smoothed distributions shown in Figure 8-1, the point of intersection is indicated as C_x and corresponds to a raw score of 32.

An Example Using the Contrasting Groups Method

A more typical result than the smooth distributions of Figure 8-1 is shown in Figure 8-2. The hypothetical distributions of masters and nonmasters shown

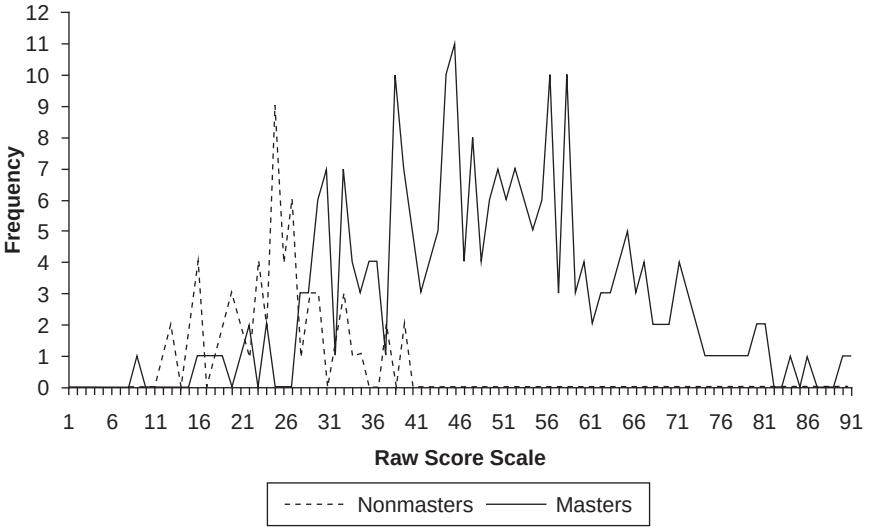


Figure 8-2 Typical "Toothy" Distributions of Nonmaster and Master Raw Score Data

in Figure 8-2 were generated using a normal distribution generator to obtain each distribution. The distribution of masters shows a plot of 250 examinees' scores with a mean of 50 and standard deviation of 15; the distribution of nonmasters is a plot of 60 examinees' scores with a mean of 25 and a standard deviation of 7. The data used to generate the figure are available at www.sagepub.com/cizek/contrastingdata.

As is evident in Figure 8-2, finding the intersection of the two distributions is not a task that can be accomplished easily via visual inspection. Instead, the cut score must be obtained analytically using fairly commonly available statistical software.

One basic method involves the use of measures of central tendency and simply finding the midpoint between the centers of the two distributions. For the distributions shown in Figure 8-2, the two medians are 24 and 52 (for the nonmaster and master distributions, respectively). The midpoint between these two medians—which would be used as the cut score—is 38. Alternatively, the midpoint between the two means (i.e., the mean of the two group means) could be used; in this case, that value would be 37.5.

Another solution involves the use of logistic regression, and this is perhaps the most common method of determining a point of intersection between the distributions formed when a Contrasting Groups method is

used. Logistic regression is used to determine the raw score point at which the probability of category membership is .50. (We recall, however, that the choice of .50 is one that sets false positive and false negative classification errors as equally serious, which again highlights that the choice of this value is a policy, rather than a technical decision.)

A data set configured for logistic regression analysis using the data analysis program SPSS is found at www.sagepub.com/cizek/contrastingdata. This data set is the same as found in www.sagepub.com/cizek/contrastingdata, which was used to produce Figure 8-2. However, the data layout has been reconfigured so that it can be easily analyzed, with the primary change being that a record is created for each examinee, and the participants' judgments for each examinee as a master or nonmaster have been coded as a 0 (nonmaster) or 1 (master). In addition, each examinee's actual test score (which was not known to participants during the judgment process) has been added to his or her record. The complete data file consists of 310 records (one for each of the 250 examinees judged to be masters and the 60 examinees judged to be nonmasters) and two variables. In the SPSS data file, the variables are named RawScore and MNStatus (for Master/Nonmaster status).

To conduct the logistic regression analysis in SPSS, "Regression" is chosen from the "Analyze" drop-down menu. From the options in the "Regression" menu, "Binary Logistic" is selected, because the outcome variable (in this case, MNStatus) is a dichotomous variable coded 0/1. In the dialog box listing the available variables, highlight MNStatus and click the arrow to place it in the box labeled "Dependent Variable." From the remaining available variables, highlight RawScore and click the arrow to place it in the box labeled "Covariates." Clicking the "OK" button produces the analysis. (The same analysis can be conducted using other commonly available software programs; detailed procedures for conducting logistic regression using Excel are described in Chapter 9.)

Selected SPSS output resulting from applying this set of procedures is shown in Figure 8-3. The first panel in the figure shows that model coefficients were successfully estimated. The second panel provides all that is necessary to compute the Contrasting Groups cut score for these data. The second panel shows the logistic regression model coefficients. For now, we proceed directly to describe how the cut score is obtained from these results; for the reader interested in interpreting all of the values shown in Figure 8-3, additional information is provided in Chapter 9 in the section "Interpreting the Logistic Regression Output."

To begin the process of obtaining a Contrasting Groups cut score for these data, we begin by referring to the simple one independent variable

Model Summary

Step	–2 Log Likelihood	Cox & Snell *R*-Square	Nagelkerke *R*-Square
1	152.696 a	.387	.619

[a]Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 a	Raw Score	.198	.027	53.598	1	.000	1.219
	Constant	–5.288	.818	41.810	1	.000	.005

[a]Variable(s) entered on step 1: RawScore

Figure 8-3 SPSS Logistic Regression Results for Contrasting Groups Data

regression model used here. That general logistic regression equation may be written as

$$y^* = a + b\ (x) \qquad \text{(Equation 8-1)}$$

where y* is the predicted value on the outcome variable (MNStatus) for an examinee, *a* is a constant, *b* is the slope of the regression function, and x is an examinee's observed score (i.e., RawScore).

Substituting into Equation 8-1 the values of *a* and *b* obtained from the analysis results shown in Figure 8-3 yields the following:

$$y^* = -5.288 + .198\ (x) \qquad \text{(Equation 8-2)}$$

In typical regression contexts, a data analyst would be interested in knowing the predicted score, y*, associated with a given value of x. In that case, the analyst would substitute a value of x into Equation 8-2 and solve for y*. However, in this case, we are interested in obtaining the value of x (i.e., the raw score) associated with an outcome that is exactly midway between the two possible classifications (i.e., nonmaster and master). Because these categories were coded 0 and 1, respectively, .50 is used as the value

that best differentiates between master and nonmaster status. In effect, this procedure yields a raw score at which examinees classified as nonmasters first reach a 50% chance of being classified as masters.

Thus the equation used to solve for the Contrasting Groups cut score becomes

$$.50 = -5.288 + .198\,(x) \qquad \text{(Equation 8-3)}$$

and solving Equation 8-3 for x yields 29.23, or a raw cut score of approximately 29. As the reader will notice, this value is appreciably less than the value obtained previously by using the midpoint between these two medians (38) or the midpoint of the two group means (37.5). Such a result—that is, one in which different cut score results are obtained via different analytical methods—is not particularly unusual, especially given the small sample sizes and "toothy" nature of the Contrasting Groups data described previously and illustrated in Figure 8-2. In addition, the use of logistic regression with small samples often yields large standard errors for model parameters, only modest model fit, and small values for *R*-squared. For these reasons, it may be preferable in many situations to use one of the midpoint methods for obtaining a Contrasting Groups cut score.

The Borderline Group Method

Although the Contrasting Groups method is often commended as presenting a more natural judgmental task to participants, to some extent that method can also be viewed as making a difficult demand on participants. That is, participants must classify persons as masters or nonmasters, passing or failing, or whatever categories are used. In reality, not all persons can be readily classified into one of those two groups; there are surely persons for whom a dichotomous judgment is difficult for participants. For such persons, it may not be easy for the participant to classify them as clearly passing or failing; the reality in the mind of the rater may be that the person is not squarely in one category or another, but on the border between the two categories, possessing not quite enough knowledge or skill (in the participant's judgment) to be classified as Passing or Master, while at the same time not being so deficient in knowledge and skill in the participant's mind as to be classified as Nonmaster or Failing.

To address this issue, an alternative to the Contrasting Groups method, called the Borderline Group (Zieky & Livingston, 1977) method, can be used. Rather than force participants into a dichotomization that they may

not feel adequately captures their judgment about an examinee, the Borderline Group method allows participants to classify persons into more than two categories, one of which is a "borderline" category. Typically, three categories are used, and participants evaluate each person as passing, borderline, or failing.

As in the Contrasting Groups method, participants selected to engage in a Borderline Group procedure are familiar with the content tested and have specific knowledge of the knowledge, skills, and abilities of individuals who are subject to the examination. To implement the Borderline Group method, participants first engage in an extended discussion to develop a description of an examinee on the borderline between mastery and nonmastery. With this conceptualization in mind, and without knowledge of examinees' test performance, participants using the Borderline Group method identify specific examinees whom they believe to lie at the borderline separating acceptable and unacceptable competence. Alternatively, participants may be asked to classify examinees into three categories: clearly competent, clearly inadequately prepared, and on the cusp between competent and not competent.

Then, using the actual test scores obtained by the examinees, a distribution of the test performance of the borderline group is formed and the median of the distribution is often used as the recommended standard. Figure 8-4 illustrates a cut score obtained using the Borderline Group method. In this case, the median (42) would be identified as the cut score.

Alternative Procedures and Limitations

The Borderline Group method is similar to the Contrasting Groups method in that participants are required to make judgments about persons that classify them into two or more categories. One way of thinking about the difference between the two methods is that the Borderline Group method requires slightly finer judgments. That is, contrary to the appearance of a single judgment implied by Figure 8-4, participants must essentially make one of three judgments as they classify an examinee as Failing, Borderline, or Passing. In fact, the Body of Work method described in Chapter 9 can be seen as a specific case of the Borderline Group approach.

An extension of the borderline group concept has been described by Plake and Hambleton (2001), which suggests an alternate procedure that has some record of successful implementation. In their work, each of three categories (*Basic, Proficient,* and *Advanced*) was subdivided into three groups (high, medium, and low) yielding categories such as low-*Basic,*

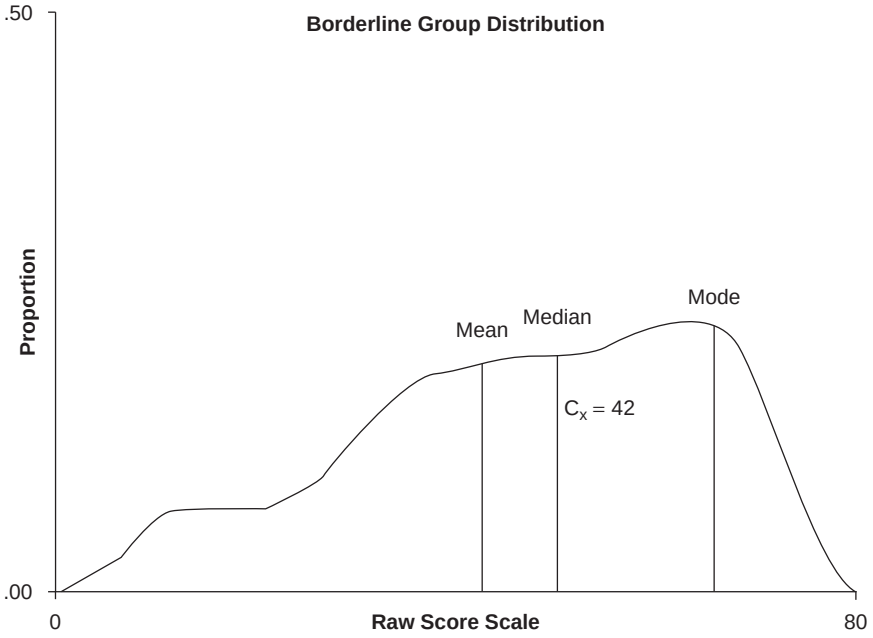


Figure 8-4 Hypothetical Distributions Illustrating Borderline Group Method

mid-*Basic*, high-*Basic*, and so on. The high-*Basic* group would be very similar in knowledge and skill to the low-Proficient group, while the high-*Proficient* group would be very similar to the low-*Advanced* group. A cut score based on this classification strategy can then be obtained either in a Contrasting Groups sense (by taking the midpoint between the two border categories or via logistic regression) or using a Borderline Group approach whereby a single borderline group is formed by combining two adjacent groups (e.g., the high-*Basic* and low-*Proficient* groups), and a measure of central tendency from this group is used as the cut score to differentiate *Basic* and *Proficient* performance levels. The mean would be most appropriate for determining this level if the number of persons in the two border categories was approximately equal.

Both the Contrasting Groups and Borderline Group methods can at least claim an advantage in the intuitive nature of the judgmental task presented to participants. Conversely, a major limitation of the Borderline Group method lies in the fact that the size of the group judged to be on the borderline may be quite small, making estimates of the cut score unstable. In fact, this same limitation is found in the Contrasting Groups method, although it is perhaps

not as obvious, because the problem of small borderline group size is not evident until the data analytic method selected—particularly logistic regression—is used to identify a cut score. A technique selected to smooth the master and nonmaster distributions may also introduce a source of error. And, because all smoothing techniques are more effective with larger samples, the problem of small samples in the Contrasting Groups or Borderline Group methods extends to that data handling technique as well.

Despite the apparent intuitive nature of these methods, they rely heavily on the assumption that a large and representative group of participants—with sufficient knowledge about a sufficiently larger group of individual examinees—can be impaneled. Further, the Borderline Group method relies heavily on meeting facilitators and training of participants to ensure that they avoid errors of central tendency and assign disproportionately large numbers of persons to the borderline group. Although such a rating tendency would have the beneficial effect of increasing the size of the borderline group and perhaps yielding a more stable estimate of C_x, such a tendency would also have the potential to bias the estimate in unknown ways.

Another limitation related to participants in a Contrasting Groups or Borderline Group approach is bias and inconsistency in judgments. Both methods essentially treat participants' judgments as a "gold standard" or criterion variable with inherent reliability and validity. Obviously, if judgments of mastery/nonmastery, passing/failing, and so on could be made consistently and without error, there would be no need for a test or other psychometric device to serve as a proxy for such judgments, except for the benefit that might be accrued in terms of efficiency. However, it is well known that humans have difficulty focusing their judgments exclusively on the construct of interest and that judgments about the same characteristic vary across occasions.

One potential threat to the validity of a Contrasting Groups or Borderline Group procedure is that participants may not base their judgments solely on an examinee's knowledge or skill. Rather, because the participants are—indeed, *must be*—familiar with the examinees they judge, their judgments may be affected by other, construct-irrelevant factors such as the examinee's personality, punctuality, helpfulness, work habits, leadership, sex, ethnicity, or any one of a large number of other variables. Clearly, as with all other methods discussed in this book, the training of participants to make the judgments required of them, combined with effective facilitation to monitor the extent to which the training has been effective, are critical to the success and defensibility of a standard-setting procedure.

Finally, it is important to note two cautions with respect to the Contrasting Groups and Borderline Group methods. First, as we have

stressed repeatedly, the strategy for identifying the cut score illustrated in Figure 8-1 makes an implicit assumption that the consequences of misclassifications are equally serious. In this case, a false negative misclassification would occur when an examinee judged to truly have mastery of the content (a "true master," TM) obtains a test score that falls below C_x. A false positive misclassification occurs when an examinee judged to have an unacceptable level mastery of the content (a "true nonmaster," TNM) obtains a test score above C_x. The cut score identified in Figure 8-1, because it was set at the point of intersection of the two distributions, treats the two misclassifications as of equal relative cost. In all cases, however, an explicit policy should be established in advance with respect to the relative costs of each type of classification error, and the actual location of C_x should be derived with respect to that policy decision.

Second, once participants have completed the judgmental tasks of the Contrasting Groups or Borderline Group methods, the actual cut scores must be obtained with knowledge of the test performance of the examinees who were the subject of those judgments. There are two issues to be considered here. The first is the issue of requiring examinees to take a test for which the performance standard to which they will be held is unknown and to be determined. The second issue is the logical difficulty of failing an examinee based on his or her test score, when expert judgments about the examinee have been recorded attesting to the examinee's mastery status. Conversely, some examinees will, based on their examination scores, pass the test for whom experts have rendered the opinion that they are not competent. As is evident in Figure 8-1, it is likely that these false negative and false positive classifications will occur in relatively few cases, although depending on the size, variability, and proximity of the master and nonmaster distributions, and the location of the final cut score, the number of such cases could be substantial.

9

The Body of Work and Other Holistic Methods

Holistic standard-setting methods comprise a family of procedures whose chief feature is the evaluation of whole sets of examinee work by one or more judges who render a single (holistic) verdict about each work sample. These judgments take the form of a rating, which may be dichotomous (such as Pass/Fail) or ordinal (such as *Below Basic, Basic, Proficient,* and *Advanced)*. These ratings, along with scores for each examinee, are then combined in one of several ways to yield one or more cut scores.

All of the following methods would be characterized as holistic methods: the Borderline Group and Contrasting Groups methods (see Chapter 8), Judgmental Policy Capturing method (Jaeger, 1995), the Dominant Profile method (Plake, Hambleton, & Jaeger, 1997), the Analytical Judgment method (Plake & Hambleton, 2001), and the Body of Work method (Kingston, Kahl, Sweeney, & Bay, 2001). Three of these methods (Judgmental Policy Capturing, Dominant Profile, and Analytical Judgment) will, because they are used relatively infrequently, be described briefly in the following paragraphs. The remainder of this chapter will then focus in detail on what is perhaps the most widely used of the holistic methods—the Body of Work approach.

The Judgmental Policy Capturing Method

The Judgmental Policy Capturing method (JPC; Jaeger, 1995) stems from research in management and other areas in which experts evaluate actual

work samples or multidimensional work profiles. In its basic form, the JPC method uses the overall ratings of the expert judges and multiple regression techniques to derive a set of weights to apply to the various dimensions of the work. When originally proposed by Jaeger, the method was applied in the context of a credentialing examination for the National Board for Professional Teaching Standards (NBPTS). In that study, expert judges of teacher performance rated the hypothetical work sample scores of candidates for certification. Looking at all possible combinations of scores on the various tasks that were judged to be important for success in the teaching profession, the participants offered an overall evaluation of the candidate's fitness to teach in a specified field. Jaeger then used the ratings to derive a multiple regression equation for each judge and a system of weights to apply to the dimensions to derive a predicted outcome for each candidate.

A simple illustration of the JPC method follows. Consider, for example, a set of work samples, each rated on five dimensions, and with each dimension rated on a 5-point scale. A panel of experts, working independently, assigns each sample to one of four categories. The four possible category assignments are treated as scores, ranging from 1 to 4. These category scores serve as the criterion variable, while the scores on the dimensions of the work sample serve as the predictors. Table 9-1 presents the hypothetical data consisting of ratings on five dimensions and an overall judgment generated by one expert rater for five work samples.

Over a large number of such data collections (200 were used in Jaeger's study), it is possible to produce a stable multiple regression equation in which the predicted score on the criterion (1 to 4) is expressed as a weighted linear sum of the component scores from the work samples. A separate regression equation is computed for each expert rater. The several equations are then merged into one, giving greater weight to the more stable raters' regressions; different approaches to obtaining the single equation are possible depending on whether the overall system is to be compensatory or conjunctive.

In summary, the distinguishing aspect of the JPC method is the use of regression techniques to discover each participant's implicit "policy" regarding proficiency or competence and the merging of those individual policies into a single, coherent policy. In effect, the application of the regression procedure allows the participants to examine score information and sort candidates into categories without explicitly articulating their rationale for doing so. Perhaps because of this comparatively lesser amount of explicit consideration of the meaning of competence or borderline performance and participant interaction, the JPC method has been used primarily in the NBPTS context and does not appear to have been implemented outside the certification field.

Table 9-1 Hypothetical Data for Judgmental Policy Capturing Method

Rater ID	*Work Sample ID*	*Dimension 1 Rating*	*Dimension 2 Rating*	*Dimension 3 Rating*	*Dimension 4 Rating*	*Dimension 5 Rating*	*Overall Classification*
1	001	1	1	2	1	1	1
	002	2	2	2	2	3	2
	003	3	4	4	4	3	3
	004	3	4	4	4	4	4
	005	5	5	5	4	5	4

The Dominant Profile Method

The Dominant Profile (DP) method introduced by Plake et al. (1997) shares the same basic goal as the JPC method but takes a more direct route. Rather than asking participants to evaluate work samples and then analyze the data from those judgments to derive implicit performance standard "policies," facilitators using the DP method ask participants to state their decision rules directly. For example, a participant reviewing a test comprising 5 performance tasks might judge that, in order to be considered Proficient or Passing an examinee "must get at least a 3 on Tasks 1 and 3, and not score lower than 2 on any of the other three tasks." It is this feature of direct judgments that is a distinguishing characteristic of the DP method. In fact, Plake et al. developed the DP method in response to the concerns of participants in Jaeger's JPC approach that a more direct approach would be preferable.

Also in the context of setting performance standards for the NBPTS certification examinations, Plake et al. (1997) asked a sample of 20 experienced teachers to evaluate work samples for candidates of an Early Adolescence Generalist credential. The content of the test was similar in nature to that of the Jaeger (1995) study, and many of the other conditions were similar. Indeed, the participants were first trained on the JPC method and then turned their attention to the DP approach. Participants examined profiles of candidate scores on several subtests of the Early Adolescence Generalist exam and placed them in one of four categories, as they had done with the JPC method. The primary distinction between the two methods was that, with Dominant Profile, the judges first classified the candidates (as with the JPC method) and then wrote down their reasons in policy form. They then worked in small groups to discuss their written policies and attempted to reach consensus on a single policy definition.

In the end, panelists in the Plake et al. (1997) study produced a series of policy statements that essentially boiled down to three, each of which contained both conjunctive and compensatory elements. For example, all three sets stated a minimum total score of 18 (out of 24) points, which reflects a compensatory model. However, all three policy statements also included a separate minimum score requirement on Task 2, reflecting a conjunctive approach (which would have disqualified several candidates who had a total of 18 points over all the tasks). Two of the policy statements also mentioned other tasks for which a low score could disqualify a candidate, regardless of total score on the other tasks.

Contrary to the hoped-for outcome, the three sets of policy statements never coalesced into a single definitive policy. A follow-up questionnaire to the participants provided some closure, at least in terms of level of endorsement

and confidence. Responding to the questionnaire, the participants strongly endorsed one of the three policy statements and provided only weak support for the other two. Plake et al. (1997) concluded that, if the procedure were to be replicated, they would recommend conducting this final round face-to-face, as opposed to through the mail. Moreover, they noted that there were some issues that might never be resolved—both in the study at hand and in future implementations of the method. They speculated that perhaps a simple vote would be the best approach, with the policy statement garnering the most votes being the one that would apply.

The Analytical Judgment Method

The Analytical Judgment Method (AJM) has its roots in the same soil as the Judgmental Policy Capturing and Dominant Profile methods. It is possible to think of the AJM approach as a multiple borderline group procedure. As in the Borderline Group method, the data of primary interest in the AJM method are the scores of examinees who are in the upper end of a lower (e.g., "failing") category and those in the lower end of an upper (e.g., "passing") category. The AJM procedure generalizes the Borderline Group approach for the $n > 2$ categories in situations where more than one cut score is required on a single test. For example, if a situation in which four performance levels existed, each level would be divided into low, medium, and high subcategories, resulting in 12 separate and distinguishable categorizations.

The AJM procedure was initially developed by Plake and Hambleton (2001) for tests that include polytomously scored performance tasks and other formats, resulting in a total test comprising different components. In such situations, a holistic method is useful for arriving at a single categorical decision for the composite measure. In Plake and Hambleton's work, four performance levels were used (Novice, Apprentice, Proficient, and Advanced), which were further subdivided to form 12 categories: low-Novice, mid-Novice, high-Novice, low-Apprentice, mid-Apprentice, high-Apprentice, low-Proficient, mid-Proficient, high-Proficient, low-Advanced, mid-Advanced, and high-Advanced.

To implement the method, participants reviewed a carefully selected set of materials for each component, representing the range of actual examinee performance on each of the questions comprising the assessment (although examinees' scores were not revealed to the panelists). Participants then classified the work samples according to 1 of the 12 performance categories. Boundary group distributions were then formed by combining the judgments for work samples in the boundary categories. For example, one boundary

group formed based on work samples rated as high-Novice or low-Apprentice; a second boundary group formed from the high-Apprentice/low-Proficient work samples, and a third boundary group formed based on the high-Proficient/low-Advanced samples.

Cut scores for the AJM method were calculated by simply computing the average (mean) of total scores for all work samples contained in a given boundary group. Thus, for example, the cut score for Apprentice was the average score for the high-Novice/low-Apprentice group, and so on. Although Plake and Hambleton (2001) suggested alternative methods for calculating the eventual cut scores, a simple averaging approach appeared to work as well as the others.

Finally, with other applications of the AJM procedure, it is likely that the system of categories could be simplified. For example, the same 12-category scheme described previously could be streamlined into a 7-category system. Using the streamlined approach, the categories low-Novice and mid-Novice could be eliminated, and the lowest category (capturing the work samples of those two groups) would be called Below Basic. Similarly, the categories of mid-Advanced and high-Advanced could be combined to form a single category labeled Advanced. Falling between the new end-point categories of Below Basic and Advanced would be the categories Borderline Basic, Basic, Borderline Proficient, Proficient, and Borderline Advanced. Computation of performance standards would be streamlined as well, with cut scores obtained directly as the means of the Borderline Basic, Borderline Proficient, and Borderline Advanced groups. In a study of a simplified AJM approach, Plake and Hambleton (2001) found little difference in the locations of cut scores when comparing the 12-group configuration with the 7-group configuration.

Summary Analysis of Three Holistic Methods

The three methods described in the preceding sections of this chapter (i.e., the Judgmental Policy Capturing, Dominant Profile, and Analytical Judgment methods) represent a special class of holistic methods. In particular, the JPC and DP methods focus not just on the establishment of cut scores but on the articulation of a set of rules for entry into a profession. Looking far beyond the standard-setting task at hand, the participants and entity responsible for setting the performance standards were also communicating to the population of applicants what was important for success in the specialty area covered by the examination.

As they begin the credentialing process, candidates are informed regarding how many points each performance task is worth and how those task

scores are combined to form a total score. The portfolios or work samples that candidates create for these credentialing assessments consume hundreds of hours of candidates' time. Knowing that certain of the tasks or portfolio elements are more important (to standard setters) than others, candidates can more intelligently allocate their finite time and other resources to building the portfolio. The rules for setting the cut scores are also important to incumbents in the field and to prospective employers of newly credentialed teachers. More than simply dividing candidates into successful and unsuccessful groups, the procedures define the entry points into the field in very concrete ways.

The AJM procedure yields cut scores but not policy statements. Its link to the other two methods is primarily historical or developmental. In the 1990s, investigators were looking for alternatives to the item-based Angoff and Nedelsky procedures, which, with their emphasis on **selected-response** format items, were not considered well suited for a new generation of assessments composed of performance-based tasks or varying proportions of items and tasks of differing formats.

The three methods described here are primarily attributable to the work of Richard Jaeger, Barbara Plake, and Ronald Hambleton who collaborated on several investigations. The fact that these methods called on independent subject matter experts to rate examinee performance sets them apart from previously existing holistic methods (Borderline Group and Contrasting Groups) that typically called for holistic evaluations of student performance by their own instructors. The rigor introduced by this group of investigators, as well as the credibility bestowed upon them by their sponsoring agencies, immediately placed these three methods in the foreground of standard-setting procedures. Anyone considering setting standards for performance-based or complex (i.e., mixed format) assessments—whether in credentialing, certification, or educational assessment contexts—should give these methods due consideration.

Overview of the Body of Work Method

The Body of Work (BoW) method (Kingston, Kahl, Sweeney, & Bay, 2001) is another member of the family of holistic standard-setting methods that includes all procedures whose chief feature is the evaluation of whole sets of examinee work by one or more judges who render a single (holistic) verdict about each work sample. In this section, we focus on the BoW method because of its increasing use as a standard-setting method in statewide high-stakes assessments and because of some of its features that may not be as well understood as the features of other holistic methods.

Three unique features distinguish BoW from the other holistic methods. First, the work samples are ordered by total score so that panelists, without knowing the exact score for any work sample, know which work samples received high scores and which received low scores and that scores generally increase (or decrease, depending on how the work samples are arranged) from the beginning of the set to the end. Having placed a work sample from the low end of the set in the Basic category, for example, a panelist would have to think twice before placing a later work sample in the Below Basic category. This feature is both efficient and effective. It permits panelists to move through the set of work samples quickly and confidently, and it provides a convenient backdrop for interround discussions of ratings.

The second unique feature of BoW is the systematic removal of work samples between rounds. Sets of samples that are clearly in a given performance level (i.e., for which there is little or no disagreement) can be taken out of the set and replaced by additional samples with scores near the score points at which there is greater disagreement. Kingston et al. (2001) refer to this process as **rangefinding** (Round 1) and **pinpointing** (Round 2). We address rangefinding and pinpointing in greater detail later in this chapter.

A final unique feature is closely tied to the second. Kingston et al. (2001) describe a data analytic technique known as **logistic regression**. This technique is a perfect adjunct to the systematic removal of work samples (and whole ranges of test scores) from the set. With the systematic removal of scores, methods that rely on calculation of means or medians (e.g., contrasting or borderline groups) will systematically over- or underestimate those measures and thus miscalculate midpoints. Logistic regression eliminates that potential problem. We will return to this issue later in the chapter as well.

Procedures for the Body of Work Method

The BoW method was introduced by Kingston et al. (2001) as an alternative for setting performance standards on complex assessments, such as those consisting exclusively or predominantly of constructed-response (CR) items (e.g., essays), performances, work samples, or a combination of item formats that would not be amenable to standard setting using typical item-based methods. The method has found wide application in setting performance standards for writing assessments, portfolios, and alternate assessments (see Chapter 16). In the following sections of this chapter, the key features of the method are summarized. These include training procedures that are unique to the method, the processes of rangefinding and pinpointing, and how to calculate cut scores using the method.

Training

As in any standard-setting activity, the facilitators provide an orientation to the task and a practice session to familiarize panelists with the method. This training includes opportunities to become familiar with the test and the performance level descriptions (PLDs). Well in advance of the meeting, the facilitators and sponsors select work samples to present during training and during the rounds of rating. Kingston et al. (2001) note that the selection process systematically eliminates work samples with unusual or conflicting score patterns. For example, they eliminate samples with high scores on some CR items and very low scores on others. The resulting set of samples will therefore present fewer challenges to panelists than, for example, a randomly selected set.

In a typical BoW session, panelists examine a small set of work samples during the practice round. Kingston et al. (2001) noted that their panelists examined six samples and assigned each to one of four performance levels (ranging from Failing to Advanced). Panelists then compare their ratings with those of other panelists and discuss similarities and differences in ratings. Regardless of the number or labels assigned to the performance levels, the task given to BoW participants is the same: They must become familiar with established performance category descriptions (i.e., PLDs), then carefully review work samples that are provided to them and assign each sample to one of the performance categories.

Rangefinding

The first round of BoW is referred to as **rangefinding.** During this round, panelists examine sets of work samples arranged from low to high raw score (or high to low). Typically, samples will be assembled in folders so that the samples are arranged in score point order within a folder and folders are arranged in score point order as well. Table 9-2 shows the results of the rangefinding round of ratings by 12 panelists of 30 student work samples organized into 10 folders. (The complete data file on which Table 9-2 is based is available at www.sagepub.com/cizek/bowdata1.)

As shown in Table 9-2, Folder 1 contains the three samples with the lowest scores (13, 15, and 16). Folder 2 contains the next three samples (scored 17, 19, and 20) and so on, to Folder 10 with samples scored 46, 48, and 50. Each line in the table shows the distribution of panelists' ratings for the work samples in a folder. For example, the first data line in the table shows the ratings of the participants for one of the entries in Folder 1—the particular work sample that had received a raw score of 13. Of the 12 participants, all 12 judged that work sample to be *Below Basic.* Similarly, the third data line in the table shows the ratings of the participants

Table 9-2 Distribution of Ratings for Work Samples in Four Performance Categories: Rangefinding Round

		Frequency of Rating in Each Performance Category				
Folder Number	*Raw Score*	*Below Basic*	*Basic*	*Proficient*	*Advanced*	*Total*
1	13	12	0			12
	15	12	0			12
	16	10	2			12
2	17	9	3			12
	19	8	4			12
	20	6	6			12
3	21	5	7			12
	22	5	7	0		12
	23	4	7	1		12
4	24	1	9	2		12
	25	0	10	2		12
	27		9	3		12
5	28		7	5		12
	29		6	6		12
	30		4	8	0	12
6	31		3	8	1	12
	32		2	9	1	12
	33		1	9	2	12
7	34		0	10	2	12
	35			9	3	12
	36			8	4	12
8	38			7	5	12
	39			6	6	12
	40			4	8	12
9	41			2	10	12
	43			1	11	12
	44			1	11	12
10	46			0	12	12
	48				12	12
	50				12	12
	Totals	72	87	101	100	360

for the third entry in Folder 1—the work sample that had received a raw score of 16. Ten of the participants judged this work sample to be *Below Basic*, and 2 participants judged the sample to reflect *Basic* performance; none of the participants judged this work sample to be *Proficient* or *Advanced*.

To begin the rangefinding process, panelists examine each sample in each folder and enter a rating (performance level) for each sample on a special form. A sample BoW rating form is shown in Figure 9-1 and is available electronically at www.sagepub.com/cizek/bowform. When participants have completed their review of one folder, they move on to the next. During rangefinding, panelists may discuss their ratings with one another and change their ratings if they wish.

In most multiround standard-setting activities, each round has as at least one goal the identification of a preliminary cut score or scores. In BoW, the rangefinding round serves primarily as a means of rough identification of cut scores, with a more refined zeroing in on cut scores taking place in a second round called pinpointing. Between rangefinding and pinpointing, meeting facilitators and data analysts examine participants' Round 1 ratings and determine which work samples to eliminate before beginning Round 2. As is evident in Table 9-2, work samples in Folder 1 produced very little disagreement—virtually everyone judged those samples to be

Language Arts Standard-Setting Rating Form: Round 1

Standard Setter ID Number_____ **Date**_____

Directions: *Enter the code (1, 2, or 3) corresponding to your judgment about the performance level of each work sample. Use the following codes:*

1 = Basic ***2 = Proficient*** ***3 = Advanced***

Work Sample ID	Rating	Work Sample ID	Rating
1	_____	27	_____
2	_____	28	_____
3	_____	29	_____
4	_____	30	_____
5	_____	31	_____
6	_____	32	_____
7	_____	33	_____
8	_____	34	_____

Figure 9-1 Sample Body of Work Rating Sheet

Below Basic. Before proceeding to pinpointing, facilitators would eliminate this folder, along with others that do not contribute to the decision-making process, and retain those that they find useful.

Usefulness in this case is characterized by observed variability with regard to the classification of one or more samples within a folder. Where there is little or no variability in judgments in a range of work samples, it is unlikely that a cut score distinguishing levels of performance is located in that score range. Where there is variability in judgments about work samples, it is possible that the variation indicates a possible cut score region. Thus, for example, if a substantial number of participants identify even one work sample in a folder as *Basic* and a substantial number identify that sample or another sample in the same folder as *Proficient,* it is likely that the cut score separating *Basic* and *Proficient* lies in the range of scores represented in that folder. Applying this concept to the data shown in Table 9-2, it is clear that Folders 1, 9, and 10 would be eliminated as not likely containing the range of raw scores in which a potential cut score would be identified; alternatively, Folders 2 through 8 would be retained and examined further in pinpointing.

It may be useful to explain in detail why Folders 1, 9, and 10 would be eliminated and Folders 2 though 8 retained. Folder 1 contains samples with scores of 13, 15, and 16. All 12 panelists rated the first two samples as *Below Basic,* and 10 out of 12 rated the third sample *Below Basic*. There is very nearly perfect agreement that all the samples in Folder 1 are in the *Below Basic* category; thus the cut score separating *Below Basic* from *Basic* is not very likely to be found in this folder. Folder 2 contains samples with scores of 17, 19, and 20. For these three samples, panelists are split 9/3, 8/4, and 6/6 with respect to their ratings (*Below Basic* vs. *Basic*). It is thus likely that the cut score for *Basic* lies somewhere in this score range. Similarly, Folder 3 has very nearly even splits at score points 21 and 22. Each additional folder, up through Folder 8, contains at least one work sample for which there is considerable disagreement among panelists and therefore the possibility of a cut score. These folders will be retained for pinpointing and likely augmented with additional work samples in the same score range. For example, a sample with a score of 18 might be added to Folder 2, and a sample with a score of 26 might be added to Folder 4, as no samples with these scores were reviewed in the rangefinding round. We just as easily might have examined Table 9-2 and selected points at which the likelihood of being in a given category shifted to favor the next category. Thus, for example, at score point 20, the likelihood of being classified as either *Below Basic* or *Basic* is 50%. At score point 29, the likelihood of being classified as *Basic* or *Proficient* is likewise 50%, and at score point 39, there is an equal likelihood of being classified as *Proficient* or *Advanced*. Choosing

just the folders at which there is a clear shift (i.e., at the 50/50 split) would produce a very limited set of work samples to evaluate in the pinpointing round. Retaining all the rangefinding folders would lead to unnecessary work for panelists. The optimal number of work samples for Round 2 lies somewhere between these two extremes.

The guiding principle in selecting work samples between rangefinding and pinpointing is to ensure that panelists have enough opportunities to pinpoint all the cut scores without overwhelming them with work samples. In a morning or afternoon or full day, panelists can review a finite number of work samples. Making sure that panelists have the most useful samples for focusing their review and judgments in the area of the eventual cut scores is one responsibility of those who conduct a BoW procedure. Although Kingston et al. (2001) have provided a very useful set of criteria for selecting work samples for the pinpointing round, experienced BoW facilitators should use their best judgment and knowledge of the test, the PLDs, and, to the extent possible, of the panelists themselves, to make the work of the panelists in Round 2 as productive and efficient as possible. We will return to the topic of selecting work samples at the end of the chapter because it is critical to all holistic methods, not just BoW.

Pinpointing

During Round 2 of a BoW session—called the **pinpointing** round—participants receive folders of samples with scores in the narrowed range of raw scores resulting from the rangefinding round. The task for participants in this round is the same as that in the first round: to review a set of work samples and assign each to one of the performance levels. Once again, the work samples are laid out in score point order, both within and across folders. Many or all of the Round 1 (i.e., rangefinding) work samples may be included (assuming they have met the Round 2 selection criteria), and they may be augmented by other samples with scores in the same ranges. Panelists review the samples, enter their ratings, and turn in their forms to the facilitators. Table 9-3 shows a sample set of ratings resulting from a pinpointing round. These data are also included in the data set found at www.sagepub.com/cizek/bowdata2.

Calculating Cut Scores

Logistic regression is one analytical procedure that can be used to calculate final cut score recommendations based on the final round of ratings obtained from a BoW standard-setting method (see Kingston et al., 2001, pp. 230–231); these data are commonly from Round 2 (pinpointing data).

Table 9-3 Distribution of Ratings for Work Samples in Four Performance Categories: Pinpointing Round

		Frequency of Rating in Each Performance Category				
Folder Number	*Raw Score*	*Below Basic*	*Basic*	*Proficient*	*Advanced*	*Total*
2	17	10	2			12
	18	9	3			12
	19	7	5			12
	20	6	6			12
3	21	5	7			12
	22	5	7			12
	23	4	8	0		12
4	24	0	10	2		12
	25		10	2		12
	26		9	3		12
	27		9	3		12
5	28		8	4		12
	29		6	6		12
	30		5	7		12
6	31		4	8	0	12
	32		2	8	2	12
	33		1	9	2	12
7	34		0	10	2	12
	35			9	3	12
	36			8	4	12
	37			7	5	12
8	38			7	5	12
	39			6	6	12
	40			4	8	12
	Total	46	102	103	37	288

The primary goal of the calculation is to identify a raw score at which the likelihood of being assigned to a given level equals or exceeds that of being assigned to the next lower level, that is, where $p = .50$.

For example, in Table 9-3, we see that at a raw score of 20 there is a 50% chance of being assigned to the *Basic* level (i.e., 6 out of 12 entries are at this level). At a raw score of 21, the likelihood of being assigned to the *Basic* level is 7 out of 12 (about 58%). On the basis of a simple visual inspection of Table 9-3, we might conclude that a raw score of 20 would be a good cut score to differentiate the *Below Basic* and *Basic* performance levels. Similarly, we can see that at a raw score of 29, a work sample has an equal chance of being assigned to *Basic* (6/12) and *Proficient* (6/12). Thus 29 might seem like a good cut score for differentiating the *Basic* and *Proficient* performance categories, because above that score a student is more likely to be assigned to *Proficient* than to *Basic*. Similarly, at a raw score of 39, a student is as likely to be assigned to *Advanced* as *Proficient,* and above 39, the student is increasingly likely to be assigned to the *Advanced* category; thus 39 might seem to be a good cut score between the *Proficient* and *Advanced* performance levels. These relationships are shown graphically in Figure 9-2.

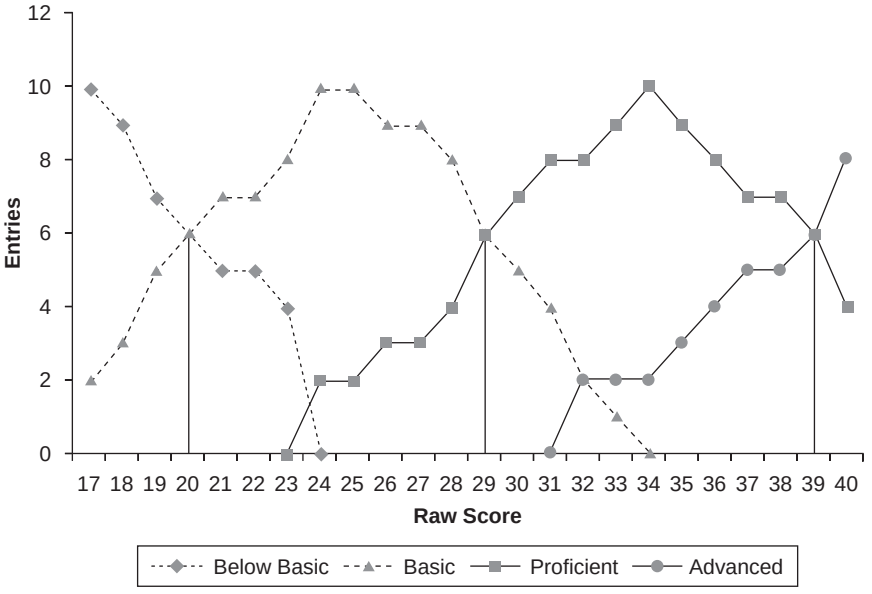


Figure 9-2 Graphical Display of Score Distributions from Table 9-3 Showing Potential Cut Scores at Intersections of Adjacent Distributions

In each instance, we have identified a cut score as the point at which the likelihood of being placed in the next higher category first reaches 50%. The 50% rule usually has the consequence of equalizing false positive and false negative classification errors. By setting the *Proficient* cut score at 29, for example, if we counted all the samples with scores of 29 or above and ratings of *Basic* or below (12 in all) and then all the samples with scores below 29 and ratings of Proficient or above (14 in all), we find a total of 26 errors out of 188 students rated as either *Basic* or *Proficient*. By moving the cut score even one point in either direction, we make a total of 36 errors (at 28) or 32 errors (at 30). Moving farther away from a cut score of 29, we see more false positive classifications or more false negative classifications, depending in which direction the cut score changes.

The preceding discussion is based on observed probabilities for a full range of raw score values. In practice, logistic regression is used to establish a model of the relationship between score and performance level in which some raw score values will be missing. It is the model-based cut scores that are of interest.

In the following paragraphs, we use the data in Table 9-3 (i.e., data from a hypothetical pinpointing round of a BoW standard-setting session) to illustrate the fundamentals of logistic regression applied to the BoW standard-setting context.

In a situation such as the assignment of work samples to a performance level, our task is to calculate or predict the likelihood that a given score will result in a specific level. As we examine Table 9-3 or any other set of standard-setting data, we quickly realize that up to a point, the likelihood of being assigned to category C increases as the raw score increases. Then the likelihood of assignment to that category begins to decrease because the sample is now more likely to be assigned to a higher category. We see, for example, that the probability of being assigned to the Basic category increases steadily from a raw score of 15, where the probability of a performance at that level being classified as *Basic* is zero, to a raw score of 25, where the probability of being classified as *Basic* is greatest. From 25 to 34, the probability diminishes, reaching zero again at a raw score of 34. Thus the relationship between raw score (S) and the probability of being assigned to a given performance category (C) generally takes an inverted U shape.

In the context of the BoW standard-setting method, the application of logistic regression changes the analysis slightly. Rather than determining the likelihood of being assigned to category C, given score S, the task is to find the likelihood of being assigned to category C *or higher,* given raw score S. Thus, in order to consider Table 9-3 in the context of logistic regression, it is necessary to consider each entry in terms of the log of the odds of assignment to category C or higher, as shown in Table 9-4.

In Table 9-4, the raw scores are shown in column 1. The next three columns show the probability of being classified as *Basic* or better in the column labeled $p(B^+)$, *Proficient* or better in the column labeled $p(P^+)$, and *Advanced* in the column labeled $p(A)$. These probabilities are followed by

Table 9-4 Data from Table 9-3 Recast in Terms of Log Odds

RS	$p(B^+)$	$p(P^+)$	$p(A)$	$ln\ \{p(B^+)/[1 - p(B^+)]\}$	$ln\ \{p(P^+)/[1 - p(P^+)]\}$	$ln\ \{p(A)/[1 - p(A)]\}$
17	0.17	0.00	0.00	–1.60944		
18	0.25	0.00	0.00	–1.09861		
19	0.42	0.00	0.00	–0.33647		
20	0.50	0.00	0.00	0		
21	0.58	0.00	0.00	0.336472		
22	0.58	0.00	0.00	0.336472		
23	0.67	0.00	0.00	0.693147		
24	1.00	0.17	0.00		–1.60944	
25	1.00	0.17	0.00		–1.60944	
26	1.00	0.25	0.00		–1.09861	
27	1.00	0.25	0.00		–1.09861	
28	1.00	0.33	0.00		–0.69315	
29	1.00	0.50	0.00		0	
30	1.00	0.58	0.00		0.336472	
31	1.00	0.67	0.00		0.693147	
32	1.00	0.83	0.17		1.609438	–1.60944
33	1.00	0.92	0.17		2.397895	–1.60944
34	1.00	1.00	0.17			–1.60944
35	1.00	1.00	0.25			–1.09861
36	1.00	1.00	0.33			–0.69315
37	1.00	1.00	0.42			–0.33647
38	1.00	1.00	0.42			–0.33647
39	1.00	1.00	0.50			0
40	1.00	1.00	0.67			0.693147

the respective log odds for each of these probabilities, that is, the natural log of $[p/(1 - p)]$. The relationship between predicted raw score ($\hat{y}$) and the log odds (logistic) function can be expressed as follows:

$$\hat{y} = \ln [p/(1 - p)] \qquad \text{(Equation 9-1)}$$

When $p = .50$, then $1 - p$ will also be .50, and $\hat{y}$ will be the natural log of 1 (i.e., .50/.50), which is 0. Thus, to find the point at which the likelihood of being classified in category C or higher is .50, we find the point at which $\hat{y} = 0$. But $\hat{y}$ can also be modeled in terms of a simple regression equation:

$$\hat{y} = a + b\text{x} \qquad \text{(Equation 9-2)}$$

where a and b represent the intercept and slope, respectively, of the regression line, and x represents a given raw score.

Table 9-5 shows the results of three regression analyses of the raw scores and logistic functions shown in Table 9-4. Using the slope and intercept values for *Basic, Proficient,* and *Advanced* in Table 9-5, we can establish $\hat{y}$ values for each as shown in the equations below, where the subscripts B, P, and A represent *Basic, Proficient,* and *Advanced:*

$$\hat{y}_B = -7.70468 + .37325 * RS_B \qquad \text{(Equation 9-3)}$$

$$\hat{y}_P = -12.63950 + .43973 * RS_P \qquad \text{(Equation 9-4)}$$

$$\hat{y}_A = -11.14140 + .28911 * RS_A \qquad \text{(Equation 9-5)}$$

Using these equations and setting each $\hat{y}$ to 0 (i.e., at the score point at which $p = .50$), we can solve for the raw cut score for the *Basic* level (RS_B) corresponding to $\hat{y}_B = 0$ as follows:

$$0 = -7.70468 + .37325 * RS_B \qquad \text{(Equation 9-6)}$$

$$7.70468 = .37325 * RS_B \qquad \text{(Equation 9-7)}$$

$$RS_B = 7.70468/.37325 \qquad \text{(Equation 9-8)}$$

$$RS_B = 20.64 \qquad \text{(Equation 9-9)}$$

And, by a similar approach, we can solve for the other cut scores (i.e., *Proficient* and *Advanced*) to obtain

$$RS_P = 28.74 \text{ and } RS_A = 38.54 \qquad \text{(Equation 9-10)}$$

Table 9-5 Results of Regression Analyses for Data in Table 9-4

	Coefficient	*Standard Error*	*t*	*p*
Basic				
Intercept (a)	–7.70468	0.952125	–8.092090	0.0004670
Slope (b)	0.37325	0.047370	7.879364	0.0005290
Proficient				
Intercept (a)	–12.63950	1.081852	–11.68320	0.0000026
Slope (b)	0.43973	0.037768	11.64278	0.0000027
Advanced				
Intercept (a)	–11.14140	0.980866	–11.35870	0.0000091
Slope (b)	0.28911	0.027176	10.63833	0.0000142

The regression analyses to obtain the cut scores were carried out in Excel using the same spreadsheets in which the original data were tallied and transformed to logistic functions. We suspect that this approach will be convenient for most readers. However, several statistical packages (e.g., SAS, SPSS, STATA) can also be used to perform logistic regression. An illustration using SAS PROC LOGISTIC is shown in Table 9-6, which contains the SAS programming code and output for the same data.

Interpreting the Logistic Regression Output

The SAS software uses a maximum likelihood procedure to carry out the logistic regression, as do SPSS and STATA. Thus some of the output may seem unfamiliar to those accustomed to least squares approaches. In the following paragraphs, we briefly describe the output shown in Table 9-6. For more detailed explanations of the concepts involved, the reader should consult more comprehensive treatments of logistic regression (e.g., Cizek & Fitzgerald, 1999; O'Connell, 2005).

A first step in evaluating the logistic regression results is to examine the model fit statistics. Three measures of fit are displayed for each model: AIC, SC, and –2 Log L. The last measure, –2 Log L (–2 times the log of the likelihood function), is a basic measure of model fit. The first two measures (AIC and SC) are functions of –2 Log L. The Akaike Information Criterion (AIC) and the Schwarz Criterion (SC) are defined as follows: AIC = –2 Log

Table 9-6 SAS Programming Code and Selected Output for Calculating Cut Scores Based on Data in Table 9-3

SAS Programming Code

```
proc logistic data = ssround descending;
model category = rs / rsq;
where category = 1 or category = 2;
output out = logout p = prob1;
title 'PROC LOGISTIC Logistic Regression Category 1 vs. 2';
run;
proc logistic data = ssround descending;
model category = rs / rsq;
where category = 2 or category = 3;
output out = logout p = prob1;
title 'PROC LOGISTIC Logistic Regression Category 2 vs. 3';
run;
proc logistic data = ssround descending;
model category = rs / rsq;
where category = 3 or category = 4;
output out = logout p = prob1;
title 'PROC LOGISTIC Logistic Regression Category 3 vs. 4';
run;
```

SAS Output

PROC LOGISTIC Logistic Regression Category 1 vs. 2
Probability modeled is Category=2.

Model Fit Statistics

Criterion	*Intercept Only*	*Intercept and Covariates*
AIC	185.445	117.668
SC	188.443	123.663
–2 Log L	183.445	113.668

Analysis of Maximum Likelihood Estimates

Parameter	*df*	*Estimate*	*Standard Error*	*Chi-Square*	*Pr > ChiSq*
Intercept	1	–11.3266	2.0169	31.5392	<.0001
	1	0.5613	0.0968	33.6066	<.0001

PROC LOGISTIC Logistic Regression Category 2 vs. 3
Probability modeled is Category=3.

Model Fit Statistics

Criterion	*Intercept Only*	*Intercept and Covariates*
AIC	286.185	139.683
SC	289.508	146.329
–2 Log L	284.185	135.683

Analysis of Maximum Likelihood Estimates

Parameter	*df*	*Estimate*	*Standard Error*	*Chi-Square*	*Pr > ChiSq*
Intercept	1	–15.1934	2.0873	52.9824	<.0001
Slope	1	0.5262	0.0720	53.3302	<.0001

PROC LOGISTIC Logistic Regression Category 3 vs. 4
Probability modeled is Category=4.

Model Fit Statistics

Criterion	*Intercept Only*	*Intercept and Covariates*
AIC	163.698	132.942
SC	166.639	138.825
–2 Log L	161.698	128.942

Analysis of Maximum Likelihood Estimates

Parameter	*df*	*Estimate*	*Standard Error*	*Chi-Square*	*Pr > ChiSq*
Intercept	1	–13.7280	2.8197	23.7029	<.0001
Slope	1	0.3583	0.0773	21.5010	<.0001

L + 2m where m is the total number of independent parameters (in this case 2, the slope and intercept); and SC = –2 Log L + ln(n)m, where m is defined in the same way and n is the number of observations (i.e., ratings entered).

The primary importance of these measures of fit is the comparison of their values for "Intercept Only" with their values for "Intercept and Covariates." This comparison shows the reduction in the magnitude of the fit measure going from the null model (i.e., no predictors) to the model of interest (i.e., with one or more predictors; in this case, raw score). Here we see that each measure shows considerable reduction from the null to the model of interest.

The next sets of values shown are the estimates for intercept and slope, along with their associated standard errors and chi-square estimates. From each of the three portions of the output titled "Analysis of Maximum Likelihood Estimates" we use the values listed in the column of "Estimates" to obtain each of three pairs of "Intercept" and "Slope" values. These values allow us to express $\hat{y}$ as a function of the intercepts and slopes for a given raw score (RS):

$$\hat{y}_B = -11.3266 + .5613 * RS \qquad \text{(Equation 9-11)}$$

$$\hat{y}_P = -15.1934 + .5262 * RS \qquad \text{(Equation 9-12)}$$

$$\hat{y}_A = -13.7280 + .3583 * RS \qquad \text{(Equation 9-13)}$$

Recalling that when p is equal to .50, $\hat{y}$ will be equal to 0, we set $\hat{y}$ equal to 0 and solve these equations for RS (20.07, 28.87, and 38.54 for the *Basic, Proficient,* and *Advanced* cut scores, respectively). These are the raw score points at which the likelihood of being in the next higher category (or higher) reaches 50%. Above that score, the likelihood of being in that category (or higher) exceeds 50%. Assuming equal costs associated with false positive and false negative errors of classification, these inflection points are taken as the cut scores.

The statistical significance of the estimates in the model is tested via a chi-square approach. The chi-square values are calculated by dividing the estimates by their standard errors and then squaring the results. The resulting values have a chi-square distribution with one degree of freedom under the null hypothesis. As we can see by examining Table 9-6, all estimates yield associated probabilities of less than .0001, indicating that the estimates for both intercept and slope are significantly different from 0.

An Application of the Body of Work Method

In the following paragraphs, we illustrate a recent application of the Body of Work method conducted for a statewide high school proficiency examination

in language arts. Three performance levels were required: *Approaching Proficient, Proficient,* and *Advanced.*

In most respects, the procedure was carried out as described in the first part of this chapter. However, as is frequently the case in the application of any method to a specific situation, there were modifications and adaptations to the circumstances of the state's testing program. In this situation, the modifications included the following two substantial modifications:

1. Although the BoW method normally involves presenting work samples in score order (lowest to highest or highest to lowest), officials within the state department of education directed that work samples be presented in random order; and
2. Although the BoW method normally involves removing some Round 1 work samples and augmenting the sample set with work samples at selected score points, officials within the department directed that all work samples evaluated by participants in Round 2 would be new (i.e., not carried over from Round 1).

Selecting Work Samples

The language arts test consisted of 20 multiple-choice (MC) reading items, 4 CR reading items, and 2 writing prompts. Each MC item counted 1 point, and the 4 CR items counted 4 points each. Both writing prompts were scored on a 6-point scale, but for one of the prompts, scores assigned by two independent scorers were summed while, for the other prompt, they were averaged. Thus it was possible to have fractional scores on the second prompt (e.g., 2.5, 3.5, 4.5, etc.). The reading portion of the test yielded a total of 36 points, while the writing portion yielded 18 points (in half-point increments). Total scores on the test therefore ranged from 0 to 54, in half-point increments. Given the possibility of half-point scores, 108 unique score points were possible (0, 1, 1.5, etc. to 54.0).

The facilitators worked directly with department of education staff and contractor scoring staff to identify work samples for both rounds of standard setting. Approximately 75,000 students had taken the test, and scoring staff identified 200 samples from an initial set of approximately 1,000. Each sample consisted of an entire answer folder for one student: that is, all MC responses, all CR responses, and both essays. The goal of the facilitators was to mirror the overall score distribution as well as the variety in approaches to the reading and writing responses and patterns of MC item responses. No student earned a score below 11 or above 53 on the test, and 99.97% earned scores of 50 or below. The initial set of 200 therefore had no sample with a score below 11 or above 50, and the final set of 33 tests had scores ranging

from 12 to 50. Every whole score point on each CR reading item and essay was represented in this set; MC total scores ranged from 6 to 20, while CR/essay total scores ranged from 6 to 30. The complete set is shown in Table 9-7. The remaining 167 tests were held in reserve for further selection of Round 2 work samples.

Training Participants

Facilitators spent a full day orienting panelists to the tests, including a full-scale test administration and scoring activity on the first day. The full agenda is shown in Figure 9-3, and an electronic version of the agenda that can be adapted to different contexts is available at www.sagepub.com/cizek/bowagenda.

For this application, qualified participants were recruited, mailed advance information, and assembled for the standard-setting session. Once on site, facilitators provided the participants with an orientation to the goals of the week, the PLDs, scoring guides, and specific training in the Body of Work method, including a practice round. During the practice round, each participant evaluated a common set of three student work samples, using a form designed especially designed for that exercise. Participants entered their ratings (1 for *Partially Proficient,* 2 for *Proficient,* and 3 for *Advanced*). At the end of the practice round, participants shared their ratings while the facilitator entered them on a flip chart at the front of the room. The purpose of the exercise was to familiarize panelists with the logic of rating and the mechanics of completing the forms as well as the fact that there would be differences of opinion within the group. Facilitators explained that the panelists' ratings would be used to derive cut scores mathematically in a way that would reflect the average response across all panelists (i.e., facilitators did *not* attempt to explain the process of logistic regression except to note that the mathematical process that would be used was one that the state had used on the elementary and middle school test standard-setting activities previously).

At the end of the practice round, the facilitators distributed an evaluation form. This form was used throughout the session, before each round of rating, and again at the end of the session to gauge participants' understanding of the process and their readiness to proceed to the next step. (See Chapter 3, Table 3-8 or www.sagepub.com/cizek/evaluationform for a sample evaluation form.)

Rangefinding

To complete the rangefinding step of the BoW method, each participant received a folder with 33 student work samples (arranged in the same way

Table 9-7 Work Samples for Language Arts Test: Rangefinding Round

Order in Folder	*Score on MC Items*	*Score on CR Items*	*Total Score*
1	17	24	41
2	17	20	37
3	20	28	48
4	16	20	36
5	17	15	32
6	13	15	28
7	6	6	12
8	13	18	31
9	19	24	43
10	17	21	38
11	20	27	47
12	10	14	24
13	13	17	30
14	18	21	39
15	8	15	23
16	19	23	42
17	18	22	40
18	14	21	35
19	19	30	49
20	20	25	45
21	20	30	50
22	18	23	46
23	20	24	44
24	13	20	33
25	10	16.5	26.5
26	6	14	20
27	15	19	34
28	8	13	21
29	9	9	18
30	13	12	25
31	15	14	29
32	12	15	27
33	6	10	16

Day 1	
8:00-8:30 A.M.	**Continental breakfast**
8:30-10:00	**1.0 Opening session**—Forms, announcements, and housekeeping 1.1 Objectives 1.2 Processes 1.3 Selection of panelists 1.4 Expectations of panelists 1.5 Cut score defined 1.6 Using cut scores
10:00-10:15	**Break**
10:15-12:00 Noon	**2.0 Review of the assessment materials** 2.1 Assessment content 2.2 Description of assessment development process 2.3 Administration of the assessment to panelists (Part I)
12:00-1:00 P.M.	**Lunch**
1:00-4:30	**Review of the assessment materials (continued)** 2.3 Administration of the assessment to panelists (Part II) 2.4 Scoring the assessment
Day 2	
8:00-8:30 A.M.	**Continental breakfast**
8:30-10:00	**3.0 Understanding the definitions of the standards** 3.1 Definitions of student performance standards 3.2 Interpretation of student performance (Proficient) 3.3 Interpretation of student performance (Advanced) 3.4 Summary of student performance levels 3.5 Midprocess evaluation
10:00-10:15	**Break**
10:15-12:00 Noon	**4.0 Introduction to standard-setting procedures** 4.1 Description of the procedure 4.2 Practice 4.3 Summary of the standard-setting procedure
12:00-1:00 P.M.	**Lunch**
1:00-4:30	**5.0 Round 1: Holistic classification** 5.1 Distribution of rating sheets and instructions 5.2 Classification of papers
Day 3	
8:00-8:30 A.M.	**Continental breakfast**
8:30-10:30	**Holistic classification (continued)** 5.3 Discussion of panelists' ratings 5.4 Review of classifications
10:30-10:45	**Break**
10:45-12:00 Noon	**6.0 Round 2: Holistic classification** 6.1 Distribution of rating sheets and instructions
12:00-1:00 P.M.	**Lunch**

Figure 9-3 Sample Body of Work Standard-Setting Meeting Agenda

1:00-4:30	**Round 2: Holistic classification (continued)** 6.2 Classification of papers
Day 4	
8:00-8:30 A.M.	**Continental breakfast**
8:30-10:30	**Round 2: Holistic classification (continued)** 6.3 Discussion of panelists' ratings
10:30-10:45	**Break**
10:45-12:00 Noon	**7.0 Review of impact data** 7.1 Introduction of impact data 7.2 Final standard determinations 7.3 Return all secure materials 7.4 Turn in expense claim vouchers
12:00-1:00 P.M.	**Lunch**
1:00-3:00	**8.0 Evaluation of the standard-setting process** 8.1 Complete evaluation forms

Figure 9-3 (Continued)

as was illustrated in Table 9-7), along with a rating sheet similar to that shown previously in Figure 9-1. Participants were instructed to work through the folder, one work sample at a time, and provide a holistic evaluation of the sample, entering a rating from 1 to 3 on the form corresponding to the performance category to which they believed the work sample was best matched. They were reminded not to attempt to score the samples but simply to render a holistic verdict about each one. Participants were also reminded that the work samples were in random order in the folders so that the overall quality of any work sample might be better or worse than the ones before or after it.

Participants spent the afternoon of Day 2 completing their rating forms. During the final hour of the day, they discussed their ratings with other members of their small groups (5–6 panelists per group) and made any adjustments to their ratings that they wished to make. At the end of the afternoon, facilitators collected all materials and dismissed for the day. Contractor staff then entered and verified panelist ratings and examined the distributions. Tests with scores below 27 and above 43.5 yielded very little information regarding the location of a cut and were therefore eliminated from consideration. Likewise, tests with scores between 33.5 and 38 had yielded a high degree of agreement and were eliminated from further consideration in Round 2. Given these findings, the facilitators selected 22 new work samples with scores surrounding each of the two preliminary cut scores; that is, for each of the two cut scores, 11 samples with scores at or

below the cut and 11 with scores above the cut were selected. The 22 new work samples for the lower cut score were referred to as the Low 22, and the 22 new work samples used for the higher cut score were referred to as the High 22. Scores for the Low 22 set ranged from 27 to 33, while those for the High 22 set ranged from 38.5 to 43.5. The work samples used for Round 2 samples are shown in Table 9-8.

On the morning of Day 3, participants reviewed their Round 1 ratings as a whole group. Facilitators provided overhead transparencies of distributions of ratings and other summaries of Round 1 results. (Electronic versions of all of these materials are available at www.sagepub.com/cizek/bowsamples.) After a break, facilitators distributed the Round 2 rating sheets and new folders with the Low 22 and High 22 sets of work samples. Participants completed an evaluation to determine their readiness to begin Round 2, then they began rating the pinpointing work samples. The pinpointing rating activities continued until noon on Day 3 and resumed after lunch through midafternoon. During the final hour of Day 3, participants discussed their Round 2 ratings in small groups, made any necessary adjustments, and completed their forms. At the end of the day, facilitators collected all materials and dismissed the group. Contractor staff then entered panelist ratings and recalculated cut scores and rating distributions for each sample.

For the *Proficient* cut score, the logistic regression using SAS yielded a cut score of 29.4 and, distressingly, a cut score of 16.7 for *Advanced*. Investigation of the model fit statistics of both models (one for each cut score) revealed that the model fit for *Proficient* was good based on the significant reduction in the log likelihood statistics and significant chi-square statistics for both intercept and slope. For the *Advanced* cut score, however, the model-to-data fit was exceptionally poor, as shown by an actual increase in the log likelihood statistics from null hypothesis to model hypothesis and insignificant chi-square values for both intercept and slope. These results are shown in Table 9-9. The analyses were replicated using Excel; regardless of the software used, the projected *Proficient/Advanced* cut score was not even in the range of scores considered, providing further evidence of the unacceptable model/data fit.

To address this anomaly, facilitators examined participants' ratings and found that the panelists could not consistently identify an *Advanced* sample. The Round 2 ratings are shown in Table 9-10. As can be seen in the table, there was general agreement that work samples with raw scores below 40 were *Proficient*. At score point 40, there were two work samples provided for participants to rate. Although both work samples had been assigned identical total raw scores of 40, the first work sample received

Table 9-8 Work Samples for Language Arts Test: Round 2

Low 22	
Code	*Score*
205721	31.5
206379	30.0
206423	28.0
206666	31.0
206728	28.5
206754	30.5
206856	31.5
241086	30.0
241141	29.5
241517	30.5
241780	31.5
241800	33.0
250068	30.5
250124	28.0
250151	29.0
284286	29.5
284321	32.0
285178	29.0
285180	31.0
285883	27.0
285919	28.5
285921	27.5

High 22	
Code	*Score*
120802	43.5
120807	43.0
120847	38.5
206212	42.5
206357	41.0
206358	39.5
206441	41.5
206445	41.5
206681	39.0
206778	42.5
206807	40.5
206827	40.5
206835	40.0
240987	43.5
241534	39.5
241797	41.5
241815	40.0
241826	41.0
242319	40.5
249204	42.0
250633	42.5
284232	39.5

13 ratings of *Proficient* and 13 ratings of *Advanced*, while the second work sample was classified as *Proficient* by 5 participants and as *Advanced* by 21 participants. Thus it would seem that the likelihood of being rated *Advanced* crossed the threshold at raw score point 40.

Table 9-9 Logistic Regression Results for Pinpointing Round

Model Fit Statistics for Low 22

Criterion	*Intercept Only*	*Intercept and Covariates*
AIC	792.693	758.337
SC	794.477	761.905
-2 Log L	790.693	754.337

Analysis of Maximum Likelihood Estimates for Low 22

Parameter	*df*	*Estimate*	*Standard Error*	*Chi-Square*	*Pr > ChiSq*
Intercept	1	–9.2961	1.6225	32.8282	<.0001
Slope	1	0.3162	0.0544	33.7988	<.0001

Model Fit Statistics for High 22

Criterion	*Intercept Only*	*Intercept and Covariates*
AIC	738.638	740.455
SC	740.352	743.882
-2 Log L	736.638	736.455

Analysis of Maximum Likelihood Estimates for High 22

Parameter	*df*	*Estimate*	*Standard Error*	*Chi-Square*	*Pr > ChiSq*
Intercept	1	0.4430	2.5394	0.0304	0.8615
Slope	1	–0.0265	0.0618	0.1833	0.6686

To investigate further, however, the three work samples that had been scored 40.5 were examined. As shown in Table 9-10, these three work samples also generated considerable disagreement. The first two received mostly

Proficient ratings (16 out of 26 and 14 out of 26, respectively), while the third received mostly *Advanced* ratings (17 out of 26). The next work sample to receive a majority of *Advanced* ratings was at score point 41.5, but the ratings soon turned back to *Proficient*. The two highest scoring samples in the set received 20 and 19 *Proficient* ratings (out of 26), respectively.

In consultation with officials and staff for the agency responsible for setting the cut scores, facilitators employed a backup plan to the use of logistic regression, which consisted of plotting the scores against ratings. At score point 40, the *Proficient* and *Advanced* curves crossed, indicating that a 50% chance of being rated *Advanced* occurred first at that score point. A similar graph was prepared for the Low 22 (*Proficient* cut) so that all results could be presented in a common format. The *Proficient* cut was rounded to 29.5, and the *Advanced* cut was presented as 40.0.

On the morning of the final day of the standard-setting meeting, participants received and discussed the Round 2 (i.e., pinpointing) results in a large-group setting. After a break, panelists then received impact data, showing the numbers and percentages of examinees that would be classified at each of the three performance levels, based on the Round 2 cut scores. Each panelist received a table and a graph depicting each raw score, the percentage of students at that score point, and a cumulative distribution for each raw score point. The data were also presented on overhead transparencies. The subsequent whole-group discussions lasted approximately 30–45 minutes, at the end of which panelists indicated their readiness for the final round of standard setting on their readiness forms. Round 3 rating sheets were then distributed, and participants were instructed to enter exactly two numbers, one representing the lowest possible total score a student could receive and be classified as *Proficient* and one representing the lowest possible score a student could receive and be classified as *Advanced*. A copy of the Round 3 rating sheet is shown in Figure 9-4. By this point in the process, the participants were conversant with the Round 2 work samples, the scores associated with each, and the relationship between cut scores and percentages of examinees in the statewide population who would be classified at each level, given a particular set of cut scores.

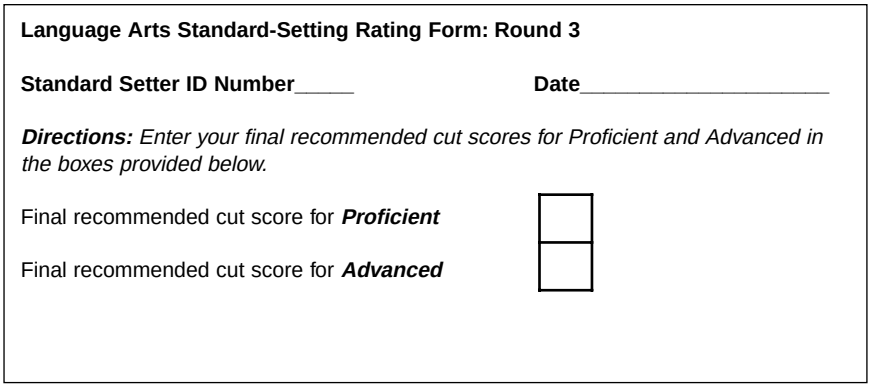
Participants completed their Round 3 forms, then completed their readiness and evaluation forms, and were dismissed. By agreement with agency officials and staff, the final cut scores for both the *Proficient* and *Advanced* levels were set at the mean of all participants' recommendations, with both the mean and standard deviation to be reported to the state board of education. Contractor staff then calculated the final cut scores and standard deviations: 29.5 for *Proficient* and 42.0 for *Advanced,* with standard deviations of .39 and .43 for *Proficient* and *Advanced,* respectively.

Table 9-10 Round 2 Ratings for High 22 Work Samples

Raw Score Assigned to Work Sample	*Frequency Work Sample Rated Proficient*	*Frequency Work Sample Rated Advanced*	*Total*
38.5	21	5	26
39.0	26	0	26
39.5	22	4	26
39.5	20	6	26
39.5	19	7	26
40.0	13	13	26
40.0	5	21	26
40.5	16	10	26
40.5	14	12	26
40.5	9	17	26
41.0	25	1	26
41.0	22	4	26
41.5	18	8	26
41.5	11	15	26
41.5	4	22	26
42.0	13	13	26
42.5	25	1	26
42.5	25	1	26
42.5	11	15	26
43.0	9	17	26
43.5	20	6	26
43.5	19	7	26

Alternative Procedures and Limitations

Each of the holistic standard-setting methods described in this chapter has advantages and limitations. The BoW method has quickly gained popularity as a standard-setting technique, particularly for statewide assessments with


Language Arts Standard-Setting Rating Form: Round 3

Standard Setter ID Number_____ **Date**_____

Directions: *Enter your final recommended cut scores for Proficient and Advanced in the boxes provided below.*

Final recommended cut score for ***Proficient***

Final recommended cut score for ***Advanced***

Figure 9-4 Body of Work Final Round Rating Sheet

large essay or CR item components. In particular, the method seems well suited to language arts tests. The problems encountered in the application described in this chapter are not peculiar to the method. Inconsistent patterns of ratings of work samples will undermine any holistic method. In this particular instance, the directive to present work samples in random, rather than score-point, order could have actually reduced the effectiveness of the BoW method. That the confusing logistic regression results actually paralleled the confusing raw data was somewhat reassuring: Results should not seem more reasonable than the raw data.

From the point of view of participants, the BoW method, like other holistic methods, seems quite natural, particularly to educators, who are accustomed to providing summary judgments (e.g., letter grades) for student work samples. The intra- and interround discussions of a common set of work samples are also quite natural for most participants. During these discussions, the facilitator is often called upon not only to mediate but also to redirect and keep the discussion on the content of the samples, as it relates to the PLDs.

From the point of view of those who are charged with planning and conducting the standard-setting meeting, the BoW and other holistic methods require extensive advance preparation and intensive effort during the meeting itself. For instance, the work samples to be used in the pinpointing round cannot be selected in advance because the precise range of raw score values to be included is not known until the rangefinding round has been completed. Ordinarily, the tasks of selecting and duplicating copies of the relevant work samples must then be undertaken during the evening between the Round 1 and Round 2 sessions. Even when the same work samples are used for Round 2 (as prescribed by Kingston et al., 2001), the selection of

the new samples to augment the initial sample requires the same degree of attention to detail that the selection of the initial set required.

One alternative to consider when implementing the BoW method involves the configuration of the rounds of judgments. In some applications, the rangefinding round is divided into Rounds 1A and 1B. In Round 1A, participants enter their ratings without discussion. In Round 1B, they discuss the ratings, make changes if they choose, and complete their forms. At the end of the round, they turn in their rating forms for Round 1B to the facilitators. As in the rangefinding round, the second round of judgments (i.e., the pinpointing round) may be subdivided into Rounds 2A and 2B. In Round 2A, panelists would enter their ratings without discussing them. They would then discuss their ratings, either in small groups or as a single large group, and make adjustments to their Round 2 ratings. They would submit these ratings as Round 2B ratings.

In the BoW method, there are also alternatives to the manner in which work samples should be presented to participants. In this chapter, we have described situations in which the work samples were provided in score-point order. Another option—and one for which additional research would aid in determining its feasibility—would be to present work samples in random order. Finally, some methods of data analysis may fail to result in an acceptable solution, as was the case in the application described here with the use of logistic regression. In such cases, an alternative procedure is needed. Reasonable options include calculating the means of adjacent performance categories and using the midpoint between the two means as the cut score. Or, as Plake and Hambleton (2001) have suggested, it is possible to simply ask panelists to vote on the final cut score(s).

In the preceding sections of this chapter, we have limited our discussion of statistical methods for calculating cut scores in the BoW approach primarily to logistic regression, one regression at a time. However, it is possible to calculate all cut scores simultaneously through ordinal logistic regression (OLR). There are certain advantages to be obtained through the use of OLR, as well as certain caveats to keep in mind. First, however, we will describe the logic of this analytical alternative and offer an example, again using the data in Table 9-3.

According to one source on ordinal logistic regression:

> If a single model could be used to estimate the odds of being at or below a given category across all cumulative splits, that model would offer far greater parsimony over the fitting of k–1 different logistic regression models corresponding to the sequential partitioning of the data as described above. The goal of the cumulative odds model is to simultaneously consider the effects of a set of independent variables across these possible consecutive cumulative splits to the data. (O'Connell, 2006, p. 28)

SAS (and other statistical packages) make this modeling possible. The SAS code for calculating the three cut scores simultaneously, along with the relevant results, are provided in Table 9-11.

It should be noted that while Table 9-11 shows three intercept values, only one slope value is given; this occurs because we have created only one

Table 9-11 SAS Programming Code and Selected Output for Calculating Cut Scores Based on Data in Table 9-3 Using Ordinal Logistic Regression

SAS Programming Code

```
proc logistic data = ssround descending;
model category = rs/rsquare;
output out = ss1 predprobs = cumulative;
title 'NEW COMMANDS 1';
run;
```

SAS Output

PROC LOGISTIC
Probabilities modeled are cumulated over the lower ordered values.

Model Fit Statistics

Criterion	*Intercept Only*	*Intercept and Covariates*
AIC	750.174	394.270
SC	761.163	408.922
-2 Log L	744.174	386.270

Analysis of Maximum Likelihood Estimates

Parameter	*df*	*Estimate*	*Standard Error*	*Chi-Square*	*Pr > ChiSq*
Intercept 4	1	–19.7624	1.5803	156.3951	<.0001
Intercept 3	1	–15.0671	1.2544	144.2762	<.0001
Intercept 2	1	–10.4522	0.9330	125.5011	<.0001
Slope	1	0.5217	0.0429	148.0579	<.0001

model. This is an important difference, relative to individual logistic regression models for each cut score. In essence, we are postulating that the relationship between raw score and classification probability remains the same throughout the score range. With three separate models, we permit the relationship to change over the raw score range. This is more than a simple procedural difference; it is one that will require some consideration of the nature of the relationship between scores and classification. If a uniform relationship can be defended, a cumulative probability model is appropriate; if not, then separate models should be worked out for each cut score.

The results from Excel, SAS individual estimates, and SAS cumulative (i.e., simultaneous) probability estimates based on the data in Table 9-7 are summarized in Table 9-12. Table 9-12 gives the slopes, intercepts, and resulting cut scores for each of the models. We can see that the results are quite similar, regardless of procedure or statistical package: All cut scores for *Basic* are near 20, all cut scores for *Proficient* are about 29, and all cut scores for *Advanced* are approximately 38–38.5. For these particular scores and ratings, the choice of procedure may be one of convenience. Excel would normally work just as well (for binary logistic regression at least) as SAS or another statistical package, allowing for differences in approach (i.e., least squares vs. maximum likelihood). In general, however, the choice of procedure should be based on *a priori* assumptions about the relationship between ratings and scores.

Table 9-12 Summary of Logistic Regression Analyses Using Excel, SAS Binary, and SAS Cumulative Procedures

Performance Level	*Method*	*Slope*	*Intercept*	*Cut Score*
Basic	Excel	.3732	–7.7047	20.64
	SAS Binary	.5613	–11.3266	20.07
	SAS Cumulative	.5217	–10.4522	20.03
Proficient	Excel	.4397	–12.6395	28.74
	SAS Binary	.5262	–15.1934	28.87
	SAS Cumulative	.5217	–15.0671	28.88
Advanced	Excel	.2891	–11.1414	38.54
	SAS Binary	.3583	–13.7280	38.31
	SAS Cumulative	.5217	–19.7624	37.88

Of the other holistic methods described in this chapter, the JPC method is perhaps the most mathematically daunting. Particularly at the weighting and combining stage, the procedure may prove too much for some potential users. The fact that the policy is derived, rather than developed directly through interaction among the judges, is also cause for some concern. Thus, in the future, the DP method may emerge as the stronger candidate for a holistic standard-setting alternative. Even this method, however, has its limitations, as Plake et al. (1997) have noted. The procedure may not always yield a clear, or at least an undisputed, policy.

The discussion of AJM procedure has also hinted at possible modifications or variations. As Plake and Hambleton (2001) observed, a shift from 12 categories to 7 had minimal impact on the placement of the cut scores. Along those same lines, a shift from 7 categories to 4 might also be effective. Plake and Hambleton actually began with four performance levels (*Below Basic, Basic, Proficient,* and *Advanced*) then expanded this to 12 categories by creating three intraclass categories (low, middle, and high) for each performance level. They then reduced this number to 7 by combining borderline categories (e.g., high proficient-low advanced) and dropping the two extreme categories (i.e., low below basic and high advanced). The next logical simplification would be to eliminate the original four categories and retain only the three borderline groups (*Borderline Basic, Borderline Proficient,* and *Borderline Advanced*). Participants would then only be required to judge which work samples should be classified into these categories. Cut scores would be obtained by calculating the mean scores for each of these three borderline groups. Another approach would be to retain the original four groups (i.e., *Below Basic, Basic, Proficient,* and *Advanced*) and simply conduct a contrasting-groups procedure, setting the three cut scores at the midpoints between the four group means. Such an alternative has, in fact, already been explored in the context of setting performance standards on alternate assessments. We describe these procedures and the results of using this approach in Chapter 15.

10

The Bookmark Method

The Bookmark procedure may be viewed as a logical successor to a series of item-mapping strategies developed in the 1990s in conjunction with standard settings carried out for the National Assessment of Educational Progress (NAEP) by researchers at American College Testing (ACT). Early item-mapping techniques were applied less as standard-setting procedures per se than as feedback mechanisms embedded in other procedures (cf. Loomis & Bourque, 2001).

In 1996, for example, researchers at ACT employed an item-mapping procedure in conjunction with a method they referred to as Mean Estimation, which was essentially an extension of the modified Angoff (1971) technique. That item-mapping procedure was applied to tests with both multiple-choice and constructed-response items (Loomis, Bay, Yang, & Hanick, 1999). Item maps were used to provide feedback after a second round of item ratings for the 1996 Science assessment and the 1998 NAEP Civics and Writing assessments. The maps showed the location of each item in relation to the NAEP-like scale score, which was also associated with the various NAEP achievement level descriptors (ALDs, which are now commonly referred to as performance level descriptors). Each multiple-choice item was mapped in accordance with its probability of correct response for each scale score, and each constructed-response item was mapped once for each score point, that is, for the probability of obtaining a score of 1, 2, 3, or higher at each scale score point.

Item-mapping techniques evolved through the course of several NAEP standard-setting studies at ACT. The Reckase chart (Reckase, 2001) was introduced as a way to simplify the task set before participants. With

Reckase charts, participants would receive their Round 2 item estimates (i.e., the probability of a correct response by a student at the cut score for multiple-choice items and estimated raw score for constructed-response items for this same student or group of students), along with a preprinted table or "map" of item probabilities.

A sample Reckase chart for an individual participant is shown in Table 10-1. A unique Reckase chart would be developed based on each participant's item ratings. The first column in the Reckase chart shown in the table presents scaled scores arranged from high to low. Scaled scores are used in Reckase charts as a measure of overall examinee competence or ability on whatever construct is measured by the test. Each of the remaining columns contains information on a single item. Table 10-1 shows information on five items with Items 1–4 being dichotomously scored multiple-choice format items and Item 5 being a constructed-response item scored on a 0–5 scale. For the multiple-choice items, the data in each column show the probability of an examinee at each scaled score answering that item correctly, based on the three-parameter item response model. For example, an examinee with an overall ability level (i.e., scaled score) of 170 has a .53 probability of answering Item 1 correctly. For constructed-response items, the values in a column show the expected item score for examinees at a given scaled score location. Again considering an examinee with an ability level of 170, the expected score of that examinee on the constructed-response item (Item 5) is 1.8 out of 5.

In Table 10-1, one value in each column appears in brackets; it is in this way that Reckase charts are individualized for each participant. When used as feedback in standard setting, Reckase charts help participants gauge how consistently they are applying their conceptualization of the minimally competent examinee, borderline candidate, or whatever hypothetical examinee is considered. The Reckase chart for a participant who is consistently applying his or her conceptualization would show brackets aligned in a single row. For example, consider the participant whose judgments resulted in the values shown in the table. In addition, let us assume that the participant held an implicit conceptualization that the minimally qualified examinee is one with an ability level (represented by a scaled score) of 170. Reading across the row in the table corresponding to a scaled score of 170, we see that the probability estimate (i.e., Angoff rating) generated by this participant was .53; this participant is saying that the probability of a minimally qualified examinee answering Item 1 correctly is .53. Now, if this participant were applying his or her conceptualization of the minimally qualified examinee consistently, he or she would have generated an Angoff rating of .83 for Item 2, .34 for Item 3, and .77 for Item 4. For the

constructed-response item (Item 5), this participant would have estimated the minimally qualified examinee's score to be 1.8 out of 5.

From the Reckase chart shown in Table 10-1, however, the participant can see that he or she is not making totally consistent judgments. For the remaining three multiple-choice items (Items 2–4), the participant has estimated the items to be more difficult than they are for an examinee of ability level 170. For example, for Item 2, the participant judged the minimally qualified examinee to have a .57 probability of success on the item when, using the standard implied by this participant's rating of Item 2, the rating for Item 2 should have been .83. For the constructed-response item, the reviewer exhibited more consistent behavior with his or her implicit performance standard as shown by the fact that his or her rating of Item 5 of 1.5 is very close to the expected constructed-response item score of 1.8 for examinees with an overall ability level of 170. If this participant were being perfectly consistent, the bracketed values would be aligned in a row corresponding to a single ability level (scaled score).

Table 10-1 can be thought of as an early item map. From this foundation, it was not a great step to refine the item-mapping procedure by reordering the items according to their difficulty. Loomis, Hanick, Bay, and Crouse (2000) reported on field trials for the 1998 NAEP Civics test in which the item maps were reordered from least to most difficult item. These item maps also included brief descriptions of item content, which permitted participants, at a glance, to summarize both the location and content of an item and to reframe their own judgments of those items. From difficulty-ordered item maps with content information and probability of correct response, the leap to an ordered test booklet with similar information was a short but significant one. Researchers at CTB/McGraw-Hill made that leap and introduced the Bookmark method (Lewis, Mitzel, & Green, 1996).

Overview of the Bookmark Method

The standard Bookmark procedure (Mitzel et al., 2001) is a complete set of activities designed to yield cut scores on the basis of participants' reviews of collections of test items. The Bookmark procedure is so named because participants express their judgments by entering markers in a specially designed booklet consisting of a set of items placed in difficulty order, with items ordered from easiest to hardest. This booklet, called an *ordered item booklet,* will be described in greater detail in the next portion of this chapter.

The Bookmark procedure has become quite popular for several reasons. First, from a practical perspective, the method can be used for complex,

Table 10-1 Example of a Reckase Chart

	Probabilities of Correct Response for Given Scale Score				
Scale Score	*Item 1*	*Item 2*	*Item 3*	*Item 4*	*Item 5*
215	.99	.99	.99	.99	4.8
212	.99	.99	.98	.99	4.7
209	.99	.99	.97	.98	4.6
206	.98	.99	.96	.98	4.5
203	.98	.99	.94	.97	4.4
200	.97	.99	.91	.97	4.3
197	.96	.98	.88	.96	4.1
194	.94	.98	.83	.95	3.9
191	.92	.97	.77	.94	3.7
188	.89	.96	.70	.92	3.5
185	.85	.95	.63	.91	3.2
182	.81	.93	.55	.89	2.9
179	.75	.91	.48	.86	2.6
176	.68	.89	.42	.83	2.4
173	.60	.86	.37	.80	2.1
170	[.53]	.83	.34	.77	1.8
167	.45	.79	.31	.73	[1.5]
164	.37	.74	.29	.69	1.3
161	.30	.69	.28	.66	1.1
158	.24	.63	.27	.62	0.9
155	.20	[.57]	.26	.58	0.7
152	.16	.52	.26	.55	0.6
149	.13	.46	.26	.52	0.5
146	.11	.41	.25	.49	0.4
143	.09	.36	.25	.46	0.3
140	.08	.32	.25	[.44]	0.2
137	.07	.29	.25	.43	0.2
134	.07	.26	.25	.41	0.2
131	.06	.24	.25	.40	0.1
128	.06	.22	.25	.39	0.1
125	.06	.20	[.25]	.38	0.1

NOTES: For multiple-choice items (Items 1–4) the values in brackets [ ] are a participant's Angoff ratings; for constructed-response items (Item 5) the value in brackets is the participant's estimated mean score for a minimally competent examinee.

Source: Adapted from Reckase (2001).

mixed-format assessments, and participants using the method consider selected-response (SR) and constructed-response (CR) items together. As the prevalence of mixed-format examinations continues to increase, it is likely that the Bookmark method will become even more widely used and that other innovative approaches for setting performance standards in such contexts will be developed.

Second, from the perspective of those who will be asked to make judgments via this method, it presents a relatively simple task to participants, and one with which, at a conceptual level, they may already be familiar. To fully grasp the extent to which the Bookmark method simplifies the standard-setting task, it is instructive to consider a test with four performance levels (*Below Basic, Basic, Proficient,* and *Advanced*), 60 SR items, and four CR items (with four score points each). If item-based standard-setting methods such as the Angoff or modified Angoff procedures were used, participants would have 192 separate tasks to perform per round of ratings (i.e., three probability judgments for each of 64 items). With the Bookmark procedure, the same participant may still consider the content covered by the items in a test but is required to make only three judgments—one for each of three bookmarks (*Basic, Proficient,* and *Advanced*) he or she will be asked to place in a difficulty-ordered test booklet (described in more detail later in this chapter). The task is perhaps even more streamlined because it would seem reasonable that the bookmark for *Advanced* should be placed after the bookmark for *Proficient,* and that the bookmark for *Proficient* should be after the bookmark for *Basic.* Thus once a participant has identified one cut score through the placement of his or her bookmark, it is not necessary for him or her to start the search for the next cut score at the beginning of the ordered test booklet. In order to make judgments about each subsequent cut score, participants can examine a relatively narrow range of items rather than reexamining each item and making a new estimate of the probability of a student just barely at a particular performance level answering correctly.

Third, in addition to being relatively easy for participants, the Bookmark method is also comparatively easy for those who must implement the procedure. Although some of the computational aspects of the method are mathematically complex, most of the intensive work is done long before the standard-setting session itself occurs. For those who conduct such sessions, this is an important feature of the procedure that helps reduce the potential for errors and the time required for the standard-setting meeting.

Finally, from a psychometric perspective, the method has certain advantages because of its basis in item response theory (IRT) analyses, and because of the fidelity of the method to the test construction techniques that spawned the assessment. With few exceptions, most high-stakes, large-scale

assessments are constructed in accordance with an IRT model, either Rasch or 3PL. The analyses normally carried out in the construction and equating of these tests make IRT-based standard-setting procedures a natural extension. Once participants provide page numbers, the associated theta values have a built-in relationship to scores, and results can be interpreted in the same manner as other procedures carried out with these tests. In the absence of other IRT-based standard-setting procedures, the bookmark procedure is a natural choice.

The Ordered Item Booklet

Perhaps the most distinctive feature of the Bookmark method is the collection of items that serves as the focus of participants' judgments. This booklet, called an *ordered item booklet* (OIB), can contain both SR format items, such as multiple-choice, and CR items intermingled in the same booklet. An SR item appears in the OIB once, in a location determined by its difficulty (usually its IRT *b* value). Each CR item appears several times in the booklet, once for each of its score points. For a typical application of the procedure, each SR item will have one associated difficulty index, and each CR item will have as many step (difficulty) functions as it has score points (excluding zero). For a given CR entry, the item prompt and the rubric for a particular score point would ordinarily also be provided to participants, along with sample responses illustrating that score point.

The OIB can be composed of any collection of items spanning the range of content, item types, and difficulty represented in a typical test and need not consist only of items that have appeared in an intact test. This booklet can have more items or fewer items than an operational test form. One advantage of permitting items beyond those included in an operational test form is the fact that gaps in item difficulty or content coverage can be filled with items from a bank. For example, if two adjacent items in the ordered booklet have difficulty indices of 1.05 and 1.25 logits, additional items with difficulty indices of 1.10, 1.15, and 1.20 could be inserted to help standard-setting participants place their bookmarks more precisely. Conversely, a clear advantage of using an intact test form for standard setting using the Bookmark method is the fact that the results can be interpreted in a straightforward manner; namely, the test booklet on which standards are set is the same set of items on which student scores and (sometimes high-stakes) decisions are based.

Figure 10-1 shows the general layout of a hypothetical OIB. As indicated previously, items in the OIB appear one per page. Each SR item appears on a single page; each CR item is included in the OIB a number of times equal to the number of possible score points (excluding zero) associated with the item, along with one or more sample responses at that score point. Thus an

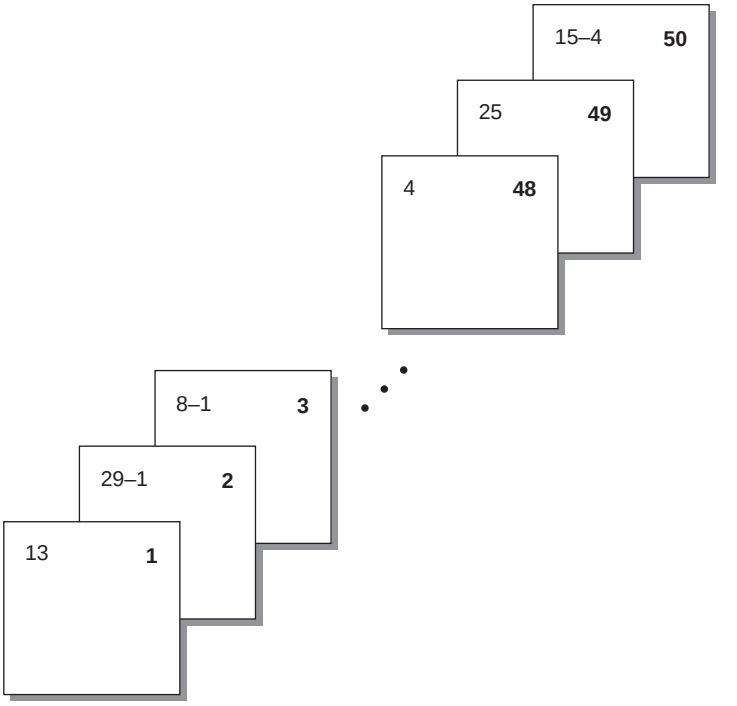


Figure 10-1 Hypothetical Illustration of an Ordered Item Booklet

item scored on a 5-point scale (0–4) and thus having four nonzero score points (i.e., 1, 2, 3, 4) would be represented on four different pages in the OIB. These configurations are shown in Figure 10-1. The bold numbers at the top right of each page illustrated in the figure simply indicate the sequence of the items in the OIB (i.e., pagination). The numbers at the top left indicate the positions in the test form upon which the OIB is based. For example, the item appearing on the first page of the OIB appeared as Item 13—an SR item—in the intact test form. It should be noticed that some of these numbers in the top left corner of the OIB pages have hyphens. These numbers refer to the original item number and the score point represented on that page. For example, the second page in the OIB represents a response earning a score of 1 to original Item 29 (a CR item); page 50 in the OIB contains the response earning the highest score (i.e., 4) on another CR item (Item 15 in the original intact test form). In an actual OIB, information beyond simple pagination and original item numbers would be included. A more detailed description and illustration of the information typically provided on an OIB page is presented later in this chapter.

The Response Probability (RP) Value

In the Bookmark procedure, the basic question participants must answer is "Is it likely that the minimally qualified or borderline examinee will answer this SR item correctly (or earn this CR item score point)?" Obviously, it is important to define "likely" or to operationalize this decision rule. In practice, the Bookmark procedure employs a 67% likelihood (or sometimes a 2/3 chance) of desired response (i.e., of getting the SR item correct or of achieving a certain CR score point or higher).

In the more than 30 years that have intervened between the introduction of the Angoff and Bookmark methods, there has been considerable experimentation with decision rules. Huynh (2000, 2006) has argued that the probability value that maximizes the information of the correct response would produce the optimum decision rule. As it turns out, for a three-parameter model with the guessing parameter removed (i.e., a two-parameter model), a 67% likelihood (i.e., a response probability [RP] of .67) optimizes this value. Thus the typical decision rule for the bookmark procedure is .67, although other percentages (ranging from .50 to .80) are also sometimes used.

In a Rasch model context, Wang (2003) has expressed a preference for a 50% likelihood (RP = .50). Indeed, the choice of .50 for the Rasch model has certain practical advantages over .67 in that the likelihood of a correct response is exactly .50 when the examinee's ability is equal to the item's difficulty. Wang pointed out, however, that the issue should not be considered resolved and urged further research into the efficacy of the .50 decision rule in Rasch applications. Although the difference may at first seem trivial, following both the suggestion of the originators of the Bookmark procedure and our own experience in implementing the Bookmark method, our tendency is to use a decision rule of 2/3. We note too that we tend to express the decision rule in this way (rather than as RP = .67). Of course, framing the issue as a decision rule of 2/3 or as an RP of .67 is (at least mathematically) nearly the same. In our experience, however, standard-setting participants seem better able to grasp and work with the notion of "two out of three" more readily than a probability of .67.

Response Probabilities and Ordered Item Booklet Assembly—Rasch Model

As may already be obvious, the choice of a decision rule (or RP value) is essential to the assembly of OIBs and to the calculation of cut scores when the Bookmark method is used. In the following description, we assume that a Rasch model has been used for test construction, item calibration, and so

on and that a decision rule of 2/3 has been incorporated into participants' training, practice, and OIB rating activities. We begin with the basic Rasch equation, set forth in Wright and Stone (1979), which expresses the probability of answering an item correctly, $p(x = 1)$, as a function of the item's difficulty (β_j) and the examinee's ability (θ_i):

$$p(x = 1|\theta_i, \beta_j) = \exp(\theta_i - \beta_j)/[1 + \exp(\theta_i - \beta_j)] \quad \text{(Equation 10-1)}$$

Now, setting p equal to 2/3 and solving for θ_i we obtain

$$\exp(\theta_i - \beta_j)/[1 + \exp(\theta_i - \beta_j)] = 2/3 \quad \text{(Equation 10-2)}$$

$$\exp(\theta_i - \beta_j) = 2/3 * [1 + \exp(\theta_i - \beta_j)] \quad \text{(Equation 10-3)}$$

$$\exp(\theta_i - \beta_j) = 2/3 + 2/3 * \exp(\theta_i - \beta_j) \quad \text{(Equation 10-4)}$$

$$\exp(\theta_i - \beta_j) - 2/3 * \exp(\theta_i - \beta_j) = 2/3 \quad \text{(Equation 10-5)}$$

$$1/3 * \exp(\theta_i - \beta_j) = 2/3 \quad \text{(Equation 10-6)}$$

$$\exp(\theta_i - \beta_j) = 2/3 \div 1/3 \quad \text{(Equation 10-7)}$$

$$\exp(\theta_i - \beta_j) = 2 \quad \text{(Equation 10-8)}$$

Finally, taking the natural log of both sides of Equation 10–8, we obtain

$$\theta_i - \beta_j = .693, \text{ and} \quad \text{(Equation 10-9)}$$

$$\theta_i = \beta_j + .693 \quad \text{(Equation 10-10)}$$

The reader who is familiar with the work of Wright and Stone (1979) will notice that we have used slightly different notation than that source. Our substitution of θ and β to represent examinee ability and item difficulty, respectively, is an attempt to make the notation used in the preceding explication more consistent with standard notion across the family of IRT models. (We should also note another small but important difference between a 2/3 decision rule and an RP67 rule. If a response probability of .67 had been used, Equation 10-10 would have been $\theta_i = \beta_j + .708$; that is, it would have been computed by taking the item's difficulty plus the natural logarithm of .67/.33.)

The use of the final result in Equation 10-10 to assemble the OIB is straightforward. For SR items, to calculate the value of θ_i needed to have

a 2/3 chance of answering a given SR item correctly, we simply add .693 to the Rasch difficulty value for that item, where the Rasch difficulty of an item is obtained by use of an IRT calibration program (e.g., WINSTEPS). As is perhaps evident, when the Rasch model is used to create an OIB with SR items, the procedure just described will result in the same ordering of items in the OIB as if the booklet had been assembled using the items' *b* values. This result would *not* likely occur, however, for items calibrated using a 2PL or 3PL model. As we will see a bit later, these same values used to determine the placement of SR items in the difficulty-ordered test booklet are also used in determining the raw score associated with setting a bookmark right after this item in the OIB.

Locating the appropriate placement of CR items in the OIB is only slightly more complicated. To locate the score points of CR items in the OIB within a Rasch framework, the Partial-Credit Model (PCM; Wright & Masters, 1982) is used. In the following discussion, the procedure is illustrated for a CR item with five score points (0, 1, 2, 3, 4); however, the logic is applied to items with any number of steps.

To begin, for CR items, the likelihood (π_{nix}) of a person with a given ability (θ_n) obtaining any given score (*j*) in any item (*i*) is shown in the following equation, taken from Wright and Masters (1982, equation 3.1.6):

$$\pi_{nix} = \frac{\exp \Sigma(\theta_n - \delta_{ij})}{\Sigma \exp \Sigma(\theta_n - \delta_{ij})} \qquad \text{(Equation 10-11)}$$

In Wright and Masters's formulation, the difficulties associated with each score point are referred to as step functions and are symbolized generally as δ_{ij}. The step function for score point 0 is set equal to 0 in Equation 10-11; that is, $\delta_{i0} \equiv 0$, such that

$$\Sigma(\theta_n - \delta_{ij}) = 0, \text{ and } \exp \Sigma(\theta_n - \delta_{ij}) = 1 \qquad \text{(Equation 10-12)}$$

The numerator values for the other steps are derived as follows:

Step 1. $\Sigma(\theta_n - \delta_{ij}) = \Sigma(\theta_n - \delta_{i0}) + \theta_n - \delta_{i1}$
$= 0 + \theta_n - \delta_{i1}$
$= \theta_n - \delta_{i1}$ (Equation 10-13)

Step 2. By similar logic: $\Sigma(\theta_n - \delta_{ij}) = 2\theta_n - \delta_{i1} - \delta_{i2}$ (Equation 10-14)

Step 3. By similar logic: $\Sigma(\theta_n - \delta_{ij})$
$= 3\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3}$, and (Equation 10-15)

Step 4. By similar logic: $\Sigma(\theta_n - \delta_{ij})$
$$= 4\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3} - \delta_{i4} \quad \text{(Equation 10-16)}$$

The exponential values of these summations shown in Equations 10-12 through 10-16 are simply the natural logarithm *e* raised to the respective values, that is:

Step 0. $\exp(0)$ (Equation 10-17)

Step 1. $\exp(\theta_n - \delta_{i1})$ (Equation 10-18)

Step 2. $\exp(2\theta_n - \delta_{i1} - \delta_{i2})$ (Equation 10-19)

Step 3. $\exp(3\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3})$ (Equation 10-20)

Step 4. $\exp(4\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3} - \delta_{i4})$ (Equation 10-21)

The denominator of Equation 10-11 now becomes the simple sum of the values yielded by Equations 10-17 through 10-21 for Steps 0–4. Finally, the desired end—the likelihood of obtaining any given score (0 through 4)—is calculated by dividing the numerator associated with that score point by this common denominator. These calculations can be carried out by hand or with various software programs, such as SPSS, SAS, or Excel. A step-by-step procedure for using Excel to compute the appropriate OIB locations for CR item score points is provided in Table 10-2. The Excel spreadsheet on which the table is based is available with the other electronic materials accompanying this book at www.sagepub.com/cizek/bookmark.

As an initial check on the accuracy of the values obtained, begin by locating the values in columns N–Q. Read down the column of values until .5000 or the closest value to .5000 is found. Then simply read across the row from this value to column A to find the corresponding value of θ_n. This value of θ_n should correspond to the Thurstone Threshold Value reported in WINSTEPS. Having verified that the RP50 value corresponds to the Thurstone Threshold Value, continue down columns N–Q (depending on the score point desired) to find the value closest to .6667, or use interpolation to obtain an exact value. Again, read across the row to column A to find the corresponding value of θ. This value is the ability (or achievement) level associated with a 2/3 chance of obtaining the particular score point or better on the CR item. These values are then used to determine the placement of the score points for CR items in the OIB and in calculating raw scores associated with setting a bookmark right after this item/score point.

Table 10-2 Excel Instructions for Calculating Ability (θ_n) for a Specified Response Probability (RP)

Column	*Excel Code/Instructions [Explanation]*
A	Enter values of θ from –4 to +4 in increments of .01 (i.e., –4.00, –3.99, –3.98, etc.).
B	Enter 1 in every row. [Numerator value for Step 0.]
C	= exp(value in Col. A – δ_{i1}). [Numerator value for Step 1.] Copy to remaining rows in this column.
D	= exp (2 * value in col. A – δ_{i1} – δ_{i2}). [Numerator value for Step 2.] Copy to remaining rows in this column.
E	= exp(3 * value in col. A – δ_{i1} – δ_{i2} – δ_{i3}). [Numerator value for Step 3.] Copy to remaining rows in this column.
F	= exp(4 * value in col. A – δ_{i1} – δ_{i2} – δ_{i3} – δ_{i4}). [Numerator value for Step 4.] Copy to remaining rows in this column.
G	= sum(values in col. B – F). [Denominator.] Copy to remaining rows in this column.
H	= (value in col. B)/(value in col. G). [Probability value for Step 0.] Copy to remaining rows in this column.
I	= (value in col. C)/(value in col. G). [Probability value for Step 1.] Copy to remaining rows in this column.
J	= (value in col. D)/(value in col. G). [Probability value for Step 2.] Copy to remaining rows in this column.
K	= (value in col. E) / (value in col. G). [Probability value for Step 3.] Copy to remaining rows in this column.
L	= (value in col. F)/(value in col. G). [Probability value for Step 4.] Copy to remaining rows in this column.
M	= sum(values in col. H – L). [Sum of the probability values.] Copy to remaining rows in this column. Note: This can be used as a check on the accuracy of calculated values. For any given value of θ_n, the sum of the probabilities should be 1.00.
N	= sum(values in col. I – L). [Probability of obtaining a score of 1 or better.] Copy to remaining rows in this column.
O	= sum(values in col. J – L). [Probability of obtaining a score of 2 or better.] Copy to remaining rows in this column.
P	= sum(values in col. K – L). [Probability of obtaining a score of 3 or better.] Copy to remaining rows in this column.
Q	= (value in col. L). [Probability of obtaining score of 4.] Copy to remaining rows in this column.

Other software programs can be used to calculate the RP50 and RP67 (or P2/3) values without displaying all the results from all the intermediate steps. In our experience, however, it is often helpful to be able to review all the intermediate values because they can be used to create item characteristic curves and to check the accuracy of results along the way. For example, as we alluded to previously, WINSTEPS produces a threshold value for each step of a CR item, which is equivalent to the RP50 value for the item. Table 10-3 shows a portion of an Excel spreadsheet for a set of calculations for a hypothetical 4-point CR item where the items for the test were scaled using the Rasch model. The step values associated with each of the four score points are provided at the bottom of the table. Figure 10-2 shows the response characteristic curves associated with each option for that item, and Figure 10-3 shows the curves associated with the probability of obtaining a given score or better on the same item.

Response Probabilities and Ordered Item Booklet Assembly—2PL Model

Mitzel et al. (2001) note that the probability of a correct response, $p(x = 1)$, to a given SR item is a function of examinee ability (θ), item difficulty (b_j), item discrimination (a_j), and a threshold or chance variable (c_j) in accordance with the fundamental equation of the three-parameter logistic (3PL) model:

$$p(x = 1|\theta) = c_j + (1 - c_j)/\{1 + \exp[-1.7a_j(\theta - b_j)]\} \quad \text{(Equation 10-22)}$$

where c_j is the lower asymptote or threshold value of the item (the likelihood that an extremely low-scoring student would answer correctly by guessing), a_j is the discrimination index of the item, and b_j is the difficulty of the item. In practice, Mitzel et al. (2001) and others using this model set the threshold or chance parameter (c_j) equal to zero, reducing Equation 10-22 to the following:

$$P_j(\theta) = 1/\{1 + \exp[-1.7a_j(\theta - b_j)]\} \quad \text{(Equation 10-23)}$$

or a two-parameter logistic (2PL) model.

In the procedure described by Mitzel et al. (2001), the basic standard-setting question is whether an examinee just barely qualified for a given performance level would have a 2/3 chance of answering a given SR item

Table 10-3 Selected Spreadsheet Entries and Calculations for a Hypothetical 4-point CR Item, Rasch Scaling

	Numerator					*Denom.*	*P*						*P*			
Theta	*0*	*1*	*2*	*3*	*4*	*Sum*	*0*	*1*	*2*	*3*	*4*	*Total*	*1 or Better*	*2 or Better*	*3 or Better*	*4*
–4.00	1.0000	0.0074	0.0001	0.0000	0.0000	1.0075	0.9926	0.0074	0.0001	0.0000	0.0000	1.0000	0.0074	0.0001	0.0000	0.0000
–3.99	1.0000	0.0075	0.0001	0.0000	0.0000	1.0076	0.9925	0.0075	0.0001	0.0000	0.0000	1.0000	0.0075	0.0001	0.0000	0.0000
–3.98	1.0000	0.0076	0.0001	0.0000	0.0000	1.0077	0.9924	0.0075	0.0001	0.0000	0.0000	1.0000	0.0076	0.0001	0.0000	0.0000
–3.97	1.0000	0.0077	0.0001	0.0000	0.0000	1.0077	0.9923	0.0076	0.0001	0.0000	0.0000	1.0000	0.0077	0.0001	0.0000	0.0000
–3.96	1.0000	0.0078	0.0001	0.0000	0.0000	1.0078	0.9922	0.0077	0.0001	0.0000	0.0000	1.0000	0.0078	0.0001	0.0000	0.0000
–3.95	1.0000	0.0078	0.0001	0.0000	0.0000	1.0079	0.9922	0.0078	0.0001	0.0000	0.0000	1.0000	0.0078	0.0001	0.0000	0.0000
–3.94	1.0000	0.0079	0.0001	0.0000	0.0000	1.0080	0.9921	0.0078	0.0001	0.0000	0.0000	1.0000	0.0079	0.0001	0.0000	0.0000
–3.93	1.0000	0.0080	0.0001	0.0000	0.0000	1.0081	0.9920	0.0079	0.0001	0.0000	0.0000	1.0000	0.0080	0.0001	0.0000	0.0000
–3.92	1.0000	0.0081	0.0001	0.0000	0.0000	1.0081	0.9919	0.0080	0.0001	0.0000	0.0000	1.0000	0.0081	0.0001	0.0000	0.0000
–3.91	1.0000	0.0081	0.0001	0.0000	0.0000	1.0082	0.9919	0.0081	0.0001	0.0000	0.0000	1.0000	0.0081	0.0001	0.0000	0.0000
–3.90	1.0000	0.0082	0.0001	0.0000	0.0000	1.0083	0.9918	0.0082	0.0001	0.0000	0.0000	1.0000	0.0082	0.0001	0.0000	0.0000
1.64	1.0000	2.0959	4.3929	5.1039	2.3164	14.909	0.0671	0.1406	0.2946	0.3423	0.1554	1.0000	0.9329	0.7923	**0.4977**	0.1554
1.65	1.0000	2.1170	4.4817	5.2593	2.4109	15.268	0.0655	0.1386	0.2935	0.3444	0.1579	1.0000	0.9345	0.7959	**0.5023**	0.1579
1.66	1.0000	2.1383	4.5722	5.4195	2.5093	15.639	0.0639	0.1367	0.2924	0.3465	0.1604	1.0000	0.9361	0.7993	0.5070	0.1604
1.67	1.0000	2.1598	4.6646	5.5845	2.6117	16.020	0.0624	0.1348	0.2912	0.3486	0.1630	1.0000	0.9376	0.8028	0.5116	0.1630
1.68	1.0000	2.1815	4.7588	5.7546	2.7183	16.413	0.0609	0.1329	0.2899	0.3506	0.1656	1.0000	0.9391	0.8062	0.5162	0.1656
1.69	1.0000	2.2034	4.8550	5.9299	2.8292	16.817	0.0595	0.1310	0.2887	0.3526	0.1682	1.0000	0.9405	0.8095	0.5208	0.1682

	Numerator					Denom.	P						P			
Theta	0	1	2	3	4	Sum	0	1	2	3	4	Total	1 or Better	2 or Better	3 or Better	4
1.70	1.0000	2.2255	4.9530	6.1104	2.9447	17.233	0.0580	0.1291	0.2874	0.3546	0.1709	1.0000	0.9420	0.8128	0.5254	0.1709
1.71	1.0000	2.2479	5.0531	6.2965	3.0649	17.662	0.0566	0.1273	0.2861	0.3565	0.1735	1.0000	0.9434	0.8161	0.5300	0.1735
1.72	1.0000	2.2705	5.1552	6.4883	3.1899	18.103	0.0552	0.1254	0.2848	0.3584	0.1762	1.0000	0.9448	0.8193	0.5346	0.1762
1.73	1.0000	2.2933	5.2593	6.6859	3.3201	18.558	0.0539	0.1236	0.2834	0.3603	0.1789	1.0000	0.9461	0.8225	0.5392	0.1789
1.74	1.0000	2.3164	5.3656	6.8895	3.4556	19.027	0.0526	0.1217	0.2820	0.3621	0.1816	1.0000	0.9474	0.8257	0.5437	0.1816
1.75	1.0000	2.3396	5.4739	7.0993	3.5966	19.509	0.0513	0.1199	0.2806	0.3639	0.1844	1.0000	0.9487	0.8288	0.5482	0.1844
1.76	1.0000	2.3632	5.5845	7.3155	3.7434	20.006	0.0500	0.1181	0.2791	0.3657	0.1871	1.0000	0.9500	0.8319	0.5528	0.1871
1.77	1.0000	2.3869	5.6973	7.5383	3.8962	20.518	0.0487	0.1163	0.2777	0.3674	0.1899	1.0000	0.9513	0.8349	0.5573	0.1899
1.78	1.0000	2.4109	5.8124	7.7679	4.0552	21.046	0.0475	0.1146	0.2762	0.3691	0.1927	1.0000	0.9525	0.8379	0.5618	0.1927
1.79	1.0000	2.4351	5.9299	8.0045	4.2207	21.590	0.0463	0.1128	0.2747	0.3707	0.1955	1.0000	0.9537	0.8409	0.5662	0.1955
1.80	1.0000	2.4596	6.0496	8.2482	4.3929	22.150	0.0451	0.1110	0.2731	0.3724	0.1983	1.0000	0.9549	0.8438	0.5707	0.1983
1.81	1.0000	2.4843	6.1719	8.4994	4.5722	22.727	0.0440	0.1093	0.2716	0.3740	0.2012	1.0000	0.9560	0.8467	0.5751	0.2012
1.82	1.0000	2.5093	6.2965	8.7583	4.7588	23.322	0.0429	0.1076	0.2700	0.3755	0.2040	1.0000	0.9571	0.8495	0.5796	0.2040
1.83	1.0000	2.5345	6.4237	9.0250	4.9530	23.936	0.0418	0.1059	0.2684	0.3770	0.2069	1.0000	0.9582	0.8523	0.5840	0.2069
1.84	1.0000	2.5600	6.5535	9.2999	5.1552	24.568	0.0407	0.1042	0.2667	0.3785	0.2098	1.0000	0.9593	0.8551	0.5884	0.2098
1.85	1.0000	2.5857	6.6859	9.5831	5.3656	25.220	0.0397	0.1025	0.2651	0.3800	0.2127	1.0000	0.9603	0.8578	0.5927	0.2127
1.86	1.0000	2.6117	6.8210	9.8749	5.5845	25.892	0.0386	0.1009	0.2634	0.3814	0.2157	1.0000	0.9614	0.8605	0.5971	0.2157
1.87	1.0000	2.6379	6.9588	10.175	5.8124	26.584	0.0376	0.0992	0.2618	0.3828	0.2186	1.0000	0.9624	0.8632	0.6014	0.2186
1.88	1.0000	2.6645	7.0993	10.485	6.0496	27.299	0.0366	0.0976	0.2601	0.3841	0.2216	1.0000	0.9634	0.8658	0.6057	0.2216

(Continued)

Table 10-3 (Continued)

	Numerator					*Denom.*	*P*						*P*			
Theta	*0*	*1*	*2*	*3*	*4*	*Sum*	*0*	*1*	*2*	*3*	*4*	*Total*	*1 or Better*	*2 or Better*	*3 or Better*	*4*
1.89	1.0000	2.6912	7.2427	10.804	6.2965	28.035	0.0357	0.0960	0.2583	0.3854	0.2246	1.0000	0.9643	0.8683	0.6100	0.2246
1.90	1.0000	2.7183	7.3891	11.134	6.5535	28.794	0.0347	0.0944	0.2566	0.3867	0.2276	1.0000	0.9653	0.8709	0.6143	0.2276
1.90	1.0000	2.7183	7.3891	11.134	6.5535	28.794	0.0347	0.0944	0.2566	0.3867	0.2276	1.0000	0.9653	0.8709	0.6143	0.2276
1.91	1.0000	2.7456	7.5383	11.473	6.8210	29.577	0.0338	0.0928	0.2549	0.3879	0.2306	1.0000	0.9662	0.8734	0.6185	0.2306
1.92	1.0000	2.7732	7.6906	11.822	7.0993	30.385	0.0329	0.0913	0.2531	0.3891	0.2336	1.0000	0.9671	0.8758	0.6227	0.2336
1.93	1.0000	2.8011	7.8460	12.182	7.3891	31.218	0.0320	0.0897	0.2513	0.3902	0.2367	1.0000	0.9680	0.8782	0.6269	0.2367
1.94	1.0000	2.8292	8.0045	12.553	7.6906	32.077	0.0312	0.0882	0.2495	0.3913	0.2397	1.0000	0.9688	0.8806	0.6311	0.2397
1.95	1.0000	2.8577	8.1662	12.935	8.0045	32.964	0.0303	0.0867	0.2477	0.3924	0.2428	1.0000	0.9697	0.8830	0.6352	0.2428
1.96	1.0000	2.8864	8.3311	13.329	8.3311	33.878	0.0295	0.0852	0.2459	0.3935	0.2459	1.0000	0.9705	0.8853	0.6394	0.2459
1.97	1.0000	2.9154	8.4994	13.735	8.6711	34.821	0.0287	0.0837	0.2441	0.3945	0.2490	1.0000	0.9713	0.8876	0.6435	0.2490
1.98	1.0000	2.9447	8.6711	14.154	9.0250	35.794	0.0279	0.0823	0.2422	0.3954	0.2521	1.0000	0.9721	0.8898	0.6476	0.2521
1.99	1.0000	2.9743	8.8463	14.585	9.3933	36.799	0.0272	0.0808	0.2404	0.3963	0.2553	1.0000	0.9728	0.8920	0.6516	0.2553
2.00	1.0000	3.0042	9.0250	15.029	9.7767	37.835	0.0264	0.0794	0.2385	0.3972	0.2584	1.0000	0.9736	0.8942	0.6556	0.2584
2.01	1.0000	3.0344	9.2073	15.487	10.175	38.904	0.0257	0.0780	0.2367	0.3981	0.2616	1.0000	0.9743	0.8963	0.6596	0.2616
2.02	1.0000	3.0649	9.3933	15.958	10.591	40.007	0.0250	0.0766	0.2348	0.3989	0.2647	1.0000	0.9750	0.8984	0.6636	0.2647
2.03	1.0000	3.0957	9.5831	16.444	11.023	41.146	0.0243	0.0752	0.2329	0.3997	0.2679	1.0000	0.9757	0.9005	**0.6676**	0.2679

NOTE: Step values are .9, .9, 1.49, and 2.43 for score points 1, 2, 3, and 4, respectively.

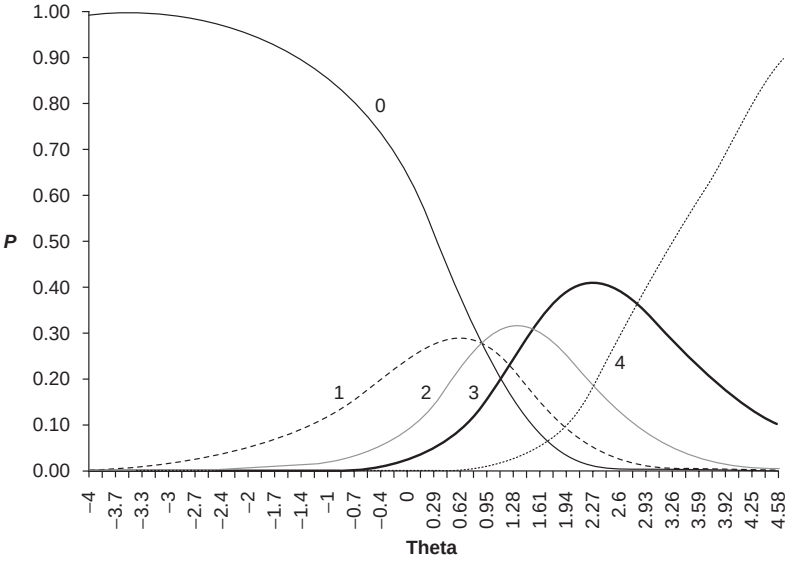


Figure 10-2 Response Characteristic Curves for Score Points 0–4 (based on data in Table 10-3, Rasch scaling)

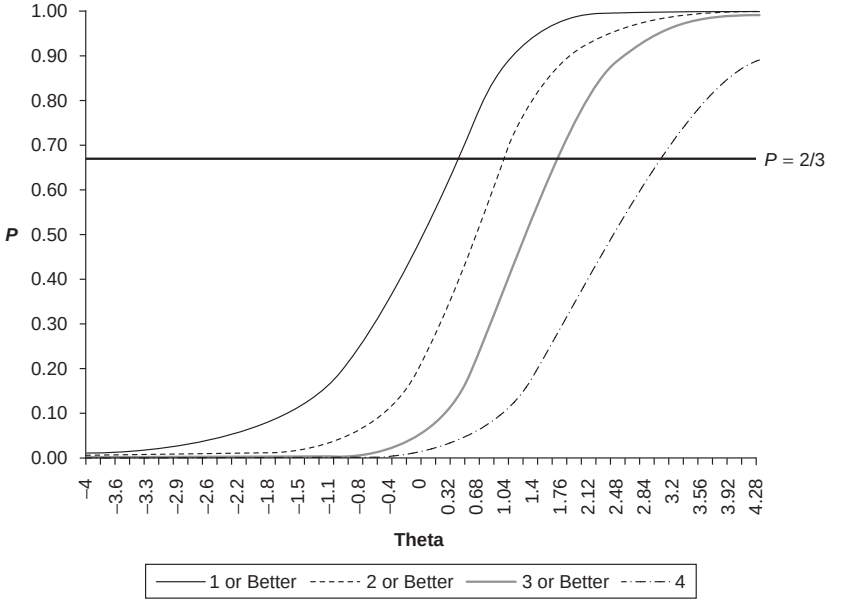


Figure 10-3 Probability of Obtaining a Given Score Point or Better as a Function of Ability (based on data in Table 10-3, Rasch scaling)

correctly. Thus, starting with a probability of 2/3 and solving for the ability (θ) needed to answer an item correctly, we obtain the following:

$$\theta = b_j + .693/1.7a_j \qquad \text{(Equation 10-24)}$$

(Again, had the RP been .67, rather than 2/3, the final result would have been $\theta = b_j + .708/1.7a_j$.)

For CR items, the situation becomes somewhat more complicated. Mitzel et al. (2001) used the two-parameter partial-credit (2PPC) model, with its fundamental equation relating the probability of obtaining score point k to student ability [$P_{jk}(\theta)$] and score point (step) difficulty (γ):

$$P_{jk}(\theta) = \exp(z_{jk})/\Sigma\exp(z_{ji}) \qquad \text{(Equation 10-25)}$$

where m_j is the number of score points or steps for item *j*,

$$z_{jk} = (k - 1)\alpha_j - \Sigma\, \gamma_{ji}; \qquad \text{(Equation 10-26)}$$

α_j is the discrimination index of item j;

k is the number of this score point or step; and

γ_{ji} is the step value for item *j* at step *i*.

Thus the probability of scoring at step *k* is a joint function of examinee ability, item discrimination, and the likelihood of obtaining any of the *k*-1 other scores. In this formulation, the value for a score of 0 (step 0) is set equal to zero; that is, $\gamma_{j0} = 0$ for all items. Procedures similar to those for establishing values of θ for each score point for each CR item within a Rasch framework can be established for the 2PL model.

As we illustrated in the Rasch context, we provide a portion of an Excel spreadsheet for a set of calculations, in this case for a hypothetical 3-point CR item, when a 2PL model is used. The spreadsheet appears as Table 10-4; the step values associated with each of the three score points are provided at the bottom of the table. And, also as before, we illustrate the response characteristic curves associated with each option for that item (Figure 10-4) and the curves associated with the probability of obtaining a given score or better on the item (Figure 10-5).

Directions to Bookmark Participants

As with other standard-setting methods, the selection and training of participants is an important aspect of the process. And, as with other

Table 10-4 Selected Spreadsheet Entries and Calculations for a Hypothetical 3-point CR Item, 2PL Scaling

	Numerator				*Denom.*	*P*				*P*			
Theta	*0*	*1*	*2*	*3*	*Sum*	*0*	*1*	*2*	*3*	*1 or Total*	*2 or Better*	*Better*	*3*
−4.00	1	0.087	0.002	0.000	1.089	0.918	0.080	0.002	0.000	1.000	0.082	0.002	0.000
−3.99	1	0.087	0.002	0.000	1.089	0.918	0.080	0.002	0.000	1.000	0.082	0.002	0.000
−3.98	1	0.088	0.002	0.000	1.090	0.917	0.081	0.002	0.000	1.000	0.083	0.002	0.000
−3.97	1	0.088	0.002	0.000	1.091	0.917	0.081	0.002	0.000	1.000	0.083	0.002	0.000
−3.96	1	0.089	0.002	0.000	1.091	0.916	0.082	0.002	0.000	1.000	0.084	0.002	0.000
0.14	1	1.371	0.549	0.074	2.993	0.334	0.458	0.183	0.025	1.000	0.666	0.208	0.025
0.15	1	1.380	0.556	0.075	3.011	0.332	0.458	0.185	0.025	1.000	**0.668**	0.210	0.025
0.16	1	1.389	0.563	0.077	3.030	0.330	0.459	0.186	0.025	1.000	0.670	0.211	0.025
0.17	1	1.399	0.571	0.078	3.048	0.328	0.459	0.187	0.026	1.000	0.672	0.213	0.026
2.15	1	5.239	8.013	4.115	18.367	0.054	0.285	0.436	0.224	1.000	0.946	0.660	0.224
2.16	1	5.274	8.120	4.198	18.593	0.054	0.284	0.437	0.226	1.000	0.946	0.663	0.226
2.17	1	5.310	8.229	4.283	18.822	0.053	0.282	0.437	0.228	1.000	0.947	0.665	0.228
2.18	1	5.345	8.340	4.370	19.055	0.052	0.281	0.438	0.229	1.000	0.948	**0.667**	0.229
2.19	1	5.381	8.452	4.458	19.291	0.052	0.279	0.438	0.231	1.000	0.948	0.669	0.231
2.20	1	5.417	8.565	4.548	19.531	0.051	0.277	0.439	0.233	1.000	0.949	0.671	0.233
4.39	1	23.342	159.047	363.925	547.313	0.002	0.043	0.291	0.665	1.000	0.998	0.956	0.665
4.40	1	23.498	161.183	371.280	556.961	0.002	0.042	0.289	0.667	1.000	0.998	0.956	**0.667**
4.41	1	23.655	163.348	378.784	566.787	0.002	0.042	0.288	0.668	1.000	0.998	0.957	0.668
4.42	1	23.813	165.541	386.440	576.795	0.002	0.041	0.287	0.670	1.000	0.998	0.957	0.670
4.43	1	23.973	167.764	394.251	586.988	0.002	0.041	0.286	0.672	1.000	0.998	0.957	0.672
4.44	1	24.133	170.017	402.219	597.370	0.002	0.040	0.285	0.673	1.000	0.998	0.958	0.673

NOTE: Step values are -0.33, 1.513, and 3.149 for score points 1, 2, and 3, respectively.

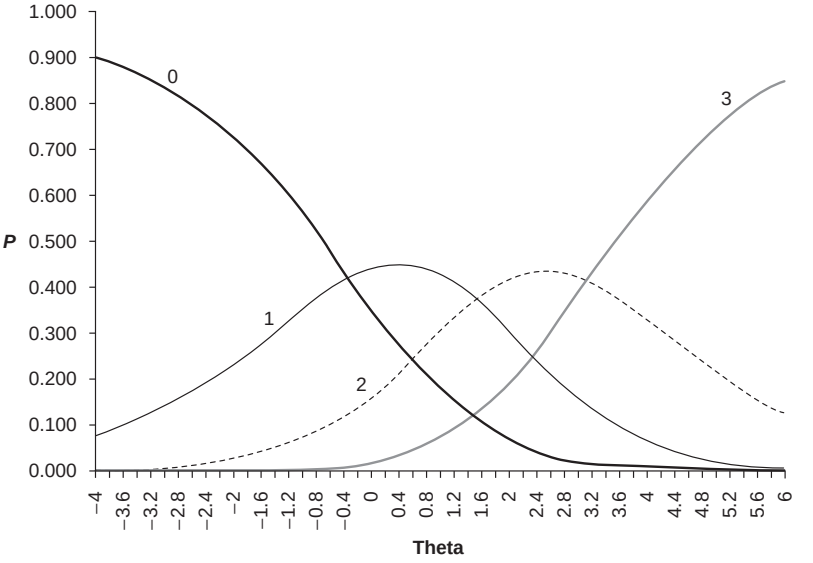


Figure 10-4 Response Characteristic Curves for Score Points 0–3 (based on data in Table 10-4, 2PL scaling)

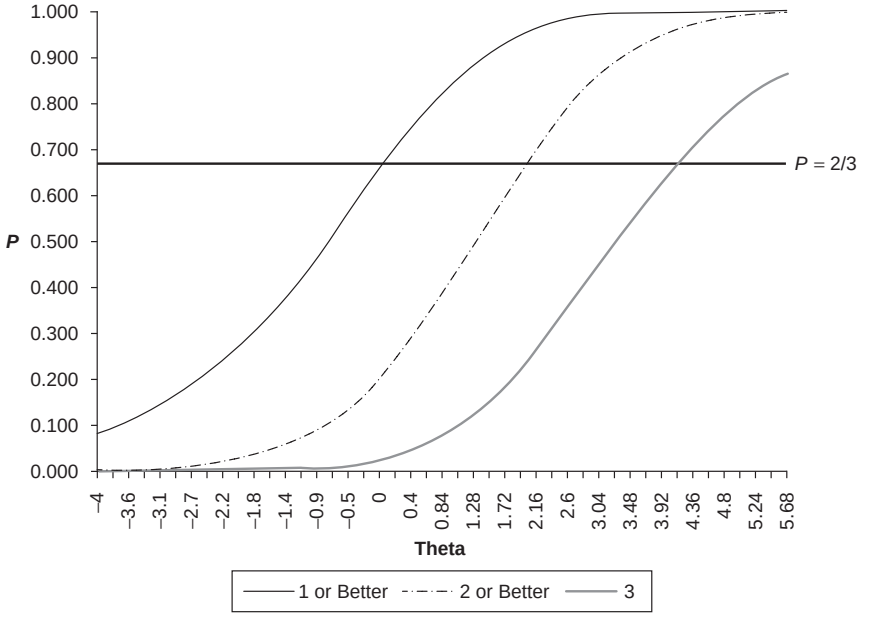


Figure 10-5 Probability of Obtaining a Given Score Point or Better as a Function of Ability (based on data in Table 10-4, 2PL scaling)

methods, when the Bookmark method is used participants must gain a clear understanding of the judgment task they are to perform.

The task presented to participants in a Bookmark standard-setting procedure is straightforward. Using the OIB assembled with one item (or score point) on each page, they are instructed to indicate the point at which they judge that the borderline or minimally qualified examinee's chances of answering the item correctly (or obtaining the score point) fall below the specified response probability or decision rule. For example, if a 2/3 decision rule is used, participants beginning to work through the OIB would ordinarily judge that the minimally qualified examinee would have better than a 2/3 likelihood of answering items at the beginning of the OIB (i.e., the easiest items) correctly. At some point in the OIB, however, participants would begin to discern that the chances of the minimally qualified examinee answering correctly approach and begin to drop below 2/3. Participants are instructed to indicate the point in the OIB at which the chances of the minimally qualified examinee answering correctly drop below 2/3. They indicate this judgment by placing a page marker—often a self-adhesive note or similar indicator—on the first page in the OIB at which the chance drops below the criterion. That is, the participants are indicating that the items prior to the marker represent content that the minimally qualified examinee would be expected to master at the RP or decision rule specified.

Standard-setting panelists generally work in small groups, evaluating the contents of small clusters of items as they appear in the difficulty-ordered test booklet. They discuss what makes one item or group of items more difficult than those that preceded it and ultimately place a bookmark at a point where they believe the difficulty of the subsequent items exceeds the ability of an identified group of students. In standard-setting contexts where more than one cut score is required (e.g., *Basic, Proficient,* and *Advanced*), participants would begin with the first item and ask themselves if a minimally qualified student (or group of students) at a particular achievement level (e.g., just barely *Basic*) would have the specified chance of answering the item correctly. They would then ask themselves the same question for each subsequent item until they reached one where they could not answer affirmatively. The final item yielding an affirmative response would mark the boundary of that performance level, and the participants would place a bookmark at that point (i.e., after the last attainable item). After making that judgment for the *Basic* category, participants would continue examining items beyond the bookmark just placed in order to identify the *Proficient* cut score, and so on for each cut score required.

Calculating Bookmark Cut Scores

Once participants have expressed their judgments by placing one (or more) bookmarks in the OIB, these judgments can be translated into cut scores. In a traditional Bookmark approach, the translation from bookmark placement to cut score is straightforward. For example, suppose that a participant has placed his or her bookmark on page 39 of a 50-page OIB to distinguish between *Proficient* and *Advanced* achievement levels. This does *not* correspond to a raw cut score of 39; rather, as we have indicated previously, this mark signals the participant's judgment that examinees classified as *Advanced* would be expected to be successful (defined in terms of whatever decision rule is being used) on the items through page 39 in the OIB. Of course, the examinee would also have some probability of success on the items after page 39. To obtain the cut score, the ability level associated with RP67 (or whatever decision rule is in place) that corresponds with the page in the OIB on which the bookmark was placed is the cut score, expressed in ability scale (i.e., theta) units. In this example, the theta associated with RP67 for the item appearing on page 39 of the OIB is the recommended cut score. From this point, it is a simple matter to transform the theta value to the raw score metric via the test characteristic curve or to another scaled score metric using the appropriate raw-to-scaled-score or theta-to-scaled-score conversion equation.

In the illustration in the preceding paragraph, the Bookmark cut score for a performance level was based only on the bookmark placement of a single participant. Obviously, in any application of the Bookmark standard-setting method, the procedure will be implemented using a large panel of participants. In the usual case, the bookmark placements of the participants will vary. In our experience, the typical method for addressing this situation is to find, for each participant, the theta (ability) level associated with the page in the OIB immediately preceding the one on which the participant's bookmark was placed in the same manner as just described. The result is a distribution of theta values, one for each participant. The overall recommended cut score in theta units is derived by taking the mean of these theta values and then obtaining the cut score in raw (or scaled score) units using one of the methods described in the preceding paragraph. We note, however, that the choice of central tendency measure most frequently used—that is, the mean—is perhaps based largely on statistical tradition and that the use of another statistic such as the median would likely be equally appropriate.

Finally, a point of clarification is in order here regarding the actual placement of the bookmark and the corresponding ability value that is used in cut score calculations. As we have described, when a Bookmark procedure

is conducted, participants are instructed to place their bookmark on the last page for which the participant could answer affirmatively the standard-setting question "Would an examinee just barely at this level have a 2/3 chance of answering this item correctly?" However, in other standard-setting sessions, participants are sometimes told to place their bookmarks on the first page for which the answer to this question is "No." If, for example, a participant answered "Yes" regarding the item on page 27 and "No" for the item appearing on page 28 of the OIB, some facilitators would have the participant place the bookmark on page 27, and some would have the participant place the bookmark on page 28. Strictly speaking, if the participant were actually placing a bookmark in a book, it would be placed between pages 27 and 28. In the example we have described here, the correct theta value for use in calculating the Bookmark cut score is the one found on page 27 of the OIB, not the one on page 28.

Our point here is that, regardless of the instructions given, it should be made clear to all involved (facilitators and participants) what is intended when a bookmark is placed in a given location: The most difficult item for which the participant can answer the standard-setting question affirmatively is the item whose values are entered into Bookmark cut-score calculations, whether participants identify that item by placing a bookmark on it, after it, or on the next page.

An Implementation of the Bookmark Procedure

Much of the terrain covered in previous sections of this chapter has outlined the mathematical foundations of the Bookmark standard-setting procedure. In this portion of the chapter, we seek to illustrate a typical Rasch-based application of the Bookmark method of the sort that is commonly used in the context of standards-referenced K–12 student achievement testing. In the illustration presented in the following paragraphs, we describe many practical aspects of the method, including training, presentation of the ordered booklet, and rounds of ratings.

Training

Training for a Bookmark standard-setting activity typically involves familiarizing participants with the performance level descriptions (PLDs), the test on which performance standards will be set, and the Bookmark standard-setting procedure. A sample agenda for a three-day session using the Bookmark method is shown in Figure 10-6. During the first day, participants

Day 1		
	8:00 A.M.	Registration, breakfast
	8:30	Introductions; distribute materials; collect security forms
	8:45	Background and overview
	10:00	Break
	10:15	Test administration
	12:30	Lunch
	1:30 P.M.	Test scoring and discussion
	3:00	Review of performance level descriptors
	4:00	Adjourn
Day 2		
	8:00 A.M.	Breakfast
	8:30	Distribute materials; introduction to the Bookmark procedure
	10:00	Break
	10:15	Practice round; evaluation of readiness
	11:00	Questions & answers
	Noon	Lunch
	1:00 P.M.	Instructions for Round 1
	1:15	Round 1
	3:45	Wrap-up
	4:00	Adjourn
Day 3		
	8:00 A.M.	Breakfast
	8:30	Distribute materials; review of Round 1 results
	9:45	Round 2
	Noon	Lunch
	1:00 P.M.	Discussion of Round 2 results
	1:30	Round 3
	3:00	Final recommendations
	3:30	Closure; evaluation
	4:00	Adjourn

Figure 10-6 Sample Agenda for Bookmark Standard-Setting Procedure

receive an overview of the purpose of the session and their objectives. This overview is followed by administration and scoring of the tests in order to give participants a clear understanding of the test contents. This activity is then followed by presentation and discussion of the PLDs. Placing the PLDs after the administration and scoring of the test helps participants view the content of the PLDs in a real-world context. Once participants understand and can articulate key components of the PLDs, they are given an opportunity to narrow the definition of each level to apply to those just barely in each performance level, that is, students at the threshold or cut score for that level.

On the morning of the second day, participants receive training in the specifics of the Bookmark method, followed by a short practice round in

which they place bookmarks for one cut score. A complete set of sample training materials that can be adapted to differing contexts is available at www.sagepub.com/cizek/bookmarktraining. After completing the practice activities, participants discuss their experiences and complete an evaluation form to assess their understanding of the training and their readiness to begin the bookmarking tasks.

Introducing the Ordered Item Booklet

As described previously, the OIB for a Bookmark standard-setting procedure consists of a series of SR and CR items in difficulty order, with the easiest item on the first page and the most difficult item on the last page. It is worth noting at this point that in the Rasch model it makes no difference whether the items are ordered by difficulty or by ability required to have a 2/3 chance of correct response; either method will result in the same ordering. If a 3PL or 2PPC model is used however, item difficulty and required ability will not necessarily order the items in the same way because the required ability is a function of both item difficulty and discrimination. Given two items of equal difficulty, the item with the lower discrimination index will require the higher ability to yield a 2/3 chance of correct response. (Recall that $\theta = b_j + .693/1.7a_j$, so that as a_j increases, the right side of the equation decreases.) Under these circumstances, a more difficult item might precede a less difficult item by several pages in an OIB ordered by theta, rather than by difficulty. In our experience, participants in a Bookmark procedure, who are usually far more sensitive to item difficulty than to discrimination, can be confused by an ordering based on theta values; thus it seems preferable to order booklets strictly by item difficulty.

Figure 10-7 shows an enlargement of a single page in an OIB. Information on the page includes page number, original item number and score point, and the Rasch achievement level required for a 2/3 chance to answer the item correctly. The key (A) is placed at the bottom of the page in a smaller font to serve as a quick check on the participant's own response to the item without interfering with the participant's estimation of the difficulty of the item. In practice, because items associated with a given stimulus (e.g., reading passages, graphics for sets of science or geography items, etc.) are likely to vary widely in difficulty and therefore be scattered throughout the test booklet, all common stimulus materials are placed in a companion booklet. The companion booklet is distributed to participants along with the difficulty-ordered test booklet.

The OIB page shown in Figure 10-7 contains all the information a participant would need to make a judgment about the item. All the information is printed at the top of the page so that it will be easily accessible to

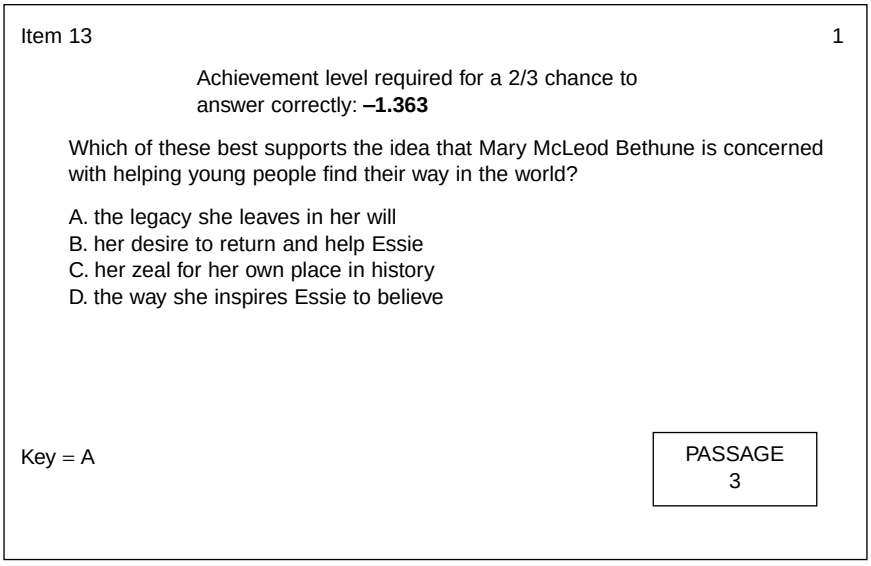


Figure 10-7 Sample Page From Ordered Item Booklet

participants. As can be seen in the figure, this item appears on page 1 of the OIB (as indicated by the numeral 1 in the box at the upper right corner of the page). This page number is boldfaced and of a larger size that makes it clearly distinguishable from other numbers on the page; this is important because participants use the page number as their indicator for a bookmark placement. The figure also shows that this item appeared as Item 13 in the actual test form, as indicated by the 13 printed in the upper left corner of the page. Also printed on the page is the achievement level (i.e., ability or theta) required for an examinee to have a 2/3 chance of answering this item correctly assuming (as is true in the sample page shown) that the item is an SR format item. In the sample page shown in the figure, the ability required (expressed in logits) is –1.363. If the item on this page had been a CR format item, the ability level expressed in logits would be the value associated with a 2/3 chance of obtaining that particular raw score point or higher. These values are obtained as described previously in this chapter.

Round One of a Bookmark Procedure

After an introduction to the procedure, each participant receives an OIB, a stimulus booklet, and a set of bookmarks. As mentioned earlier, the OIB

has one item per page, starting with the easiest item in the test booklet; each page contains information like that shown in Figure 10-7. Each CR item is represented once for each of its score points, as noted previously. Each CR page contains the item and one or more sample responses that are exemplars of the particular score point. Because there are several different ways to earn each score point, it is often a good idea to select sample responses that cover a broad range of possibilities across the various CR items.

For tests that have common stimuli (e.g., reading passages, maps, graphs), a separate stimulus booklet is prepared and distributed to participants. In an OIB, items for a given scenario, map, case, chart, or other stimulus are scattered throughout the booklet. To simplify the task participants face in matching items in the OIB with their associated stimuli, it is helpful to create a code for each stimulus and then repeat that code at the beginning of the corresponding item in the OIB. The box in the bottom right corner of Figure 10-7 provides a correspondence between that item and its associated stimulus (in this case, Passage 3).

Each participant also receives a printed form on which to enter his or her bookmarks (page numbers). The forms are printed on one side of a piece of card stock. Each form is similar to the one shown in Figure 10-8. In Rounds 1 and 2, participants enter the page number for each bookmark. At Round 3, participants will be familiar with the relationship between page number and cut score. At this stage, participants may enter page numbers and associated cut scores, as well as the impact data. The purpose of asking each participant to also enter the impact data is to help ensure that each participant is fully cognizant of the consequences that his or her recommendations will have in terms of the percentages of examinees that would be classified into each of the performance categories if the participants' cut scores were applied to actual test results.

During Round 1, participants usually work in small groups of three to five individuals. While they discuss the item contents among themselves, each participant completes his or her own Bookmark recording form like the one shown in Figure 10-8. As they complete Round 1, participants review their forms to make sure they are complete, return all materials to the facilitator, and are dismissed for the day.

Obtaining Preliminary Bookmark Cut Scores

At the end of Round 1 (and following rounds), standard-setting staff collect participants' bookmark cards and enter the values from the cards into a spreadsheet similar to the one shown in Table 10-5. After verifying the accuracy of the results, meeting facilitators return the cards to the

Panelist Number _____

Directions: Enter your Bookmark page numbers for each performance level in the spaces below.

ROUND 1

	Basic	***Proficient***	***Advanced***
Page Numbers			

ROUND 2

	Basic	***Proficient***	***Advanced***
Page Numbers			

ROUND 3

	Basic	***Proficient***	***Advanced***
Page Numbers			
Cut Scores			
% At or Above			

Notes:

Figure 10-8 Sample Bookmark Participant Recording Form

participants, along with the results. The sample information shown in Table 10-5 allows participants to see where their bookmarks fall relative to those of other participants. It also gives them a sense of where the group average lies, as well as how far their own bookmarks fall from the group average.

Table 10-5 provides a summary of bookmark placements, in addition to the resulting cut scores. Also shown are the mean cut score (along with its standard deviation), the minimum and maximum recommended cut scores for each performance level, and cut scores one standard deviation above and one standard deviation below the mean recommended cut scores. Individual cut scores in raw score units are not shown, but means, medians, minimum, and maximum cut scores in raw score units are provided.

Table 10-5 Sample Output From Round 1 of Bookmark Standard-Setting Procedure

	Basic		*Proficient*		*Advanced*	
Participant ID No.	*Page in OIB*	*Theta @ Cut*	*Page in OIB*	*Theta @ Cut*	*Page in OIB*	*Theta @ Cut*
1	5	–0.334	12	0.286	46	1.627
2	8	0.082	22	0.600	46	1.627
3	8	0.082	22	0.600	47	1.650
4	6	–0.243	22	0.600	46	1.627
5	10	0.270	18	0.551	38	1.333
6	6	–0.243	16	0.493	39	1.340
7	9	0.193	21	0.579	40	1.489
8	6	–0.243	16	0.493	39	1.340
9	7	–0.176	16	0.493	40	1.489
10	8	0.082	16	0.493	40	1.489
11	7	–0.176	16	0.493	43	1.586
12	8	0.082	16	0.493	41	1.510
13	9	0.193	32	1.046	42	1.580
14	9	0.193	23	0.616	46	1.627
15	9	0.193	26	0.891	39	1.340
16	9	0.193	14	0.440	42	1.580
17	13	0.420	19	0.558	38	1.333
18	8	0.082	13	0.420	22	0.600
19	10	0.270	17	0.540	39	1.340
20	11	0.272	17	0.540	39	1.340
Summary Statistics in Theta (Ability) Metric						
Mean cut		0.060		0.561		1.442
Median cut		0.082		0.540		1.489
SD		0.217		0.161		0.233
Minimum		–0.334		0.286		0.600
Maximum		0.420		1.046		1.650
Mean – 1SD		–0.158		0.401		1.209
Mean + 1SD		0.277		0.722		1.676
Summary Statistics in Raw Score Metric						
Mean cut		22.04		28.73		39.87
Median cut		22.31		28.44		40.32
Minimum		18.00		25.00		30.00
Maximum		27.00		36.00		42.00
Mean – 1SD		19.46		26.51		37.34
Mean + 1SD		24.83		30.99		42.00

The translation of cut scores in the theta metric to a cut score in raw score units is a relatively straightforward process. Programs such as WINSTEPS or other IRT-based programs (e.g., PARDUX, PARSCALE, etc.) produce a conversion table showing raw scores and associated theta values. Using the calculated mean values for thetas at the three performance levels illustrated in Table 10-5, each of the three thetas is located in the conversion table and the closest raw score is obtained (or interpolated). Because a precise correspondence between the exact theta cut and raw cut will almost never be observed, the board or entity responsible for the performance standards, in advance of standard setting, will need to make a policy decision regarding whether to take the closest raw score, the raw score with an associated theta value just below the calculated mean theta, the raw score with associated theta value just above the calculated mean theta, or some other value. As we have urged previously, such decisions should also be documented, along with the rationale behind them.

A Caveat and Caution Concerning Bookmark Cut Scores

We digress for a moment from our description of this specific Bookmark implementation to offer a clarification and caution regarding how Bookmark cut scores are obtained. Indeed, we have seen a variety of applications of the Bookmark procedure in which alternative mechanisms for calculating a cut score have been employed. For example, in some applications of the Bookmark standard-setting procedure, the cut scores have been obtained by simply taking the mean recommended page number in the OIB and translating that number into a raw score. For instance, if the mean page number were 29, 29 could be taken as the cut score. The rationale for doing so would be that, on average, participants thought the minimally qualified examinee at that level would have a 2/3 chance of answering the first 29 items correctly. Such a procedure is ill-advised, however, and closer examination of the logic behind the appropriate procedure seems warranted.

The logic of setting a cut score at the raw score associated with the mean theta identified by participants is this: Participants place their bookmarks on the last item in the OIB for which they believe a minimally qualified examinee has a 2/3 chance of answering correctly. Minimally qualified examinees will still have some chance of answering subsequent items correctly, of course; right up to the end of the OIB, minimally qualified examinees (indeed nearly all examinees) will have some (very small) chance of answering each item correctly. Moreover, these examinees will have a greater than 2/3 chance of answering items correctly that appear prior to the location of their bookmarks.

IRT models are based on the notion that each examinee has a calculable probability of answering each item correctly (or obtaining any given score point on a CR item). The estimated raw score for a given theta is the sum of these probabilities and expected values. Thus, for example, if the mean theta value for *Proficient,* based on the estimates of the 20 participants represented in Table 10-5, is .561, then the Bookmark-based cut score for *Proficient* is 28.73, which is the interpolated value from the WINSTEPS output showing theta values associated with raw scores of 28 and 29. If we simply took the average page number as our cut, we would get a cut score of 18.7. If this value were used as the cut score (rounded to either 18 or 19; for purposes of this point, it doesn't matter which), there would be approximately a 10-raw-score point difference between this value and the correct value of 28.73—clearly a practically significant difference.

Round One Feedback to Participants

An example of one type of normative information provided to participants in a Bookmark standard-setting procedure is shown in Figure 10-9. At this point, only the page numbers bookmarked by the participants are shown. In this way, participants get a graphic view of how their bookmarks compare to the bookmarks of the other participants. The figure helps illustrate where there are gaps, that is, page ranges in which no participant chose to place a bookmark for any cut score. In subsequent rounds, the page ranges will typically not be the focus of attention; rather, discussion and consideration will center on the range of pages in which Round 1 judgments have indicated that the eventual cut score recommendations will likely be located.

Interestingly, Figure 10-9 also shows where there are overlaps in individual judgments. For example, one participant placed his bookmark at page 37 for the *Advanced* level, whereas another participant placed her bookmark for *Proficient* on page 39. In effect, one participant would set the cutoff for *Proficient* higher than at least one participant would set the cutoff for *Advanced*. Visualizations such as Figure 10-9 are excellent mechanisms for promoting the important discussions that will characterize Rounds 2 and 3.

In addition to normative information, impact information is also usually provided to participants in any standard-setting procedure. The juncture at which such information is provided varies, however. In this case, we illustrate the provision of impact information at the end of Round 1, although it can be introduced at the end of Rounds 1, 2, or 3. We note, however, that in our experience the later that impact information is presented to participants, the less an impact on participants' judgments it appears to have. The purpose of impact data is to allow participants to see how many (or what

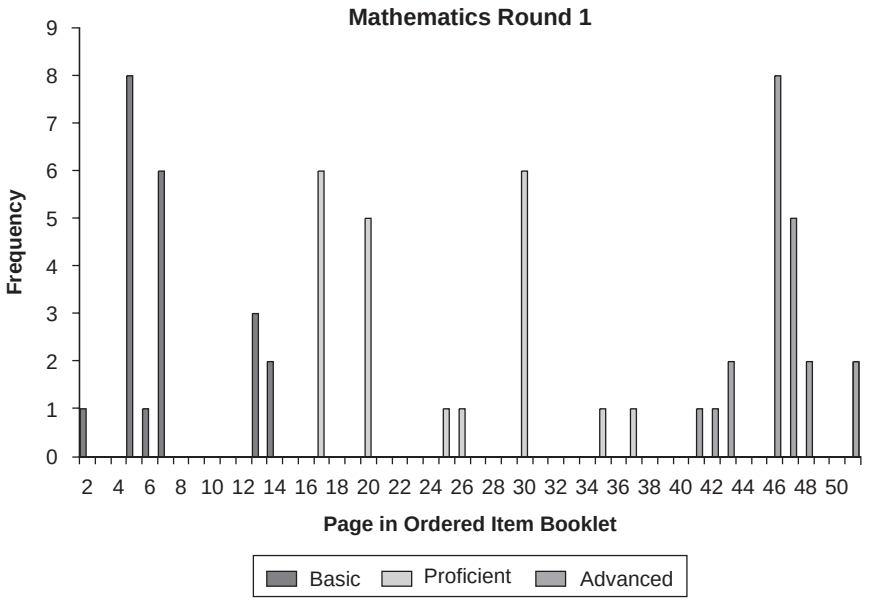


Figure 10-9 Sample Display of Round One Bookmark Placement Feedback

percentage of) examinees would be classified at each performance level if the mean cut scores from that round were implemented. An example of impact information is presented later in this chapter (see Table 10-6).

Round Two of a Bookmark Procedure

Following the schedule shown previously in Figure 10-6, the third day of the standard-setting session begins with participants receiving their OIBs and other materials from Round 1 plus the bookmark summary data and impact information from Round 1. The first activity is a discussion, led by meeting facilitators and centered on the Round 1 ratings and impact data. This discussion generally focuses on range of cut scores, areas of particular disagreement, and concerns about difficulty location of individual items. As part of this discussion, it is sometimes helpful for participants to explicitly address differences between their perceived difficulty of a particular item and the placement of that item relative to others in the OIB.

Once participants have discussed the results of Round 1 as a total group, they continue their work in small groups of three to five members to begin Round 2. The reassignment of participants to smaller groups may be random, or it may be done purposefully in order to bring divergent points

of view together at the same table. In either event, reassignment between rounds maximizes opportunities for participants to express their own—and hear others'—points of view. The participants' task for Round 2 is essentially identical to that of Round 1, consisting of (re)consideration of bookmark placements and the content of items captured by performance levels and discussion of those judgments with other small group members. The primary difference between Rounds 1 and 2 is the amount of information available to each participant. At the end of the second round of ratings, facilitators collect all materials and dismiss participants for lunch, during which facilitators again analyze the bookmark placements and prepare reports similar to those shown in Table 10-5 and Figure 10-9. This information is provided to participants at the beginning of Round 3.

Round Three of a Bookmark Procedure

To begin Round 3 of a Bookmark standard-setting procedure, participants again use their OIBs and are provided with all of their other Round 2 materials plus a summary of the Round 2 judgments. In our experience, it is at this point that a special version of Table 10-5 appears to be quite helpful to participants. An example of this version is shown in Table 10-6. The distinctive feature of Table 10-6 is that it includes actual raw score equivalents associated with the theta values that are the recommended cut scores. This feature helps clarify for participants the relationship between their bookmark placements, the theta values associated with those placements, and the impact that a bookmark placement (or changing a bookmark placement) will have on both the raw cut score and the percentages of examinees classified at or above a given performance level.

Round 3 begins with facilitators' leading a discussion of the impact data and other topics of concern from Round 2. At the end of this discussion, participants are asked to evaluate all of their previous ratings and all information at hand and to simply enter three bookmarks and the associated cut scores on their recording form (see Figure 10-8). At this stage, participants are actually asked to enter several pieces of data on their recording forms. Reviewing Figure 10-8 reveals that, in addition to the page number at which they have placed bookmarks for each performance level, participants are asked to enter the raw cut score associated with the page number and the corresponding percentage of examinees that would be classified at or above that level. The requirement that participants enter all three of these values for each cut score is an attempt to verify participants' understanding of the final task, to highlight the impact of the judgments, and to provide a check on the accuracy of the participants' intentions. The final task of

Table 10-6 Round 3 Feedback for Bookmark Standard-Setting Procedure

Page No. in OIB	*Original Item No.*	*IRT Item/Step Difficulty*	*Theta @ RP*	*Raw Cut Score*	*% At or Above*
1	6	–2.305	–1.612	8	99.61
2	2	–1.986	–1.293	10	99.11
3	3	–1.950	–1.257	10	99.11
4	12	–1.304	–0.611	15	95.66
5	14	–1.027	–0.334	18	92.13
6	11	–0.936	–0.243	19	90.62
7	1	–0.869	–0.176	20	88.86
8	28	–0.611	0.082	23	82.34
9	24	–0.500	0.193	24	79.66
10	15-1	0.480	0.270	25	76.47
11	18	–0.421	0.272	25	76.47
12	17	–0.407	0.286	25	76.47
13	26-1	0.790	0.420	27	68.80
14	5-1	0.650	0.440	27	68.80
15	8-1	0.350	0.440	27	68.80
16	7	–0.200	0.493	28	64.77
17	29-1	0.240	0.540	29	60.69
18	27	–0.142	0.551	29	60.69
19	34	–0.135	0.558	29	60.69
20	20	–0.124	0.569	29	60.69
21	13	–0.114	0.579	29	60.69
22	37-1	1.250	0.600	30	56.48
23	21	–0.077	0.616	30	56.48
24	15-2	0.750	0.740	32	47.81
25	26-2	0.320	0.810	33	43.56
26	16	0.198	0.891	34	39.34
27	33-1	0.620	0.900	34	39.34
28	37-2	0.090	0.910	34	39.34
29	30	0.252	0.945	35	35.18
30	36	0.295	0.988	35	35.18
31	38	0.305	0.998	35	35.18
32	4	0.353	1.046	36	31.15
33	15-3	0.320	1.090	36	31.15
34	35	0.464	1.157	37	27.20
35	9	0.498	1.191	38	23.39
36	5-2	0.050	1.200	38	23.39
37	26-3	1.650	1.290	39	19.64
38	10	0.640	1.333	39	19.64
39	32	0.647	1.340	39	19.64
40	19	0.796	1.489	41	12.45
41	37-3	1.530	1.510	41	12.45
42	8-2	0.630	1.580	42	9.61
43	22	0.893	1.586	42	9.61
44	31	0.896	1.589	42	9.61
45	26-4	–0.010	1.590	42	9.61
46	23	0.934	1.627	42	9.61
47	25	0.957	1.650	42	9.61
48	15-4	1.130	2.040	45	3.73
49	37-4	0.860	2.080	45	3.73
50	29-2	1.280	2.120	46	2.69
51	33-2	1.560	2.410	47	1.71

Round 3 occurs as participants complete evaluation forms (see Chapter 3) and are dismissed. Facilitators then check each completed bookmark for accuracy, tally these final ratings, and calculate the mean recommended cut score for each achievement level.

Alternative Procedures and Limitations

Given the time that has elapsed since its introduction, the limited amount of published research evidence about the Bookmark procedure is somewhat surprising (see Karantonis & Sireci, 2006). Then again, given the number of times the Bookmark procedure has been used, it is not surprising that, for the most part, many initial concerns about the method have been addressed via procedural changes based on these experiences. Yet one fundamental aspect of the procedure must be reckoned with each time it is employed: The cut score is absolutely bound by the relative difficulty of the test. It is this limitation that future research on the Bookmark method must address.

To see the impact of this limitation, we can consider the relative alignment of the difficulty of a test and the ability of the population of examinees who will take the test. If the test is easy, relative to the examinee population, it will be impossible to set a cut score below a certain point, no matter what any of the participants may wish. For example, let us assume that every participant placed a bookmark for Basic on page 1 of an OIB. On this (relatively) easy test, the ability level associated with an RP67 will very likely yield a cut score of two (or more) on the raw score scale. In a recent application, we found that the theta level for page 1 in a certain OIB yielded a raw score of 10! Similarly, a bookmark placed even on the last page of an OIB will not necessarily yield a raw cut score of 100% correct. We have even conducted bookmark procedures in which some participants would have preferred to go beyond the last page in the booklet for their final bookmark, claiming that the most difficult item in the booklet was not sufficiently difficult to distinguish the highest category of performance.

Of course, tests that are too easy or too difficult for the population of examinees present problems that are not unique to the Bookmark standard-setting method, but would pose problems for other methods as well. Indeed, this limitation may actually be a credit to the method in that it brings the limitation to light. To address the limitation, careful item writing and test construction procedures must be in place; standard-setting methods cannot compensate for weaknesses in content coverage, performance characteristics of items, and so on.

A second difficulty that arises in Bookmark applications has to do with unusually large gaps in difficulty between items. When the OIB comprises

items selected from a deep item pool (as opposed to a specific test form), this problem can generally be avoided. However, when the ordered item booklet is created directly from the operational test, it is likely that the dispersion of item difficulties will be uneven. This problem may cause difficulty for participants in the process, particularly when it is apparent that one of the cut scores falls in one of the gaps. For example, let us assume that five contiguous items in the OIB have the following RP67 theta values: Item 21 (1.41), Item 22 (1.53), Item 23 (1.62), Item 24 (2.04), and Item 25 (2.17). Participants examining this string of items may judge that Item 23 is well within the grasp of a barely *Proficient* examinee, but that Item 24 is far beyond the grasp of such an examinee. In such a case, participants may not wish to place their bookmark on either item, preferring instead—if it were possible—to place a bookmark somewhere between Items 23 and 24. Their predicament is to settle for placing a bookmark on Item 23, which may yield a cut score that is lower than the panel as a whole can support, or placing the bookmark on Item 24, which could result in a cut score higher than participants are comfortable recommending.

For these reasons, we recommend that special care be given to the development of the operational test if it will be used for Bookmark standard setting and that standard-setting issues be carefully considered early in the test-development process. It may be possible to forestall both problems (difficulty/ability mismatch and item difficulty gaps) through targeted test design. We recognize, of course, that final values for examinee ability and item characteristics can be known only after operational administration, making it especially critical that practitioners be aware of this potential problem and plan to avoid its consequences in advance.

Alternative procedures can be conceived to address these potential limitations, however. With regard to the item difficulty/examinee ability mismatch leading to unanticipated cut score placements, one simple approach essentially ignores the *b*-theta relationship. If this alternative is used, the page number in the OIB is taken directly as the raw cut score; that is, if a participant puts a bookmark on page 10, the recommended cut score is 10 points. Precisely this strategy was implemented in a study by Buckendahl, Smith, Impara, and Plake (2002); the authors reported that it worked well in the context of setting standards for a seventh-grade mathematics assessment in a midwestern school district.

Earlier in this chapter we described such an approach as an incorrect implementation of the Bookmark method. However, this alternative (a "modified Bookmark" method?) does provide a simple strategy and a reasonable alternative provided that all related training, materials, feedback, and so on are similarly realigned. The study cited in the previous

paragraph provides only limited support for this alternative; however, more research would be required before its use can be recommended.

Another alternative involves using classical items statistics (i.e., *p* values) instead of IRT values to order the OIB. Then, for each page in the OIB, a scale score can be assigned to each page in the OIB such that, for example, page 10 would have a scale score equivalent to 10 raw score points. Although this strategy appears to remedy the difficulty-ability mismatch, it also raises new questions. In essence, this approach resets the scale scores in ways that can have unforeseen consequences, and before recommending this strategy we await the results of research that will uncover the intended and unintended consequences of this ordering strategy. Another alternative—and one that research is needed to address—also involves the ordering of the OIB. In all applications of the Bookmark method we are aware of, the items in the OIB are compiled in increasing difficulty order. Sequencing items in the opposite order (i.e., from hardest to easiest) seems like a plausible alternative; research evidence that either ordering produces similar cut scores would add validity support for the method.

With regard to the item difficulty gap problem, as we indicated previously, a specially constructed OIB created from a bank with an abundance of items at every difficulty level is technically preferable. However, as we also mentioned, grounding standard setting in an actual operational test is also highly desirable. Between these options, there may be a midpoint. Should the operational test yield gaps that are likely to interfere with setting standards via the Bookmark procedure, it would seem prudent to identify the location of those gaps and insert a small number of items from the bank to supplement the operational test form and ameliorate the gaps. The key consideration here to be weighed is the tradeoff between measurement precision and fidelity to the operational form. Particularly if the OIB only added (a small number of) items to and took no items away from the operational booklet, the objection to this practice might be easily overcome. As with many of the other decision points we have illustrated, this is a policy issue that would need to be addressed early in the planning of the standard setting.

11

The Item-Descriptor Matching Method

In Chapter 3, we noted the increasing centrality of performance level descriptions (PLDs) in contemporary standard setting, particularly in the context of the standards-referenced tests (SRTs) now administered in every state to students in U.S. elementary and secondary schools. Although performance level labels (PLLs)—usually just one or two words—might have great rhetorical value, they do not contain as much specific information as PLDs regarding the knowledge and skills expected of examinees classified into a particular performance category. In contrast to PLLs, PLDs often consist of two or more paragraphs of detailed information about what examinees know and can do, and (sometimes by inference, sometimes explicitly) what they do not yet know or cannot do. PLDs form the foundation of many modern standard-setting methods and are one of the key referents that participants rely on when making whatever judgments a particular method requires.

In a sense, it may not be an exaggeration to claim that standards are set more by the panels who craft the PLDs than by those who rate items or performances. This claim is most defensible under two very common conditions:

1. when PLDs are highly detailed and include very specific statements about examinee abilities at the given performance levels; and
2. when a standard-setting panelist, in the course of making a judgment about an item or task in a test, relies—as he or she should—on the PLDs for a dispositive indication of how performance on the item or task relates to the performance levels.

It seems appropriate, then, that a standard-setting method that explicitly linked the PLDs and the judgmental task should be developed to meet the evolving challenges of standards-referenced testing, and it seems apparent that a standard-setting method that directly linked the PLDs to the resulting cut scores would present a strong validity argument for the results. It was in the context of SRTs and PLDs that the Item-Descriptor Matching (IDM; Ferrara, Perie, & Johnson, 2002b) method was developed.

According to the developers of the IDM method, the specific historical context for the development of the procedure began in 1991 with collaboration between a state testing agency (the Maryland State Department of Education) and a testing contractor (CTB/McGraw-Hill) to set standards for the state's performance assessment program. The procedure was revised in 1993, again involving collaboration between the state and another contractor working on the performance assessment system. IDM was further refined in 1999 when the American Institutes for Research (AIR) implemented the procedure in much the same form as it exists currently to set performance standards for high school end-of-course examinations in Philadelphia. Subsequently, the procedure has been used in other smaller scale applications in the United States, as well as in larger scale applications in Puerto Rico and Brazil.

In general, the IDM method requires participants to become familiar with the finalized descriptions of the performance levels into which examinees are to be classified (e.g., *Basic, Proficient,* and *Advanced*). Then, using the test form on which cut scores are needed, participants match the items (or tasks, or responses at various score points for constructed-response items) to those performance levels. The method shares much in common with other procedures, however. For example, the IDM procedure uses an ordered item booklet as does the Bookmark method; both methods can be thought of as special cases of item-mapping approaches. IDM focuses participants' judgments in areas of uncertainty about classifications, so it has something in common with the Borderline Groups methodology. Finally, the analytic procedures applied to the data yielded by an IDM method are similar to those used to identify cut scores in many of the holistic methods (see Chapter 9). Nonetheless, the distinctive features and procedures of IDM are sufficiently different to warrant individual treatment.

Procedures for the IDM Method

As with all other methods for setting performance standards, the defensibility and validity of the results of the IDM method depend on aspects of standard setting that occur in advance of the actual meeting. For example, participants

must be carefully selected for representativeness and familiarity with the examinee population, and they must be acquainted with the item and task formats and the content standards or objectives covered by the test on which standards will be set. Perhaps more so than other methods, however, high-quality, clear PLDs must be developed, and participants in an IDM procedure must be thoroughly grounded in those expectations for examinee performance in each of the performance levels. Indeed, according to the developers of the method, "The effectiveness of the ID Matching procedure depends heavily on the quality of the performance level descriptions" (Ferrara, Perie, & Johnson, 2002a, p. 12).

Implementing the IDM method begins when participants analyze items presented to them in an ordered item booklet (OIB). The OIB is prepared in the same way as it would be prepared if implementing a Bookmark approach. That is, the items in an operational test are ordered from easiest to most difficult, usually based on their *b* values when items have been calibrated using item response theory (IRT). If the test contains a mix of selected-response (e.g., multiple-choice) and constructed-response (e.g., essay) items, the multiple-choice items are assembled into the OIB in ascending *b* value order, and responses representing each score point for each constructed-response item are interspersed among the multiple-choice items at whatever points are dictated by the threshold difficulty of each score point. (See Chapter 10 for more detailed information on the assembly of an OIB.)

The essence of the judgmental task—and the aspect that makes the IDM method unique—lies in the specific directions given to participants. Participants begin at the beginning of the OIB and are instructed to carefully review each item (and, if relevant, constructed-response) and determine which performance level represents the best match to the item's "response requirements" (Ferrara et al., 2002a, p. 2). According to the developers of the IDM method,

> Item response requirements are the content area knowledge, content area skills, and cognitive processes required to respond successfully to an item. In ID Matching, panelists start with the first (easiest) item and identify which performance level its requirements most closely match. Panelists match each item's response requirements to a performance level description. They are trained to answer the following question for each item-descriptor match: "To which performance level description are the knowledge, skills, and cognitive processes required to respond successfully to this item most closely matched?" (p. 10)

On a special form provided to them with the OIB, participants are asked to indicate their matches in one column of the form. A sample form with hypothetical data and other information is shown in Figure 11-1. An electronic version of the form can be found at www.sagepub.com/cizek/IDMform.

The form shown in the Figure 11-1 was used in the context of a mathematics placement test to identify fourth-grade students in need of special summer remediation. The form shows that the test consisted of 33 items, 26 of which were multiple-choice format, and 7 were constructed-response items scored according to a rubric on which a range of 0 to 3 points could be awarded. It is also apparent that all of the items in the form are relatively easy (as might be expected given the purpose of the test), with *b* values for the items ranging from –2.13 to –0.11.

An essential characteristic of the form illustrated in Figure 11-1 is that each row of the form provides information about items in the OIB, and the rows of the form are ordered from top to bottom in the same way as the OIB. That is, the first row of the form provides information about the easiest item in the test form, the next row provides information about the next harder item, and so on. Multiple-choice format items can be recognized in the table because the number for an item is listed only once in the column labeled "Item Location in Test Form." Constructed-response format items can be recognized because the item number associated with each constructed-response item appears multiple times on the rating form. The number of times a single constructed-response item appears on the form is equal to one fewer time than the number of possible score points associated with the scoring rubric for that item. Thus, for the constructed-response items on this test, each of which is scored on a rubric with four score points (0, 1, 2, 3), each item appears in the OIB three times. For example, in Figure 11-1, Item 16 is a constructed-response item. Score points 1, 2, and 3 for Item 16 are located in the OIB positions numbered 9, 28, and 36, respectively, as 16-1, 16-2, and 16-3. The exact location of each score point in the OIB is determined by calculating the point at which there is a .50 probability of attaining a score of at least the given score point or higher.[1] For example, Item 9 in the OIB (constructed-response Item 16-1) appears at the scale location at which there is a .50 probability of attaining a score on the item of 1 or higher; the second score point (i.e., Item 16-2) appears at the scale location at which there is a .50 probability of attaining a score on the item of 2 or higher, and so on. In addition to providing general information about item order, a typical IDM form provides participants with other data about the items in the form, such as the difficulty of the item (or its threshold) and the subtest area into which the item has been classified.

At the beginning of an IDM procedure, the last column of the IDM item information form shown in Figure 11-1 would be blank. It is this column that participants would use to record their judgments about the performance level that best represents the knowledge, skills, and cognitive processes demanded by the item. Figure 11-1 illustrates an information and

Item Location in OIB	*Item Location in Test Form*	*Item Difficulty (b value)*	*Content Area Strand*	*Item-Descriptor Match*
1	9	–2.13	Data, Statistics, and Probability	BB
2	19	–2.11	Number Systems	BB
3	24	–1.96	Measurement	BB
4	13	–1.80	Geometry	BB
5	33-1	–1.80	Data, Statistics, and Probability	BB
6	27	–1.80	Patterns, Algebra, and Functions	BB
7	20	–1.74	Number Systems	BB
8	32	–1.70	Patterns, Algebra, and Functions	BB
9	16-1	–1.65	Measurement	B
10	10	–1.62	Number Systems	BB
11	31	–1.59	Geometry	B
12	25-1	–1.59	Patterns, Algebra, and Functions	B
13	28	–1.55	Data, Statistics, and Probability	BB
14	2	–1.55	Number Systems	B
15	21	–1.47	Data, Statistics, and Probability	B
16	5-1	–1.47	Geometry	B
17	4	–1.47	Number Systems	B
18	6-1	–1.46	Patterns, Algebra, and Functions	B
19	15-1	–1.44	Measurement	B
20	14	–1.40	Geometry	P
21	5-2	–1.35	Geometry	B
22	8	–1.32	Number Systems	P
23	1	–1.32	Patterns, Algebra, and Functions	P
24	23	–1.32	Number Systems	P
25	26-1	–1.26	Geometry	P
26	33-2	–1.26	Data, Statistics, and Probability	P
27	30	–1.23	Number Systems	P
28	16-2	–1.23	Measurement	P

Figure 11-1 Hypothetical IDM Item Information and Participant Response Form

29	7	–1.03	Data, Statistics, and Probability	P
30	6-2	–1.03	Patterns, Algebra, and Functions	P
31	15-2	–1.03	Measurement	P
32	5-3	–0.99	Geometry	P
33	3	–0.94	Patterns, Algebra, and Functions	A
34	18	–0.94	Measurement	P
35	33-3	–0.94	Data, Statistics, and Probability	P
36	16-3	–0.88	Measurement	A
37	11	–0.83	Number Systems	A
38	26-2	–0.77	Geometry	P
39	22	–0.77	Patterns, Algebra, and Functions	A
40	25-2	–0.71	Patterns, Algebra, and Functions	A
41	29	–0.71	Geometry	A
42	12	–0.66	Patterns, Algebra, and Functions	A
43	6-3	–0.66	Patterns, Algebra, and Functions	A
44	26-3	–0.59	Geometry	A
45	17	–0.48	Data, Statistics, and Probability	A
46	15-3	–0.33	Measurement	A
47	25-3	–0.11	Patterns, Algebra, and Functions	A

Figure 11-1 (Continued)

NOTE: Adapted from Ferrara, Perie, and Johnson (2002a). Used with permission.

participant response form that has already been completed by a participant. For the hypothetical IDM procedure shown in the figure, participants used a set of four abbreviations for *Below Basic* (BB), *Basic* (B), *Proficient* (P), and *Advanced* (A) to indicate their judgments as to the best item-to-PLD match. For example, the participant who completed the form in Figure 11-1 judged Item 17 in the OIB (which appeared as Item 4 in the actual test booklet as administered to students) to be well matched to the description of the *Basic* performance level. The same participant judged that a score of 3 on constructed-response Item 5 (which appeared as the 32nd entry in the OIB) to be well matched to the *Proficient* performance level.

Another key component of the IDM method, called the *threshold region,* is apparent in Figure 11-1. Although the items are ordered by difficulty in the OIB, participants who are directed to explicitly consider content, instructional practice, student development, and so on may not always match items sequentially into categories. For example, as can be seen in the judgments for OIB item numbers 9 through 14, the participant who filled in this response form equivocated in this range of item difficulty (–1.65 to –1.55) as to whether the items in this range better matched the *Below Basic* or the *Basic* PLD. After an unbroken string of items classified as clearly matching the *Below Basic* PLD (OIB Items 1–8), the rater judged some items as best matched to *Below Basic* and others as best matched to *Basic,* until reaching OIB Item 15, at which point the participant recorded an unbroken string of matches to *Basic* PLD. It is the items between these unbroken strings of matches that constitute a threshold region, that is, a region within which the participant implicitly has indicated that the cut score separating two performance levels should be located. Also, the flexibility that participants have to match items to PLDs without regard to strict sequencing (such as is inherent with the Bookmark procedure) is considered a benefit of the IDM approach. In fact, groups of items in an OIB often have scale locations that are quite similar. To the extent that those scale locations are only estimates derived from IRT calibrations (i.e., they could vary given the standard errors of the location parameter values), their locations in the OIB cannot be considered immutable. Permitting participants to alternate in matching some items to a lower performance level and some to a higher performance level may more accurately reflect the underlying dimension tapped by that group of items.

It is also this alternating matching that reveals the presence of a threshold region. According to Ferrara et al., the threshold region is the area within the OIB consisting of a run of non-systematically alternating categorizations between runs of consistently categorized items (2002a, p. 14). The authors elaborate:

> These threshold regions represent places . . . where the matches between the items and the descriptors are not clear. Several factors can account for this lack of clarity, including peculiarities in the model or in the items themselves, estimation error, or vagueness in the performance level descriptions. . . . Moreover, performance level descriptions generally do not include clear borders from one level to the next; there may be some overlap. Thus, we would expect some overlap or general fuzziness in distinguishing the matches at the borderline between item-descriptor matches. (p. 14)

Intuitively, this way of conceptualizing the threshold region is appealing and perhaps helpful for use in explaining the procedure to participants.

However, a more precise definition of the threshold region is necessary to actually implement the IDM method and derive cut scores. In two implementations of the IDM method, Ferrara, Perie, and Johnson have used a starting/stopping rule that identifies a threshold region as beginning the first item after a run of at least three consistent classifications at a lower performance level and ending with the item just before a run of at least three consistent classifications at the next (i.e., higher) performance level. For example, from the information provided in Figure 11-1, it can be seen that this participant's threshold region between *Below Basic* and *Basic* is fairly wide, comprising Items 9 through 13 in the OIB. The lower bound of the threshold region for the cutoff between *Below Basic* and *Basic* begins with Item 9 in the OIB (Item 16-3 in the test form), which is the first item after a long run of consistent classifications of items at the *Below Basic* level. The upper bound of the threshold region for the cutoff between *Below Basic* and *Basic* ends with Item 13 in the OIB (Item 28 in the test form), which is the item immediately preceding the run of at least three consistent classifications as matching the *Basic* performance level. In comparison, the threshold region for the *Basic/Proficient* cutoff is quite narrow for this participant, consisting only of OIB entries numbered 20 and 21. The threshold region for this participant's *Proficient/Advanced* cutoff is broader, extending from Item 33 to Item 38 in the OIB. In Figure 11-1, the rows of information for items falling within three threshold regions are highlighted in bold type.

Once the threshold regions have been identified by each participant, initial cut scores within each threshold region can be determined. The procedure for deriving the initial cutoffs is one aspect of the IDM procedure for which there has not yet been developed a standard practice. In early implementations of the IDM method, participants were simply directed to mark the location in the threshold region where they believed the cut score for each boundary should be located—an activity very similar to the placement of a bookmark. In subsequent implementations of the IDM method, participants have been instructed to place a mark at the midpoint of the threshold region, though this instruction can be somewhat difficult for participants to implement when a threshold region comprises an odd number of items. For example, if the midpoint is used for the items shown in Figure 11-1, the initial cut score dividing the *Basic* and *Proficient* performance levels is easily located between OIB entries 20 and 21. However, the midpoint of the *Below Basic/Basic* region is somewhat more difficult to identify precisely, falling approximately at Item 11 in the OIB.

While participants in an IDM procedure would use individual items as reference points for making decisions about the locations of cut scores, the actual cut points can be precisely determined by calculating the medians or

means of the scale values for threshold items. For example, if means were used, an initial cut score dividing the *Below Basic* and *Basic* performance levels would be obtained by taking the average of the scale values beginning and ending the threshold region. In this case, the initial cut score in logits would be $(-1.65 + -1.55)/2 = -1.60$. By similar computations, the cut score dividing *Basic* and *Proficient* would be –1.375, and the cut score separating the *Proficient* and *Advanced* categories would be –0.855. Medians could also be used to locate the midpoint of the threshold regions; the use of medians would reduce the influence of an extreme scale value at the border of a threshold region.

We have referred to the cut scores derived this way as "initial" values because, like other standard-setting methods, the IDM procedure is typically conducted in various rounds. In Round 1, participants would determine their matches between items and PLDs, independently completing the last column in the form shown in Figure 11-1. When all participants have completed their forms, each participant has his or her threshold regions determined by those conducting the standard-setting meeting, and group discussions (either as a whole group or as subgroups, depending on the size of the panel) take place. These discussions focus on items about which there is disagreement among participants concerning the appropriate performance level match. Simple frequency distributions of category identifications for each item in the OIB are one way to provide this normative feedback to participants.

In Round 2, participants are again directed to review their item-descriptor matches individually, changing the performance level in which they have classified an item if they choose, although participants are also reminded that their own judgments—grounded in the content of the item and the PLDs—should be the dominant factor in their item-descriptor matches. At the end of Round 2, the threshold regions for each participant are again obtained and initial cut scores for each participant are calculated. A set of overall cut scores is also derived; these are obtained by using the mean of the individual participants' cut scores for each performance level.

Round 3 consists of a second opportunity for participants to discuss their threshold regions and items for which disagreement exists regarding the match to a PLD. In this round, participants may also receive impact information to help them understand the consequences (in terms of percentages of examinees in categories) of their individual cut score recommendations. Alternatively, impact based on the group mean recommended cut score could also be provided. At the end of this round, participants would make their final judgments regarding item-descriptor matches, changing (or not) their categorizations of items on the recording forms as they judge

appropriate. A final recommended cut score would be calculated by meeting facilitators based on these data and presented to the participants in a whole-group setting for their review. The final step in the IDM procedure would be the administration of a summative evaluation designed to obtain participants' overall reactions to various aspects of the meeting and to measure their confidence and support for the final group recommendations.

Alternative Procedures and Limitations

Although the IDM method has only recently been introduced and applied only in comparatively few contexts, research on the procedure suggests its viability for use in a broad array of applications, including assessment of educational achievement and licensure and certification testing programs. Also, despite its short history of use, a number of alternatives have been suggested or can be envisioned as promising improvements in the procedure.

First, in the implementations of the IDM procedure documented by Ferrara et al. (2002a, 2002b), participants themselves are trained to locate their threshold regions using the rules for consistent and alternating item-descriptor matches described previously. The authors note that "[standard-setting] workshop leaders can either teach the panelists how to determine threshold regions or . . . identify the threshold region without the panelists' participation" (2002a, p. 15). However, if participants themselves are permitted to identify their own threshold regions, it would seem necessary to institute a quality control check at these points to ensure that the regions were identified correctly. Consequently, simply having meeting facilitators or support staff determine the threshold regions would appear to be the most efficient and accurate approach.

Second, as mentioned earlier, when determining initial cut scores (or final cut scores) either the mean or the median scale locations of the boundary items in participants' threshold regions can be used. In the work of Ferrara et al. (2002a), medians are typically used to decrease the influence of outliers. However, means are equally appropriate, and in most cases the difference between the two summary statistics is likely to be minimal. For example, the midpoints (i.e., the cut score) for the *Proficient/Advanced* boundary illustrated in Figure 11-1 would be –0.91 if the median were used and -0.88 if the mean were used. Regardless of whether means or medians are used, it is advisable that participants in the standard-setting process be made aware during training that it is the group's overall standard that will serve as the recommended cutoff; if any outlying individual recommendations are to be "trimmed" in computing the group recommended standard,

the nature and circumstances of any such adjustment should be described to participants before they begin the item-descriptor matching process.

Third, although the developers of the IDM method refer to the form shown in Figure 11-1 as an "item map," it is probably best thought of as simply a single-sheet summary of the contents of an OIB. Extensive work on item-mapping approaches has been done, and one particularly promising item-mapping application, called a "mapmark" procedure has been developed by Schulz and his colleagues (Schulz, Kolen, & Nicewander, 1999, Schulz & Lee, 2002; Schulz, Lee, & Mullen, 2005). Using Schulz's method, items are grouped by subtest or content domain area and displayed simultaneously on a single sheet with the cut score(s) indicated across all subareas. In this way, participants are truly able to see the mapping of items on the underlying scale, the relative difficulty of the subareas, and how the cut score(s) they are recommending imply relative examinee performance in each of the areas.

One example of such an item map developed for setting performance levels on the Grade 12 mathematics test of the National Assessment of Educational Progress (NAEP) is shown in Figure 11-2. The figure shows a 20-item test comprising four 5-item subtests. The items are displayed grouped according to their content domains, labeled Domain A, Domain B, Domain C, and Domain D. The difficulty of the item (or the latent ability required to have some established probability of answering the item correctly) is plotted on the y-axis. The cut score on the test is shown as a horizontal line on the graph and is labeled C_x. Figure 11-2 shows the relative difficulty ordering of the domains, with Domain B comprising relatively easier items, all of which examinees above the cut score would likely have mastered, and with Domain C comprising relatively harder items and revealing that examinees just at the cut score would not be likely to have mastered the content captured by that domain.

A number of limitations of the IDM method exist. For one—and like many other standard-setting methods—the IDM procedure requires performance data. This is a limitation for those who are introducing a testing program and who wish to set standards in advance of the first operational test administration.[2] The scale locations for items necessary for assembling the OIB should be based on large, representative samples of examinees who responded to the items under actual testing conditions (i.e., motivated) so that the item location parameter estimates are stable.

Second, from descriptions of IDM implementations in the literature, it is clear that additional research and experience are necessary for users of the IDM method to have confidence regarding the starting and stopping rules for determining the beginning and ending of threshold regions. Although the developers of the procedure assert that "the rule of using a run of three

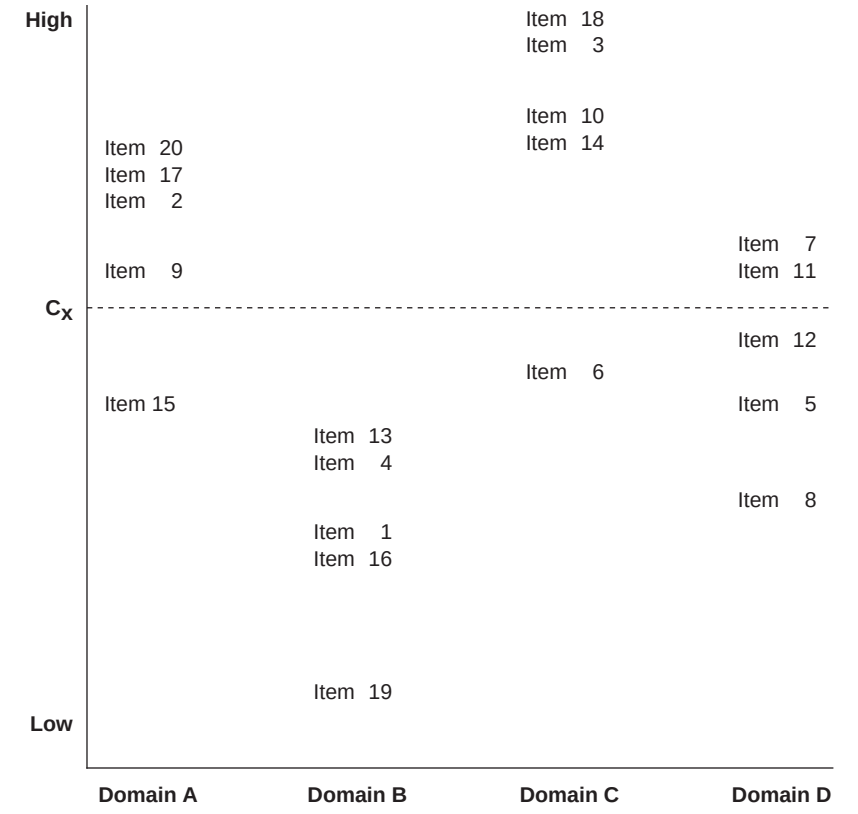


Figure 11-2 Hypothetical Item Map With Relative Domain Difficulties

to define the beginning of a new level is similar to the stopping rule used in individual IQ, achievement, and diagnostic testing" (Ferrara et al., 2002a, p. 14), the rule seems somewhat arbitrary in any of those contexts. Additional experience with this rule should help illuminate its usefulness in the standard-setting context. As well, additional experience with the IDM method is needed to help derive practical solutions to the situation when a participant fails to produce clear patterns of "runs." For short tests, this eventuality seems particularly likely—and particularly threatening to the viability of the IDM method in that situation.

Finally, and perhaps most importantly, the validity of the IDM procedure hinges as much as or more so than other methods on the clarity, completeness, and utility of the PLDs. As the developers of the procedure note: "The effectiveness of the ID Matching procedure depends heavily on

the quality of the performance level descriptions" (Ferrara et al., 2002a, p. 12). It seems an advisable first step in any implementation of the IDM method to focus on the development of the PLDs and to ensure that they have been pilot tested for utility by a small sample or focus group of persons similar to those who will use them in the actual standard setting. Once refined in this way, the PLDs should be reviewed and approved by those responsible for the eventual performance standards.

Notes

1. We use RP50 for this example, although other response probabilities are possible. For additional information on RP values, see Chapter 10.

2. We recognize that the necessity of having operational data, while mentioned as a limitation here, is most often thought of as an advantage in standard setting, particularly to the extent that accurate operational data on item and examinee performance are highly desirable for considering the impact of any performance standards being established.

12

The Hofstee and Beuk Methods

All of the methods for setting performance standards described in the preceding chapters of this book have in common, to a greater or lesser degree, grounding in the content covered by a test. In addition, all of the methods described thus far can be classified as "absolute," invoking the term used by Nedelsky (1954) to distinguish his method from the norm-referenced methods that were prevalent in his time. The phrase "absolute standard setting" is not used much anymore, although the concept of establishing a fixed level of performance, based on explicit articulation of performance expectations, has remained. Alternate terms, such as "criterion referenced" or "standards referenced" are now more commonly used. In theory at least, such terms are intended to connote that an evaluation of the knowledge or skill necessary for competent practice or success in some area has been performed and that a performance standard has been established that reflects *only* the translation of that performance standard into a cut score. An examinee's performance on an examination—such as passing or failing—is dependent only on the examinee's knowledge, skill, or ability and the level at which the criterion for passing (i.e., the cut score) is set. In that sense, the standard (in Nedelsky's terms) is absolute, and not dependent on the performance of other examinees or related to normative expectations, social or economic factors, and so on.

In a strict sense, however, it is impossible for humans to establish a purely criterion-referenced or "absolute" standard. Rather—and particularly in standard setting on examinations—participants always bring normative

information to bear whenever a putatively criterion-based judgment is required. To illustrate that a mix of norm- and criterion-referenced information is always, albeit often tacitly, considered, it may help to refer to a practical situation that nearly all parents have experienced: the day that they are finally able to proudly proclaim that their child has been potty trained. One of the authors of this book (GJC) uses a fictitious—though presented as truthful—account of the potty training event with classes of introductory measurement students, announcing the great news to the class, accepting the congratulations, and noticing the vicarious joy of many of the students who, as parents themselves, know that a major childhood hurdle has been overcome. Then, to the surprise of the students, he announces that the potty-trained child has also just turned 9 years old.

Of course, the purpose of the surprise ending to the story is to illustrate that all of the hearers of the good tidings unknowingly interpreted the criterion-referenced information (i.e., successful potty training) in light of normative expectations; that is, they tacitly brought to bear an expectation of what age would be appropriate for success on such a task. When confronted with information about the age of the child, many of the hearers of the story were obviously less joyful: What had seemed initially to be a normal accomplishment was now considered to be a long overdue behavior.

The point of the fictitious experiment was to demonstrate that the kinds of human judgments made in standard-setting contexts can never be strictly absolute or criterion referenced; rather, participants in any procedure will always bring to bear, to some extent, norm-referenced information. The realization of this fact has led many psychometricians to suggest that no standard-setting method can be developed that is solely criterion- or standards-referenced and that standard-setting methods should be developed that explicitly take into account the blending of criterion- and norm-referenced information processing that comprise the judgmental tasks engaged in by participants.

The term **compromise method** was coined by a developer of one such method, Willem Hofstee (1983), to capture the blended nature of setting performance standards. According to Hofstee, the process of standard setting inevitably involves the balancing of standards-referenced perspectives, political agendas, economic pressures, policy concerns, and other factors. In Hofstee's words, "A [cut score] solution satisfactory to all does not exist and . . . the choice between alternatives is ultimately a political, not a scientific matter" (p. 109). The following portions of this chapter describe two such compromise methods—the method suggested by Hofstee (1983) and a method developed by Beuk (1984).

The Hofstee Method

Hofstee's (1983) compromise method began as one invented to address the practical problem of conflicting criterion- and norm-referenced expectations. In a paper describing the development of the method, Hofstee described a situation he faced in 1979 in the course of administering an examination in research methods to 160 second-year psychology students. According to Hofstee, in that year,

> the passing score on the test had to be lowered to 45 percent mastery, and even then only 55 percent of the students passed. In 1980, the passing score was set at close to 60 percent; and over 90 percent of the students passed. The learning materials were essentially the same; the teachers and the test items were essentially the same; and in view of the large numbers [of students in each group] it would be difficult to ascribe the discrepancy to a cohort effect. (p. 117)

Thus Hofstee was faced with a situation in which normative expectations established in the first year clashed with actual group performance information observed in the second year. The result apparently called into question the "correctness" of the performance standard that had been established in the first year and called for some kind of balance to be struck in the second year between the criterion-based and normative information.

The general approach developed by Hofstee treated both situations as independent. According to Hofstee, his model treated the standard to be set for the second year as if it were a "situation in which a cutoff score on an achievement test is set for the first time . . . [and when] no agreed-upon prior or collateral information is available on the difficulty of the test, the quality of the course, or the amount of preparation by the students" (1983, p. 117).

Procedures for Implementing the Hofstee Method

As with all other standard-setting methods, the Hofstee approach begins with identification of the purpose of standard setting; recruitment, selection, and training of participants; and orientation of participants to the judgmental task they will be expected to perform. Hofstee himself did not explicitly address two issues, but the method he proposed presumes that participants are familiar with the examination that serves as the basis for examinee classifications and that participants have key conceptualizations regarding minimal competence that they can translate into percentage estimations.

Under the assumption that the training, content familiarization, and review of key conceptualizations can be accomplished in approximately a one half-day session, the remainder of the Hofstee method—that is, the

completion of the actual judgmental tasks required of participants—ordinarily requires less than one additional hour, making the Hofstee method one of the most time-efficient ways of setting cut scores. The Hofstee judgmental task requires each standard-setting participant to respond to four questions and to assume that examinees are first-time test takers. Two of the questions focus on the acceptable level of knowledge that examinees should possess (for which Hofstee used the symbol k); the other two questions focus on the tolerable examination failure rates (for which Hofstee used the symbol f). The four questions asked of participants follow:

1. "What is the highest percent correct cut score that would be acceptable, even if every examinee attains that score?" This value is a participant's estimate of the maximum level of knowledge that should be required of examinees and is symbolized as k_{max}.
2. "What is the lowest percent correct cut score that would be acceptable, even if no examinee attains that score?" This value represents a participant's judgment of the minimum acceptable percentage of knowledge that should be tolerated and is symbolized as k_{min}.
3. "What is the maximum acceptable failure rate?" This value represents a participant's judgment of the highest percentage of failing examinees that could be tolerated and is symbolized as f_{max}.
4. "What is the minimum acceptable failure rate?" This value represents a participant's judgment of the lowest percentage of failing examinees that could be tolerated and is symbolized as f_{min}.

A simple rating form is used to collect the four values that each participant generates in response to these questions. Participants should be instructed that all values must be between 0 and 100, inclusive, and that each participant's "max" value must exceed its corresponding "min" value. Alternatively, a form could be used on which participants simply indicate a value for each question from a limited number of choices provided (e.g., values from 0 to 100 in multiples of 10). The mean values of the responses to these questions, across participants, are then calculated. These values—normative expectations—are then used in conjunction with actual performance information (in the form of a distribution of test scores resulting from administration of the examination) to obtain a cut score.

Figure 12-1 illustrates a hypothetical application of the Hofstee method in which the mean of participants' judgments about k_{min}, k_{max}, f_{min}, and f_{max} are approximately 32.5, 80, 12.5, and 90, respectively. Ordered pairs of the points (k_{min}, f_{max}) and (k_{max}, f_{min}), which in this case correspond to the points

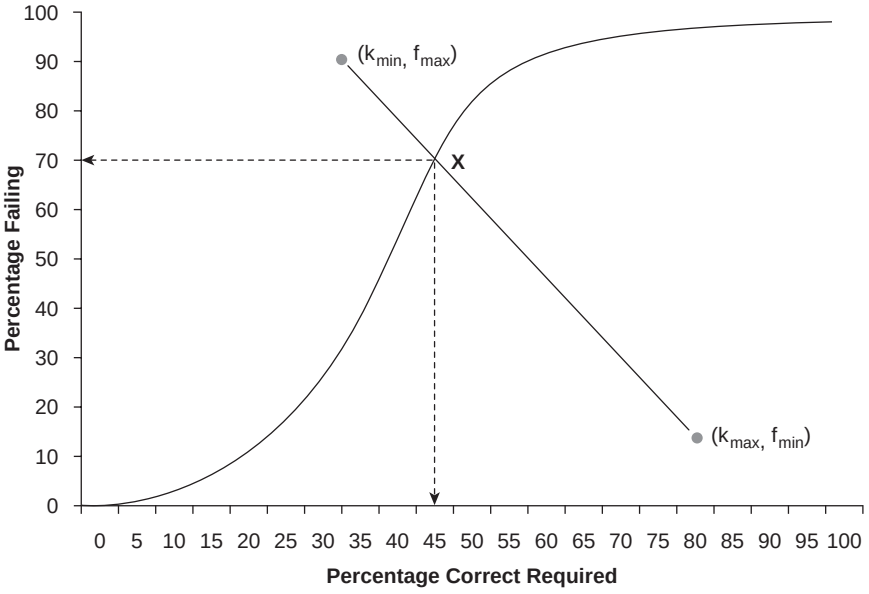


Figure 12-1 Illustration of Hofstee Cut Score Derivation

(32.5, 90) and (80, 12.5), are used to plot a line. These coordinates have been identified in the figure, and the resulting straight line between these points (running from the upper left to lower right portion of the graph) has been plotted. According to Hofstee, it is along this line that all "compromise" solutions to the cut score may be found.

In addition to this line, the observed test score distribution is required. According to Hofstee, the optimal compromise cut score solution is identified by locating the intersection of the line and the observed test score distribution. The observed test score distribution is presented in Figure 12-1 as a cumulative ogive, which shows the functional relationship between level of performance required (as a percentage of total raw score) and percentage of the examinee group that would be classified as failing if that level of performance were used as the cut score. As the percentage correct required of examinees (shown on the abscissa) increases, the cumulative function reveals that the corresponding failure rate (shown on the ordinate) increases, as would be expected.

To obtain a final recommended cut score using the Hofstee method, the line established from the ordered pairs (k_{min}, f_{max}) and (k_{max}, f_{min}) is projected onto the cumulative distribution. The point at which the line and the function intersect is shown in Figure 12-1 as point X. The coordinates of

this point are then used to obtain the compromise percentage correct required (i.e., the cut score for the test) and the corresponding failure rate that would be observed if that cut score were used. Figure 12-1 shows these values to be 70 and 45, respectively.

The Beuk Method

Like Hofstee, Cees Beuk realized that setting performance standards "is only partly a psychometric problem" and that "good measurement is a necessary, but not a sufficient condition to solve the problem of standard setting" (1984, p. 147). In developing a slightly different compromise standard-setting method, he also suggested that standard-setting procedures should take into account the absolute level of content mastery judged to be essential for passing, credentialing, or other examination purposes, but also take into account relevant comparative (i.e., normative) information about examinees and participants' values. According to Beuk, "A very important question in this respect is the way in which relevant information is obtained, processed, and interpreted" (p. 147).

The Beuk method rests on two assumptions: one about the participants in the standard-setting process and one about how the process should be configured. According to Beuk, first, it must be assumed that "each member of the standard setting committee has an opinion of (a) what passing score should be required, and (b) what pass rate can be expected" (1984, p. 148). Second, Beuk assumed that "the relative emphasis given to the two types of judgments should be in proportion to the extent to which members of the committee agree with each other" (p. 148). As can be seen from these assumptions, and as will be seen in the brief description of his method that follows, Beuk's approach is essentially a special case—a simplification—of the Hofstee procedure.

Procedures for Implementing the Beuk Method

As with other standard-setting methods, appropriate selection, training, and orientation of participants must be provided. And as in the Hofstee method, participants using a Beuk approach are familiarized with the content of the test for which standards will be set and the population of examinees to which the standards will be applied, and they consider the purpose of the examination as they form key conceptualizations about the level of knowledge, skill, or ability necessary for performing the job or obtaining whatever credential is offered by the testing program.

Then, to implement Beuk's (1984) method, participants holistically consider the examination on which decisions are to be made and record their responses on a form provided to only two questions:

1. "What should be the minimum level of knowledge required to pass this examination?" This judgment is to be expressed as percentage correct for the total test and is symbolized as the variable *x*.

2. "What passing rate should be expected on this examination?" This judgment is expressed as a percentage of the examinee population and is symbolized as the variable *y*.

The quantitative answers to these questions are summarized across participants, yielding means of x and y. Then, after the examination has been administered, the judgments can be compared with the actual performance of the examinees, and participants' judgments can be used to effect an adjustment or compromise between normative expectations and observed test performance.

Figure 12-2 shows a hypothetical application of the Beuk method. Suppose, for example, that participants, on average, judged that the minimum percentage correct should be 69%, and they judged, on average, that 89% of examinees should pass. In Figure 12-2, these points are labeled $\bar{x}$ and $\bar{y}$, respectively. Then the intersection of points $\bar{x}$ and $\bar{y}$ is determined; it is labeled "A" in the figure. Further suppose that the examination was administered and that the functional relationship between all possible cut scores and the percentage of examinees that would pass if each of the possible cut scores was applied was plotted. The result is represented in Figure 12-2 by the monotonically decreasing function shown as a solid, slightly choppy line. The line indicates (as is necessarily true) that the percentage of examinees who would pass the examination decreases as the percentage correct required to pass increases.

The next step in applying the Beuk method is to calculate the standard deviations of participants' answers to the two questions related to expected percentage correct and passing rates. The ratio of these two standard deviations (S_x/S_y) is used to construct a line with a slope equal to this ratio and that, originating at point "A," is projected onto the distributional curve. The point at which the line intersects the curve is labeled "B" in the figure. The adjusted values of $\bar{x}$ and $\bar{y}$ are then obtained by projecting point "B" onto the two axes. The result is the determination of an adjusted or "compromise" percentage correct (i.e., the cut score) and the associated passing rate. These points are labeled $\bar{x}'$ and $\bar{y}'$, respectively, in the figure. In this hypothetical illustration, the projection yields a recommended percentage correct (i.e., cut score) of 51% correct and a corresponding passing rate

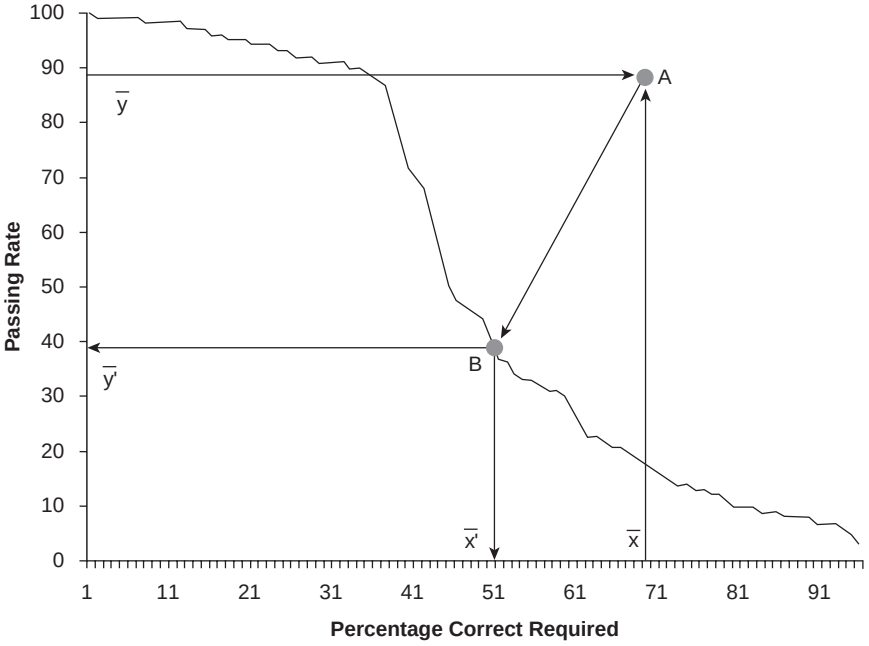


Figure 12-2 Illustration of Beuk Cut Score Derivation

of 39% (very strict judgments, indeed!). As with any other method, the cut score in raw score units would be obtained by multiplying the adjusted percentage correct required ($\overline{x}'$) by the total number of possible points on the examination.

Alternative Procedures and Limitations

Both the Hofstee and Beuk methods are often suggested as secondary or "fall-back" methods to be used in conjunction with other standard-setting approaches. Thus, one alternative manner of applying these methods is to include the four (for the Hofstee method) or two (for the Beuk method) questions in a final procedural evaluation form that would be administered to participants at the completion of a standard-setting meeting in which a more traditional, criterion-referenced method had been implemented. At that point, because all participants would already have become familiar with the examination purpose and content, and because they would have solidified any key conceptualizations required of them, it would be a simple and quick matter

to gather the additional data. However, as we discuss in greater detail later in this book (see the topic titled "Using Multiple Methods of Standard Setting" in Chapter 16), there are good reasons why it may not be advantageous to use the Hofstee or Beuk methods in conjunction with another standard-setting procedure.

Both of these methods were developed in the context of examinations for which only a pass/fail decision was necessary. No record exists of either method being used in contexts where more than one cut score was required for a single examination (as is increasingly the case where multiple performance levels such as *Basic, Proficient,* and *Advanced* must be established for a single test); it would seem difficult though not impossible to extend either method in this way.

One alternative procedure that would seem fairly easy to implement would be to substitute information gathered via a traditional method to answer one (or more) of the questions required by those using either the Beuk or Hofstee approaches. For example, the mean Angoff rating across participants using that method could be substituted as the mean value in response to the question "What should be the minimum level of knowledge required to pass this examination?" for which respondents using the Beuk method essentially supply an estimated percentage correct.

Possible alternative ways of implementing the Beuk and Hofstee methods are few, however. In fact, although both of these compromise methods are often mentioned as candidates for potential use, neither method appears to be used nearly as extensively as the strictly criterion- or standards-referenced methods described in the preceding chapters of this book. It is not clear why this would be the case, particularly given the comparatively much greater efficiency of the methods. One reason may have to do with the relatively little research that has been done on these methods. Especially problematic is the fact that few studies have been published in which results from Beuk or Hofstee methods are compared with results from more traditional methods. In part, the relatively infrequent use of the methods may also be attributable to the fact that both methods require that observed score distributions for a test be available in order to apply the method. This means that the final cut score for a test could not be known until after the test had been administered and the standard-setting meeting had taken place. A final limitation—a very practical one—is that computer programs do not exist to accept data from either method and to analytically derive a cut score. The computational work to derive cut scores using these methods must be accomplished by hand, introducing a source of potential error when cut scores must be approximated or visually estimated, or a specialized program would need to be written and tested to ensure accuracy of results.

SECTION III

Challenges and Future Directions in Standard Setting

In the preceding section of this book, we focused on describing individual methods for setting performance standards. The challenges of standard setting extend, however, beyond instances in which a single cut score or multiple cut scores must be established for a fairly traditional test. And there are aspects of standard setting that go beyond the particular method selected, but that are equally—if not more—important toward ensuring that the implementation of a specific method will be successful.

For example, the first chapter in this section, Chapter 13, addresses the practical reality of scheduling a standard-setting procedure. Included in this chapter are sample timelines, needed materials, and guidelines on conducting the standard-setting meeting itself so that the process is orderly, smooth, and permits—as the *Standards for Educational and Psychological Testing* requires—that participants be able to bring to bear their judgments in reasonable ways.

Chapter 14 addresses the standard-setting reality that sometimes one or more cut scores must be set on more than a single test. Increasingly, multiple cut scores must be set on multiple tests that span a range of levels and content areas. The individual cut scores set on one test cannot be established in isolation. To address this challenge, a family of new standard-setting procedures called "vertically-moderated standard setting" has been developed to promote coherence and consistency in the system of performance standards across levels and content areas.

Another increasingly common standard-setting context involves establishing performance standards on tests that are nontraditional in their characteristics. In particular, in the area of educational achievement testing, it is increasingly necessary to establish performance standards for tests administered to students with special needs, such as students with severe cognitive or physical impairments. Such tests, called "alternate assessments" often cannot comprise typical test formats such as multiple-choice items, but instead consist of collections of evidence such as portfolios, videotapes, observational records, or other nontraditional assessment information. Standard-setting methods have recently been invented or adapted to meet this challenge, and these procedures for setting standards on alternate assessments are the focus of Chapter 15.

The concluding chapter in this section covers a potpourri of special standard-setting topics. The topics in Chapter 16 are not arcane or academic, however. Rather, as we have done in the rest of this book, we have tried to identify the practical and vexing issues that are often faced by those who actually conduct standard setting. For example, among the topics addressed in Chapter 16 are methods of adjusting cut scores, deciding how to incorporate the uncertainty that accompanies all standard-setting results; how to deal with the potentially consequential issue of rounding when the quantitative results of standard setting require it; how, when, or if using multiple methods of standard setting is desirable; and how a critical aspect of all standard setting—the training of participants in the judgmental tasks they must perform—can be improved.

13

Scheduling Standard-Setting Activities

Standard setting is not an activity occurring in isolation. Whether for a certification examination or a statewide K–12 student achievement test, standard setting is an integral part of a larger enterprise that requires careful planning, coordination, and communication to complete successfully. In this chapter, we offer some suggestions for scheduling a standard-setting activity, drawing primarily on our experiences in large-scale credentialing programs and educational assessment and providing examples of each.

Scheduling Standard Setting for Educational Assessments

Table 13-1 provides an overview of the main activities to be completed, along with a timetable for their completion. The following discussion traces the steps of a facilitator who plans and carries out a standard-setting activity the week of May 20–24, 2006. A generic version of the form shown in Table 13-1 can be found at www.sagepub.com/cizek/edschedule. That version can be easily adapted to any standard-setting activity simply by entering different dates for the test administration. All other dates are automatically updated.

Table 13-1 includes several test-development activities not directly relevant to standard setting in order to show the overall place of standard

Table 13-1 Generic Standard-Setting Calendar for an Educational Assessment

	Timing (Days Out)*		*Schedule*	
Event	*Start*	*Finish*	*Start*	*Finish*
Conduct design conference	–730	–700	03/20/04	04/19/04
Develop specifications	–700	–669	04/19/04	05/20/04
Develop PLLs and PLDs (Draft)	–700	–669	04/19/04	05/20/04
Review PLLs and PLDs	–669	–660	05/20/04	05/29/04
Develop test items	–669	–547	05/20/04	09/19/04
Review test items	–547	–487	09/19/04	11/18/04
Prepare for field testing	–487	–426	11/18/04	01/18/05
Print/distribute field tests	–426	–380	01/18/05	03/05/05
Field test items	**–365**	**–350**	**03/20/05**	**04/04/05**
Score field tests	–350	–319	04/04/05	05/05/05
Analyze field test data	–319	–304	05/05/05	05/20/05
Review field test item data	–304	–274	05/20/05	06/19/05
Draft standard-setting plan	–304	–274	05/20/05	06/19/05
Select items for operational tests	–274	–243	06/19/05	07/20/05
Present standard-setting plan to stakeholders	–273	–213	06/20/05	08/19/05
Review standard-setting plan	–273	–213	06/20/05	08/19/05
Prepare for operational testing	–243	–182	07/20/05	09/19/05
Finalize standard-setting plan	–213	–182	08/19/05	09/19/05
Set dates for standard setting	–200	–185	09/01/05	09/16/05
Identify standard-setting candidates	–200	–140	09/01/05	10/31/05
Secure meeting site	–185	–180	09/16/05	09/21/05
Review operational tests	–182	–122	09/19/05	11/18/05
Select standard-setting participants	–140	–100	10/31/05	12/10/05
Prepare final PLLs and PLDs	–122	–91	11/18/05	12/19/05
Print and distribute operational tests	–122	–15	11/18/05	03/05/06
Notify standard-setting participants	–100	–95	12/10/05	12/15/05
Finalize contract with meeting site	–60	–50	01/19/06	01/29/06
Prepare standard-setting materials	–30	60	02/18/06	05/19/06

	Timing (Days Out)*		*Schedule*	
Event	*Start*	*Finish*	*Start*	*Finish*
Administer operational test	**0**	**12**	**03/20/06**	**04/01/06**
Score operational tests	15	61	04/04/06	05/20/06
Follow-up letter to participants	20	25	04/09/06	04/14/06
Create on-site data analysis programs	32	43	04/21/06	05/02/06
Create training presentations	32	43	04/21/06	05/02/06
All panelists return housing forms	35	39	04/24/06	04/28/06
Complete hotel arrangements	39	40	04/28/06	04/29/06
Send rooming list to hotel	46	46	05/05/06	05/05/06
Analyze operational test data**	46	61	05/05/06	05/20/06
Arrange for time on board agenda	47	49	05/06/06	05/08/06
Rehearse presentations	49	50	05/08/06	05/09/06
Print all rating forms	54	54	05/13/06	05/13/06
Prepare participant materials (see sample materials list)	54	56	05/13/06	05/15/06
Purchase all supplies (see supplies/ equipment)	55	55	02/25/00	05/14/06
Route copies of final plan to all presenters	56	56	05/15/06	05/15/06
Assemble panelist packets	56	56	05/15/06	05/15/06
Secure equipment	56	56	05/15/06	05/15/06
Conduct final in-house dry run	56	56	05/15/06	05/15/06
Check meeting rooms	60	60	05/19/06	05/19/06
Attend to participant check-in issues	60	60	05/19/06	05/19/06
Conduct standard setting	**61**	**65**	**05/20/06**	**05/24/06**
Present results to stakeholders	68	72	05/27/06	05/31/06
Review/adopt cut scores	76	91	06/04/06	06/19/06
Print/distribute score reports	**91**	**102**	**06/19/06**	**06/30/06**

NOTE: * Relative to start of operational testing

** Although less than 100% of data are analyzed by this date, it may be sufficient for standard-setting purposes if the sample is well defined, representative, reviewed, and approved in advance by the technical advisory committee and the sponsoring organization.

setting in the larger scope of test development. It shows the planning for standard setting beginning two years before the actual standard-setting session. There is a practical reason for this amount of lead time: This schedule assumes a new testing program. Part of the planning process is establishing the number and nature of the performance levels to be set. If these are established by state law or board action, then some of the work has already been done. However, even in such instances, it will still be necessary to begin to bring some precision to the performance level labels (PLLs) and performance level descriptions (PLDs).

Overall Plan

Why start so early with the PLLs and PLDs? If this is a new testing program, no test items have been written yet, and certainly none has been reviewed or field tested. By establishing PLLs and PLDs at the outset, it will be possible to ensure that there are test items that will support these levels. That is, item writers will be well served by clearly defined PLDs as they write items that will be used to categorize examinees into these levels. Item reviewers will have a better understanding of the necessity for a range of item difficulty as well as the need for items with specific characteristics. More than one standard setter has been frustrated by having to work with test items that didn't seem to fit the performance levels because the PLLs and PLDs emerged long after the tests were developed. Drafting a standard-setting plan before item writing begins is one way to make sure the test supports the standard-setting activity that is eventually carried out.

The schedule provided in Table 13-1 also shows a field test exactly one year prior to the first operational administration of the test. Testing programs often have a fixed administration schedule that varies little from year to year. Thus, during the first year, a regular testing window would be reserved for field testing. In subsequent years, the operational test and parallel field test would be conducted during the same window, perhaps with some field testing taking place shortly after or at the end of the testing window.

Planning the standard-setting activities for a large-scale assessment may begin as much as a year before the actual standard-setting meeting takes place; Table 13-1 shows a start date of May 20, 2005, one full year before the standard-setting meeting is scheduled to occur. The planning for standard setting should specify a method (e.g., Angoff, Body of Work, Bookmark), an agenda, training procedures, and analysis procedures. It should be appropriate to the content and format of the tests, the nature of the standards to be set, and the pool of potential panelists.

Table 13-1 also refers to review of the standard-setting plan by a **technical advisory committee** (TAC). Many assessment programs employ a panel of nationally recognized assessment experts to advise them on technical issues related to those programs. In some instances, TACs also advise on policy and practical matters, depending on the membership of the TAC and their charge. Given the high profile of standard-setting activities and the number of both technical and practical considerations, review and approval of standard-setting plans are almost always a central focus of the TAC.

Other groups also have an interest in the plans for standard setting. Table 13-1 refers to stakeholder review. Stakeholders are individuals and groups with a particular interest in the testing program and may include elected or appointed officials (state superintendent, commissioner, board members, etc.), oversight committees (test steering committee, content review committee, bias/sensitivity review committee, examination construction committee), educator organizations, professional associations, parent organizations, business groups, media representatives, and others. The standard-setting process should be considered one of the high-profile, consequential events associated with a testing program and approached accordingly. Although stakeholders will vary from state to state and test to test, it is always a good idea to know who they are and obtain their input as early in the process as possible.

One very special stakeholder group is the policy board that will actually make the decision to adopt, modify, or reject the cut scores. For licensure and certification testing programs, the policy entity is usually the professional association or credentialing board. For statewide assessments, the policy board is usually the state board of education. Rules and procedures for board actions vary from state to state, and it is usually the responsibility of the agency staff to know the rules for their state and to make sure those rules are followed.

For example, in the context of educational achievement testing programs, in some states a resolution is introduced for discussion at one meeting of the state board of education, and then the resolution is voted on at the next meeting. If board approval for cut scores is required before score reports can be produced, then it will be necessary to work backward from the first board meeting (introduction of a resolution) to schedule the standard setting, allowing sufficient time for review of results, calculation of cut scores, review by state education agency staff and other stakeholder groups, and whatever lead time the board requires for accepting agenda items. Thus, for example, if the board meets monthly, it may introduce an intent to adopt cut scores at the June meeting and approve the resolution in July. If score reports are due by late July, this schedule may work. If score reports

are due by June 30, however, it will be necessary to introduce the resolution at the May meeting and wait for approval at the June meeting. Assuming the board approves the cut scores at its June meeting, score reports can be printed and shipped by June 30. Depending on when the board meets in May, it may also be necessary to move the standard-setting session up to early May or late April, again with implications for availability of student scores and any panelist materials that depend on the availability of valid scores (e.g., ordered booklets, work samples).

Working with a state board of education or a credentialing board requires a clear understanding of protocol. Boards accept agenda items under specific conditions. The lead facilitator and the agency official responsible for the standard setting should know the conditions for forwarding cut scores and supporting information and follow them. If the board requests clarification or additional information, that information should be provided quickly and succinctly. Reports sent to boards should be clear and to the point, devoid of psychometric jargon. When in doubt, examine previous reports to the board for wording, tone, and length.

The hypothetical schedule shown in Table 13-1 shows a request going to the board about two weeks before the standard-setting activity. The board meeting at which the cut scores are to be presented (Review/adopt cut scores) is about three to four weeks after the standard setting, allowing four to five weeks lead time between the request and the actual meeting. The requirement will vary from board to board. This schedule also assumes that the resolution will be presented and voted on in the same meeting, permitting the printing and distribution of score reports by June 30. Note also that there is a period of time between standard setting and the board meeting during which the cut scores and supporting information are shared with stakeholder groups. In turn, these groups often provide their own reports to the board, either as adjuncts to the state education agency report or as separate reports.

It should also be clear that the planning process is an iterative one. Sometimes, several drafts of a plan may be necessary. As Table 13-1 shows, planning starts two years out and continues until about six months out. Ultimately, the agency, the TAC, the standard-setting facilitator, and other stakeholders will be involved in the refinement of the plan. Successful planning and execution of the standard-setting activity will, of necessity, include the identification of and communication with as many of these stakeholder groups as possible.

Although it would be preferable to have all aspects of the plan approved by all interested parties six months before standard setting begins, practical realities mean that this is rarely the case. In some instances, fine tuning of

the plan continues through the first day of the on-site standard-setting activity. Careful planning that starts early can usually prevent the need for last-minute changes and surprises.

Participants

As the overall standard-setting plan is being reviewed by various advisory committees and stakeholder groups, we move to the next key phase of the plan: identifying and recruiting the individuals who will participate in the standard-setting activity (i.e., the panelists). For credentialing examinations, it is important that the participants be representative of important perspectives, have expertise in the area covered by the examination, and be able to make judgments as appropriate to the purpose of the standard setting. For example, for a recertification examination, credentialed participants with many years of professional experience would be appropriate. For an examination to permit entry-level practice, credentialed participants with fewer years of experience—that is, those who are likely to have a greater understanding of the nature and demands of entry-level practice—would be desirable. For either situation, familiarity with the practice analysis or role delineation study and daily job responsibilities of the positions in which credentialed candidates will practice is also essential.

For statewide assessments, it is preferable that the panelists be as representative of the state as possible. In many instances, parents, business and community leaders, and representatives of identified stakeholder groups have been included as potential panelists. Table 13-1 shows the process of identifying these individuals about nine months before standard setting begins.

For a credentialing examination, the process of recruiting and selecting participants is less complicated than that often used in educational testing contexts. For professional licensure or certification programs, the process often begins by contacting local jurisdictions and soliciting nominations. Or a general call for indications of interest in serving on the standard-setting panel may be sent to credentialed, practicing members of the profession. Members of the board may also be asked to nominate qualified individuals.

Although more complicated than the credentialing context, the procedure for recruitment and selection of participants in education contexts has some elements that may be useful for those in credentialing contexts to consider. Identifying potential panelists for a statewide assessment program usually involves working through local education officials, usually local superintendents. Local superintendents identify individuals either in accordance with a set of specifications from the chief school officer of the state

[Date]

Dear [Name]:

You have been nominated by your superintendent or designee to serve on a panel of educators to consider performance standards for the [Test Name]. This preliminary letter is to determine your interest in and availability for a three-day meeting to be held on [Date] in [Location]. This is not an invitation to participate in standard setting. If you are selected, you will be notified well in advance of the meeting so that you will have time to prepare to participate.

If selected, you will receive training in the specific standard-setting procedure we will use. You will work with a group of other educators to review the test that will be administered this spring and consider each item in terms of how likely a "just-proficient" student would be to answer it correctly.

Please complete the attached information form and return it to [Name] by [Deadline] using the postage-paid envelope enclosed with this note. As soon as we receive your information, we will complete the selection process and notify you.

Sincerely,

[signature]

Director, Assessment Division

cc: Superintendent

Enclosure: Return Envelope

Figure 13-1 Sample Initial Contact Letter

or, more often, his or her designee. These specifications are a crucial part of the plan. Thus, even though the overall plan is still under review at this point, this particular portion should be resolved before any letters go out to local superintendents. A sample letter of initial contact and a sample follow-up letter are shown in Figures 13-1 and 13-2. Electronic versions of this correspondence that can be tailored to the reader's specific circumstances can be found at www.sagepub.com/cizek/letters.

Creation of the standard-setting panels is a three-step process. First, local superintendents or their designees identify potential panelists in accordance with specifications provided by the state education agency. Local education officials forward to the state education agency the names and brief resumes of their candidates.

The second step involves notification of the candidates. When asked to notify candidates prior to submitting their names, some local education officials occasionally fail to do so. Thus, primarily as a safeguard and in an effort to convey a consistent message to all candidates, the state education

To: Standard-Setting Meeting Participants

From: [Contractor Contact Person]

Date: [Month, Day, Year]

Subject: Arrangements for Standard-Setting Meeting

Thank you for agreeing to serve on the standard-setting committee for [Test Name] to be held on [Dates] at [Location]. Enclosed you will find an agenda, along with some basic information about the goals for the meeting and the tasks you will perform. We will provide a complete introduction to your tasks and continue to provide support to you throughout the meeting.

I look forward to seeing you on [Date]. A map showing how to reach [Meeting Location] is enclosed. If you requested overnight lodging, a room has been reserved for you, although please note that the hotel may not permit check-in before 3:00 P.M. If you have any further questions about this meeting or your role in it, please contact me by telephone at [Phone Number] or by e-mail at [E-mail Address].

cc: Director, Testing Division

Figure 13-2 Sample Participant Follow-Up Correspondence

agency sends an initial letter to all candidates. This letter informs the candidate that he or she has been nominated and makes it clear that final selection has not yet occurred. The letter should include a form on which the candidate can indicate interest in and availability for a standard-setting meeting on specific dates in a specific city. Specific hotel information should not be given at this time because such information could easily lead the candidate to believe that he or she is being invited to a specific site. That information will be included in the follow-up letter. The form also includes a request for demographic information that will be used in the final selection process to ensure balance and representativeness. An example of this form is shown in Figure 13-3. An electronic version of the form can be found at www.sagepub.com/cizek/letters.

Once potential participants return their interest/availability forms, the third step begins. Those who are not interested or unavailable are eliminated from further consideration, but their forms are retained for later reference. State education agency staff sort the nominations to create the required number of panels with the approved numbers of panelists. After considerable sifting and sorting, the panels are formed, and agency staff prepare a follow-up letter to inform all candidates of their selection or nonselection. The nonselection letter is an important part of the process,

Standard-Setting Information Form

In order to make sure the educators selected to recommend performance standards are as representative as possible of the population of [State], we are asking you and all other candidates to provide some information about yourself. Please complete this form and return it to [Contractor] in the prepaid envelope provided. Please print or type all information.

Name _____

School _____ **District** _____

1) Years of teaching experience _____

2) Subjects certified to teach (check all that apply)

❑ Language Arts ❑ Mathematics ❑ Science ❑ Social Studies

3) Primary current teaching assignment (check all that apply):

❑ Regular Classroom ❑ Gifted/Talented ❑ ESL/ELL ❑ Special Education (specify) _____

4) Gender (check one): ❑ Female ❑ Male

5) Race (check one): ❑ Asian/Pacific Islander ❑ Native American/Alaskan Native ❑ African American/Black ❑ Multiethnic ❑ Caucasian ❑ Other ❑ Hispanic

6) I am interested in serving on the standard-setting committee for (check one)

❑ Language Arts ❑ Mathematics ❑ Science ❑ Social Studies

7) I am available to attend a meeting in _____ **(location) on** _____ **(date).**

❑ Yes ❑ No

_____ Signature _____ Date

Figure 13-3 Standard-Setting Participant Information Form

perhaps even more so than the selection letter. Samples of these letters are shown in Figures 13-4 and 13-5, respectively, and electronic versions can be found at www.sagepub.com/cizek/letters. In both cases, it is important to send the letter well in advance of standard setting so that those selected can begin making arrangements and so that those not selected do not wait and wonder what happened to their candidacy.

[Date]

Dear [Participant Name]:

Congratulations! You have been selected to serve on the [Subject Area] standard-setting committee that will meet in [Location] on [Month, Day, Year]. As you know, the purpose is for you and selected educators from across the state to recommend a cut score to identify students who should receive a high school diploma. This will be a unique opportunity for you to participate in a decision that will significantly affect [State] students and is likely to have important consequences for [State] schools.

I look forward to seeing you on [Day, Date] at [Time], when you will register and pick up the materials you will need. A housing request form is enclosed. Please complete this form and return it to me in the postage-paid envelope or fax it to me at [Fax Number]. The form includes space for you to indicate any special housing or dietary needs you may have. As soon as we receive your housing request, we will make arrangements with the hotel and send you a confirmation and final instructions.

Again, thank you in advance for your service to the profession and your participation in this important endeavor.

Sincerely,

[Contractor Contact Person]

cc: Superintendent

Testing Division Director

Enclosure

Figure 13-4 Sample Participant Selection Notification

Before any final selection decisions are made about panelists, it may be useful to consider historical no-show rates for various activities and committee meetings. For example, if the state agency needs three panels with 25 members each, and if actual participation in similar functions over the years has been 90%, it may be well to recruit 27 participants, with the expectation that 25 would present themselves for duty. This issue merits considerable discussion because absenteeism affects overall demographic makeup of the panels, and nonrepresentativeness can potentially present a validity concern and cast doubt on the credibility of the process and acceptability of the results.

All activities pertaining to the identification, selection, and notification of panelists should take place between about nine and five months before standard setting. For a spring testing program, this timetable fits nicely with

[Date]

Dear [Participant Name]:

Thank you for offering to help establish performance standards for the [Test Name]. We have completed the selection process and regret that we will not be able to offer you a position on one of the committees.

We appreciate your interest in the standard-setting process. There will be additional activities such as item review and test review that will require the assistance of [State] educators in the coming year. We would like to keep your information on file so that we might contact you about serving on one of those committees in the future.

Again, thank you for your willingness to serve the profession and improve education for the students of [State].

Sincerely,

[Contractor Contact Person]

cc: Superintendent

Testing Division Director

Figure 13-5 Sample Nonselection Notification

the beginning of the school year, training program, and so on. In education contexts, teachers and other school personnel who will constitute the largest group of panelists will be in place. For fall standard-setting activities, it is necessary to contact panelists during the winter or spring of the preceding school year, and the loss of potential panelists is greater than for a spring event.

After these three steps have been completed, there is still a need for follow-up. The invitation letter is sent out about five months prior to standard setting in order to allow panelists ample time to schedule the event on their calendars. In the case of professions where highly qualified participants have more demanding schedules, an even earlier notification may be required. Even if the exact location where standard-setting activities will be held is known well in advance, it is best to withhold that information until four to six weeks prior to the event. At that time, a final letter to all panelists should reconfirm their participation and provide the location and driving directions, a reminder of the purpose of the meeting, and contact telephone numbers in case of emergencies. This letter should include detailed information about accommodations or a housing form, with a submission deadline and complete directions for completion and submission.

Upon receipt of housing forms, the sponsoring agency or contractor makes final arrangements with the meeting site facility (usually a hotel or convention center). Once rooms are confirmed, the sponsoring agency or the contractor may send a housing confirmation to each panelist. In some situations, this entire process is conducted with paper (as opposed to e-mail or Web-based registration, for example). However, we have successfully used Web-based procedures for housing and registration, and such procedures are often used by licensing and certification agencies. In the future, electronic procedures will almost certainly be used exclusively, but whatever procedures are used will need to align with the capabilities and preferences of the sponsoring organization.

Of course, just as qualified panelists must be identified, the personnel who will actually conduct the standard setting must be carefully chosen and trained. Such personnel include facilitators, monitors, data entry specialists, and support staff. One person should be designated as the lead facilitator, and a different person should be designated as the logistical coordinator. The lead facilitator will be responsible for all training and other matters directly related to standard setting. The logistical coordinator will be responsible for anything related to hotel guest rooms, meeting rooms, catering, copying, and related matters.

Materials

Every standard-setting method involves training materials, forms, and data analysis programs. The timing of the preparation of these materials is crucial. Some materials can be prepared well in advance; others cannot be prepared until almost the commencement of the standard-setting meeting. The distinction between these two categories of materials should become clear by the end of this section. The timeline shown in Table 13-1 assumes a standard-setting activity based on an operational administration of the test (typically the first such administration), using live, scored data from that administration. Thus some of the programming for standard setting will take place as scoring nears completion. First, however, let us consider those tasks that can be completed well in advance.

Anything generic can and should be prepared as far in advance as possible, to leave time for completion of those materials that have to wait until nearer the time of the standard setting. For example, every application of a particular standard-setting method follows a fairly well-prescribed pattern. Thus the general overview of the procedure to be provided to participants can be prepared in advance: This includes printed materials that will be provided to participants as hard copies, visual materials that will be presented

Table 13-2 Sample Materials List for Standard-Setting Meeting

Participant Materials	*Description*
Agenda	List of major activities, showing start and stop times
Note-taking guide	Usually miniature reproductions of PowerPoint slides with lines for notes
Security agreement form	Binding agreement not to divulge secure contents of tests or other documents
Readiness form	Interim session evaluation form indicating panelist understanding of processes and readiness to participate in standard-setting activities
Expense form	Form for reimbursement of all participant expenses not direct billed to agency or contractor
Bookmark	Form on which panelist enters page numbers for each performance level, by round
Evaluation form	Form used by panelists to evaluate process and outcomes of the session

as PowerPoint or other presentation, and scripts for training that will be presented orally and rehearsed in advance. Table 13-2 contains a list of materials that would be required to conduct a Bookmark standard-setting activity, and every item in Table 13-2 can be created well in advance of the meeting. Electronic examples of these materials are available at www.sagepub.com/cizek/materials1.

Conspicuously missing from this list is the ordered item booklet, that is, the test booklet with items reordered in terms of difficulty. If the test contains both multiple-choice and constructed-response items, it will be necessary to score a large sample of both types of responses, merge the scores, and analyze the results to order the items by difficulty and select scored responses to include in the ordered booklet. (For details of this process, see Chapter 10.) Similarly, for a Body of Work or other holistic procedure, it will be necessary to score responses, select representative student work samples, and assemble those work samples for the various rounds or to use as replacements. (See Chapter 9 for details on this procedure.) For any standard-setting activity that requires panelists to deal with test scores, preparation of those materials will have to wait until tests have actually been scored and analyzed.

As reporting deadlines for educational assessment programs become tighter and tighter, and as examinee expectations rise regarding immediacy of results from credentialing examinations, the window of opportunity for scoring, analyzing, setting standards, and inserting cut scores into the score-reporting programs is shrinking. For some credentialing programs with on-demand, computer-based testing, one response has been to delay score reporting for initial examinees only as long as necessary to obtain a sufficient quantity of scored examinations. In education contexts, one response has been to identify a sample of districts representing the entire state. Those districts are identified in advance, with the approval of the state education agency and review by the TAC. All responses (or even selected responses) from the representative sample of districts are scored first, and those data are analyzed while scoring of the remaining responses continues. For a Bookmark procedure, for example, item difficulty indices can be calculated on the sample in time for facilitators to create ordered booklets, get them approved, and make copies for standard setting. For holistic procedures, the initial responses can be used to generate sufficient numbers of student work samples at various score points to satisfy the requirements of the specific method used.

For all procedures employing impact information (see Chapter 3), those data can be generated either from the initial representative sample or from a larger data set available as standard setting begins. In some instances, if impact information is introduced on the second or third day of a standard-setting activity, it is possible to download and run the latest version of the database the night before presenting impact data so that it reflects the largest possible number of examinees.

Although it is preferable to have all scoring completed and every examinee represented in the impact information that is provided to participants, standard setting sometimes must occur before the end of scoring or data merging. When that is the case, carefully crafted advance agreements as to what should and should not be included in impact information are in order. These advance agreements (among the state education agency, TAC, and contractors) should also include specific actions to be taken in the event of differences between the sample data and the population data when they become available. Although all methods can be affected by differences in score distributions, the Bookmark procedure is particularly sensitive to relatively small differences in item difficulty indices, as these differences can easily translate into a different ordering of the items, potentially invalidating the ordering of items in booklets constructed on the basis of sample item difficulty data. Thus extreme caution must be exercised when sample data are used for standard setting.

Other analytic programs specific to the standard-setting method can be prepared in advance, up to a point. If panelists are to generate ratings, programs to process those ratings can be prepared in advance. In a Bookmark procedure, for example, panelists enter page numbers. If feedback to participants will involve asking them to examine the distribution of page numbers, the spreadsheet or other programs to produce those distributions should be prepared in advance. For feedback that shows how those page numbers translate into cut scores, it will be necessary to wait until the necessary data are available, although the basic structure of that spreadsheet or other program can also be developed in advance. For holistic procedures, specific packet or sample code numbers will have to be entered after all samples have been selected, which may be quite late in the process. In all instances, it is helpful to develop and pilot test a set of generic data processing programs that can be modified to fit each application and updated easily when situation-specific data become available.

Final Preparations

Standard setting is almost always a high-profile activity. Everyone involved needs to be thoroughly prepared. All presentations (e.g., method-specific training) should be scripted and rehearsed. All rating forms should be double-checked and printed in sufficient quantities to complete the activity with extra materials on hand to accommodate unforeseen circumstances. All participant materials should be produced, duplicated, collated, and assembled into easily used sets. For example, materials should appear in packets in the order in which they are to be used or referred to. Table 13-3 provides a sample list of materials. Examples of some of these materials can also be found at www.sagepub.com/cizek/materials2.

If participants or staff will need special equipment or materials (e.g., flip charts, markers, blank overhead transparencies), those materials should be purchased, logged in, and prepared for shipment to the standard-setting site. Everyone involved in the conduct of the standard-setting meeting (e.g., presenters, facilitators, data entry operators, monitors) should be familiar with the final, approved plan, which should also be copied and distributed for reference to those involved. As a final part of preparation, the entire standard-setting staff should conduct a dress rehearsal (in-house dry run), making sure that timing of presentations is consistent with the agenda, that all forms are correct and usable, and that the flow of events is logical. (Sometimes it is helpful to have other staff who are generally familiar with the testing program but unfamiliar with the specific content of the session to serve as panelists.) Confirmation should be obtained that an on-site supplier will have major equipment ready (such as computers, projectors,

Table 13-3 Additional Materials and Supplies for Standard-Setting Meeting

Test booklet(s)	Tape
Answer document(s)	Copy paper
Scoring guide(s)	Whiteboard markers (1 set per room)
Achievement level descriptors (3)	Whiteboard erasers (1 per room)
Difficulty-ordered test booklets	Flip charts (1 per room)
Passage/stimulus booklets	Flip chart markers (1 set per room)
Replacement materials for packets	Airbills for returning materials
Round 1 Summary tables	
Round 1 Summary graphs	*Equipment*
Round 1 Impact data tables	Computers (1 per room)
Round 1 Impact data graphs (2)	LCD projectors (1 per room)
Round 2 Summary tables	Printers (1 per room)
Round 2 Summary graphs	Overhead projector
Round 2 Impact data tables	Disks or USB drives for storing data
Round 2 Impact data graphs (2)	Cables and connectors
Round 3 Summary tables	Pencil sharpeners (1 per room)
Round 3 Summary graphs	Photocopier (or arrangements for copying)
Round 3 Impact data tables	Extension cords
Round 3 Impact data graphs (2)	Power strips
Supplies	Spare bulbs
Self-adhesive notes	Batteries
Pencils	Calculators (for math tests, expense forms)
Pens	
Blank transparencies (printable)	*Other*
Blank transparencies (write-on)	Secure storage site
Overhead markers	Workroom
Boxes	

printers, and copiers), or these items should be inspected and made ready for shipment to the site and rechecked on-site for possible damage in shipment.

At the Standard-Setting Site and Following Up

When they finally begin, the actual standard-setting activities unfold with amazing speed during the meeting. It is at this point that the value of advance preparation becomes apparent. The lead facilitator should attend to matters related to conduct of the sessions only, while the logistics coordinator deals with everything else. For most standard-setting sessions, the day does not end when panelists complete their tasks and turn in their materials at the end of the day. For data-entry staff, the day may be just beginning. For those who will analyze and present results the next morning, it will be a long day—or, perhaps a series of long days and nights. Sometimes staff from the sponsoring agency will want an early morning briefing session prior to the meeting. Whether such a meeting occurs or not, it is clearly beneficial to all facilitators to become familiar with the results they will be presenting as soon as those results become available, which may be at some point during the evening following a session or early the next morning.

All data entry should be verified by a second person before data analysis begins. All analyses should be checked by a person other than the person who ran them. Results should be presented to the facilitators and checked and discussed before making copies, slides, or transparencies for sharing with panelists. Facilitators should discuss results among themselves and with representatives of the sponsoring agency (assuming a common method for all facilitators) to ensure that they have accurate, consistent understandings of the results before presenting those results to panelists. In short, it will be necessary to maintain a single voice throughout the session in order to maintain the confidence of the panelists.

As noted earlier in this chapter, if the standard setting is conducted for an educational assessment, the state education agency responsible for the standard setting should have arranged time on the agenda of the state board of education as soon as possible after standard setting in order to have cut scores approved. If conducted for a licensure or certification board, the agency will need to ensure that time has been allocated for consideration of the results at the appropriate board meeting. Regardless of the context, in the interim between the conclusion of standard setting and the board meeting, the agency staff and the facilitators should prepare summary reports and executive summaries to forward to the board to support the adoption of the cut scores.

During this same interim, it will be helpful to share the results with the other stakeholder groups as appropriate, obtain their reactions, and enlist their support. Presentations to these groups should be as straightforward and nontechnical as possible. If any of the stakeholders have reservations about the recommended cut scores or their impacts, those reservations should be noted in appendices to the main report in support of the cut scores. If one or more authorized groups have specific objections or would recommend alternative cut scores, those cuts should also be appended for the board's consideration, along with an explanation of who recommended them and why.

Once cut scores are adopted by the board, it will be possible to include them in the score reporting programs and produce score reports. Once again, caution is urged. Score reporting programs can be long and complex. Cut score values may need to be inserted at more than one location; for example, cut score values may be required in one location that translates raw scores into a performance level and in another location where an examinee's subtest scores are compared with subtest scores of examinees in different performance levels. Every point in every score reporting program that requires a cut score should be identified well in advance. Programming supervisors and quality assurance staff should know these locations. Before producing live score reports, it is imperative to produce a test batch and check all output against a printed set of approved cut scores. Once the program updates pass this test, it will be possible to produce score reports. If approved score reports refer to performance levels, explanations of the performance categories should be provided for all performance levels, preferably in some location on the examinee score reports themselves, and in whatever additional interpretative aids, manuals, or aggregated score reporting materials are produced (e.g., training program, jurisdiction, building, district reports, etc.). These explanations should be the same ones that standard-setting panelists used in their deliberations.

Scheduling Standard Setting for Credentialing Programs

Scheduling standard setting for credentialing programs poses different challenges than those just described. For example, educational assessment programs are often bound to specific times of the academic year, and tests are typically given in the spring or fall, during the same time period (usually the same week or weeks) every year. All activities, from initial planning to distribution of score reports, must be scheduled to fit into this fixed, recurring time frame.

In contrast to educational assessments, credentialing programs are often not bound by the same constraints. While credentialing programs face unique challenges, these challenges often introduce some degree of flexibility. For example, computer-adaptive testing (CAT) or computer-based testing (CBT) may permit test administration on virtually any day of the year. Some systems may provide for instant score reporting to candidates for a credential. Even when an organization has only a single administration per year, there may be some flexibility in its timing. And, for example, in cases such as when an organization tests only during (or immediately before or immediately after) the annual meeting, that set of circumstances fixes only one point—namely, the administration date—on the schedule, and other dates may still be flexible.

Table 13-4 provides an overview of the major tasks for a credentialing testing program. It is assumed that the program includes both a written examination and a performance assessment component (as well as whatever additional qualifications or other eligibility requirements are in place). As was provided to illustrate scheduling for an educational assessment program, a generic version of the schedule for certification testing shown in Table 13-4 is available in electronic form; the electronic version can be easily adapted to any standard-setting schedule by entering different dates for the test administration, and all other dates are automatically updated. This version is available for download at www.sagepub.com/cizek/certscheduling.

The standard-setting schedule illustrated in this section is actually fairly complicated; it comprises two separate standard-setting activities performed in sequence. We recognize that not all certification programs will have a two-component qualification procedure, but many will. The schedule calls for one activity to be completed in March (for the written examination) and the other in May (for the performance assessment). All activities relevant to a two-component certification assessment program are described in the following portions of this chapter. For readers interested in credentialing programs with only one assessment component, it is a simple matter to eliminate the irrelevant portions from the electronic schedule illustrated in Table 13-4.

Overall Plan

For the hypothetical certification testing program illustrated here, we assume that a national credentialing agency or board decided at their 2001 annual meeting to institute a certification test. The test was to have two assessment components: written and performance. To qualify for the

Table 13-4 Generic Certification Test Planning Calendar

	Timing (Days Out)*		*Schedule (Dates)*	
Event	*Start*	*Finish*	*Start*	*Finish*
Conduct design conference	–730	–700	03/20/01	04/19/01
Conduct job analysis	–336	–304	4/2/2001	5/4/2001
Set dates for standard setting (Written and performance)	–301	–300	5/7/2001	5/8/2001
Secure annual meeting site (One year out)	–301	–300	5/7/2001	5/8/2001
Analyze job analysis results	–301	–297	5/7/2001	5/11/2001
Conduct design conference	–287	–283	5/21/2001	5/25/2001
Develop test and item specifications	–273	–248	6/4/2001	6/29/2001
Develop test items	–245	–171	7/2/2001	9/14/2001
Finalize contract with annual meeting site	–210	–207	8/6/2001	8/9/2001
Review test items	–161	–157	9/24/2001	9/28/2001
Identify standard-setting participants	–161	–136	9/24/2001	10/19/2001
Prepare for field testing	–154	–136	10/1/2001	10/19/2001
Select standard-setting participants	–133	–129	10/22/2001	10/26/2001
Field test items	**–133**	**–108**	**10/22/2001**	**11/16/2001**
Notify standard-setting participants	–109	–105	11/15/2001	11/19/2001
Score field tests	–105	–101	11/19/2001	11/23/2001
Analyze field test data	–98	–87	11/26/2001	12/7/2001
Review field test item data	–84	–80	12/10/2001	12/14/2001
Select items for written tests	–77	–73	12/17/2001	12/21/2001
Notify candidates of written test administration	–61	–52	1/2/2002	1/11/2002
Process applications for written test	–58	–28	1/5/2002	2/4/2002
Draft standard-setting plan	–42	-38	1/21/2002	1/25/2002
Present standard-setting plan to stakeholders	–35	–35	1/28/2002	1/28/2002
Review operational tests	–35	–31	1/28/2002	2/1/2002

(Continued)

Table 13-4 (Continued)

	Timing (Days Out)*		*Schedule (Dates)*	
Event	*Start*	*Finish*	*Start*	*Finish*
Review standard-setting plan	–34	–10	1/29/2002	2/22/2002
Secure standard-setting sites (Both)	–28	–24	2/4/2002	2/8/2002
Print and distribute written tests	–21	–20	2/11/2002	2/12/2002
Prepare standard-setting materials	–21	–17	2/11/2002	2/15/2002
Follow-up letter to panelists (Written test)	–14	–12	2/18/2002	2/20/2002
Finalize standard-setting plan	–7	–3	2/25/2002	3/1/2002
Administer written test	**0**	**4**	**03/04/2002**	**03/08/2002**
Score written tests	8	9	3/12/2002	3/13/2002
Analyze written test data	9	10	3/13/2002	3/14/2002
Set standards for written test	**10**	**11**	**3/14/2002**	**3/15/2002**
Present results to stakeholders	14	18	3/18/2002	3/22/2002
Review/adopt cut scores	21	22	3/25/2002	3/26/2002
Notify candidates of results of written test	23	25	3/27/2002	3/29/2002
Notify candidates of performance test administration	28	32	4/1/2002	4/5/2002
Process applications for performance test	35	46	4/8/2002	4/19/2002
Administer performance test	**64**	**65**	**5/7/2002**	**5/8/2002**
Follow-up letter to panelists (Performance test)	70	71	5/13/2002	5/14/2002
Score performance test	98	100	6/10/2002	6/12/2002
Analyze performance test data	99	100	6/11/2002	6/12/2002
Set standards for performance test	**101**	**102**	**6/13/2002**	**6/14/2002**
Present results to stakeholders	103	112	6/15/2002	6/24/2002
Review/adopt cut scores	115	116	6/27/2002	6/28/2002
Notify candidates of results	**117**	**117**	**6/29/2002**	**6/29/2002**
Prepare final report	119	130	7/1/2002	7/12/2002

NOTE: * Relative to administration of written test

performance component, candidates would first have to pass a written component. The board further decided that the performance component of the first administration of the certification test would take place in conjunction with the association's 2002 annual meeting. That decision gave test developers and standard setters one year to have a performance test in place to administer to candidates who had already passed a written test. This in turn meant that test developers and standard setters would have just less than 10 months to develop the first form of the written test for operational administration.

Fortunately, the first three tasks on the schedule had already been performed—namely, the association had already conducted a job analysis, staff members working for the association had secured the next year's annual meeting site, and the date had been set for performance test component. That left test developers and standard setters the rest of the tasks to complete.

As with the implementation of educational assessments discussed in the first half of this chapter, test development and planning for standard setting go hand in hand. The review of the job analysis focused not only on the tasks performed by incumbents but also on the degree of difficulty of those tasks and the degree of accuracy and completeness clients and the profession expect of a qualified practitioner. The association contracted with a testing company to develop written and performance items to reflect not only the knowledge, skills, and abilities of incumbents but an overall performance level befitting a certified practitioner. Item review by members of the association's examination review committee (experienced practitioners in the field) had a similar focus. Each item was evaluated in terms of its fit to the specifications and its ability to discriminate between qualified and unqualified practitioners. Discussions regarding the essential differences between these two groups were not deferred until late in the test development cycle; consideration of the key hypothetical conceptualization—the minimally qualified candidate—began well in advance of standard setting.

Development of a standard-setting plan formally began about four months before the standard-setting meeting for the performance test and three months before the standard-setting meeting for the written test. Unofficially, the standard-setting planning began to take place as soon as the May 2001 meeting was over (one year prior to the actual standard setting). The board and its standard-setting contractor made several key decisions in May and June 2001. The first was the sequential nature of the two standards. The second was to have a single cut score separating certified from noncertified examinees. Finally, whereas ultimate certification would be configured in a conjunctive manner (i.e., candidates would have to pass

the written exam to qualify for the performance exam), standards on each individual component were to be set using a compensatory approach (i.e., once a cut score was set for each test, any combination of points that met that cut score would be sufficient to pass that particular component).

Other elements of the standard-setting plan began to take shape as the board reached decisions about the job analyses and test and item specifications. Once the test and item specifications were in place, it was possible to determine the basic approach for each standard-setting activity. All that remained to be done in January and February 2002 was to document all of the decisions in writing and organize them around a set of established standard-setting procedures and the schedule of the association.

Participants

As with many certification testing programs, standard-setting participants included a mix of board members, committee chairs, past board members and officers, and other experienced subject matter experts (SMEs) selected from within the membership of the association. There were two standard-setting panels, and some members served on one or the other, and some served on both. As is also common in many certification testing programs, the panels were relatively small; each panel comprised 8–10 members instead of 20–25 participants, as is more common to high-stakes educational assessment standard-setting activities.

Even when the pool of potential participants is small, it is a good idea to establish objective criteria for selection and to make those criteria, as well as the overall selection process, widely known within the association. Having articulated these criteria and procedures, it is then helpful to have in place a formal notification and follow-up procedure to make sure that all panelists are in place and prepared to participate in the standard setting.

Particularly when an agency is certifying candidates for the first time, there may be questions about the qualifications of the standard setters. It may therefore be a good idea to ensure that the initial panel be limited to board members and others who are highly regarded within the association, who have substantial experience and contributions in the field, or who have parallel certification in other organizations. It would be inappropriate for anyone to sit for an examination they have reviewed or for which they have served as a standard-setting participant. (Board members and others who are intimately involved in the development of the first examination form can wait until the second or a subsequent administration and be examined for certification using a test form they have not seen. This approach seems preferable to simply waiving the examination requirement for those

involved in the initial test development process, an approach sometimes referred to as "grandfathering.")

Materials

In this section on materials, we refer principally to the standard-setting materials themselves, as opposed to the field test or operational test materials. For this particular series of standard-setting activities (one for the written test and one for the performance test), it was necessary to prepare two sets of materials, and due to the tight schedule for completion of the activities, the timing was critical. The written qualifying examination comprised selected- and constructed-response format items. A modified Angoff procedure (see Chapter 6) was used for the selected-response (i.e., multiple-choice item) portion of the written examination. Thus some of the materials for that standard-setting activity could be prepared well in advance. For the constructed-response portion of the written examination, actual candidate work samples were used in a generalized holistic procedure (see Chapter 9). Thus it was necessary to wait until those work samples had been scored. Cost and logistical considerations were also involved, as there was an opportunity to score written tests and conduct standard setting in conjunction with a board meeting to be held in March 2002. Board members and other experienced members of the professional association served as scorers, and some of the scorers also served as standard-setting participants. In certification testing, particularly with small organizations, this is frequently the case.

Given the date of the board meeting, the scheduling of other activities was accomplished by working backward through scoring and arriving at a suitable time to administer the written test: about two weeks before the board meeting. Written tests were administered to association member volunteers in a variety of locations, with each location monitored by a member of the contractor staff. Contractor staff scored the multiple-choice portions of the tests and prepared the constructed-response portions for scoring by board members and other members of the association.

Working further back chronologically from the administration date, dates for administering and scoring written and performance field tests were derived. These field tests were conducted over a period of three to four weeks in a variety of locations with association member volunteers. As an adjunct to data analysis, contractor staff asked examinees to provide relevant professional information such as years in practice, other certifications held, practice setting, education level and continuing education units (CEUs) completed, full-time/part-time status, professional achievements, and so on.

Final Preparations

Final preparations were needed for both the written test and the performance assessment. For the written test, after that component had been administered and scored, contractor staff prepared modified Angoff rating sheets for the multiple-choice items and holistic rating sheets for the work samples. They also prepared a brief overview of the modified Angoff procedure to be used for the multiple-choice items and the holistic procedure to be used for the constructed-response items. Copies of the test specifications, job analysis, and other pertinent materials were also prepared for the panelists to review and have available during standard setting.

In addition, contractor staff made arrangements for scorers and standard-setting panelists to travel to a central site for those activities. These arrangements included flight arrangements, hotel accommodations, ground transportation (to and from the airport as well as to and from the hotel each day), meals, and entertainment. (In some ways, these arrangements can be as important to the success of the standard-setting meeting as data analysis or training in the standard-setting method.)

As with the written test, contractor staff prepared holistic rating sheets for participants' use with the performance assessment component. That component consisted entirely of open-ended tasks, and a generalized holistic approach was used. Again, contractor staff made all travel and lodging arrangements for scorers and standard-setting panelists.

Of particular importance in preparing for standard setting for the performance test was the selection of work samples. In Chapter 9, we noted the importance of selecting work samples that represent the full range of performance while avoiding samples that might mislead or confuse participants. The same is true in certification testing. After scoring, contractor staff examined every performance test to select a sample that represented the full range of scores, giving particular emphasis to exams of consistent quality (i.e., uniformly poor, mediocre, or outstanding, or very nearly so).

At the Standard-Setting Site and Following Up

Once on site for the standard-setting meeting, activities for the written and performance components of the test are remarkably similar and fast paced. It is important to remember that members of most professional societies are members of their profession first and standard-setting participants second. Explanations of standard-setting procedures and psychometric concepts have to make sense in their terms—which may differ from perspectives that can often be assumed in a measurement-oriented audience—and

presentations may need to deviate from the standard ways in which testing specialists typically explain concepts. In addition, leading a group of high-profile participants through a standard-setting activity will require a considerable amount of skill on the part of the meeting facilitator. We found it helpful for the facilitator to work with the board on several aspects of test development and planning during the year leading up to standard setting. By the time of standard setting, the panelists and facilitator had established a solid working relationship and mutual respect.

In addition to orientation to the process, which is best provided by an officer or senior member of the credentialing board, training in the standard-setting methods to be used is critical. Because many volunteers serving as participants are busy professionals, the training must be streamlined and time-efficient. And perhaps the most critical aspect of the training phase will be introducing participants to the concept of the minimally qualified or "borderline" candidate. For one thing, this is often not a familiar notion to those in a profession. And, for many specialists (particularly in health-related professions), there is some degree of discomfort in wrestling with the notion of credentialing a "minimally qualified candidate": Members of a profession often believe that minimally qualified is not good enough. Obviously, the training should focus on whatever level of competence the participants (and the board) have agreed upon as adequate for receiving a credential. To accomplish this, the creation of the PLD is critical. If this is done in advance by policymakers or those in leadership roles at the organization, it can be provided to participants in the standard-setting meeting and serve as a basis for discussion. If not already created, a generous amount of time may need to be allocated for developing the PLD(s) during the standard-setting meeting.

Following the standard-setting meeting, there are two final activities. First, the board or other authority must review and approve the cut scores. Even when every member of the board has served as a participant or in some other capacity in the development of the certification test, it will be necessary for them to meet officially and adopt the cut scores. The contractor, too, makes important determinations at this point, as is standard procedure when important decisions are based at least in part on test results. It is then necessary for the contractor to check every examination, particularly those close to the cut score, to make sure each one has been properly scored and to correct any mistakes or oversights. Having certified the results, the contractor provides to the board a list of names and scores, noting those candidates who have met the performance standard. The board, or its designated committee, then notifies the successful and unsuccessful candidates of their results, or the board may have arranged for the

contractor to send out notifications, score reports, and perhaps credentials on behalf of the board as part of the contracted slate of services.

Finally, for the integrity of the certification testing program (or, as we have noted previously, for the integrity of *any* program), it will be necessary to document all aspects of the development of the tests and all activities related to standard setting. Because the development and standard-setting processes play out over a considerable period of time, we recommend documenting in stages, for example, job analysis, item development, test construction, planning for standard setting, implementing standard setting, test administration, and so on. Each of these phases has a built-in review cycle. As standard setters plan and carry out these interim activities and provide the necessary documentation for their review and approval, it is helpful to consider that the same reports (or portions thereof) might be used in the final report if they are planned with that eventuality in mind.

Conclusions and Recommendations

To our knowledge, this chapter provides the first formal treatment of scheduling activities for standard setting. We have gone into some detail on the specifics of scheduling; however, we have given less attention to another important topic: the interactions between licensure and certification boards and testing contractors. On the one hand, we recognize that some organizations do not contract for many (or any) testing-related services. Some associations register candidates, develop and administer examinations, produce score reports, and continually monitor, document, and improve the quality of their examination services. On the other hand, many organizations—perhaps due in part to resources and size of the organization—contract for some or all of the critical examination services they provide to their members. Readers who are interested in a more comprehensive treatment of the intricacies of contracting for testing services are referred to the work of Trent and Roeber (2006) on that topic.

We are uncertain why the topic of scheduling has not received much attention, particularly given the facts that the activity is so essential and that problems with scheduling can be so consequential. Perhaps the topic has escaped notice because it is such a simple matter or perhaps because it is viewed as a matter of personal style or mere logistic necessity: There is a due date, and we work backward from that point to figure out when and how to complete all the intermediate steps.

We hope this chapter has helped highlight that scheduling is indeed more than mere logistics. With both educational assessments and certification

examination programs, there are opportunities to shape the final product from conception to delivery. We believe that standard setters typically become involved in the process after tests are developed and it is time to set standards. However, we also believe that—in order for standard setting to be accomplished as effectively as possible—it is ill-advised to delay consideration of standard setting or to fail to fully integrate it into the examination development and scheduling of activities.

As we have attempted to show in this chapter, there are key decisions to be made about performance standards (cut scores) from the moment a credentialing testing program is considered. With educational assessments, for example, an understanding of the eventual number and general nature of the performance levels will greatly enhance item-development efforts. For instance, knowing that there will be two groups above the *Proficient* level, rather than one, will help item-development staff plan for upper-level items. Knowing that professionals seeking certification will be required to demonstrate a particular level of sophistication with and command of the basic concepts of their field will guide performance task developers more clearly than simply knowing that the test will measure, say, accounting skills. In other words, PLDs, whether formally or informally articulated, are as integral a part of the test-development process as they are of the standard-setting process.

Our recommendations are therefore straightforward. We propose that planning for standard setting be made an integral part of planning for test development. If a credentialing board and contractor are involved, and if different staff will be involved in each activity, it may be helpful to have an initial joint meeting, periodic meetings, or at least exchanges of reports over the life of the test-development activity. Plans of the standard-setting facilitators should be reviewed by test development staff, and vice versa. Moreover, one person—preferably one with authority over both item developers and standard setters, if they are different—should have informed oversight over both activities. By informed oversight, we mean the ability to see potential problems, conflicts, or disconnects between test development and standard setting and the authority to correct them.

Finally, we recommend attention to scoring. Particularly with open-ended or constructed-response items, scheduling of both development and standard-setting activities must include a large portion of time for scoring of the constructed responses by human beings. In addition to these time considerations, however, time must be allocated for the development of rubrics for individual items in light of the global performance standards. As we noted in our example in the previous section with reference to the certification examination, board members were fairly clear about the level of

competence expected of a certified member of their profession. This expectation was translated into test items addressing skills at the expected competence level and, ultimately, into a final cut score that distinguished between candidates at or below the desired competence level.

But it is also necessary to translate this expectation into the scoring rubrics for individual items. An item rubric that is more lenient or more stringent than the overall performance standard can lead to confusion. A scoring rubric that awards maximum points for a response that a standard would not consider typical or even marginal for a person who otherwise demonstrates the desired competency level creates problems at more than one level. Standard-setting participants are susceptible to the urge to apply scoring rubrics to individual items; many will try to rescore responses instead of viewing them holistically in the context of the standard-setting task they are asked to perform. Even those who resist the temptation to rescore constructed responses will usually be stymied by a rubric that is out of phase with the overall performance standard. Thus we recommend that attention to how performance tasks and item rubrics are developed, how constructed responses will be scored, and how the rubric development and item scoring processes will be described to participants in standard setting should be considered early in the test development process and in the overall plan for standard setting.

In summary, we have tried to convey that test planning, test development, scoring, and standard setting are interlinked parts of a single enterprise. In this chapter, we have attempted to show how planning for one must of necessity carefully consider the others. With examples and interactive calendars, we have illustrated how these activities can be carried out logically and effectively. We note that the calendars we have provided are based on actual test programs. However, we also note that they are merely illustrative, not prescriptive. Each standard setter (or test developer) faces a different set of constraints and logistical challenges. We hope that the examples and recommendations we have provided here will help others plan and schedule all test-related activities with confidence.

14

Vertically-Moderated Standard Setting

In the chapters comprising Section II of this book, each of the methods of setting performance standards we described is routinely applied to contexts in which the need is for a single cut score or set of performance levels for a single test. These contexts have included setting a single cut score on a licensing examination, deriving three performance levels (e.g., *Basic, Proficient, Advanced)* on a single end-of-grade subject area test, and obtaining multiple cut scores on a set of four or five different subject tests used to determine eligibility for high school graduation. In each instance, there was no apparent need to link cut scores on one test to those on another. Indeed, until fairly recently, there had been little or no need for a method to link standards across tests.

In at least one context, however, that has changed. Federal legislation now requires states to administer tests in Grades 3–8 and in high school to measure the annual progress of students in reading and mathematics. (Additional tests in science are also required, but not in a contiguous set of grade levels.) According to the law, a key purpose of the legislation is

> to determine the success of children served under this part in meeting the State student academic achievement standards, and to provide information to teachers, parents, and students on the progress being made toward meeting the State student academic achievement standards described in section 6311(b)(1)(D)(ii) of this title. (NCLB, 2001, Sec. 6312 [b][1][A][I])

As is well-known, the legislation contains key language regarding progress. According to the *No Child Left Behind* (NCLB) *Act:*

> "Adequate yearly progress" shall be defined by the State in a manner that:
>
> (i) applies the same high standards of academic achievement to all public elementary school and secondary school students in the State; (ii) is statistically valid and reliable; (iii) results in continuous and substantial academic improvement for all students; (iv) measures the progress of public elementary schools, secondary schools and local educational agencies and the State based primarily on the academic assessments described in paragraph (3); (v) includes separate measurable annual objectives for continuous and substantial improvement for each of the following: (I) The achievement of all public elementary school and secondary school students; (II) The achievement of: (aa) economically disadvantaged students; (bb) students from major racial and ethnic groups; [and other specified subgroups]. (NCLB, 2001, Sec. 6311[b][2][C])

The requirement for the measurement of adequate yearly progress (AYP) immediately raises the issue of grade-to-grade growth, relative to the standards for each grade. If performance standards on reading and mathematics tests are set independently for each of the grades and subjects, the ability to make claims about AYP is severely weakened; it is challenging, if not impossible.

As Cizek (2005) has observed, the passage of NCLB has forced those interested in standard setting to confront this challenge. One of the most substantial of the challenges has been that of creating a coherent system of performance standards across grades and subjects that will make inferences about progress meaningful and as accurate as possible. Or it is perhaps more accurate to say that there exists a system of interrelated challenges that center on how to link one test to another.

The Interrelated Challenges

Psychometricians have developed sound methods for linking tests across grades within the comparatively friendly context of norm-referenced testing (NRT). NRTs have typically been created by commercial test publishers who can impose stringent controls over the content and the statistical characteristics of test forms. The inferences traditionally supportable based on NRT performance (i.e., relative status of students) were also amenable to cross-grade

linkages because a within-grade NRT could be designed to span a comparatively broad, cross-grade range of knowledge and skills. In the context of NRTs, the technology to accomplish the linking of tests across grades (sometimes called **vertical equating** when the constructs measured are the same and other conditions are met) represents reasonably well-developed psychometrics.

In contrast, the relatively newer assessment type to which NCLB applies—standards-referenced testing (SRT)—requires tests built to statistical specifications that routinely call for comparatively narrower targets and content specifications that are also narrower and tightly matched to specific within-grade content standards that often do not have considerable across-grade overlap. The content standards upon which SRTs are based can thus militate against the construction of traditional cross-grade scales; vertically linking SRTs requires strong assumptions about the equivalence of constructs being assessed at different levels. The less stringent form of equating appropriate for such situations (actually not equating at all) is sometimes referred to as vertical scaling or vertical linking.

A second issue has to do with the existence (empirically determined or theoretically assumed) of a continuous, developmental construct across grade levels. Reasonable arguments have been made for the existence of an underlying developmental construct in, say, mathematics; equally well-reasoned arguments have been proffered to support the position that grade-to-grade mathematics achievement differs not only in quantity but also in the nature of the construct assessed. Even linking two different tests that are designed to measure the same variable and express their scores on the same scale is viewed as a difficult psychometric challenge (see, e.g., Cizek, Kenney, Kolen, Peters, & van der Linden, 1999; National Research Council, 1999).

Other very specific challenges center on the sheer number of performance levels that NCLB requires. According to the Act, two levels representing higher achievement (*Proficient* and *Advanced*) are required, as is a lower level (*Basic*). This need for multiple levels is compounded by the requirement of performance standards on three different tests (reading, mathematics, and science) at several grade levels (Grades 3–8, plus one secondary for reading and mathematics and three grade levels for science). Thus, if a state implemented only the minimal requirements of NCLB, a total of 51 cut scores would be needed to delineate the boundaries of 68 performance categories.

But, as Cizek (2005) has noted, it is not simply the sheer number of performance standards required that makes cross-grade linkage a challenge. The fact that the tests span such a wide grade and developmental range introduces another layer of complexity. It would be considerably less of a technical feat to link SRTs in such a way as to permit fairly confident

statements about progress between two adjacent elementary grades. It is quite a different matter to construct a scale that permits accurate inferences about progress between, say, Grade 3 and the senior year of high school.

The need for establishing so many performance standards can also be contrasted with several realities: (1) the fact that different standard-setting methods may be required depending on the nature, format, and other characteristics of a test; (2) the research evidence in the field of standard setting that different standard-setting methods often produce different results, and (3) the finding that, even when the same method is used on a test by equivalent groups of participants, resulting cut scores can vary, sometimes a great deal. These realities mean that a system of cut scores *across grades within one content area* could vary in ways that would be illogical if, in fact, there existed a continuously developed ability being measured and if standard setters were consciously trying to maintain a consistent standard of performance across grade levels.

Further, the panels of educators who participate in standard setting are only one source of potential variability in performance standards across grade levels. It is almost certain that the panels of educators and others who participated in the development of the *content standards* for those grade levels would not have been instructed, intended, or even able to create content standards of equal difficulty across the grades. In fact, it is a reasonable hypothesis that even if a consistent standard-setting procedure were faithfully executed for each grade level, observed variation in recommended performance standards across grade levels would be more attributable to the content-standards development process than to the standard setting.

The same realities are reflected in the problem of coherence of standards across subject areas. A system of cut scores *within a grade level across different areas* could also vary. For example, standard setters could recommend cut scores for *Proficient* that would result in 80% of students being classified in that performance category in reading, 30% in mathematics, and 55% in science. Such a result would be illogical if, for example, there were no educational reason to suspect differential achievement in the subjects, if the content standards for the areas were of approximately equal difficulty, if instruction were equally allocated to the respective content standards, and if standard setters were consciously trying to maintain a consistent standard of proficiency across the subjects. And, as noted previously, there is often no reason to assume that the panels of educators and others who developed the content standards for these three subject areas were necessarily focused on creating content standards that were of equivalent difficulty across subjects. Thus, in this example, it may well be that the relatively lower performance of students in mathematics compared to

reading can in part be attributed to the fact that the mathematics content standards are much more challenging.

Addressing the challenges, it would seem, would involve developing and implementing standard-setting methods that set performance levels in concert, that is, across all affected grade levels (and perhaps subject areas) with some method for smoothing out differences between grades. One approach to the challenge is found in what has come to be known as **vertically-moderated standard setting** (VMSS). In this chapter, we describe the general concept of VMSS, some specific approaches to conducting VMSS, and a specific application of VMSS. We conclude with an assessment of the current state of the art and a call for additional research as well as standardization of goals, terms, and methodology.

A Brief History of Vertically-Moderated Standard Setting

As is perhaps evident at this point, the challenges to creating meaningful systems of performance standards across grades and subject areas are many. The psychometric technology to address the challenges is in its infancy, however. One approach has been proposed by Lissitz and Huynh (2003b), who introduced the concept of VMSS in a background paper prepared for the Arkansas Department of Education. The paper, titled "Vertical Equating for the Arkansas ACTAAP Assessments: Issues and Solutions in Determination of Adequate Yearly Progress and School Accountability," spelled out the problem (demonstration of AYP) and proposed a solution: VMSS.

In that paper, Lissitz and Huynh noted, "While expectations for the *Proficient* level will vary by state, AYP is based on the percent of students meeting *Proficient* and the expected percentage increases over time" (2003b). In essence, this refines the purpose of VMSS as one deriving at a set of cross-grade standards that realistically tracks student growth over time and provides a reasonable expectation of growth from one grade to the next. The heart of the matter, then, is defining reasonable expectations.

Lissitz and Huynh (2003b) first explored the possibility of setting reasonable expectations through vertical scaling or equating. After citing various psychometric and practical problems with the applicability of vertical scaling to most statewide achievement exams, they concluded that vertical scaling would generally not produce a satisfactory set of expectations for grade-to-grade growth, noting, "The result of examining these many issues, is that the construction of a vertical scale is difficult to accomplish, difficult to justify, and difficult to utilize productively" (p. 12).

As an alternative to vertical scaling (equating), Lissitz and Huynh (2003b) recommended VMSS. Specifically, they recommended that "new cut scores for each test be set for all grades such that (a) each achievement level has the same (generic) meaning across all grades, and (b) the proportion of students in each achievement level follow a growth curve trend across these grades." They then offered a list of specific recommendations to carry out the vertical moderation of a set of performance standards across Grades 3–8, including an annual validation of the process.

Shortly after Lissitz and Huynh's (2003b) paper, an entire issue of *Applied Measurement in Education* (2005) was devoted to the topic of VMSS. An introduction and six articles addressed the issue from both theoretical and practical perspectives, with a particular focus on how state and local education agencies can align performance standards across subjects and grades. Although introduced in the context of NCLB, the use of VMSS is not necessarily limited to that situation, and may be adaptable to use whenever it is desired to establish meaningful progressions of standards across levels or to enable reasonable predictions of student classifications over time when traditional vertical equating is not possible. Of course, it remains to be seen if VMSS will provide satisfactory solutions to the current contexts to which it is being applied, as well as how adaptable VMSS will be to other contexts.

What Is VMSS?

VMSS—also sometimes referred to as a process of **vertical articulation** of standards—is a new and developing concept. Lissitz and Huynh have defined VMSS as

> [A] judgmental process and a statistical process that, when coupled, will enable each school to project these categories of student performance forward to predict whether each student is likely to attain the minimum, or proficient, standard for graduation, consistent with NCLB requirements. (2003a)

Simply put, VMSS is a procedure or set of procedures, typically carried out after individual standards have been set, that seeks to smooth out the bumps that inevitably occur across grades. Reasonable expectations are typically stated in terms of percentages of students at or above a consequential performance level, such as Proficient.

To illustrate VMSS, consider the following scenario for a typical state testing program. Six groups of standard setters have gathered to set

Table 14-1 Hypothetical Results of Independent Standard Settings Across Six Grade Levels

Grade	*Percentage of Students Classified as At or Above Proficient Performance Level*
3	37
4	41
5	34
6	43
7	29
8	42

standards for the state mathematics tests in Grades 3–8. Given the placement of the cut scores for the *Proficient* category, percentages of students that would be classified as *Proficient* or above are shown, by grade level, in Table 14-1. In this example, we see that the percentages of students considered *Proficient* or better (e.g., *Proficient* or *Advanced*), goes up from Grade 3 to Grade 4, drops back down at Grade 5, goes up again at Grade 6, drops again at Grade 7, and rises again at Grade 8.

To grasp the practical implications of VMSS, imagine that we were discussing these test results with a group of parents of fourth or sixth graders whose children had scored at the *Proficient* level. If we believed the groups of students on whom these results were based were typical (i.e., we would expect similar results next year with a new group of students), we would probably need to point out to the parents that their currently *Proficient* children would only have about a 75% chance of scoring at the *Proficient* level the next year. Why? Because the standards have been set in such a way that fifth and seventh graders have a lower probability of scoring at the Proficient level than do fourth and sixth graders. This means that many Proficient fourth and sixth graders are going to appear to lose ground in the subsequent grades (17% and 33%, respectively), much to the consternation of this group of parents, the fifth- and seventh-grade teachers, school administrators, and others.

To remedy this situation, VMSS requires reexamination of the cut scores and percentages in light of historical or other corollary information available at the time of standard setting. For example, if a state or district also

administers an NRT in some or all of these grades, how have students historically done? Do they lose ground—relative to the national norm group—over time, hold steady, or gain ground? If cohorts of students have been tracked over time, what has been their direction on the NRT? If the state standards-based test has been in place for some time at some or all of these grades, how have students performed, either over time for a single cohort or by grade within each year? And, of course, how have the students typically performed on the National Assessment of Educational Progress (NAEP)? Have fourth and eighth graders scored similarly, or has one grade consistently done better than the other? Answers to these questions play a major role in suggesting appropriate adjustments to the cut scores for the six grades so that, whatever happens to our hypothetical groups of fourth and sixth graders, we have a reasonable expectation of what *should* happen.

Approaches to VMSS

VMSS can be thought of as one alternative to vertical scaling and methods for aligning scores, scales, or proficiency levels. VMSS requires few assumptions about the scalability of the scores across grades or even comparability of scores across grades. Indeed, rather than address scores per se, VMSS typically focuses on percentages of students at various proficiency levels.

A simple solution to the problem of different percentages of students reaching a given performance level—say, the *Proficient* cut score—at different grades would simply be to set all standards at the same score point or such that equal percentages of students would be classified as *Proficient* at each grade level, by fiat. An alternative would be to set standards only for the lowest and highest grades and then align the percentages of *Proficient* students in the intermediate grades accordingly. In the preceding example, we would simply take the 37% figure for Grade 3 and the 42% figure for Grade 8 and set cut scores for Grades 4–7 so that their resulting percentages of students at or above *Proficient* would fall on a straight line between 37% and 42% as shown in Table 14-2.

If sound standard-setting procedures were used for establishing cut scores for Grades 3 and 8, then there is some reason to have confidence in at least the 37% and 42% values for the anchor grades. In fact, in some instances, VMSS has assumed precisely this form; that is, a linear trend has been imposed on intervening grade levels to obtain cut scores for those grades. In the situation just described, a linear trend seems plausible. However, in all cases, VMSS is based on assumptions about growth in achievement over time.

Table 14-2 Results of Smoothing Standard-Setting Results Across Six Grade Levels

Grade	*Percentage of Students Classified as At or Above Proficient Performance Level*
3	37
4	38
5	39
6	40
7	41
8	42

Lewis and Haug (2005) have identified such assumptions regarding growth over time. According to these scenarios, the percentage of students classified as at or above Proficient would be expected to be (1) equal across grades or subjects, (2) approximately equal, (3) smoothly decreasing, or (4) smoothly increasing. Lewis and Haug also assumed a consistent scale across grades, with higher and higher cut scores for *Proficient* at each succeeding grade.

Another constellation of growth possibilities was posited by Ferrara, Johnson, and Chen (2005). According to these authors, assumptions for standard setting are based on the intersection of three growth models and four expected growth amounts as depicted in Table 14-3. The three growth models may be summarized as follows: (1) *Linear growth,* which assumes that the proficiency of all examinees increases by a fixed amount, and examinees retain their positions relative to one another; (2) *Remediation,* which assumes that the proficiency of examinees at the lower end of the score distribution increases more than those of examinees at the upper end; and (3) *Acceleration,* which assumes that the proficiency of examinees in the upper portion of the score distribution increases at a greater rate than that of examinees at the lower end of the score distribution.

Applications of VMSS

Because of the newness of VMSS, only recently have results of VMSS approaches been presented in the psychometric literature. As we noted

Table 14-3 Matrix of Growth Patterns by Model Type and Growth Trajectory

Growth Trajectories	*Growth Model Types*		
	Linear	*Remediation*	*Acceleration*
Negative growth	All groups show decline over time.	Overall group shows negative growth, but remedial group fares better than nonremedial group.	Overall group shows negative growth, but nonremedial group fares better than remedial group.
No growth	All groups show no growth over time.	Overall group shows no growth, but remedial group fares better than nonremedial group.	Overall group shows no growth, but nonremedial group fares better than remedial group.
Low growth	All groups show low growth over time.	Overall group shows low growth, with most gain coming from remedial group.	Overall group shows low growth, with most gain coming from nonremedial group.
Moderate growth	All groups show moderate growth over time.	Overall group shows moderate growth, with most gain coming from remedial group.	Overall group shows moderate growth, with most gain coming from nonremedial group.

NOTE: Adapted from Ferrara, Johnson, and Chen (2005).

earlier, several studies were published in a special issue of *Applied Measurement in Education* (2005). In one of those studies, Lewis and Haug (2005) used VMSS to set standards in Colorado (Grades 3–10 in writing and Grades 5–10 in mathematics). Buckendahl, Huynh, Siskind, and Saunders (2005) applied VMSS for standard setting on a science examination (Grades 3–6) in South Carolina. Ferrara et al. (2005) set standards for reading in Grades K–3 in another (unnamed) state. These three applications had many common elements, as well as some intriguing differences.

The VMSS procedure used by Lewis and Haug (2005) was characterized by a considerable amount of pre–standard-setting preparation. Working with committees of Colorado educators, they developed models of

performance for Colorado students over time. Using this information, they established preliminary cut scores and made this information available to standard-setting participants. Participant training was extensive: Panelists took the actual tests for which they were to recommend cut scores and received a thorough introduction to the Bookmark procedure. Within a given subject area (writing or mathematics), panelists worked in a single room at grade-specific tables. There were frequent all-room discussions, such as after each round of bookmark placement, so that by the end of the process, each panelist was fully aware of the set of recommended cut scores across the complete grade spectrum. The preliminary cut scores, based on the models developed previously, were introduced during these discussions. The discussions led to compromises such that, by the end of the standard-setting session, the cut scores had a more or less smooth trajectory across grades, with writing following an equal percentage trajectory (at around 50% *Proficient* or above at each grade) and mathematics following a smoothly decreasing trajectory (declining from 54% at Grade 3 to 24% at Grade 10).

Ultimately, the Colorado State Department of Education (CSDE) made some modifications to these moderated cut scores. For the writing tests, the CSDE adopted standards that resulted in 50% of students at or above *Proficient* in each grade. These standards were generally within one or two percentage points of the panel's recommendations, except at fifth grade (7% above panelists' recommendation) and eighth grade (4% below the recommendation). In mathematics, CSDE adopted the panelists' recommended performance standards for *Proficient* at every grade level.

Buckendahl et al. (2005) took a somewhat different approach. In South Carolina, standard-setting participants met to recommend standards for science tests in Grades 3–6. Previously, South Carolina had administered science tests in Grades 3, 6, and 8, as well as tests in mathematics in Grades 3–6 and NAEP science tests in Grades 4 and 8. Thus there was considerable information available from which to construct a model of cross-grade performance. The South Carolina Department of Education (SCDE) assembled this information; the information was not presented to the standard-setting panelists, but it was provided to the Technical Advisory Committee (TAC) for the state's assessment programs.

Standard-setting participants worked in two-grade groups: 3–4 and 5–6. Within each group, panelists examined test items from two grades and made their recommendations for cut scores for these two grades. After the final round of bookmarks, participants completed questionnaires about their satisfaction with the process. Their cut score recommendations, along with calculated standard errors, were then forwarded to the SCDE and the

TAC. These two groups considered the recommended cut scores, standard errors, and historical data (mathematics and science scores from 1996 through 1999). In the end, the SCDE adopted a set of standards that had an essentially equal percentage (at the *Proficient* level or above) projectile. Interestingly, the final cut scores for *Proficient* matched perfectly the recommendation of participants representing Grades 4 and 5 and differed by less than three percentage points at Grade 3. At Grade 6, the participants recommended a cut score that would have classified 43% of students at or above *Proficient;* the final adopted cut score resulted in 20.1% of sixth graders being classified at this level. As a system of performance standards, the adopted cut scores followed a smoothly decreasing performance trajectory (with 23.2%, 21.9%, 21.4%, and 20.1% of third, fourth, fifth, and sixth graders, respectively, scoring at or above *Proficient*). Compared to the final adopted cut scores, participants' original cut score recommendations represented slight underestimations of the performance of third graders (with 18.9% scoring at or above *Proficient* by their reckoning) and considerable overestimation of the performance of sixth graders.

Ferrara at al. (2005) followed procedures similar to those of Lewis and Haug (2005). However, the context in which they applied VMSS was that of a new testing program at grades where there were no historical data to permit confident assumptions about appropriate growth models, either with or without the help of statewide committees of educators. Thus there were few external data to inform standard-setting participants. The authors did, however, share results of bookmark standard setting across grades. Perhaps the most unique part of this application of VMSS—at least regarding the future of VMSS—was that Ferrara and colleagues collected information *from* the panelists. At the end of the process, panelists responded to a questionnaire regarding the value of cross-grade articulation information and their satisfaction with the process—data that can be used to refine future applications of VMSS procedures.

Each of the three previously described studies was groundbreaking in its own way: Each made important contributions. The studies can also be analyzed as a group to synthesize the essential components of a set of procedures that will likely serve as the nucleus of VMSS applications in the future. Six such components are described in the paragraphs that follow.

1. *Grounding in historical data.* Lewis and Haug (2005) and Buckendahl et al. (2005) collected and used historical performance data to prepare for and interpret results of standard setting. Wherever such historical data are

available or obtainable, they can be and should be used in VMSS. Collection of historical data and planning for their use may include discussions with stakeholders and content experts in advance of standard setting.

2. *Establishment of performance models.* Whatever assumptions are made regarding performance, the performance models should be based on the historical evidence. Where such evidence is unavailable, models might instead rely on theories of cognitive development, discussions with content experts and stakeholders, or generalization from other tests. For example, South Carolina used data from their statewide mathematics tests to project growth curves for science test performance.

3. *Consideration of historical data.* When available, historical data should also be presented to those involved in setting standards. Importantly, we include in that group the participants who work through the multiple rounds of a standard-setting procedure, cross-grade or cross-subject articulation panels or "meta-panels," as well as subsequent groups such as TACs, state department of education staff, and state board of education members who ultimately vote on the standards.

4. *Cross-grade examination of test content and student performance.* Each of the previously described studies included some degree of cross-grade review by standard-setting participants. If cut scores are to be articulated across grades, it seems reasonable that the cut scores for a given grade be considered by individuals with strong interests in the performances of students in at least the adjacent grades. Where possible, all-grade review should be included in a full-scale VMSS for at least one round, either the final round or at some point just prior to the final round.

5. *Polling of participants.* Two studies of VMSS (Buckendahl et al., 2005; Ferrara et al., 2005) included the collection of data from participants at the end of the standard-setting activity. Such information is essential not only as vital validity evidence for the standard-setting activity at hand but also for future standard-setting activities. For the immediate application, these data may be used to support the cut scores or to show that support is weak, paving the way for more aggressive modifications by the responsible agency to fit a model.

6. *Follow-up review and adjustment.* In each of the studies described previously, it is likely that facilitators of the standard-setting meetings reminded participants that their work would be reviewed by technical experts and was subject to review and possible modifications by state officials. This notification was specifically stated in the Buckendahl et al. (2005) study. Such

follow-up and adjustment are important for two reasons. First, elected or appointed state officials are responsible for the successful implementation of the performance standards. Such responsibility should be undergirded by commensurate authority. Second, even with the best intentions and earnest application of standard-setting techniques, participants may still hold fairly disparate notions with regard to where cut scores should be set. South Carolina's application of the standard error of the mean for the final round of standard setting is fairly typical of even single-test standard-setting activities and has been employed for many years as an adjustment tool for state performance standards (see Chapter 16 for further discussion of such adjustments). This tool should have a place on the VMSS practitioner's workbench.

An Illustration of VMSS Procedures

To illustrate VMSS methods, the steps followed in a recent application of VMSS for establishing cut scores on an English language learners' (ELL) test will be described. The test, the English Language Development Assessment (ELDA) spans Grades 3 through 12 and was developed by the Council of Chief State School Officers (CCSSO) and a collaborative of 16 member states. The following sections describe the ELDA assessment, preparations for the VMSS approach, training of participants in the procedure, facilitation of the standard-setting meeting, the actual vertical articulation of cut scores, and final review and adoption of cut scores.

The Assessment Context

The ELDA is actually a complex assessment system consisting of four separate components: Reading, Writing, Listening, and Speaking. The Reading and Listening tests comprise 50 to 60 multiple-choice items. The Writing tests include both multiple-choice and constructed-response items. The Speaking tests consist of 16 speaking prompts to which students respond, either on tape or in the presence of a trained scorer who scores the response as the student takes the test. The ELDA forms were developed for three grade ranges (Grades 3–5, Grades 6–8, and Grades 9–12), and five performance level categories were desired (necessitating the setting of four cut scores for each of the grade ranges). The five performance level labels (PLLs) adopted for the ELDA were, from lowest to highest levels of performance: Level 1 (Pre-functional), Level 2 (Beginner), Level 3 (Intermediate), Level 4 (Advanced), and Level 5 (Fully English Proficient).

Items for the tests administered in 2005 were field-tested in 2004, and preliminary standards were set in the fall of 2004. This initial activity provided some of the historical information needed for vertical moderation. The field-testing protocol also featured cross-grade embedding of items, permitting Rasch scaling of the tests, and a continuous score scale spanning Grades 3 through 12. The availability of such a vertical scale is seen as advantageous to VMSS, but not necessary.

Preparing to Implement a VMSS Approach

First it is relevant to describe the preliminary standard setting for the ELDA, conducted in 2004. Preparations for and conduct of that standard setting provided considerable guidance for the 2005 standard setting. During administration of the ELDA field tests in the spring of 2004, teachers completed student background questionnaires that included spaces to indicate the teachers' evaluations of the students' proficiency levels on each of the four components. The five proficiency levels were defined in a manner consistent with the performance level descriptors (PLDs) used by the standard setters. Thus, in the 2004 standard setting, panelists were able to compare their ratings to teacher judgments and modify them accordingly if they were so inclined. The 2004 articulation committee considered the teacher judgments, presented as cut scores based on a contrasting-groups procedure, as they deliberated the final cut scores for each test.

Agency staff, representatives of the member states, and the contractor studied the 2004 contrasting-groups outcomes and the 2004 cut scores in preparation for the 2005 standard setting. Although they did not present all the 2004 data to the 2005 standard-setting panelists, they did include a summary overview in the orientation of the 2005 panels.

Prior to the 2005 standard setting, student responses from the 2005 operational test were analyzed, and scores were placed on a common scale. This cross-grade scaling permitted facilitators to see the range of performances within a grade cluster and, importantly, to compare performance across grade clusters on a common scale. As would be expected, performance was generally better at higher grades than at lower grades, in terms of both common items and scale scores, although there were interesting departures from this general finding. For example, two of the writing items yielded higher scores for students in Grades 6–8 than for students in Grades 9–12. Even with these two items in the mix, however, Writing test scale scores for students in Grades 9–12 were higher than those for students in Grades 6–8.

The facilitators prepared ordered item booklets for each of the 12 tests (3 grade clusters × 4 test components) to be used in a bookmark standard-setting procedure. For the preliminary standard setting in 2004, a holistic approach (similar to the one described in Chapter 15 of this volume) was used for the Speaking tests. In 2005, owing to a change in the method of scoring the Speaking tests, the Bookmark method was used for all ELDA component tests. Pertinent information from the 2004 standard setting, along with scaling information for the 2005 operational tests, was included in the overview presentation to the 2005 standard-setting participants. The information was not extensive or technical in nature but did provide a historical context for the 2005 standard-setting activity. In the course of the 2005 standard setting, additional information about the cross-grade differences (such as the differences on the two writing test items noted earlier) was shared during the between-rounds discussions.

To organize the initial standard-setting activity, participants were assigned to one of four groups: (1) Grades 3–5 (all tests except Speaking); (2) Grades 6–8 (all tests except Speaking); (3) Grades 9–12 (all tests except Speaking); and (4) Speaking (all grades). Each group had 8–10 members. In addition, two to three members from each of the four groups were selected to form a fifth group, the articulation committee or "meta-panel," which had 10 members. Panel assignments were made well in advance, and the participants in each group selected to serve on the articulation committee were identified to those groups to ensure that they received as much pertinent information as possible to take forward to the articulation committee. The four groups met August 15 through August 17, 2005, to complete two rounds of bookmark standard setting. The articulation committee met on August 18–19, 2005, to review and modify as necessary the cut scores recommended by the four standard-setting panels. A complete agenda for this VMSS procedure is shown in Figure 14-1.

Training VMSS Participants

Training of participants began with an overview of the purpose of the standard setting and a whole-group introduction to the activity. The introduction included an extensive review and discussion of PLDs for five performance levels. This orientation was followed by an introduction to the Bookmark standard-setting procedure. At the conclusion of the introduction to the Bookmark method, participants practiced the method by setting one bookmark in a six-page reading test booklet. At the end of the practice round, participants completed the first section of a readiness evaluation questionnaire. A copy of this form is shown in Figure 14-2.

Monday, August 15—Grade Span Standard-Setting Committees

8:15 A.M.	Continental breakfast and materials
8:45	Introductions and overview
9:30	Performance level descriptors
10:45	Break
11:00	Orientation to standard-setting procedures
12:00 Noon	Lunch
1:00 P.M.	Reconvene in separate rooms; Standard setting Round 1
4:00	Wrap-up
4:30	Adjourn

Tuesday, August 16—Grade Span Standard-Setting Committees

8:15 A.M.	Continental breakfast and materials
8:45	Continuation of Round 1
12:00 Noon	Lunch
1:00 P.M.	Continuation of Round 1
4:00	Wrap-up
4:30	Adjourn

Wednesday, August 17—Grade Span Standard-Setting Committees

8:15 A.M.	Continental breakfast and materials
8:45	Round 2
12:00 Noon	Lunch
1:00 P.M.	Continuation of Round 2
4:00	Wrap-up
4:30	Adjourn

Thursday, August 18—Articulation Committee

8:15 A.M.	Continental breakfast
8:45	Orientation to articulation committee tasks
10:00	Break
10:15	Reading (all grades)
12:00 Noon	Lunch
1:00 P.M.	Speaking (all grades)
2:30	Listening (all grades)
4:30	Adjourn

Friday, August 19—Articulation Committee

8:15 A.M.	Continental breakfast and materials
8:45	Writing (all grades)
12:00 Noon	Lunch
1:00 P.M.	Comprehension and composite
4:00	Wrap-up
4:30	Adjourn

Figure 14-1 Agenda for VMSS Standard-Setting Procedure

Facilitating the VMSS Standard-Setting Meeting

After training, the four groups adjourned to separate rooms, each with a trained Bookmark facilitator. Observers from the CCSSO and the member states also rotated from room to room to monitor the progress of standard setting and field any policy questions that might arise. Impact data were presented between Rounds 1 and 2.

Between rounds, the four facilitators and the observers reviewed and discussed the bookmark placements, resulting cut scores, and impact data. These between-round discussions were made easier by the fact that each

Participant ID Number_____

1. Training:	I have completed the training, and I understand what I need to do to complete Round 1.	
(Circle one):	**Yes**	**No**
2. Round 1:	I have discussed the results of Round 1, including my ratings, the ratings of others, and the impact data, and I understand what I need to do to complete Round 2.	
(Circle one):	**Yes**	**No**
3. Round 2:	I have completed my ratings, and I believe that the ratings I provided fairly represent the performances of students entering the *Beginner, Intermediate, Advanced*, and *Fully English Proficient* performance levels.	
(Circle one):	**Yes**	**No**

Figure 14-2 VMSS Evaluation Questionnaire

panel actually examined three different tests in each round, and each round lasted a full day. Thus there was ample time at the end of each day to analyze the data, discuss the results, note any developing cross-grade discrepancies, and discuss strategies for bringing those discrepancies to light in the panels and dealing with them. Because the tests had been created on a common-score scale, the focus was on the scale value of each cut score rather than the percentages of students at each grade cluster scoring at or above a cut score. Thus, for example, if the Level 3 cut scores for listening were 1.7 for Grades 6–8 and 1.5 for Grades 9–12, the facilitators would be able to focus attention on this discrepancy during Round 2 with both panels.

After participants discussed the results of Round 1 and received impact data as well as information about their own distributions of bookmarks and resulting cut scores, they completed the second section on the Evaluation Questionnaire (see Figure 14-2) and began Round 2. Round 2 proceeded in much the same manner as Round 1. At the end of the second round, participants turned in their bookmarks, responded to the final question on the Evaluation Questionnaire, and were dismissed.

Vertical Articulation of Cut Scores

The following morning (August 18), the 10 members of the articulation committee met for an orientation meeting. Customizable copies of the

Table 14-4 Reading Cross-Grade Cut Scores Presented to Articulation Committee

Grades 3–5				*Grades 6–8*			*Grades 9–12*		
Level	*Cut*	*% At or Above*	*Theta @ Cut*	*Cut*	*% At or Above*	*Theta @ Cut*	*Cut*	*% At or Above*	*Theta @ Cut*
5	44	23.6	2.00	45	11.9	2.49	51	24.5	2.19
4	38	44.4	1.00	38	37.3	1.35	44	42.1	1.58
3	33	57.5	0.45	32	56.9	0.66	36	58.0	0.86
2	26	80.7	–0.24	20	86.4	–0.57	27	74.4	0.16
1	—	100.0		—	100.0		—	100.0	

orientation presentation for this meeting, as well as the training materials for the VMSS procedure, are available at www.sagepub.com/cizek/vmss.

The lead facilitator presented the 12 cut scores for each set of tests (4 cut scores × 3 performance levels) in both tabular and graphic form, as shown in Table 14-4 and Figure 14-3. The task of the articulation committee was then to examine these cut scores, determine if they were reasonable, and if necessary, modify one or more of them to make them more reasonable.

One of the chief ground rules for the articulation committee was that any modification of any cut score would have to be grounded in the content of the respective test and the PLDs. Thus it was not permissible to raise or lower a cut score simply to smooth a line; the new cut score would have to be traced to a page in the ordered item booklet and justified on the basis of test content and the PLDs.

As can be seen in Table 14-4 and perhaps more readily in Figure 14-3, there was a steady progression from grade cluster to grade cluster with respect to placement of the cut scores for Levels 3 and 4. At Level 2, however, the cut scores for grade clusters 3–5 and 6–8 are almost identical. At Level 5, the cut score for grade cluster 6–8 was actually higher, in theta (i.e., student ability) units, than the cut score for Grades 9–12. The articulation committee first acknowledged these facts, examined the PLDs for reading, examined the bookmarks underlying these cut scores, and made a final determination of cut score level by level and grade cluster by grade cluster.

The remainder of this discussion focuses specifically on the three Reading tests, which were the first to be considered by the articulation

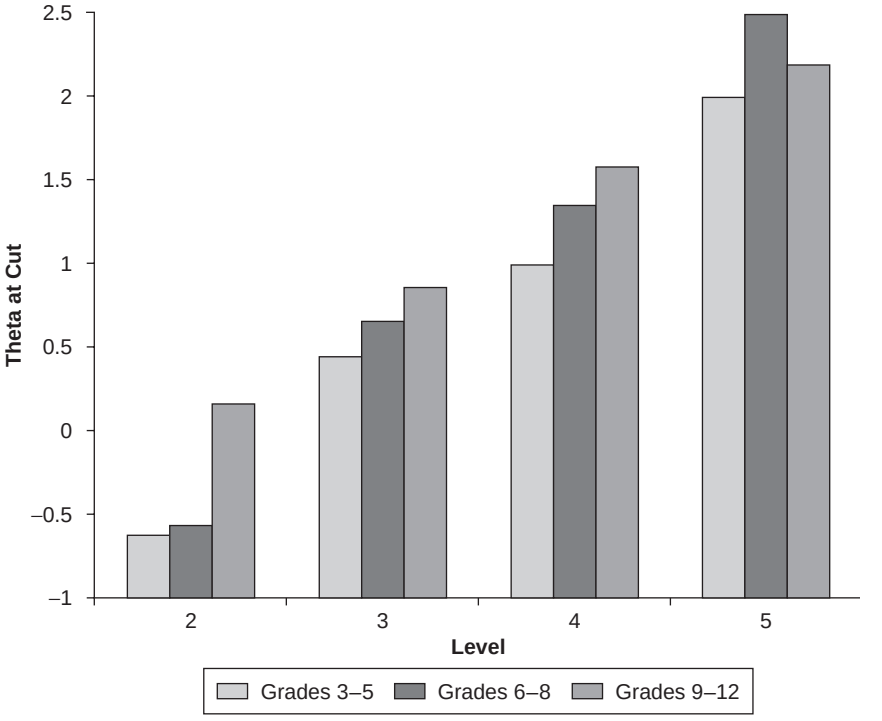


Figure 14-3 Reading Cross-Grade Cut Scores Presented to Articulation Committee

committee. Their review started with an overview of all the cuts and progressed to a comparison of Level 5 cuts across all three grade clusters, ending with the remaining cuts. The committee worked collaboratively, and consensus was not imposed by the facilitator. Any participant could recommend a review of any cut for any grade cluster. At some point in the discussion, the facilitator would ask if the group were now satisfied with all cuts for all grade clusters for that test. The committee would first approve any modification and eventually the complete set of cut scores for each test. Votes were taken after all discussions and adjustments had taken place. The revised Reading cuts were approved by the articulation committee by an 8-0 margin with two abstentions; for each of the other tests, the committee approved the final recommendations by a 10-0 vote with no abstentions.

Table 14-5 shows the Reading cut scores after two rounds of Bookmark standard setting, as well as the final recommended (i.e., vertically moderated) cut scores. The articulation committee altered 5 of the 12 cut scores

Table 14-5 Round 2 and Final Recommendations of the Articulation Committee (Reading) by Grade Range and Performance Level

	Pre-Functional (1)		*Beginner (2)*		*Intermediate (3)*		*Advanced (4)*		*FEP (5)*	
	Raw	*Theta*	*Raw*	*Theta*	*Raw*	*Theta*	*Raw*	*Theta*	*Raw*	*Theta*
Grades 3–5										
Round 2	NA		26	–.24	33	.45	38	1.00	44	2.00
Final	NA		**22**	**–.62**	33	.45	38	1.00	44	2.00
Grades 6–8										
Round 2	NA		23	–.31	33	.80	40	1.61	45	2.50
Final	NA		**20**	**–.54**	**32**	**.71**	**38**	**1.35**	45	2.50
Grades 9–12										
Round 2	NA		27	.16	36	.86	44	1.58	51	2.31
Final	NA		27	.16	36	.86	44	1.58	**51**	**2.66**

for Reading, which was the largest number of changes for any of the four tests. Across the other three tests, the articulation committee made only three changes, two in Speaking, one in Listening, and none in Writing. In each instance, the committee justified the modification on the grounds of test content and the specifications of the PLDs. Each move of a cut score was preceded by a recommendation by a member of the committee that the cut score be examined, followed by a review of the content of the bookmarked pages that would have led to the cut.

A review of Table 14-5 shows that most of the Reading cut score changes (three) came in Grades 6–8. In general, the articulation committee believed that the lower and higher grades were well separated, but that the intermediate grades were too close to the higher grades. Thus they lowered the cut scores for Levels 2, 3, and 4. These alterations, as all others, were strictly based in a review of content. In each instance, committee members examined the contents of the ordered item booklet and justified the reductions on the basis of test content and the PLDs. At Grades 3–5, they lowered the cut score for Level 2 in the same manner. Indeed, the revision of the Level 2 cut score for Grades 3–5 was prompted by the drop in the Level 2 cut score for Grades 6–8. At Grades 9–12, they moved the collective bookmark

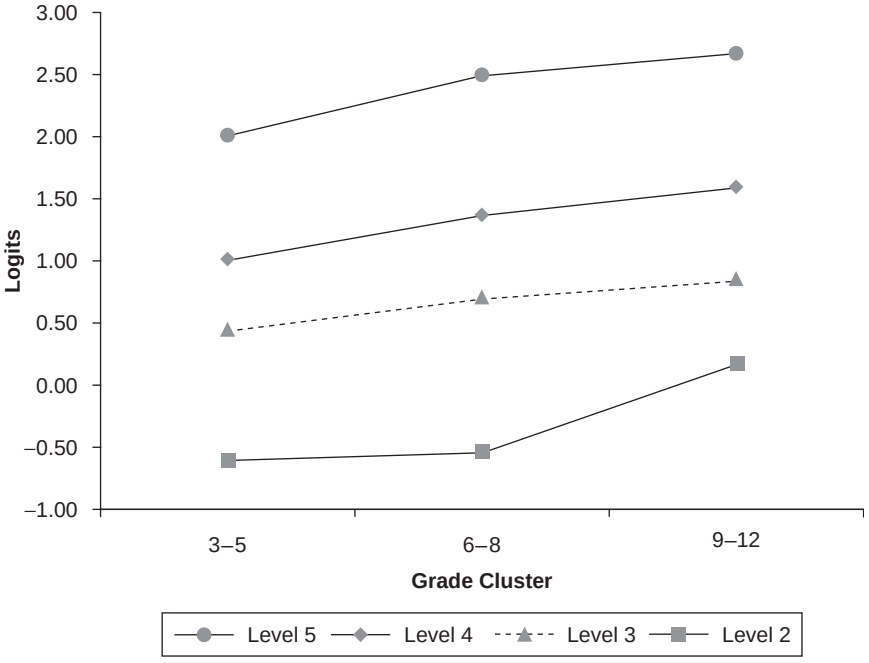


Figure 14-4 Final Cut Scores for Reading

up by a page, increasing the theta value of the cut but leaving the raw cut score unchanged, as is frequently the case in Bookmark standard setting.

In many instances, because the full documentation of the first two rounds was available, the committee examined all bookmarked pages for a particular cut before making their final recommendations. In the end, the recommendation was put to a vote, with a simple majority prevailing. As noted earlier, the final votes were never simple majorities, but unanimous decisions. Figure 14-4 shows the final, articulated cut scores for the three reading tests.

A review of Figure 14-4 reveals steadily increasing cut scores both within and across grade clusters, but the lines are not perfectly straight. At Level 2, there is very little difference between the cut scores for Grades 3–5 and Grades 6–8, and then a sharp upturn to Grades 9–12. For the other levels, there is a more orderly progression from grade cluster to grade cluster.

Final Review, Adoption, and Conclusions

For the final review and adoption of the ELDA cut scores, the articulation committee's recommendations went directly to the sponsoring

organization, CCSSO. The first step of the final review consisted of examination of the results by the TAC and representatives of the cooperating states. After considerable discussion and review of some individual cut scores, these groups adopted all cut scores recommended by the articulation committee.

In conclusion, it would seem that the VMSS process implemented for the ELDA was very successful. We must note, however, that the application presented here departs from the others we have described in this chapter in that the ELDA standard-setting process made use of a vertical scale that had been incorporated into the assessment. The articulation committee and subsequent review bodies focused on alignment of cut scores more in terms of theta than percentages of students in each category (although they did examine these percentages). As Table 14-4 shows, the final distributions of students were not strictly equal across grade clusters, nor did they conform to a steadily increasing or increasing projection across grade clusters. Figure 14-5 summarizes these projections graphically. As the figure shows, the percentages are more similar at two of the performance levels (3 and 4) than they are at the other two (2 and 5).

These differences draw attention to the fact that this particular student population is different from the populations of interest in other VMSS applications. Specifically, ELL students do not generally make up long-term cohorts. Indeed, the goal for ELL programs is to get students to Level 5 and exit. Thus, for any group represented in this standard setting, there is little reason to expect students in later grades to have had more of some instructional impact than students in earlier grades. The results observed here are consistent with what CCSSO found in the 2004 standard setting based on field testing, namely, that cumulative effects of instruction do not hold in ELL programs because students do not remain in the program the way they do in programs assessed by a statewide mathematics test, for example. In such a program, it is expected that students enter at Grade 3 (or the lowest grade tested) and remain through Grade 8 (or the highest grade tested). Such is not the case with the programs addressed by ELDA.

Alternative Procedures and Limitations

Because of the stage of development of VMSS, limitations and alternative procedures for VMSS are somewhat difficult to pinpoint. Although VMSS procedures are increasingly being used, to date there have been few thoroughly documented applications of VMSS. In addition, each application has been slightly different from the others—perhaps attributable to the many and

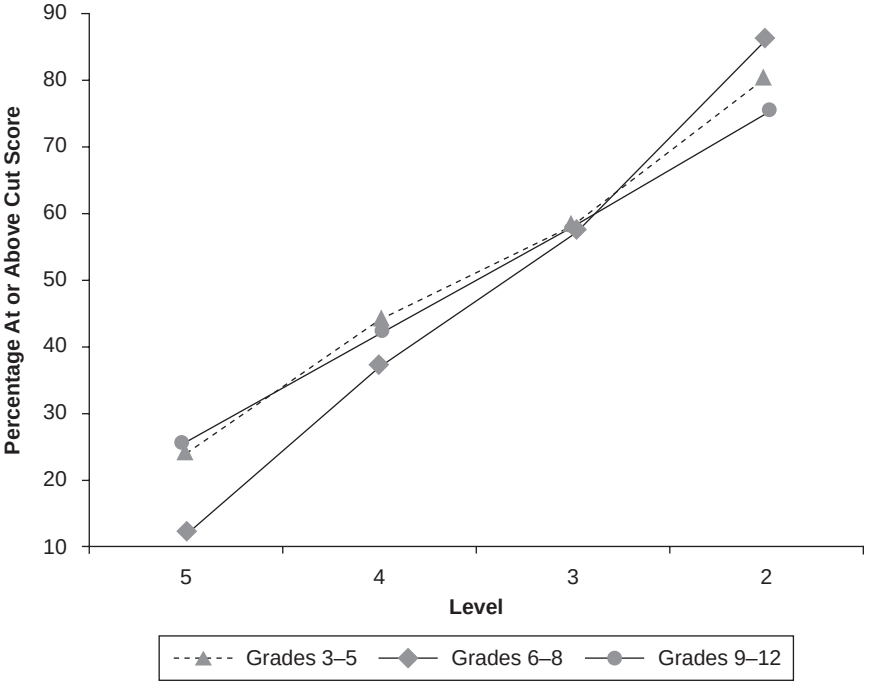


Figure 14-5 Percentages of Examinees at or Above Reading Cut Scores

diverse contexts in which the procedure has been applied. We have suggested a common core of elements to VMSS. These include (1) grounding in historical data, (2) establishment of performance models, (3) consideration of historical data, (4) cross-level examination of test content and examinee performance, (5) polling of participants, and (6) follow-up review and adjustment. However, we must also note that no fixed set of steps has emerged in applications of VMSS so far. Thus nearly every aspect of any application might be thought of as an alternative procedure. At least one alternative is clear—the incorporation of vertical scaling to support VMSS procedures—and research is needed to help illuminate how VMSS results might differ in the contexts of within-grade scaling and cross-grade (i.e., vertical) scaling.

By the same token, any limitations observed so far may well be limitations of the specific procedures employed rather than of VMSS in general. However, specific limitations can be seen. For example, from a global perspective, it would seem that lack of historical perspective or context would be one limitation of any application of VMSS. If the focus of VMSS is

percentages of students at or above a particular proficiency level, lack of historical perspective would seem not just limiting but positively debilitating.

Second, any application of VMSS is hampered if it is not supported by a theoretically or empirically sound model of achievement growth. Without a rational basis for a model of student performance, the enterprise reduces to guesswork, and the best one can hope for is that the cut scores at least follow some kind of curve that does not double back on itself.

Third, even when the work of an articulation committee is grounded in original performance standards set by grade and subject level committees that were obtained via an accepted content-referenced procedure (e.g., Angoff, 1971; Bookmark [Mitzel et al., 2001], or similar procedure), the difficulty in maintaining a content anchoring is obvious. From both theory-development and practical perspectives, the issue of how to maintain the meaning of cut scores and fidelity to PLDs is probably one of the most fundamental for future research to address.

It is clear that, as a line of inquiry for measurement specialists, VMSS research and development is a growth industry. As Cizek has noted, "VMSS is really in its infancy; contradiction and complications abound" (2005, p. 7). Just as the 1980s and 1990s witnessed the introduction of a new type of achievement test (SRTs) and assessments comprising mixed-item formats, psychometricians responded with new standard-setting techniques and embarked on a period of research and refinement that eventually yielded dependable and defensible procedures that have been found acceptable and useful to practitioners and stakeholders. In the late 1980s and early 1990s, there were no established procedures for setting standards for mixed-format SRT with multiple performance levels. Educational necessity prompted measurement specialists to develop a host of procedures for setting performance standards on these tests. Today, the measurement challenge is the need to combine those cut scores in ways that permit meaningful and confident inferences about student growth and serve longitudinal program evaluation purposes. We are encouraged that, in the form of VMSS, necessity has again given birth to appropriate psychometric procedures and that measurement specialists will refine these procedures and continue to develop others that help answer important educational and policy questions.

15

Standard Setting on Alternate Assessments

The standard-setting methods described thus far in this book have at least two things in common. First, they are intended to be applied to tests of fairly uniform composition in terms of item formats, whether it be multiple-choice, constructed-response, essay, performance, or some combination of these item formats. Second—and perhaps most relevant to the content of this chapter—the methods described so far are customarily applied to a test form that is common across test takers; that is, in most cases all examinees take the same test form and face the same items or tasks on which the performance standards have been set. However, not all assessments are so uniform. Many statewide educational achievement testing programs include a so-called **alternate assessment** component. Alternate assessments have been mandated by a collection of federal legislation, beginning with the *Individuals with Disabilities Education Act* (IDEA, 1997). The mandates have been refined via the *No Child Left Behind Act* (NCLB, 2001) and the *Individuals with Disabilities Education Improvement Act* (IDEIA, 2004). Under these laws, most students with disabilities or other special conditions must participate in the *regular* statewide assessment program and take the same test as all other students, although they may be granted some test accommodations deemed appropriate to their special needs (e.g., extended time, large print, scribe services, etc.). Table 15-1 provides a classification scheme for test **accommodations** and limited examples of each type.

Table 15-1 Categories and Examples of Test Accommodations

Accommodation Type	*Example*
Setting	Provide accessible furniture, individual or small-group administration
Timing	Allow extra time, frequent breaks during testing
Scheduling	Administer test over several sessions, different days or times
Presentation	Provide test in audiotape or large-print version
Response	Allow scribe to record student's answers, permit oral responses
Other	Permit use of highlighters, "reading rulers," or other aids

SOURCE: Adapted from Thurlow and Thompson (2004).

The Unique Challenges of Alternate Assessment

Because some students have particularly severe disabilities, however, it may not be possible for them to participate in a state's regular assessment program, even with accommodations. Instead, these students may be best measured using an entirely different assessment, called an alternate assessment. As used in educational achievement testing contexts, alternate assessments refer to tests that are specially developed for those students—usually not more than 1%–2% of all students—who have significant cognitive impairments or extremely limited English language skills. This fact suggests the first challenging aspect of alternate assessments: The total sample size for a given grade or subject is likely to be far below sample sizes that are recommended for applying traditional statistical or psychometric procedures. Even smaller sample sizes result when scores on alternate assessments are disaggregated separately within grade level by gender, ethnic group, or other variables for purposes of demonstrating adequate yearly progress (AYP).

The size of the alternate assessment population gives rise to another challenge: variability from year to year and from program to program. Cut scores established one year with a given sample of students may not offer the same promise of generalizability to the next year's population as might be afforded by the cut scores derived for the state's general assessment program.

The greatest challenge to setting standards for alternate assessments may be the nature of the assessments themselves. Alternate assessments are usually configured based primarily (or entirely) on a **collection of evidence** (COE) of student performance gathered over time. Sometimes, these student work samples are also referred to as **portfolio assessment** systems. It is increasingly common that alternate assessment systems are guided by a set of criteria and specifications concerning the number, type, and focus of work samples that may be placed in the collection of evidence. However, because of the individualized nature of alternate assessments, no two collections of evidence focus on precisely the same learning outcomes, contain the same number of work samples, and/or present information about student achievement in exactly the same medium. This characteristic presents yet another measurement challenge. COEs may be highly structured and feature some elements common for some students, although it is typical for COEs to be highly idiosyncratic.

The determination of whether a student should participate in a state's regular or alternate assessment program is made by the **individualized education plan** (**IEP**) team impaneled to direct the educational experiences and expectations of a particular student. IEPs specify long- and short-term goals for students and their teachers to pursue over the course of an academic year. Typically, those goals are based on the content standards of the state's official curriculum. However, IDEIA 2004 also permits states to develop alternate content standards, linked to the state's regular content standards, and upon which the alternate assessments are based. These alternate content standards are modifications or extensions of the regular content standards adapted to address the needs of a student with severe mental or physical impairments. Thus, for example, a high school mathematics objective in numerical operations might be translated into correctly counting change from a dollar.

By extension, it can be seen that even outstanding performance on so-called extended content standards may not satisfy the requirements of even the lowest performance level (e.g., *Basic*) or be adequately captured by the performance level description (PLD) for that level. Thus, in addition to the obvious challenge of creating assessments on which severely impaired students can demonstrate their levels of knowledge and skill, the way in which data from alternate assessments can or should be aggregated into the performance levels that exist for the standard assessment program presents a daunting measurement challenge.

Another significant measurement challenge is the interpretation or inference that can be supported based on the COE. For example, teachers who have considerable experience in creating COEs may present an entirely

different picture of the progress of a given student than would be provided for that student by a teacher who has less experience. Thus the inference that the COE reflects solely (or even primarily) achievement of the examinee is often only weakly supported.

In a related vein, many alternate assessment scoring systems include a component that is essentially an evaluation of the appropriateness of the tasks and standards-related experiences that the child's teacher has provided. Thus, to a greater or lesser extent, alternate assessment systems can be perceived as personnel evaluations by teachers or other education professionals or paraprofessionals who assemble the COE. This aspect of alternate assessments also threatens the validity of scores on the assessments and can induce a considerable amount of gamesmanship in the creation, maintenance, and scoring of a COE. As one step toward increasing the clarity and supportability of the intended inference, we strongly recommend that in the setting of performance standards for and scoring of COEs, it is always preferable to adopt a system that distinguishes between the achievements of the students and those of their teachers.

In conclusion, just as it is difficult to develop assessments that are individually tailored to students with special needs, it is clear that establishing performance standards (i.e., cut scores) for alternate assessment systems presents an array of unique challenges. Perhaps the most common strategy for setting performance standards on alternate assessments involves a holistic rating approach. In the following sections of this chapter, we describe a generalized holistic method for setting standards for alternate assessments and provide a practical example from a statewide assessment program.

Necessary Conditions for Alternate Assessment Systems

There are several necessary conditions that must be in place before performance standards can be established for alternate assessments. First, it is necessary to establish a mechanism whereby scores can be assigned in a reasonably consistent fashion to any portfolio or COE, regardless of who completes it or whose performance it describes. This condition can be satisfied through the creation of a blueprint or matrix describing essential elements that must be included in every portfolio, along with guidelines for other materials that can and cannot be included. Figure 15-1 depicts such a blueprint developed for the Virginia Alternate Assessment Program.

To appreciate the blueprint shown in Figure 15-1, it is first necessary to realize that quantification in portfolio assessment is not necessarily a

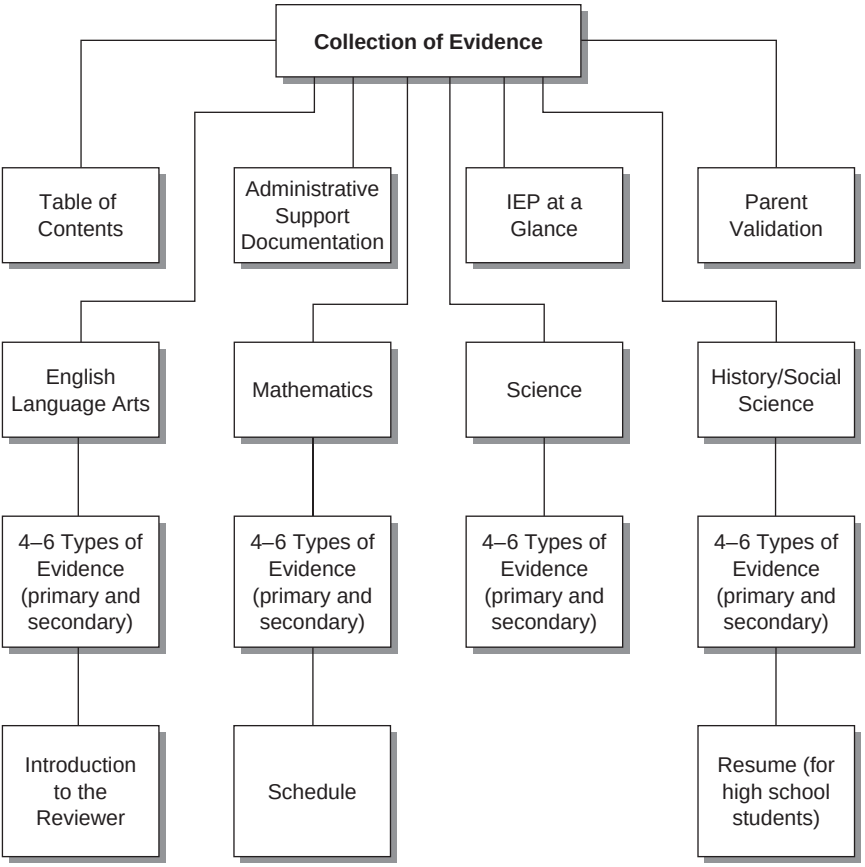


Figure 15-1 Blueprint for a Collection of Evidence (COE)

universally embraced concept. To some, quantification is regarded as, at best, a distortion of and, at worst, a barrier to accurate evaluation of the progress of students. Even when quantification is accepted, the notion of uniform and replicable quantification of student progress has taken some getting used to. For example, the idea of imposing a uniform set of rules for building a portfolio, quantifying student progress, and getting at least two people to agree on what a particular COE represents is sometimes viewed as antithetical to the aims of gathering individualized collections of student work (i.e., portfolios) in the first place. These objections notwithstanding, a uniform system such as the one employed in Virginia is essential for successful (or at least replicable) standard setting.

Once a general blueprint for creating the COE or portfolio has been developed, the next condition to be met is the establishment of a consistent scoring guide. Figure 15-2 shows a typical scoring guide generalized from many such scoring guides in use across the states. The collection of evidence is evaluated on five dimensions. Each of the five dimensions is rated on a 3-point scale. The total score for the collection thus ranges from 5 to 15 points.

For the scoring guide shown in Figure 15-2, the five dimensions are defined as follows: (1) Linkage to standards—the degree to which goals addressed by the COE relate to the state's content standards, (2) Performance—evidence of the student's mastery of one or more of the goals, (3) Settings—evidence of generalization of mastery in a variety of settings and social situations, (4) Context—the degree of meaningfulness and age-appropriateness of the tasks, and (5) Independence—evidence of use of appropriate supports for student performance.

The second necessary condition is the provision of high-quality training for those who will score the COEs. Prior to standard setting, the COEs must be evaluated by trained scorers. In the example depicted in Figure 15-2, the total range of scores is 5 to 15. Even if two scorers are used and their scores are combined, the net range of scores is only 10 to 30 points. It is therefore crucial that each score point be consistently and accurately rendered.

Typically, scorer training focuses on defining the boundaries between score points for each dimension. Trainers present selected COEs they believe represent solid 1s, 2s, and 3s and later COEs that illustrate performance on the borderlines between 1 and 2 and between 2 and 3 in order to sharpen scorers' discriminating abilities. To qualify to score, candidates typically are required to score a screened set of COEs and match the preassigned scores (i.e., assigned by master scorers or managers) at a specified level of consistency.

As may be apparent in Figure 15-2, only one of the five dimensions in this scoring guide refers to what the student actually did (i.e., the Performance dimension). The other four dimensions relate more directly to things the teacher did (e.g., provide a diversity of Settings in which the student could demonstrate the learning). Moreover, the five dimensions are equally weighted; that is, each contributes equally to the final total score. Thus, in one sense, the student's performance contributes only 3 points, while the teacher's or paraprofessional's actions contribute up to 12. During training for scoring of COEs, it is important to address this issue and head off any tendency that scorers might have to overcompensate for it. For example, some scorers will want to give the student the benefit of the

Scoring Dimension	Score Point Values and Descriptions		
	1	*2*	*3*
Linkage to Standards	No outcomes appropriate to the content area are listed, but neither the targeted IEP goals nor the tasks relate to the content standards.	Outcomes appropriate to the content area are listed, with either targeted IEP goals or tasks related to the content standards.	Outcomes appropriate to the content area are listed, and targeted IEP goals and tasks relate to the content standards.
Performance	There is little or no evidence of student performance of tasks related to the targeted IEP goals.	There is some evidence of student performance of tasks related to targeted IEP goals.	There is considerable evidence of student performance of tasks related to targeted IEP goals.
Settings	Student performs tasks primarily in one classroom with limited social interactions.	Student performs tasks in a limited variety of settings with opportunity for some interaction with peers.	Student performs tasks in a variety of settings and engages in social interactions with a diverse range of peers.
Contexts	Student performs tasks that are not meaningful or uses instructional materials that are not age-appropriate.	Student uses age-appropriate materials to perform some meaningful tasks that lead to real-world application.	Student uses age-appropriate materials to perform meaningful tasks in real-world context.
Independence	Limited or no use of appropriate supports as specified in COE documentation.	Some use of appropriate supports as specified in COE documentation.	Consistent use of appropriate supports as specified in COE documentation.

Figure 15-2 Sample Scoring Guide for Collections of Evidence (COEs)

doubt if the COE is poorly constructed or if the teacher or paraprofessional has failed to make entries for multiple contexts or require the student to demonstrate the desired behavior in a variety of settings and social interactions, reasoning that it is not the student's fault that the teacher or paraprofessional shortchanged him or her. It is our experience that if scorers (and, later, standard-setting participants) are allowed to give general (not COE-specific) feedback on how teachers and paraprofessionals can improve next year's COEs, they have a greater tendency to adhere faithfully to the scoring rubric.

COEs for alternate assessments often consist of a variety of media. For example, it is not uncommon for a single COE to include a videotape, a CD/DVD, an audiotape, photographs, drawings, craft items, and worksheets in addition to other paper-based materials. Thus it is necessary to train scorers to evaluate media elements efficiently and accurately. At a minimum, each scorer needs to know how to load a tape, play it, rewind it, and return it properly to its case or perform similar tasks with a CD or DVD. Beyond that, scorers need to know what to look for in a photograph, videotape, or DVD and what to listen for in a tape or CD, as these entries are likely to cut across dimensions within a subject and perhaps even across subjects.

The third essential condition for setting standards on alternate assessments is accurate scoring of COEs. Once scorers have been trained and qualified, scoring can begin. Throughout scoring, it will be necessary to keep COEs flowing efficiently through the scoring center (assuming centralized, as opposed to distributed, scoring). It will also be necessary to keep the physical requirements of scoring working efficiently. It is unusual for each scorer to have his or her own personal set of audiovisual equipment. These items are typically shared by a room full of scorers. It is therefore necessary to establish procedures for sharing these resources so that bottlenecks do not arise. Fortunately, COEs containing media-dependent entries also contain other entries that scorers can evaluate while waiting for a machine to become available.

As a way of foreshadowing an issue we will consider later in this chapter, the process of assigning scores to COEs naturally causes important questions to arise, such as "What do these scores actually mean?" and "What do the numbers represent?" As we indicated previously, four of the five scores have more to do with the actions of a teacher or paraprofessional than with the student. The adult chooses the tasks, contexts, and settings. The linkage to state curriculum standards may be tight or loose, the contexts may be real-world or contrived, the settings may be diverse or limited, supports may be nearly complete or entirely absent. The student has no opportunity to affect those dimensions of his or her score. In a very real

sense, the assigned score represents an interaction between the performance of the student and the teacher in a far more obvious way than, for example, a score on a 50-item reading test does.

The final necessary condition for setting performance standards on alternate assessments is the presence of performance level labels (PLLs) and PLDs. Standard setting for alternate assessment is no different from most other standard-setting endeavors with regard to the development and use of PLDs. A key difference, however, is the meaning of the terms, particularly in the all-too-common situation in which the same PLLs are used for the standard and alternate assessments. For example, many statewide assessments have the classification *Advanced* on both the regular assessment and on the alternate assessment. In the regular assessment, *Advanced* performance in English Language Arts may imply, for example, the ability to detect subtle nuances in an author's tone and the ability to identify and successfully defend a persuasive position on a controversial topic. For the alternate assessments, however, *Advanced* performance may mean the ability to read a daily schedule for chores, academic assignments, and extracurricular activities. In other situations, a different set of PLDs applies to the alternate assessment. Figure 15-3 shows the PLLs and associated PLDs from an alternate assessment program in Arkansas. These PLLs and PLDs were designed to describe categories of performance in either Literacy or Mathematics. We note that, in this case, the PLLs differ from those used for the regular assessment program, eschewing the confusion with identical PLLs just noted. However, this difference also introduces additional issues associated with the problem of how to combine results from the two systems for overall reporting purposes. Regardless of these differences, a common feature of all alternate-assessment standard setting is that PLDs are required, and interpretation of the results for alternate assessments will be just as bound to those PLDs as is the case for the regular assessment.

A Generalized Holistic Method

In this section, we describe a generalized holistic (GH) method for setting standards for alternate assessment. The method was first applied to a statewide alternate assessment program in Virginia, and it has since been used successfully in other states. The method described here depends on a clear set of PLDs and a well-defined and consistently applied scoring rubric for its successful implementation.

The method described here combines elements of the Contrasting Groups method (see Chapter 8), the Bookmark method (see Chapter 10),

Performance Level Label	*Performance Level Description*
Not Evident	Students do not demonstrate evidence of performance toward the literacy or mathematics skills being assessed.
Emergent	Students do not sufficiently demonstrate the literacy or mathematics skills needed to attain the Supported Independence level. They are just beginning to show understanding or use of these skills; however, they are unable to perform these skills accurately without continuous support and assistance.
Supported Independence	Students are attempting to meet authentic, age-appropriate challenges but have limited success. They demonstrate a partial or minimal ability to apply literacy or mathematics skills and require frequent prompting or support. They make errors but occasionally perform these skills accurately.
Functional Independence	Students frequently meet authentic, age-appropriate challenges. They demonstrate reasonable performance in multiple settings and are prepared for more challenging tasks. They can apply established literacy or mathematics skills to real-world situations but may require occasional prompting or support. They perform these skills accurately in most instances but make occasional errors.
Independent	Students demonstrate performance well beyond the Functional Independence level. They demonstrate mastery of authentic, age-appropriate, and challenging tasks in multiple settings. They can apply established literacy or mathematics skills to real-world situations on their own. They can generalize learned skills to solve new challenges.

Figure 15-3 Generalized Performance Level Descriptions (PLDs) for Alternate Assessment

and the Body of Work and Analytical Judgment methods (see Chapter 9). Similar to the Body of Work method, the GH method relies on holistic judgments about entire student work samples; like the Bookmark method, this method simplifies participants' judgmental task because, from a large body of COEs, facilitators select a representative sample of prescored COEs and present them in score-point order (lowest to highest score), both on a rating sheet and in physical arrangement. At this point, the GH method begins to resemble the Analytical Judgment procedure in that participants evaluate the COEs, discuss their findings, and proceed to a second round of review of the same COEs, not a pinpointing set that consists of a different collection from the first-round or rangefinding set. Data analytic techniques are similar to those of the Contrasting Groups procedure: A mean is calculated

for each group, and then the facilitator calculates the midpoint between adjacent group means. These midpoints serve as the cut scores.

Overview of an Application of the GH Method

On September 18–21, 2001, 41 educators from the Commonwealth of Virginia met in Richmond to examine COEs for four levels of the Virginia Alternate Assessment Program (VAAP). The four levels and their applicability to corresponding student ages are Elementary I (students who are 8 years old), Elementary II (students who are 10 years old), Middle School (students who are 13 years old), and High School (one year prior to the student's exit year). For each level, the assessment is completed during the school year in which the student reaches the specified age or before September 30.

Three cut scores were needed for each level of the VAAP to categorize performance into one of three performance levels: Needs Improvement, Proficient, and Advanced. The initial PLDs used to define those performance categories are shown in Figure 15-4.

Performance Level Label	*Performance Level Description*
Needs Improvement	Evidence is not relevant to specific IEP content area goals, or IEP goals are not related to SOL content area objectives; student shows little or no evidence of performance of IEP-related goals; student performs tasks in a limited range of contexts, performs tasks that are not meaningful or are not age-appropriate, fails to use appropriate supports.
Proficient	Evidence of relevant IEP goals is present; IEP goals are relevant to SOL content-area objectives; the student shows some evidence of performance of those goals, performs in a limited variety of settings with opportunity for some interactions with peers, uses age-appropriate materials to perform some meaningful tasks with real-world applications, uses some appropriate supports.
Advanced	There is ample evidence of relevant IEP goals, which are clearly related to SOL content area objectives; the student shows considerable evidence of performance related to those goals, performs tasks in a variety of settings, engages in social interactions with a diverse range of age-appropriate peers, uses age-appropriate materials to perform meaningful tasks in a real-world context, and consistently uses appropriate supports.

Figure 15-4 Initial PLDs Used for the VAAP Standard Setting

Day 1		Day 2	
8:30 A.M.	Registration and continental breakfast	8:30 A.M.	Continental breakfast
9:00	Welcome and introductions	9:00	Discussion of Round 1
9:15	Overview/orientation	9:45	Round 2
10:15	Break	12:00 Noon	Lunch
10:30	Practice session	1:30 P.M.	Discussion of Round 2
12:00 Noon	Lunch	2:00	Final cut score recommendations
1:00 P.M.	Round 1	3:00	Other recommendations
3:45	Wrap-up	3:45	Wrap-up
4:00	Adjourn	4:00	Adjourn

Figure 15-5 Sample Agenda for Alternate Assessment Standard-Setting Meeting

Contractor staff (including one of the authors of this book), provided training and assistance for the standard-setting effort, which used a GH method. Standard setters were divided into three groups of 10 (Elementary I, Middle School, and High School) and one group of 11 (Elementary II). The Elementary I and Elementary II groups met on September 18–19; the Middle School and High School groups met on September 20–21. A sample agenda for the two-day meeting is shown in Figure 15-5. Each group participated in training, a practice round, two rounds of rating, and final discussions of the potential cut scores. At the conclusion of the process, the groups also made recommendations regarding the PLDs and regarding creation of COEs in the future.

During the 2000–2001 school year, educators completed COEs for more than 2,000 students throughout the Commonwealth of Virginia. During June and July 2001, a team of scorers employed by the contractor and trained with the assistance of Virginia educators scored the COEs using the rubric shown in Figure 15-2. Once the COEs had been scored, the facilitators of the standard-setting activity selected representative COEs to present to the participants. The selected COEs had the same general distribution of scores and contents as the full range of COEs.

Procedures for Implementing the GH Method

The task of participants was to evaluate a sample of previously scored COEs in terms of a more global set of criteria in order to assign one of three categories described in Figure 15-4. These ratings could then be cross-referenced to the previously assigned scores in order to develop a rating

scale that would assign one of the three designations to each collection on the basis of its score from 10 to 30. Panelists rated COEs one at a time, entered global judgments (i.e., Needs Improvement, Proficient, or Advanced) on special forms. A sample form is reproduced in Figure 15-6.

The individual COEs used in this standard setting consisted of a large amount of student work compiled into 4-inch binders with an assortment of support materials. Due to the nature of individualized alternate assessment, the COEs that resulted were of varying length and complexity. Some had a variety of nonprint material (videotapes, audiotapes, or other media) while others had none.

Because of the sheer amount of information, it was decided that each participant would not examine every COE. Instead, facilitators rotated the COEs in such a way that each one was examined by approximately an equal number of participants. In a given round, each participant examined an average of eight to nine COEs, with some examining more and some examining fewer. Each COE was examined an average of about three times (Elementary I, Elementary II, and Middle School) or two times (High School). Because the total score range was narrow, it was possible to have most score points represented. Therefore, even though the number of reviews of a given COE was fairly limited, the number of reviews in even a very narrow score range was sufficient to yield stable cut score estimates.

Training of all participants was provided by two facilitators who presented an overview of the alternate assessment program, the rating task, and the standard-setting procedure to be used. Sample materials that can be adapted to different applications are provided at www.sagepub.com/cizek/ghm. At the conclusion of the training, panelists took part in a practice exercise in which they evaluated one component of three different COEs and compared their ratings. After verifying their readiness to rate whole COEs and enter ratings on special forms (see Figure 15-6), participants began Round 1.

For Round 1, facilitators divided the participants into two groups by grade level (Elementary I and Elementary II on the first two days, and Middle School and High School on the final two days). Each group met in a separate room. The facilitators set out the COEs (each in a large binder with corresponding support materials) on tables at the back of the room. Each participant took a rating sheet, selected a COE, rated it, entered the rating on the sheet, returned the COE to its original location, selected another, and repeated the process. Participants worked individually throughout the afternoon of Day 1 of the standard-setting meeting. Each room had an audiotape player, videocassette player, and television for participants' use in listening to or viewing a work sample. Media stations were

Participant ID No. _____ Round Number (circle one): 1 2

Directions: Each Collection of Evidence you evaluate will contain evidence of proficiency in four content areas. Evaluate each section, and place one of the following letters in each column:

N (Needs Improvement) P (Proficient) A (Advanced)

Collection of Evidence ID No.	*English*	*Mathematics*	*Science*	*History/Social Sciences*
15				
23				
28				
31				
49				
53				
56				
59				
101				
110				
132				
140				
153				
157				
192				
284				
301				
317				
322				
340				
399				
406				
410				
466				
471				
507				
523				
541				
548				
551				
556				
572				

Figure 15-6 Sample Generalized Holistic Rating Sheet

set up with earphones in different parts of the room to reduce congestion and noise. Throughout the afternoon, the facilitators monitored the selections of COEs and periodically removed some COEs from circulation to make sure that all COEs were examined at least once. At the end of the afternoon, participants returned all materials and were dismissed for the day.

Data analysis and feedback for Round 1 consisted of constructing distributions of ratings across the score range and presenting the data in both tabular and graphic form to the participants on the morning of Day 2. Cut scores for Proficient and Advanced were calculated by taking the midpoint between adjacent category means; that is, the midpoint between Needs Improvement-Proficient was used to estimate the Proficient cut score, and the midpoint between Proficient-Advanced was used for Advanced. On the morning of the second day, participants were provided feedback that included not only the preliminary cut scores but also their own ratings relative to those of the other members of their group.

Round 2 began with a discussion of the Round 1 results, with emphasis on variability of ratings for either the same COE or COEs with equal or comparable scores. As expected, there was considerable discussion of the differences among COE preparers with regard to selection and presentation of material, completely apart from the accomplishments of the students. The facilitators reminded participants that the COEs were effectively joint projects of the students and their teachers and that their ratings needed to reflect the whole COE without attempting to separate the accomplishments of the student from those of the teacher or paraprofessional. Following this discussion, participants resumed the task of rating COEs as they had done on Day 1. They worked through the morning as they had the afternoon before, turned in all their materials, and adjourned for lunch.

While the panelists had lunch, the facilitators analyzed Round 2 data, calculating the cut scores and distributions of ratings. For illustrative purposes, Round 2 cut scores for one of the assessments (Middle School History/Social Studies) are presented in Table 15-2. The table shows the frequencies with which COEs of varying point totals were classified by participants into the performance levels. For example, two COEs that had achieved a total score of 15 were reviewed; one was classified as Proficient, one as Needs Improvement. Of the seven COEs that had been assigned a total score of 19, all were considered by participants as belonging in the Needs Improvement category.

After lunch, the facilitators presented the Round 2 results to participants, along with impact information. At this point, the discussion shifted from

Table 15-2 Summary of Round 2 for Middle School History/Social Studies

	Score Points and Associated Frequencies of Ratings															
Performance Category	*15*	*19*	*20*	*21*	*22*	*23*	*24*	*25*	*26*	*27*	*28*	*29*	*30*	*Total*	*Mean*	*Midpoint*
Advanced					2		1	1	4	2	1	1	3	15	26.5	26.0
Proficient	1		1	1	3	5	2	4	6	2	2	6	4	37	25.5	22.9
Needs Improvement	1	7	6	3	4	3								24	20.3	
Number of COEs at Score Point	2	7	7	4	9	8	3	5	10	4	3	7	7			

interparticipant differences in ratings to consequences for individual students and programs.

Table 15-2 also shows cut scores of 22.9 (subsequently rounded to 23) for *Proficient* and 26.0 for *Advanced* for Middle School History/Social Studies. A closer examination of Table 15-2, however, shows considerable variability around these two cut points. At score point 22, for example, there were already two ratings of *Advanced*, and while there were no ratings of *Needs Improvement* above score point 23, there were clearly two schools of thought with regard to distinguishing between *Proficient* and *Advanced* from that score point all the way through the top end of the score range. While the midpoint of the two means would technically have served as an acceptable cut score, it was clear that the two means were only one raw score point apart.

By agreeing in advance (with the consent of the department) to permit a final vote by panelists, the facilitators were able to bring some resolution to the matter. Participants discussed their reasons for assigning either *Proficient* or *Advanced* ratings for COEs with score points 24 through 30. After all participants were afforded the opportunity to explain their classification rationales, they agreed to put the matter to a vote, as suggested by Plake et al. (1997). By show of hands, participants selected score point 22 to serve as the lowest point at which a COE would receive the *Proficient* rating and score point 27 to serve as the *Advanced* cut score.

An examination of Table 15-2 supports the reasonableness of these decisions. First, both cut scores were within one point of the calculated cut scores. For the Proficient cut, the participants selected a point that seemed to fall precisely at the crossroads of the three levels. There were entries for all three performance categories at score point 22 (2 ratings of *Advanced,* 3 ratings of *Proficient,* and 4 ratings of *Needs Improvement*). No other score point generated ratings at more than two performance categories. Participants seemed pleased with 22 as a sort of "compromise" cut score. Similarly, at 27 points, there were exactly two ratings each for *Proficient* and *Advanced*. Remarkably, above score point 27, the *Proficient* ratings outnumbered the Advanced ratings at every score point; thus participants seemed amenable to score point 27 as another candidate for compromise.

Cut scores were set in like manner for the other content areas and levels. For all the remaining assessments, setting the upper cut score (separating *Proficient* from *Advanced*) was usually more challenging. In most instances, it involved discussing a "ratio of regret" (Emrick, 1971), that is, a comparison of the seriousness of assigning a rating of *Proficient* to a COE that is more properly classified as *Advanced* versus the seriousness of assigning a COE to the *Advanced* performance level that was more properly classified

as *Proficient*. Ultimately, the standard-setting participants were able to make these distinctions, generally on the basis of the quality of the COE they had seen, but sometimes relying on the logic of the scoring process and other considerations of fairness and equity not directly related to the scoring. Nearly all final recommendations were based on unanimous decisions of the groups.

Conclusions and Discussion Regarding the GH Method

The generalized holistic method employed with the VAAP was designed to address a very specific combination of circumstances: a two-day time frame for standard setting, massive amounts of material for panelists to review, and complete or near-complete confounding of student and teacher effects. On the positive side, supporting the success of the procedure were clear PLDs, a straightforward and workable plan for assembling COEs, statewide training of teachers and paraprofessionals who assembled the COEs, and a clear and consistent scoring rubric that took into account the likely diversity of the COEs. Standard-setting participants—most of whom were special educators—learned the COE rating task quickly (although some did require occasional reminders to rate the entire COE holistically when they would rather have rescored the components).

The most difficult aspect of this or any other standard-setting activity for alternate assessment was unrelated to the choice of method. Namely, the current application—like most current systems for applying standard setting to the context of alternate assessments—faced the troubling reality that teacher and student effects are confounded. What does it actually mean that one student receives a rating of *Proficient*, while another receives a rating of *Advanced* or *Needs Improvement?* Given the way that alternate assessment systems are configured in nearly all states, the rating is partly (if not mostly) of the teacher or paraprofessional.

In one sense, the situation is entirely appropriate. Prior to the passing of the NCLB (2001) legislation, alternate assessment received little, if any, attention. NCLB focused attention on special education students and special educators. In the end, it is the educators who are held accountable, more than the students. Thus it seems appropriate that their efforts be clearly reflected in the COEs. When we note that a high percentage of students in one school district have met the proficiency goal for that district or state, we are really saying that the teachers in that district have done what was expected of them with regard to these special education students. In another district, where the percentage is lower, the teachers have had the same opportunity as those elsewhere to select educational outcomes, match

appropriate tasks to those outcomes, and present the achievements of their students.

Alternative Procedures and Limitations

At the beginning of this chapter, we described some of the challenges of alternate assessment. We turn now to one of those challenges in particular—one that we foreshadowed as a seemingly intractable problem: the problem of combining alternate assessment results with those of other assessments. A standing question has been "How does one combine the data (i.e., the proportion of students classified) from an alternate assessment as *Proficient* with classifications of *Proficient* resulting from a regular assessment in any way that makes logical sense and seems psychometrically sound?"

In response to this question, there are dramatically different answers about which equally thoughtful measurement specialists may disagree. For example, strictly from the perspective of sound measurement and validity of score inferences, it is entirely reasonable to conclude that the data from the two systems cannot—and should not—be aggregated. An equally thoughtful observer might also conclude precisely the opposite: From an inferential standpoint, the two ratings mean exactly the same thing (that a student and his or her teachers have done what was expected of them), and little stands in the way of combining the data.

Or, from a more purely policy perspective, the mixing of regular and alternate assessment results may be defensible in light of the purposes of NCLB. Under that legislation, the issue is how many students meet a particular proficiency goal. If 800 out of 1,000 students in the regular assessment program met the goal of *Proficient* and 12 out of 20 met the goal of *Proficient* on the alternate assessment, the combined percentage is 812 out of 1,020, or just under 80% (exactly 80% in the regular assessment group and 60% in the alternate assessment group). Those holding such a viewpoint would not consider it necessary to try to merge scale scores or any other scores if the target is proficiency level: A goal met is a goal met.

Although there may be differences of perspective on the issue of aggregation, other limitations of standard setting for alternate assessments are more clear-cut. For example, as nearly all current alternate assessment programs are configured, they are limited primarily by the nature of the materials available for review. The nature of alternate assessment is such that standard setting will almost certainly be based on idiosyncratic collections of evidence gathered in individualized ways, rather than on standardized combinations of multiple-choice and constructed-response items gathered

in uniform ways across students. Thus any standard-setting approach for alternate assessments will necessarily be holistic in nature. To ameliorate this limitation to even a modest degree, it is necessary that standard setting for alternate assessments be guided by clearly articulated and measurement-based principles and carried out by well-trained and motivated professional staff. Where little thought has been given to quantifying teacher observations of student achievements in a consistent manner over time or circumstances, standard setters will be extremely limited in what they can accomplish, regardless of method used.

The GH method for setting performance standards on alternate assessments is prone to the same limitations as other holistic methods. For example, there are often massive amounts of student work to evaluate, odd or inconsistent patterns of student performance, unresolved and intractable differences of opinion among participants, and so on. The generalized nature of the GH method described here, however, permits a variety of approaches and techniques tailored to the specific situation (assuming the facilitators and planners understand the situation). As in the example cited in this chapter, cut scores can be derived mathematically (e.g., using logistic regression; see Chapter 10 in this volume) or by other, less mathematical methods (e.g., by simple majority vote).

Some of these limitations can also be alleviated. For example, the massive amount of student work can be cut down to size by adding participants, yielding more observations per COE or a wider range of COEs with the same number of observations. Others, such as differences among participants, may prove more difficult to resolve. Conversely—and we suspect that this observation is generalizable to other groups of committed educators—when given an opportunity to put cut scores to a vote, participants in our example were remarkably statesmanlike, forging sensible compromises nearly every time. This approach holds promise, not just for the GH procedure but for standard setting in general. We have asked participants to vote in the final round of other meetings in which other standard-setting methods were used and achieved very similar results. Although we have not studied participants' reasons for their willingness to compromise in the interest of consensus, we suspect that there may be at least three reasons for this phenomenon: (1) Participants realize that, by the end of the final round of ratings, the magnitude of the differences is usually quite small; (2) participants might prefer recommending a final cut score that clearly reflects their own preferences to one set by an outside agent; and (3) an outside agent may be less likely to alter or disregard their recommendations if they are unanimous.

As this is a generalized procedure, any variation might be considered part of the standard procedure, although sensible variations can easily be identified. For example, in our illustration, the procedure included providing COEs in score-point order. As we indicated in a previous chapter (see Chapter 9), it is possible to present work samples in random order. Other variants would include extended time (highly recommended whenever possible, to permit a broader review of work samples) and either intra-round discussion or small-group work. Small groups would be particularly advisable if overall panel size were increased.

At a more general level, it would seem that there would be a relatively limited set of alternative methods for standard setting for alternate assessment. Any of the holistic methods might work, even Borderline Group or Contrasting Groups (see Chapter 8). Other combinations of elements from various holistic methods may also be useful. What we have described here as the GH method for setting standards on alternate assessments is admittedly a hybrid of various holistic methods. More research on standard setting for alternate assessments is clearly needed and would be welcomed to help identify the extent to which other combinations—or entirely new procedures—will work equally successfully.

16

Special Topics and Next Steps

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

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

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

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

Rounding

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

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

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

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

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

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

Methods of Adjusting Cut Scores

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Deciding How to Incorporate Uncertainty

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Generalizability of Standards

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

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

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

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

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

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

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

Decision Consistency and Decision Accuracy

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

SOURCE: Subkoviak (1988).

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

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

SOURCE: Subkoviak (1988).

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

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

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

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

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

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

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

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

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

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

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

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

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

SOURCE: Adapted from Brennan (2004b).

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

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

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

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

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

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

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

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

Other Decision Consistency Procedures

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

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

16-5.a Accuracy Relative to Actual Observed Scores

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

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

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

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

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

SOURCE: Adapted from Brennan (2004b).

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

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

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

Summary and Future Directions

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

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

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

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

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

Table 16-6 Example of Decision Matrix

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

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

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

Using Multiple Methods of Standard Setting

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

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

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

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

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

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

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

Improving Participant Training

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

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

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

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

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

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

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

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

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Andrew, B. J., & Hecht, J. T. (1976). A preliminary investigation of two procedures for setting examination standards. *Educational and Psychological Measurement, 36,* 45–50.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 508–600). Washington, DC: American Council on Education.

Angoff, W. H. (1988). Proposals for theoretical and applied development in measurement. *Applied Measurement in Education, 1*(3), 215–222.

Applied Measurement in Education. (2005). Special issue on vertically-moderated standard setting. *18*(1).

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

Berk, R. A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education, 45,* 4–9.

Berk, R. A. (1984). *A guide to criterion-referenced test construction.* Baltimore, MD: Johns Hopkins University Press.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56,* 137–172.

Beuk, C. H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. *Journal of Educational Measurement, 21,* 147–152.

Bloom, B. S. (1974). An introduction to mastery learning. In J. H. Block (Ed.), *Schools, society, and mastery learning* (pp. 3–14). New York: Holt, Rinehart & Winston.

Brennan, R. L. (1983). *Elements of generalizability theory.* Iowa City, IA: ACT.

Brennan, R. L. (2002, October). *Estimated standard error of a mean when there are only two observations* (Center for Advanced Studies in Measurement and Assessment [CASMA] Technical Note Number 1). Iowa City: University of Iowa, CASMA.

Brennan, R. L. (2004a). *BB-CLASS v. 1.1* [Computer program]. Iowa City: Center for Advanced Studies in Measurement and Assessment, University of Iowa.

Brennan, R. L. (2004b). *Manual for BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy, Version 1.1.* Iowa City: Center for Advanced Studies in Measurement and Assessment, University of Iowa.

Brennan, R. L., & Wan, L. (2004). *Bootstrap procedures for estimating decision consistency for single-administration complex assessments* (CASMA Research Report No. 7). Iowa City: University of Iowa.

Buckendahl, C. W., Huynh, H., Siskind, T., & Saunders, J. (2005). A case study of vertically moderated standard setting for a state science assessment program. *Applied Measurement in Education, 18,* 83–98.

Buckendahl, C. W., Smith, R. W., Impara, J. C., & Plake, B. S. (2002). A comparison of Angoff and Bookmark standard setting methods. *Journal of Educational Measurement, 39,* 253–264.

Bunch, M. B. (1978, April). *Making decisions about adult learners based on performances on functional competency measures.* Paper presented at the annual meeting of the Adult Education Research Conference, San Antonio, TX.

Burton, N. (1978). Societal standards. *Journal of Educational Measurement, 15,* 263–271.

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

Cizek, G. J. (1996a). Setting passing scores. *Educational Measurement: Issues and Practice, 15*(2), 20–31.

Cizek, G. J. (1996b). Standard setting guidelines. *Educational Measurement: Issues and Practice, 15*(1), 12–21.

Cizek, G. J. (2001a). Conjectures on the rise and call of standard setting: An introduction to context and practice. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 3–17). Mahwah, NJ: Lawrence Erlbaum.

Cizek, G. J. (2001b). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice, 20*(4), 19–27.

Cizek, G. J. (2005). Adapting testing technology to serve accountability aims: The case of vertically-moderated standard setting. *Applied Measurement in Education, 18*(1), 1–10.

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

Cizek, G. J., Bunch, M., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice, 23*(4), 31–50.

Cizek, G. J., & Fitzgerald, S.M. (1999). An introduction to logistic regression. *Measurement and Evaluation in Counseling and Development, 31,* 223–245.

Cizek, G. J., Kenney, P. A., Kolen, M. J., Peters, C., & van der Linden, W. J. (1999). *An investigation of the feasibility of linking scores on the proposed Voluntary*

National Tests and the National Assessment of Educational Progress. Washington, DC: National Assessment Governing Board.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

Colton, D. A., Gao, X., Harris, D. J., Kolen, M. J., Martinovich-Barhite, D., Wang, T., & Welch, C. J. (1997). *Reliability issues with performance assessments: A collection of papers* (ACT Research Report Series 97-3). Iowa City, IA: ACT.

Colton, D. A., & Hecht, J. T. (1981, April). *A preliminary report on a study of three techniques for setting minimum passing scores*. Symposium presentation at the annual meeting of the National Council on Measurement in Education, Los Angeles, CA.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Holt, Rinehart and Winston.

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

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

CTB/McGraw-Hill. (2001). *Kentucky core content tests: Standard setting technical report*. Lexington: Kentucky State Department of Education.

Downing, S. M., Lieska, N. G., & Raible, M. D. (2003). Establishing passing standards for classroom achievement tests in medical education: A comparative study of four methods. *Academic Medicine, 78*(10), S85–S87.

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

Ebel, R. L. (1972). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice Hall.

Emrick, J. A. (1971). An evaluation model for mastery testing. *Journal of Educational Measurement, 8,* 321–326.

Ferdous, A. A., & Plake, B. S. (2005). Understanding the factors that influence decisions of panelists in a standard-setting study. *Applied Measurement in Education, 18,* 257–267.

Ferrara, S., Johnson, E., & Chen, W. L. (2005). Vertically articulated performance standards: Logic, procedures, and likely classification accuracy. *Applied Measurement in Education, 18,* 35–60.

Ferrara, S., Perie, M., & Johnson, E. (2002a, December). *Matching the judgmental task with standard setting panelist expertise: The item-descriptor (ID) matching procedure*. Washington, DC: American Institutes for Research.

Ferrara, S., Perie, M., & Johnson, E. (2002b, April). *Setting performance standards: The item descriptor (ID) matching procedure*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Giraud, G., Impara, J. C., & Plake, B. S. (2005). Teachers' conceptions of the target examinee in Angoff standard setting. *Applied Measurement in Education, 18,* 223–232.

Hambleton, R. K. (1998). Setting performance standards on achievement tests: Meeting the requirements of Title I. In L. N. Hansche (Ed.), *Handbook for the development of performance standards* (pp. 87–114). Washington, DC: Council of Chief State School Officers.

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

Hambleton, R. K., Brennan, R. L., Brown, W., Dodd, B., Forsyth, R. A., Mehrens, W. A., Nellhaus, J., Reckase, M. D., Rindone, D., van der Linden, W. J., & Zwick, R. (2000). A response to "Setting Reasonable and Useful Performance Standards" in the National Academy of Sciences' Grading the Nation's Report Card. *Educational Measurement: Issues and Practice, 19*(2), 5–14.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport, CT: American Council on Education/Praeger.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8*, 41–56.

Hambleton, R. K., & Slater, S. C. (1997). Reliability of credentialling examinations and the impact of scoring models and standard setting policies. *Applied Measurement in Education, 10*(1), 19–38.

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

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement, 27*, 345–359.

Helfman, D. E. (2002). Memorandum to Dr. Carolyn C. Dumaresq re: PDE's Response to PSEA Cut Score Analysis. PSEA Interactive. Retrieved November 26, 2005, from http://league.psea.org/article.cfm.SID=145

Hofstee, W. K. B. (1983). The case for compromise in educational selection and grading. In S. B. Anderson & J. S. Helmick (Eds.), *On educational testing* (pp. 109–127). San Francisco: Jossey-Bass.

Huynh, H. (2000, April). *On item mappings and statistical rules for selecting binary items for criterion-referenced interpretation and Bookmark standard setting*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Huynh, H. (2006). A clarification on the response probability criterion RP67 for standard settings based on Bookmark and item mapping. *Educational Measurement: Issues and Practice, 25*(2), 19–20.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*, 353–366.

Impara, J. C. & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35*(1), 69–81.

Individuals with Disabilities Education Act. (1997). Public Law 105-17. (20 U.S.C. 1412a, 16–17).

Individuals with Disabilities Education Improvement Act. (2004). Public Law 108–446.

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

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

Jaeger, R. M. (1995). Setting performance standards through two-stage judgmental policy capturing. *Applied Measurement in Education, 8*, 15–40.

Jaeger, R. M., & Mills, C. N. (2001). An integrated judgment procedure for setting standards on complex, large-scale assessments. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 313–338). Mahwah, NJ: Lawrence Erlbaum.

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

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

Kane, M. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53–88). Mahwah, NJ: Lawrence Erlbaum.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Karantonis, A., & Sireci, S. G. (2006). The bookmark standard setting method: A literature review. *Educational Measurement: Issues and Practice, 25*(1), 4–12.

Kingston, N. M., Kahl, S. R., Sweeney, K., & Bay, L. (2001). Setting performance standards using the body of work method. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 219–248). Mahwah, NJ: Lawrence Erlbaum.

Langsley, D. G. (1987). Prior ABMS conferences on recertification. In J. S. Loyd & D. G. Langsley (Eds.), *Recertification for medical specialists* (pp. 11–30). Evanston, IL: American Board of Medical Specialties.

Lee, W. (2005). *A multinomial error model for tests with polytomous items* (CASMA Research Report No. 10). Iowa City: University of Iowa, Center for Advanced Studies in Measurement and Assessment. (Available from www.education.uiowa.edu/casma)

Lerner, B. (1979). Tests and standards today: Attacks, counterattacks, and responses. In R. T. Lennon (Ed.), *New directions for testing and measurement: Impactive changes on measurement* (pp. 15–31). San Francisco: Jossey-Bass.

Lewis, D. M., & Green, D. R. (1997, June). *The validity of performance level descriptors.* Paper presented at the annual CCSSO Conference on Large Scale Assessment, Colorado Springs, CO.

Lewis, D. M., & Haug, C. A. (2005). Aligning policy and methodology to achieve consistent across-grade performance standards. *Applied Measurement in Education, 18,* 11–34.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard setting procedures utilizing behavioral anchoring*. Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

Linn, R. L. (2006). The *Standards for Educational and Psychological Testing:* Guidance in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 27–38). Mahwah, NJ: Lawrence Erlbaum.

Lissitz, R. W., & Huynh, H. (2003a). Vertical equating for state assessments: Issues and solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research & Evaluation, 8*(10). Available online at http://www.pareonline.net/getvn.asp?v=8&n=10

Lissitz, R. W., & Huynh, H. (2003b). *Vertical equating for the Arkansas ACTAAP assessments: Issues and solutions in determination of adequate yearly progress and school accountability*. Little Rock: Arkansas Department of Education.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32,* 179–197.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores*. Princeton, NJ: Educational Testing Service.

Loomis, S. C., Bay, L., Yang, W. L., & Hanick, P. L. (1999, April). *Field trials to determine which rating method(s) to use in the 1998 NAEP achievement levels setting process for civics and writing*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Loomis, S. C., & Bourque, M. L. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 175–218). Mahwah, NJ: Lawrence Erlbaum.

Loomis, S. C., Hanick, P. L., Bay, L., & Crouse, J. D. (2000). *Setting achievement levels on the 1998 National Assessment of Educational Progress in Writing: Field trials final report*. Iowa City, IA: ACT.

Meara, K. P., Hambleton, R. K., & Sireci, S. G. (2001). Setting and validating standards on professional licensure and certification exams: A survey of current practices. *CLEAR Exam Review, 12*(2), 17–23.

Mehrens, W. A., & Cizek, G. J. (2001). Standard setting and the public good: Benefits accrued and anticipated. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 477–485). Mahwah, NJ: Lawrence Erlbaum.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology*. Fort Worth, TX: Holt, Rinehart and Winston.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York: Macmillan.

Millman, J. (1989). If at first you don't succeed: Setting passing scores when more than one attempt is permitted. *Educational Researcher, 18*(6), 5–9.

Mills, C. N., & Jaeger, R. M. (1988). Creating descriptions of desired student achievement when setting performance standards. In L. Hansche (Ed.), *Handbook for the development of performance standards* (pp. 73–86). Washington, DC: Council of Chief State School Officers.

Mills, C. N., & Melican, G. J. (1988). Estimating and adjusting cutoff scores: Features of selected methods. *Applied Measurement in Education, 1,* 261–275.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–281). Mahwah, NJ: Lawrence Erlbaum.

National Research Council. (1999). *Uncommon measures: Equivalence and linkage among educational tests.* Washington, DC: Author.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14,* 3–19.

No Child Left Behind Act. (2001). Public Law 107–110. (20 U.S.C. 6311).

O'Connell, A. (2006). *Logistic regression models for ordinal response variables.* Thousand Oaks, CA: Sage.

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

Plake, B. S., & Hambleton, R. K. (2001). The analytic judgment method for setting standards on complex performance assessments. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 283–312). Mahwah, NJ: Lawrence Erlbaum.

Plake, B. S., Hambleton, R. K., & Jaeger, R. M. (1997). A new standard-setting method for performance assessments: The dominant profile judgment method and some field-test results. *Educational and Psychological Measurement, 57,* 400–411.

Popham, W. J. (1978). As always, provocative. *Journal of Educational Measurement, 15,* 297–300.

Putnam, S. E., Pence, P., & Jaeger, R. M. (1995). A multi-stage dominant profile method for setting standards on complex performance assessments. *Applied Measurement in Education, 8,* 57–83.

Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 119–157). Mahwah, NJ: Lawrence Erlbaum.

Reckase, M. D. (2001). Innovative methods for helping standard-setting participants to perform their task: The role of feedback regarding consistency, accuracy, and impact. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 159–174). Mahwah, NJ: Lawrence Erlbaum.

Schulz, E. M., Kolen, M., & Nicewander, W. A. (1999). A rationale for defining achievement levels using IRT-estimated domain scores. *Applied Psychological Measurement, 23,* 347–362.

Schulz, E. M., & Lee, W. (2002, April). *Describing NAEP mathematics achievement using domain scores*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA. (ERIC Document Reproduction Service No. ED464917)

Schulz, E. M., Lee, W., & Mullen, K. (2005). A domain-level approach to describing growth in achievement. *Journal of Educational Measurement, 42,* 1–26.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Shepard, L. (1979). Setting standards. In M. A. Bunda & J. R. Sanders (Eds.), *Practices and problems in competency-based education* (pp. 59–71). Washington, DC: National Council on Measurement in Education.

Shepard, L. (1980). Standard setting issues and methods. *Applied Psychological Measurement, 4,* 447–467.

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

Shimberg, B., Esser, B. F., & Kruger, D. H. (1973). *Occupational licensing and public policy*. Washington, DC: Public Affairs Press.

Sireci, S. G., & Biskin, B. J. (1992). Measurement practices in national licensing examination programs: A survey. *CLEAR Exam Review, 3*(1), 21–25.

Sireci, S. G., Hambleton, R. K., & Pitoniak, M. J. (2004). Setting passing scores on licensure examinations using direct consensus. *CLEAR Exam Review, 15*(1), 21–25.

Skorupski, W. P., & Hambleton, R. K. (2005). What are panelists thinking when they participate in standard-setting studies? *Applied Measurement in Education, 18,* 233–256.

Subhiyah, R. G., Featherman, C. M., & Hawley, J. L. (2002, November). *How to set pass/fail standards on examinations*. Presentation at the annual meeting of Generalists in Medical Education, San Francisco.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement, 13,* 265–276.

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability for mastery tests. *Journal of Educational Measurement, 25,* 47–55.

Thorndike, R. L. (Ed.). (1971). *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

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

Trent, E. R., & Roeber, E. (2006). Contracting for testing services. In T. Haladyna & S. Downing (Eds.), *Handbook of test development* (pp. 39–60). Mahwah, NJ: Lawrence Erlbaum.

Wainer, H. (2006). Review of *Defending standardized testing*. *Journal of Educational Measurement, 43,* 77–84.

Wang, N. (2003). Use of the Rasch IRT model in standard setting: An item mapping method. *Journal of Educational Measurement, 40,* 231–253.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: MESA.

Zieky, M. J. (2001). So much has changed: How the setting of cut scores has evolved since the 1980s. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 19–52). Mahwah, NJ: Lawrence Erlbaum.

Zieky, M. J., & Livingston, S. A. (1977). *Manual for setting standards on the Basic Skills Assessment Tests.* Princeton, NJ: Educational Testing Service.

Glossary

accommodations changes in the way a test is administered or scored that do not alter the construct being measured or introduce construct-irrelevant variance into examinees' scores

achievement test a kind of test, the purpose of which is to measure the extent of an examinee's attainment of knowledge, skill, or ability across a well-specified domain of interest

alternate assessment an assessment designed for examinees for whom a standard assessment has been deemed inappropriate due to the students' severe cognitive or physical impairments. Although they are typically based on the same or similar content standards, alternate assessments address either downward extensions of the content standards (e.g., precursor skills) or life-skill parallels (e.g., counting coins as a parallel to a fifth-grade numerical operations content standard) to the standard assessment content standards. Alternate assessment differs from modified (e.g., Braille or large print) or accommodated (e.g., extended-time) assessments in that construct equivalence between the standard and alternate assessments cannot be assumed.

collection of evidence (COE) the complete body of work gathered to demonstrate the achievement or progress of an examinee, typically in an alternate assessment, although the term could conceivably refer to any type of portfolio assessment. COEs may contain written work, video- or audiotapes of an examinee performing some task, checklists, worksheets, and any other evidence deemed relevant to documenting the examinee's progress or achievement vis-à-vis a set of content standards.

compensatory a model for determining overall pass/fail or category status on a test comprising two or more components (e.g., items, tasks, subtests, etc.) in which only a specified level of total score performance is considered when the overall classification is made. The practical effect of use of a

compensatory model is that an examinee's comparatively stronger performance on one or more of the components can compensate for comparatively weaker performance on one or more of the other components.

compromise method a type of standard-setting method which recognizes and explicitly incorporates both norm- and criterion-referenced values in deriving a cut score

conjunctive a model for determining overall pass/fail or category status on a test comprising two or more components (e.g., items, tasks, subtests, etc.) in which a specified level of performance on each individual component is required in order to be judged as successful on the total test or composite

construct in the social sciences, a label used to describe a characteristic on which persons vary in their observed behavior. Characteristics measured by many tests are referred to as constructs because they are usually not directly observable, but are "constructed." For example, while a characteristic such as "honesty" does not exist in a physical sense, it is a characteristic on which people are observed to vary along a continuous dimension, with some people behaving regularly in ways regarded as more ethical, and others behaving regularly in ways regarded as less ethical. We affix the label "honesty" to describe these regularities for the purpose of measuring and clearly communicating about them.

constructed-response an item format in which the test taker must create a response. Examples of constructed-response formats include essays, short-answer items, speeches, projects, and so on.

content standards statements that describe the specific knowledge, skills, or abilities in a subject area that are addressed in a test and that are expected to be mastered by examinees at a given performance level

criterion-referenced a kind of inference or a type of test designed to yield an inference regarding whether an examinee knows or can do specific things. A test designed to yield a criterion-referenced inference (called *criterion-referenced tests* or *CRTs*) would be constructed to assess highly specific content, would be linked to specific objectives or outcomes, and would be associated with a criterion or set of criteria for judging success on the test that would have been specified a priori.

The simplest illustration of a CRT is the road portion of a driver's license test. In the parallel parking portion, the candidate for a license must meet certain *criteria:* for example, park the car within a marked area, in four minutes or less, without knocking over more than one orange pylon.

A person's success or failure—manifested in whether or not the person gets a driver's license—does not depend on how well other candidates perform; that is, it is not norm-referenced. Potentially all applicants could pass or all could fail a CRT. There is no distinction between one candidate who parks the vehicle perfectly in the middle of the space, in only two minutes, with no pylons knocked over, and the candidate who parks awkwardly within the space, in just less than four minutes, and knocks one pylon over. Both candidates meet the criteria.

cut score a point on a score scale, usually identified via a standard-setting procedure, which creates categories representing two or more states or degrees of performance. A cut score is the numerical operationalization of a performance standard.

false negative a classification error made when the application of a cut score classifies as Failing an examinee who truly possesses the level of knowledge, skill, or ability determined to be required for Passing

false positive a classification error made when the application of a cut score classifies as Passing an examinee who truly does not possess the level of knowledge, skill, or ability determined to be required for Passing

individualized education plan (IEP) a written course of action developed for students with special educational needs that details specific educational interventions appropriate for the student and designed to bring the student's level of educational achievement up to a prescribed level. The specified interventions may also include appropriate testing formats or accommodations. IEPs are mandated under various federal laws reauthorizing the Elementary and Secondary Education Act of 1965 (P.L. 89-10). The course of action specified in an IEP is developed by a team of classroom teachers, counselors, school psychologists, administrators, and parents or guardians, as appropriate.

inference an interpretation, conclusion, or meaning that is drawn regarding some underlying, usually unobservable characteristic, based on a sample of information, behavior, actions, or responses that is observed. Often, the sample of information is collected using a test.

item scoring criteria item scoring criteria specify what an examinee must do to earn a particular score point on a polytomously scored test item or task. Item scoring criteria may be generic and apply to a class of constructed-response items or be specific to a single item. Criteria may be developed in a "top-down" fashion (i.e., by describing the requirements of the top score and then describing the remaining score points in terms of

what is absent) or "bottom-up" manner (i.e., by describing the requirements of the lowest nonzero score point and then describing the remaining score points in terms of what additional knowledge or skill must be present at each). The term *rubric* is sometimes used synonymously with item scoring criteria; however item scoring criteria are not synonymous with performance level descriptors.

logistic regression a nonlinear regression function used with one or more (usually continuous) independent variables and a dichotomous dependent variable. For example, if all examinees are divided into two categories (Qualified/Unqualified) and category membership is determined by an examinee's score on a certification test, that relationship can be represented in terms of a logistic function. The logistic function expresses the probability of group membership in terms of test score (e.g., pQ|x, meaning the probability of being in group Q (qualified) given score x). That probability, when expressed as a log-odds ratio, $\ln[pQ/(1 - pQ)]$, forms the basis of a logistic regression in which test score (x) is treated as the predictor and the log-odds ratio (y) is treated as the criterion.

norm-referenced a kind of inference or a type of test designed to yield an inference regarding the relative status of an examinee. A test designed to yield a norm-referenced inference (called *norm-referenced tests* or *NRTs*) may be constructed to reflect more variable difficulty and discrimination characteristics and to assess more heterogeneous content compared to CRTs. The primary purpose of an NRT is to provide information about how an examinee's performance compares with the performance of a reference group of test takers. The reference group is often referred to as the "norm group" and the data on norm group performance are referred to as **norms**. Familiar examples of NRTs are the *SAT,* the *Wechsler Adult Intelligence Scale* (WAIS), and the *Graduate Record Examinations* (GRE).

norms test results and other characteristics of test takers, usually drawn from a representative sample of some population called a *norm group*. The test results and characteristics of the norm group are used to create comparison data useful for interpreting the performance of examinees who take the same test (or an equivalent form of the test) at a later date. Test results from subsequent test administration can be compared to the data from the norm group to determine the relative standing and relationship of individuals or groups of test takers to the norm group.

performance level description (PLD) brief operational definitions of the specific knowledge, skills, or abilities that are expected of examinees whose performance on a test results in their classification into a certain

performance level; elaborations of the achievement expectations connoted by performance level labels

performance level label (PLL) a hierarchical group of single words or short phrases that are used to label the two or more performance categories created by the application of cut scores to examinee performance on a test

performance standard the abstract conceptualization of the minimum level of performance distinguishing examinees who possess an acceptable level of knowledge, skill, or ability judged necessary to be assigned to a category, or for some other specific purpose, and those who do not possess that level. Also sometimes used interchangeably with cut score.

pinpointing in the context of a holistic (e.g., Body of Work) standard-setting procedure, the second or subsequent round of ratings of examinee work that yields final cut scores. During this round of rating, participants usually examine additional work samples with scores that are clustered within preliminary cut score ranges identified during the first round (i.e., rangefinding).

portfolio assessment an assessment comprising the collection and analysis of examinee work samples, typically consisting of constructed-response or performance tasks gathered over a specified period of time, the purpose of which is (ordinarily) to permit inferences about an examinee's progress on a specified set of skills that would not be well measured by selected-response assessments. Useful portfolio assessments are characterized by fixed scoring guides or rubrics for classes of work samples in order that all examinees may receive a score on some fixed scale that has common meaning over settings and time.

rangefinding in the context of a holistic (e.g., Body of Work) standard-setting procedure, the first round of ratings of examinee work that yields preliminary cut score ranges. During this round, panelists may evaluate examinee work samples with scores representing the full range of performance but often with some gaps (e.g., in a 100-point scoring system, 20 samples at 5-point intervals). At the end of this round, panelists will have identified one or more cut score regions (e.g., 20–25 points, 45–50 points, and 75–80 points) for which more focused samples of work would be examined in one or more subsequent (pinpointing) rounds.

reliability the characteristic, of test scores, of being dependable. Because a test, performance, or observation consists only of a sample of questions or tasks and because both the examinees who respond and those who score examinees' responses are susceptible to various unpredictabilities in their performance (called *random errors*), no score can be considered to be a

perfectly dependable representation of the examinee's knowledge, skill, or ability. Various methods can be used to quantify the degree of confidence that can be placed in students' scores. All of the methods result in a number, called a *reliability coefficient,* that can take on any value from zero (0.0) to one (1.0), with a coefficient of 1.0 indicating perfect dependability (the complete absence of random errors) and the potential for the test scores to be used with great confidence for decision making, and a coefficient of 0.0 indicating completely undependable data.

response probability (RP) criterion in the context of Bookmark and similar item-mapping standard-setting procedures, the criterion used to operationalize participants' judgments regarding the probability of a correct response (for dichotomously scored items) or the probability of achieving a given score point or higher (for polytomously scored items). In practical applications, two RP criteria appear to be used most frequently (RP50 and RP67); other RP criteria have also been used although considerably less frequently.

rubric a scoring overview that explains in brief terms the criteria for awarding each score point in a polytomous scoring system. A rubric may be generic (adaptable to a class of test items) or specific to one test item. The rubric provides the foundation—but not the totality—of a scoring system. The complete scoring system would include examples of actual graded work, as well as annotations explaining why a particular score is given for a particular response.

selected-response an item format in which the test taker must choose the correct answer from alternatives provided. Examples of selected-response formats include multiple-choice, matching, and true/false formats.

standard setting a measurement activity in which a procedure is applied to systematically gather and analyze human judgment for the purpose of deriving one or more cut scores for a test

standards-referenced a kind of inference or a type of test designed to yield an inference regarding an examinee's standing with respect to a set of content standards. Standards-referenced tests (SRTs) are similar to CRTs in that both attempt to describe the knowledge, skill, or abilities that examinees possess. Whereas CRTs express standards in terms of quantity and category (e.g., a percentage correct and passing/failing), SRTs link examinees' scores to concrete statements about what performance at the various levels means. Familiar examples of SRTs would include state-mandated student testing for students in Grades K–12 in areas such as

English language arts and mathematics to the extent that those tests are aligned with a state's content standards in those subjects.

technical advisory committee (TAC) an independent review and advisory group, impaneled by an entity responsible for a testing program. Typically, the panel comprises members with expertise in various aspects of psychometrics, testing policy, statistics, research methods, and evaluation. TACs serve the organization by providing critical review, input, and advice on program planning and by providing review and quality control for tasks contracted by the organization with other, external individuals or companies.

validity the degree to which the conclusions yielded by any sample of behavior (e.g., a test, observation, interview, etc.) are meaningful, accurate, and useful. Validity is the degree to which an examinee's performance results in decisions about the examinee that are "correct" or defensible, or the extent to which logical and empirical evidence support the inferences about the examinee's level of knowledge, skill, or ability intended to be made based on the examinee's test performance.

vertical articulation see vertically-moderated standard setting

vertical equating the strongest form of linking two tests so that scores on the tests can be reported on a single, common scale. Vertical equating is used when two forms measure the same construct (e.g., reading comprehension) but at different levels (e.g., third grade and fourth grade).

vertically-moderated standard setting (VMSS) a procedure or set of procedures typically carried out after standards have been set by independent standard-setting panels for individual tests. VMSS is applied to address perceived inconsistencies in test results across multiple levels and/or across multiple content areas. Technically, VMSS is not standard setting in the traditional sense, but a method of adjusting the individual grade or content area standards recommended by discrete standard-setting panels. VMSS provides a procedure for adjusting the unexpected, inconsistent, or implausible fluctuations in the individual standards. The goal of VMSS is to yield a system of coherent and consistent cross-level and/or cross-content area performance standards. VMSS is also sometimes referred to as *vertical articulation* of standards.

Author Index

Subject Index

About the Authors

Gregory J. Cizek is Professor of Educational Measurement and Evaluation at the University of North Carolina-Chapel Hill. He teaches courses in psychometrics, statistics, and research methods. His interests include standard setting, test security, and testing policy. He is the author of over 250 books, chapters, professional articles and presentations. He is a contributor to the *Handbook of Classroom Assessment* (1998); editor and contributor to the *Handbook of Educational Policy* (1999) and *Setting Performance Standards: Concepts, Methods, and Perspectives* (2001). He is the author of *Filling in the Blanks* (1999), *Cheating on Tests: How to Do It, Detect It, and Prevent It* (1999), and *Detecting and Preventing Classroom Cheating* (2003). He provides expert consultation on standard setting, large-scale achievement testing, and testing policy at the state and national level.

Prior to his appointment at UNC, Cizek taught for 8 years at the University of Toledo (OH). Previously, he has managed national licensure and certification testing programs at ACT (Iowa City, IA) and worked on test development for the statewide testing program in Michigan. He began his career as an elementary school teacher, and he has served as an elected member of a local board of education. He earned his doctorate in Measurement, Evaluation, and Research Design from Michigan State University.

Michael B. Bunch is Senior Vice-President of Measurement Incorporated, a test development and scoring company serving the statewide assessment community. He joined MI in 1982 and has built the research and development division of that organization into a team of more than 80 project directors, psychometricians, editors, content specialists, and support staff. At any given time, he and his staff oversee about ten large-scale assessment projects for state departments of education. In addition to corporate management responsibilities, he remains directly involved in applied testing practice, including specialization in standard setting.

Prior to joining MI, Bunch was a senior professional with NTS Research Corporation, where he provided ESEA Title I evaluation technical assistance to state education agencies. From 1976 to 1978, he was a research psychologist with the American College Testing Program (ACT). He has authored dozens of professional and technical reports on a variety of topics—many on standard setting—and has presented his work at national conferences, including annual meetings of the American Educational Research Association (AERA), National Council on Measurement in Education (NCME), American Psychological Association (APA), Council of Chief State School Officers (CCSSO), Education Commission of the States (ECS) and Adult Education Research Conference (AERC). He received his Ph.D. in psychological measurement from the University of Georgia.