Methodology of Educational Measurement and Assessment

Sigrid Blömeke
Jan-Eric Gustafsson *Editors*

Standard Setting in Education

The Nordic Countries in an International Perspective

Springer

Methodology of Educational Measurement and Assessment

Series editors

Bernard Veldkamp, Research Center for Examinations and Certification (RCEC), University of Twente, Enschede, The Netherlands
Matthias von Davier, National Board of Medical Examiners (NBME), Philadelphia, USA[1]

[1] This work was conducted while M. von Davier was employed with Educational Testing Service.

This new book series collates key contributions to a fast-developing field of education research. It is an international forum for theoretical and empirical studies exploring new and existing methods of collecting, analyzing, and reporting data from educational measurements and assessments. Covering a high-profile topic from multiple viewpoints, it aims to foster a broader understanding of fresh developments as innovative software tools and new concepts such as competency models and skills diagnosis continue to gain traction in educational institutions around the world. Methodology of Educational Measurement and Assessment offers readers reliable critical evaluations, reviews and comparisons of existing methodologies alongside authoritative analysis and commentary on new and emerging approaches. It will showcase empirical research on applications, examine issues such as reliability, validity, and comparability, and help keep readers up to speed on developments in statistical modeling approaches. The fully peer-reviewed publications in the series cover measurement and assessment at all levels of education and feature work by academics and education professionals from around the world. Providing an authoritative central clearing-house for research in a core sector in education, the series forms a major contribution to the international literature.

More information about this series at http://www.springer.com/series/13206

Sigrid Blömeke • Jan-Eric Gustafsson
Editors

Standard Setting in Education

The Nordic Countries in an International Perspective

Springer

Editors
Sigrid Blömeke
Centre for Educational Measurement
at the University of Oslo (CEMO)
Oslo, Norway

Jan-Eric Gustafsson
Department of Education and Special
Education
University of Gothenburg
Gothenburg, Sweden

ISSN 2367-170X ISSN 2367-1718 (electronic)
Methodology of Educational Measurement and Assessment
ISBN 978-3-319-50855-9 ISBN 978-3-319-50856-6 (eBook)
DOI 10.1007/978-3-319-50856-6

Library of Congress Control Number: 2017933955

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Chapter 1
Introduction

Sigrid Blömeke and Jan-Eric Gustafsson

Abstract This introduction explains why a particular need exists to discuss standard-setting in education with respect to the Nordic countries. The objectives of the book are described, and short summaries of all 17 chapters are provided. The book consists of three major parts: The international evidence on methodological issues in standard-setting is summarized and fresh lenses are given to the state of research. After that, the standard setting practices in the Nordic countries are documented and critically discussed. Finally, new methodological approaches to standard setting are presented. In many standards-based testing systems the question of how to reconcile the two logics of accreditation (grading) and diagnostics (testing) is still an unresolved one so that countries can benefit from the approaches presented.

Keywords Standard-setting • Cut score • Validity • Denmark • Norway • Sweden

1.1 Standard Setting in Education

Standard setting targets ambitious and crucial societal objectives by defining benchmarks at different achievement levels. Thus, feedback to policy makers, schools and teachers is provided about strengths and weaknesses of a school system as well as about school and teaching quality including which individual students are at risk to fail. Standard setting consists of procedures to establish conceptual frameworks for different achievement levels per subject and to operationalize these in terms of passing scores defining cut points on the score scale that are used for the classification into the levels. Candidate-centered and test-centered procedures exist.

S. Blömeke (✉)
Centre for Educational Measurement (CEMO), University of Oslo, Oslo, Norway
e-mail: sigrid.blomeke@cemo.uio.no

J.-E. Gustafsson
Faculty of Education, University of Gothenburg, Gothenburg, Sweden
e-mail: jan-eric.gustafsson@ped.gu.se

S. Blömeke, J.-E. Gustafsson (eds.), *Standard Setting in Education*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-50856-6_1

Given that consequences of the outcomes of standard setting may be dramatic on the system, institutional and individual level, quality of standard setting has to be an issue of great concern when applying this methodology. If it fails, significant repercussions can be expected in terms of arbitrary evaluations of educational policy, wrong turns in school or teacher development or misplacement of individual students. Therefore, standard setting needs to be accurate, reliable, valid, useful, and defensible, which is not an easy challenge due to the mix of content expertise, judgment, policy intentions, measurement and statistical expertise necessary.

The experiences with standard setting in the Nordic countries in fact reveal these implications. The mean achievement and the proportion of students that fails on national tests vary substantially in some subjects from one year to another. Similarly, the mean achievement and the proportion of students that fail vary substantially across subjects in one given year. These problems may be a result of varied outcomes of standard setting processes and/or of variation in test difficulty, both types of problems indicating that quality control does not work out as expected. It may not have been accomplished to sufficiently include the different expert groups necessary or to provide them with sufficient understanding of what the different achievement levels actually mean. At the same time, the discussion about methodological problems in standard setting needs to be carried out under awareness of the limitations and drawbacks of traditional approaches to formulating performance standards.

Internationally, a long tradition of methodological research on standard setting exists, in particular in the US and a few European countries. A lot of time and careful thought have been spent on improving the methods—50 states in the US alone have worked on this. In addition, credentialing agencies exist, several of which have made research contributions.

However, specific evidence on the benefits and limits of different approaches is rare and scattered. A particular lack exists with respect to research about standard setting in the Nordic countries (and with a few exceptions in Europe generally) which is problematic given that the number of national tests is increasing here as well while at the same time serious concerns increase at schools about the time and effort spent on national tests without receiving much helpful feedback or support in case of weaknesses. Thus, closely related to clarifying the methodological issues of standard setting is the issue how to transform these into valuable and easy-to-use opportunities to learn for schools and teachers. In this context, a major policy question is what can be done to mitigate the severe problems that standards-based reporting creates such as undesirable incentives for educators.

Against this background, this book has three main objectives: in Part I, the international evidence on methodological issues in standard setting is summarized, and previous research is approached with a fresh outlook. In Part II, the standard setting practices in the Nordic countries are documented and critically discussed. Part III presents new methodological approaches to standard setting. The contributing authors are among the most renowned experts on the topic of standard setting worldwide. All chapters provide therefore a profound and innovative discussion on fundamental aspects of standard setting that hitherto has been neglected. New methodological perspectives combined with a Nordic focus and an inclusion of a

broad range of European authors thus complement the only other existing book on standard setting, edited by Cizek (2012). All chapters provide conclusions for future methodological and policy-related research on standard setting.

1.2 The Chapters in this Book

In Chap. 2 following this introduction, Michael T. Kane discusses the validity of standard setting as the most fundamental quality criterion of a policy measure in education. He shows that standard setting is a type of policy *formation* and that, as such, there is no single "correct" cut score but the *reasonableness* both of the performance standard and the associated cut score is the appropriate criterion of quality. Kane uses an analogy to setting standards in the medical context to underscore his point. Even in such a rather "fact-based" science, there will necessarily be some arbitrariness. Kane introduces the idea of upper and lower bounds wherein a standard could be set. Although, these boundaries will be prone to ambiguity, the process of establishing them and making them transparent enables one to engage in a fruitful discussion about standards. Moreover, the boundaries make the intended interpretation of a score visible. In addition, if the use and interpretation of the score is sufficiently described, it makes it easier to show possible positive and/or negative effects of decisions based on that score.

Mark Wilson and Maria Veronica Santelices continue this fundamental validity discussion in Chap. 3 by expanding the traditionally dominating perspective on standard setting by including conceptual antecedents. The authors criticize the post-hoc nature of current technical practices that would often only start when a test has already been developed and scaled, and thus is taken as a given. Wilson and Santelices argue instead for a more content-focused and criterion-referenced process of standard setting rooted in qualitative evaluations of where thresholds should be by experts before a test is developed. In addition, they argue for a (developmental) learning progression perspective on standards that provides meaningful formative feedback (for teachers) and summative feedback (for policy makers) on a common basis instead of an isolated stand-alone standard at a given point in time. They demonstrate their validity concerns with respect to the Angoff and the Matrix methods, before they illustrate their approach through an expert committee's work on standard setting.

In Chap. 4, Hans Anand Pant, Simon P. Tiffin-Richards and Petra Stanat continue the validity discussion by applying Kane's interpretive argument approach to standard setting in Germany. They discuss in particular the role of standard setting procedures which define *minimum passing scores* on test-score scales. After explaining the German assessment system as a whole, a state-wide assessment of English as a foreign language is used as an example. The authors identify the cut scores as the weakest link in the validity chain, and the gradual widening of the use of a test beyond the purpose for which it was originally intended (i.e., *function creep*) as another severe threat to validity.

Rolf Vegar Olsen and Trude Nilsen contribute in Chap. 5 to the discussion of fundamental issues in the context of standard setting by comparing similarities and differences in the way the two most prominent large-scale international studies PISA and TIMSS set and formulate performance level descriptors. Although the two studies make use of similar methods, different decisions have been made regarding the nature and properties of the finally derived descriptors. PISA and TIMSS are thus cases that illustrate a less researched area in standard setting, namely different approaches to developing level descriptors (cf. Perie 2008; Egan et al. 2012). The authors provide in addition a discussion about ways in which the different approaches may be used both to improve national grading systems and to formulate national curriculum goals, thus demonstrating how the procedures applied by TIMSS and PISA may have relevance in the formulation of national standards.

Barbara S. Plake focuses in Chap. 6, the last chapter of the first part of the book, on where additional research is needed to support the many practical decisions to be made during standard setting. With the authority of someone who has been in the field for a long time, Plake provides multiple examples of standard setting procedures. She criticizes weak practices and suggests practical improvements and research directions. "Operational ratings" are used as a case to demonstrate these needs, because only some of the standard setting decisions have been based on scientific studies, whereas most have been based on human judgment, or for streamlining the process without research that supports the decisions.

Part II of the book is about the specifics of standard setting in the Nordic countries. Peter Allerup and Christian Christrup Kjeldsen present in Chap. 7 the national assessment system in Denmark. This is not only a very interesting case of standard setting practices in the Nordic context, but presents in addition the generic challenge of how computer-based adaptive testing challenges current views on how to perceive, set and work with standards in educational settings. Implementing testing at a national and system level in a computer-based and adaptive way is an innovative and, until now, only infrequently used way. The chapter presents thus for the first time the implications connected to adaptive testing, both positive and negative, for how standards are developed, understood and used.

Gudrun Erickson presents in Chap. 8 the Swedish case, which is another educational system with an elaborate standard setting system. However, the system has been developed in a decentralized way, and different procedures and practices have been established in different subject matter areas. This has created a need to develop a common framework for test development, including procedures for setting standards.

Chapter 9 by Anna Lind Pantzare describes an approach to validating Angoff-based cut scores using equating procedures in Sweden. Only few studies have so far investigated the validity of cut scores, so that this chapter closes a serious research gap by comparing a teacher-ratings driven classification system with a student-response driven classification system. The two approaches converge well in this case, which is linked to the nature of the topic (highly structured) and of the teacher involvement with the actual test (high).

Idunn Seland and Elisabeth Hovdhaugen cover the Norwegian case. This Chap. 10 presents a case of standard setting that is elaborate in practice but undeclared in theory. The authors draw on a complex set of quantitative and qualitative data from teachers, principals and school owners (municipalities), so that a description of the network of actors and how they interpret the national assessments and their interaction with other curriculum defining instruments and documents emerges. It seems as if curricular and assessment standards are widely disregarded by teachers and downplayed by educational authorities, so that the potential of national tests cannot fully be utilized for the development of educational objectives or for strengthening pedagogical efforts.

Eli Moe and Norman Verhelst applied such a modification of the Cito standard setting method to identify cut scores for a multistage reading and listening test in Norwegian for adult immigrants. Test scores are mapped onto the levels of the Common European Framework of Reference for Languages (Council of Europe 2001). The authors faced specific challenges regarding setting standards for this unique population. Thus, Chap. 11 contributes substantially to other accounts of the use of standard setting in the CEFR context (e.g., Martyniuk 2010; Tannenbaum and Cho 2014).

Chapter 12 by Ove Edvard Hatlevik and Ingrid Radtke presents an application of standard setting to recommend cut scores; however, in this case it is for a formative assessment of digital responsibility. The two standard setting methods applied (Angoff and Bookmark) are well-established, and so the value of this chapter lies not only in the domain which is complex and only recently upcoming, but also in how decision-makers negotiated the differences in recommendations from the two standard setting methods.

Finally, in Chap. 13 Gustaf B. Skar, Ragnar Thygesen, and Lars Sigfred Evensen take on the challenge of setting standards with the objective of contributing to assessment for learning in Norway. Based on a conceptual framework that elaborates on this concept of assessment for learning, the authors present two studies, namely of how assessments for learning can be developed in a bottom-up process, and how consistency can be assured in the process of standard setting. Analyses of item-characteristic curves (time series as well as comparative analysis across contexts) demonstrate that a considerable increase in reliability develops over time, but simultaneously imply a number of remaining challenges, and that further refinements will be needed in order to reach satisfactory levels.

In Chap. 14 , the final chapter of this second part of the book, Mari-Pauliina Vainikainen, Helena Thuneberg, Jukka Marjanen, Jarkko Hautamäki, Sirkku Kupiainen and Risto Hotulainen present the Finish case. This country succeeds in education without a formalized standard setting approach. However, educational monitoring happens continuously at the local level and through a national model for sample-based curricular and thematic assessments. The chapter presents this system. It turns out that the screening of support needs and the evaluation of the effectiveness of the provided support are crucial for explaining Finland's success in international comparisons.

The third and last part of this book presents new methodological approaches to standard setting. Jos Keuning, J. Hendrik Straat, Remco C.W. Feskens and Karen Keune propose in Chap. 15 an extension of the Direct Consensus approach as one of the best-known procedures for establishing performance standards (Sireci et al. 2004). Their extension includes clustering items and using cut scores applied to those clusters to predict the cut score for the full-length test, thus bringing the strengths of the traditional standard setting procedures together. This is a substantial extension of the existing approach and thus a unique contribution to the methodological discussion.

Chapter 16 by Allistair Pollitt describes the use of teacher judgment as a form of equating to maintain comparability of cut scores across test forms. This is a unique addition to the field of standard setting and measurement, where standard setting is, at times, used as a proxy for equating, when test volumes are too low for a formal equating to occur. Pollitt illustrates his Thurstone-based approach, that is applicable in various scenarios, with four examples. One surprising finding is, for example, that comparisons between (performance-wise) more heterogeneous scripts are associated with less consistent judgments.

In Chap. 17, Amanda A. Wolkowitz, James C. Impara and Chad W. Buckendahl reinforce the notion that standard setting should begin at the outset of test development—that performance level descriptors (PLD) should inform specification and item construction. This recommendation is in line with the chapters in Part I of the book, which is also consistent with Evidence-Centered-Design practices and principles (e.g., Mislevy and Haertel 2006). The authors provide an extended case study of how item writers make use of the performance level information when constructing items. The paper argues that it is advantageous to develop PLDs prior to item writing, because it yields items which are better aligned to the cut scores of the different levels, making the job of the standard setting panels easier and more consistent, and the test more efficient in targeting the different levels. A case study is presented to illustrate and support the points made.

Linda I. Haschke, Nele N. Kampa, Inga Hahn, and Olaf Köller propose in Chap. 18 an application of the item-descriptor (ID) matching method to a test on adults' competencies in the domain of science, thus addressing not only a unique population, but also covering an under-researched domain, and applying a method only infrequently used so far. The authors describe how they developed the ID method further and provide insights into its application. On the basis of a validity framework presented in Chap. 4 of the first part of the book, they address different aspects of validity to obtain evidence on the appropriateness of this standard-setting method.

Combining a methodological perspective with a policy and practice perspective on standard setting, as it is done in this book, is an infrequent approach. Moreover, the focus on the Nordic countries adds specific value to the discussion about standard setting, since research in this specific field regarding the Nordic region is scarce. Looking at standard setting in the Nordic countries opens up a specific opportunity to compare the status and function of standard setting procedures among differently evolved systems of standards-based assessment. In addition, the

discussion of how to link grading and standard setting is taken up. In many standards-based testing systems the question of how to reconcile the two logics of accreditation (grading) and diagnostics (testing) is still an unresolved one, so that countries can benefit from the approaches that are presented in this book.

References

Cizek, G. J. (Ed.). (2012). *Setting performance standards foundations, methods, and innovations*. New York/London: Routledge.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Press Syndicate.

Egan, K. L., Schneider, M. C., & Ferrera, S. (2012). Performance level descriptors: History, practice and a proposed framework. In G. J. Cizek (Ed.), *Setting performance standards* (pp. 79–106). New York: Routledge.

Martyniuk, W. (Ed.). (2010). *Relating language examinations to the Common European Framework of Reference for Languages: Case studies and reflections on the use of the Council of Europe's draft manual*. Cambridge: Cambridge University Press.

Mislevy, R., & Haertel, G. (2006). *Implications of evidence-centered design for educational testing* (Draft PADI Technical Report 17). Menlo Park: SRI.

Perie, M. (2008). A Guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practices, 27*, 15–29.

Sireci, S. G., Hambleton, R. K., & Pitoniak, M. J. (2004). Setting passing scores on licensure exams using direct consensus. *CLEAR Exam Review, 15*(1), 21–25.

Tannenbaum, R. J., & Cho, Y. (2014). Critical factors to consider in evaluating standard-setting studies to map language test scores to frameworks of language proficiency. *Language Assessment Quarterly, 11*, 233–249.

Part I
Fundamental Questions in Standard Setting

Chapter 2
Using Empirical Results to Validate Performance Standards

Michael T. Kane

Abstract Standard setting extends the interpretations of scores by adding a standards-based inference (from test scores to performance levels) to the interpretation/use argument (IUA) for the underlying score scale. For standards-based interpretations and uses to be valid, this additional inference needs to be justified. The supporting evidence can be procedural, internal, and criterion-based. Criterion-based evidence is especially important in high-stakes contexts, where the standards tend to be contentious. Standards are inherently judgmental, and therefore, to some extent, arbitrary. The arbitrariness can be reduced to some extent by employing empirical relationships (e.g., dosage-response curves) to estimate upper and lower bounds on the cut score. In evaluating standards, the question is not whether we got it right, but rather, whether the decisions based on the cut scores are reasonable, broadly acceptable, and have mostly positive consequences (which outweigh any negative consequences).

Keywords Standard setting • Validity • Criterion-based validation • Dosage-response curves

2.1 Introduction

On June 17, 1998, overnight, almost 30 million Americans became clinically overweight and several million became clinically obese. This apparent public-health crisis was not caused by an epidemic of overeating, but rather, by changes in the cut scores for these clinical categories on the body mass index (BMI), a measure of percentage body fat. The changes in the standards were made by the National Institutes of Health (Greenberg 1998; Shapiro 1998) and were based on research linking higher BMIs to various health problems (particularly cardiovascular disease and diabetes). Changes in standards can have dramatic effects. An increase in the

M.T. Kane (✉)
Educational Testing Service, Princeton, NJ, USA
e-mail: mkane@ets.org

S. Blömeke, J.-E. Gustafsson (eds.), *Standard Setting in Education*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-50856-6_2

passing score on a test will decrease the pass rate, and a decrease in the passing score will increase the pass rate. Once the distribution of scores is known (or predicted), the pass rate is an entirely predictable function of the passing score. Depending on where the passing score falls in the score distribution (e.g., on a certification test), even modest changes in the passing score could produce dramatic changes in pass rates, and these changes can vary substantially across groups (e.g., race/ethnicity, gender). In contrast, the impact of changes in test design (e.g., changes in test length, format, or content specifications) is less predictable and usually far less dramatic.

2.2 Standards, Fairness, and Arbitrariness

Standard setting is difficult, and it can have serious consequences, but it can also have substantial advantages. By setting a standard that yields a cut score on a test-score scale, we can change a subjective evaluation of a person's performance level in some domain into a simple, objective comparison of a test score to the cut score. This kind of standard-based decision rule tends to provide an efficient way to make decisions, but more important, it tends to promote transparency, fairness, and perhaps as important, the perception of fairness (Porter 1995):

> Scientific objectivity thus provides an answer to a moral demand for impartiality and fairness. Quantification is a way of making decisions without seeming to decide.

All standard setting methods are subjective to some extent. They all involve judgments about how much is enough or how much is too much. But once the standard is set, the operational subjectivity is eliminated, or at least, enormously reduced. Once the BMI guidelines were set, a decision about a person's weight status could be made by consulting the guidelines.

However, the consistency and appropriateness of judgmental standard setting in education has been repeatedly questioned. Different methods tend to give different results, and there has been no obvious way to choose among the conflicting results. Glass (1978) suggested that the results of educational standard setting tend to be arbitrary, and that it is "... wishful thinking to base a grand scheme on a fundamental unsolved problem." Since 1978, many writers have acknowledged that standards are arbitrary in the sense of being judgmental, but also that they need not be arbitrary in the sense of being unjustified or capricious (Hambleton and Pitoniak 2006). The final decisions about the BMI cut scores were made by a committee, but the committee relied on an extensive body of clinical research. The exact values of the cut scores were a bit arbitrary, but their general locations were supported by a wealth of empirical data.

It is possible to set clear, defensible standards in many contexts, but standard setting is difficult in most contexts (Glass 1978), and all standards have a large element of subjectivity. The extent to which the arbitrariness is a problem depends on how much it interferes with the intended use of the standard. An effective response to

charges of arbitrariness is a demonstration of an appropriate relationship between the standards and the goals of the testing program in which they function.

2.3 Educational Standards as Policies

Educational standard setting is designed to address a basic policy question about how good a performance must be in order to be considered good enough for some purpose. It adds a layer of interpretation (involving one or more performance levels) to the assessment scores, and it replaces subjective evaluations with objective, score-based decisions. The goal is to establish a reasonable basis for score-based decisions. The issue is not whether the standards are accurate, but rather, whether they are appropriate, in the sense that they achieve their intended purpose at acceptable cost. Policy making generally involves balancing of competing goals.

In evaluating standard setting efforts, it is useful to draw a distinction between a *cut score*, which is a point on the score scale for the assessment, and a *performance standard* that specifies a particular level of performance. For standards-based interpretations, it is claimed that test takers with scores above the cut score have generally achieved an appropriate performance level and that those with scores below the cut score have not achieved the performance level.

Standards are "set," and to be widely accepted, they have to meet certain criteria. First, they have to be reasonable in the sense that they are neither too low nor too high; the standard should be high enough to achieve its intended goal, but not so high as to cause serious side effects. Second, they have to support the claims included in the performance-level descriptions; in general, the students assigned to a performance level should be able to perform the tasks associated with the level (as described in a performance-level description) and should not be able to perform the tasks associated with the next-higher level. Third, the standards should be applied consistently across students and contexts, and until they are revised, across time.

2.4 Overview

In the next section, I will outline an argument-based approach to validity, which requires, first, that the claims based on the test scores and the assumptions inherent in these claims be explicitly stated, and, second, that the plausibility of these claims be evaluated using relevant evidence. Of particular interest in standard setting is a claim that test takers with scores above the cut score generally have achieved some performance level, and those with scores below the cut score generally have not achieved that level. The plausibility of this claim is the central concern in validating the standard-setting process.

I will then discuss standard setting in broad terms, and in particular, empirically-set standards based on dosage-response relationships, and judgmental standards setting procedures in education. I will focus on the use of empirical relationships to

establish upper and lower bounds on cut scores that are to be deemed reasonable. By establishing such bounds, it is possible to evaluate the validity of the cut scores and performance standards, and to characterize the level of arbitrariness (in terms of the range of possible cut scores between the greatest lower bound and the least upper bound) in the final cut score. Finally, I will draw some general conclusions and a "take away" message.

2.5 Validity

An argument-based approach to validation (Cronbach 1988; Kane 2013) focuses on the evaluation of the claims based on test scores and makes use of two kinds of arguments, an *interpretation/use argument* (IUA) that specifies what is being claimed and a *validity argument* that evaluates the plausibility of the IUA. A proposed interpretation or use of test scores is considered valid to the extent that the IUA is coherent and complete (in the sense that it accurately represents the proposed interpretation and use of the test scores), and its assumptions are either highly plausible a priori, or are adequately supported by evidence. It is the proposed score interpretation and uses that are validated, and not the test itself or the test scores, and the validity of the claims being made depends on how well the evidence supports these claims.

By specifying the claims being made, the IUA provides guidance on the kinds of evidence needed for validation. Once the IUA is developed, it provides a framework for collecting validity evidence, as well as criteria for evaluating the overall plausibility of the proposed interpretation and use of scores. If the IUA is coherent and complete and all of its inferences and assumptions are well supported, the interpretation/use can be considered valid. If any part of the IUA is not plausible, the interpretation/use would not be considered valid.

The validity argument subjects the IUA to critical evaluation. It is contingent, in the sense that it depends on the proposed interpretation and uses of the test scores. If the IUA makes only modest claims (e.g., that the scores indicate a test taker's competence in performing the kinds of tasks on the test), the validity argument can also be modest. If the IUA is ambitious (e.g., that the scores reflect a theoretical construct, or can be used to predict some future performance), the validity argument would need to provide support for these claims. The argument-based approach can be applied to a range of possible interpretations and uses, but in all cases, the claims being made need to be clearly stated and evaluated.

2.6 Interpretation/Use Arguments (IUA)

The IUA provides an explicit statement of the reasoning inherent in a proposed interpretation/use of test scores, and typically includes a number of linked inferences (Kane 2013; Toulmin 2001). The inferences take the general form of "if-then" rules that allow us to make a *claim* based on some *datum*. The if-then rule constitutes a

warrant for asserting the claim based on the datum for specific test takers. For the warrant to be accepted, it must be supported by adequate *backing*, or evidence that supports the if-then rule. Arguments (e.g., an IUA) are constructed using networks (or sequences) of inferences that are linked by having the claims resulting from earlier inferences serve as data for later inferences. For example, a score interpretation in terms of expected performance in some domain might be specified in terms of three main inferences: scoring, generalization, and extrapolation.

The scoring rule, or scoring inference, takes a test taker's responses to test tasks as its datum and generates an observed score as its claim. The scoring rule might be a simple sum of scores on test tasks/items, based on a scoring key or scoring rubrics, or it might employ statistical models (e.g., equating/scaling) to generate the scores. The backing for the scoring inference typically involves expert opinion for the appropriateness of the scoring rules, empirical evaluations of statistical assumptions, and in the case of extended-response tasks, empirical support for rater consistency and accuracy.

A generalization inference takes the observed score as a datum and makes a claim about expected performance over replications of the testing procedure. The generalization inference extends the interpretation from an evaluation of performance on a particular instance of the assessment to expected performance over a universe of replications of the assessment procedure (e.g., a universe score in generalizability theory). The backing for this inference is generally derived from empirical estimates (reliability or generalizability studies) of the score consistency across replications of the assessment.

An extrapolation inference extends the interpretation from test performances to some broader domain of "real-world" performances that are of interest, or to claims about a trait. If the interpretation is extrapolated to some kind of non-test performance (e.g., in college or on the job), the backing might involve empirical (e.g., regression) analyses and or qualitative analyses of the commonalities in the knowledge, skills, and abilities required by the assessment and by the non-test performances. For traits, the backing would include empirical evidence that the assessment scores have the properties expected, given the definition of the trait (Messick 1989).

Standard setting adds an additional layer of meaning to a proposed interpretation, involving a claim that test takers with scores at or above a cut score are different in some way from those with scores below the cutoff; in most cases, it is claimed that test takers with scores at or above the cut score are probably prepared for some activity (e.g., for college or a profession) and that those with scores below the cutoff are probably not adequately prepared for the activity. The additional inferences and assumptions associated with this claim need to be evaluated in order for the overall IUA to be considered valid.

2.7 Validity Argument

The validity argument is to evaluate the IUA in terms of its clarity, coherence, and plausibility. The proposed interpretations and uses are valid to the extent that the IUA reflects the interpretation and uses, and the warrants for all of the inferences in

the IUA are either inherently plausible or are supported by adequate evidence. In this chapter, the focus is on how to evaluate the claims introduced by standard setting. Before discussing how one might evaluate the standards-based claims, over and above the underlying interpretation, it is helpful to be clear about what standard setting claims to do, and what it is capable of doing.

2.8 Standard Setting

All standard setting has some characteristics in common. First, the standard is set or established by some authority (e.g., a government, a professional or scientific organization). The standard is not discovered or estimated; it is set, and it does not exist until it is set. Second, the standard is definite, and more or less objective, in the sense that it can be consistently applied to a range of cases without much ambiguity. Standard setting aims to replace some kind of subjective decisions with objective, score-based decisions. There is much value in this kind of objectivity, especially if the standard is justified, or validated, and commands general acceptance.

Third, the standards-based decisions assign each test taker to one of a sequence of categories. In the simplest case, there is one standard, and there are two categories (e.g., pass/fail); the standard is either satisfied or not. In other cases, a set of n standards is used to define $n+1$ categories (e.g., below basic, basic, proficient, advanced). Fourth, once established, the rule or standard is to be applied consistently in making the categorization decisions. It provides a way of automating these decisions, and thereby, making the decisions more transparent and fair.

2.9 The Goldilocks Criteria

In practice, standards-based decisions are generally implemented to achieve some goal, while avoiding serious side effects. The goal can suggest a general level for the standard, even though it does not generally specify a precise value. In the context of licensure testing, Kane et al. (1997) proposed "Goldilocks Criteria" for evaluating passing scores and the standard-setting methods used to generate them:

> The ideal performance standard is one that provides the public with substantial protection from incompetent practitioners and simultaneously is fair to the candidate and does not unduly restrict the supply of practitioners. We want the passing score to be neither too high nor too low, but at least approximately, just right.

The standard should not be too low (i.e., below reasonable lower bounds) and not be too high (i.e., above some reasonable upper bound). The exact placement of the standard between the bounds would be a matter of judgment, and in that sense, arbitrary, but this arbitrariness is not necessarily a problem. As long as the standard

is high enough to achieve the goals of the program and not so high as to cause serious problems, the standard can be considered reasonable. Standard setting tends to be easiest and most defensible when we have clearly defined goals and a good understanding of potential side effects.

For example, a requirement that a ferry have enough life jackets for its passengers and crew has an obvious purpose and an obvious justification in terms of the purpose. The number of passengers and crew sets a lower bound on the number of life jackets, but it would probably be reasonable to have extra life jackets in various locations on the ship, so that a lifejacket will be readily available to everyone on board if needed. However, we do not want so many lifejackets that they interfere with the functioning of the ship or add so much cost that they make the running of the ship prohibitively expensive. So the number of passengers and crew provides a clear lower bound, but it does not provide a point estimate of the number of lifejackets.

In setting standards for jobs requiring physical strength as a major requirement, it is possible to estimate the strength requirements of the job (e.g., in terms of the heaviest object to be lifted by hand) and set cut scores on strength assessments at or somewhat above the maximum requirements of the job (Campion 1983). The lower bound is grounded in the requirements of the job, and therefore, does not seem arbitrary. There is some uncertainty, or arbitrariness, in estimating the strength requirements and in deciding the safety margin to include, but the legitimacy of the lower bound for the strength requirement can be justified by the nature of the work to be done. A clear upper bound might also be available, if regulations limit the weight of the objects that need to be handled (e.g., weight limits on packages that can be mailed). Setting reasonable upper bounds is especially important in such employment contexts, because setting the requirement too high could unnecessarily exclude women and other protected groups (Campion 1983). The Goldilocks Criteria suggest that standards need to be set high enough to achieve the goal of standard setting (in this case to prevent injury), but not so high as to cause serious side effects (e.g., adverse impact).

The target performance levels on most educational tests are not so well defined. It is clearly better for high school graduates to know more mathematics, rather than less mathematics, but how much is enough? Should the target performance level in mathematics on a high school graduation test be set at a level appropriate for college-bound students (and if so, should the focus be on those planning to major in engineering or in sociology), or should the focus be on those planning to go directly into the world of work. To the extent that the goal of the standard setting can be specified, it may be possible to set lower bounds for the standard, and to the extent that potential side effects can be specified and estimated, it may be possible to set upper bounds. To the extent that the upper and lower bounds are close to each other, the resulting standard in not very arbitrary.

2.10 Empirical Standard Setting Based on Dosage-Response Curves

The organizations that promulgate health and safety guidelines, or standards, generally rely on accumulated research describing relationships between input variables and various outcomes, and the resulting recommendations get respect and acceptance (if not compliance), because they have empirical support. The BMI standards are based on extensive data relating BMI scores to outcomes like heart disease and diabetes.

In cases where some treatment (e.g., a drug) is intended to produce some response or effect (e.g., alleviation of pain), the relationship between level of treatment (or dosage) and the outcome can often be examined empirically, and the resulting dosage-response curves are generally not linear. Assume, for example that a new drug has been shown to be effective for some clinical purpose. Before using the drug on a large scale, studies are typically carried out to examine the relationship between clinical effect and dosage. Such a study might yield something like the dose-response curve in Fig. 2.1 or Fig. 2.2. For low dosages, the effect is negligible, and it does not increase much as a function of dosage until it gets into a critical range where the effect increases fairly quickly as a function of dosage. The effect then levels off, or "plateaus." Dose-response curves do not generally have this simple logistic shape, but some do, and I will use this simple model as the basis for discussion.

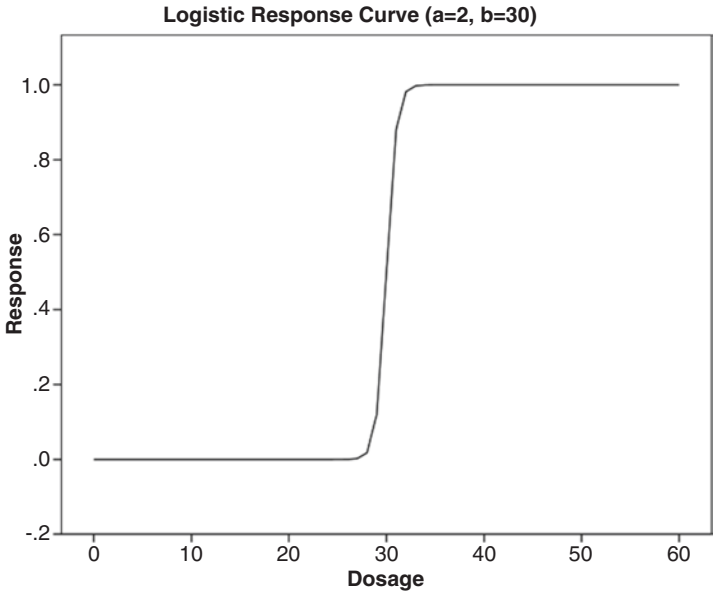

Fig. 2.1 Dose-response Curve with a Sharp Transition

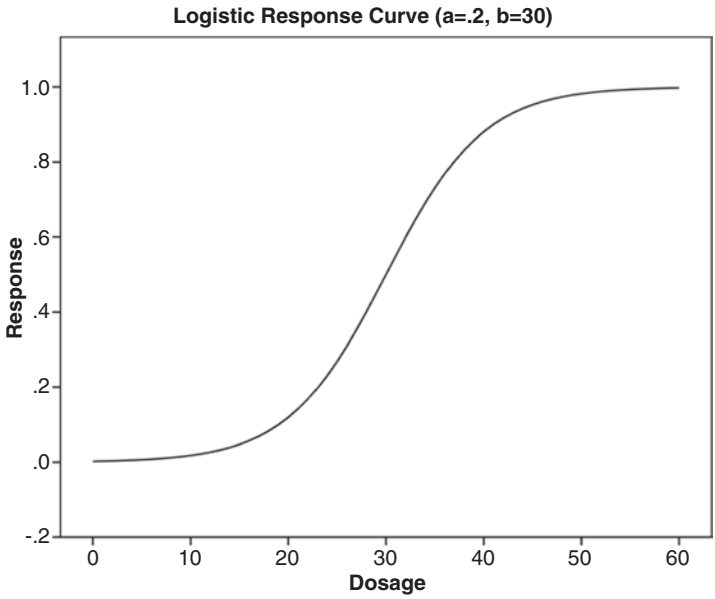

Fig. 2.2 Dose-response Curve with a well-defined Critical Range

This kind of quantitative model is very helpful to clinicians, who want to be able to prescribe a dosage that is high enough to have the desired effect, but not too high. Dose-response curves like those in Figs. 2.1 and 2.2 suggest the general location for a standard dosage; in order to achieve a high response, the dosage should be at or above the high end of the critical range, but going beyond the critical range does not add much to the expected response, and in many cases, using high dosages may lead to toxic side effects.

If the dosage-response curve approximates a step function (as in Fig. 2.1), for which there is little or no response for low doses followed by a rapid increase in the response to some maximum value, a standard dosage would be well defined. For the curve in Fig. 2.1, a dosage of about 30 or a little higher (e.g., 31 or 32) would seem to be an optimal choice in terms of achieving the intended response, without unnecessarily high dosages.

More commonly, the dosage-response curve is similar to the logistic curve in Fig. 2.2, with a very low response for low dosages, and then a gradual increase in the expected response and a flattening out for the higher dosages. Assuming the outcome is important, clinicians would prefer a lower bound that corresponds to a response that is above .5; if the treatment has no serious side effects, the minimal response might be set well above .5. In this case, the dosage-response curve could suggest a lower bound, but not say much about an upper bound. The general location of the standard dosage is indicated by a *critical range* between the upper and lower bounds, but there is a lot of room for debate about the exact location (which is one major reason why standards are set by committees).

In many cases, we do not have nice, smooth curves like that in Fig. 2.2, but rather, some general information about how the response changes as a function of dosage. In discussing the health benefits of exercise, the Tufts health-and-nutrition letter (Tufts University 2015) described two large-scale studies designed to find the "sweet spot" for the health benefits of exercise. Tufts University (2015) found that some activity was better than no activity and that meeting the pre-established guidelines of 150 min of moderate activity or 75 min of vigorous activity per week was associated with a 31% decrease in mortality, and reported that:

> Risk continued to drop with ever-increasing activity levels: 37% lower at two to three times the minimum guidelines and 39% lower at three to five times. But at that point … the association plateaued. There was no additional mortality benefit for even more exercise, but neither were there any negative associations.

The critical range indicated is pretty broad, stretching from 150 min to 750 min (or 12.5 h or more) of moderate activity.

The standard dosage can often be made more precise by considering multiple outcomes. Most treatments will have some side effects at high dosages, and this tends to be a major consideration in determining the standard dosages. For example, in the case represented in Fig. 2.2, if some serious, negative side effect (e.g., death) begins to occur at dosages around 40 and the incidence increases fairly rapidly as dosage increases above 40, it would make sense to set the upper bound at 40 or a bit lower. Note however, that if the intended effect of the drug is important enough (curing an otherwise incurable disease), the upper end of the critical range might be allowed to go above 40. Again, these decisions generally rely on the collective judgment of committees, because they often involve difficult tradeoffs, but they are not arbitrary; they are based on empirical studies of intended outcomes and side effects.

It is generally desirable to consider as many relevant outcomes (intended and unintended) as possible, because each significant outcome may be helpful in defining an upper or lower bound, or both. The committee responsible for setting the standard can then develop an overall critical range by identifying a greatest lower bound and a least upper bound. At some point, the committee responsible for setting the standard will run out of additional criteria that can be used to narrow the critical range, and at that point the committee will turn to more loosely defined criteria that are relevant, but not as well defined or generally accepted as the criteria used to constrain the critical range. For example, the importance of the intended response and the seriousness of the side effects can play a major role. If the intended response is very important (e.g., treating a fatal disease) and the side effects are not too serious (e.g., pain, nausea), the standard is likely to be set near the top of the critical range. If the intended response is less important (e.g., pain control) and the side effects are serious (e.g., death), the standard is likely to be set near the bottom of the critical range.

In cases where the intended effect and the potential negative side effects are comparable in their seriousness, deciding on standard dosages involves serious tradeoffs that are not easily resolved. In these cases, the committee is expected to use its collective wisdom to choose a point in the critical range that optimizes the tradeoff in some sense. Although the committee members could achieve agreement on the upper and lower bounds, which are strongly dependent on empirical results, it may be harder to achieve consensus of the choice of standard within the critical range.

The general methodology employed in using the dosage-response curves to set standards involves the use of various relevant empirical relationships to put bounds on the standard dosage, with the aim of identifying a fairly tight critical range, followed by a subjective judgment about exactly where to put the standard within that range. The critical range is not arbitrary, because it is determined by the empirical relationships, and the empirical results provide pretty compelling support for the general location of the standard (i.e., for the critical range), but not for a precise value.

This residual uncertainty is not necessarily a major problem. As noted above, there is no correct value for the standard, and much of the benefit of the standard is derived from having a well-defined, objectively applied standard in more-or-less the right place. Given the potentially strong empirical support for the critical range, any point in the critical range could be considered to be in more or less the right place (especially, if the critical range is fairly narrow), and for policy making, this can be good enough. Standard setting always has a goal. The goal may be to cure patients or to have students achieve some level of competence in some area. In setting the standard, we want to make it likely that we will achieve the goal (a positive consequence), without major negative consequences. So standard setting is necessarily a balancing act, and goals and side effects are easier to evaluate and compare, if they are well defined and specific.

2.11 Judgmental Standard Setting

Judgmental standard setting involves the use of a panel (or panels) of judges to set cut scores on a score scale to represent certain performance levels (Hambleton and Pitoniak 2006; Zieky et al. 2008). The goal of standard setting is to identify *cut scores* on the score scale that correspond to the performance levels, and to the extent necessary to expand or clarify the performance level descriptors. The number and nature of the performance levels depend on the intended purpose of the standards-based interpretation. In many cases, a single performance level and a single cut score are used to distinguish between acceptable and unacceptable performance (i.e., for pass/fail decisions).

Some policy-making group decides on the number of levels, on their labels, and on preliminary descriptions of the levels. For example, the National Assessment of Educational Progress reports on the performance of students at various grade levels in the United States in terms of three performance levels (“basic,” “proficient,” “advanced”). The National Assessment Governing Board, which develops the performance level descriptors, defined the proficient level, in general, as “solid academic performance exhibiting competency over challenging subject matter” (Loomis and Bourque 2001). For each grade level and subject area, each proficient level is specified in more detail, and for 12th-grade students, the proficient level in mathematics has been specified by Loomis and Bourque (2001) as:

> Twelfth graders performing at the proficient level should demonstrate an understanding of algebraic, statistical, and geometric and spatial reasoning. They should be able to perform algebraic operations involving polynomials; justify geometric relationships and judge and

> defend the reasonableness of answers as applied to real-world situations. These students should be able to analyze and interpret data in tabular and graphical form; understand and use elements of the function concept in symbolic, graphical, and tabular form; and make conjectures, defend ideas, and give supporting examples.

This performance-level descriptor clearly reflects a high level of academic performance, and is quite specific in the areas of mathematics included in the descriptor, but it allows for judgment of what constitutes "solid academic performance" in demonstrating an understanding of these topics. The performance-level descriptions define the proposed interpretation for standards-based reporting of test results. The cut score is the operational version of the target performance level. To validate the use of the standards-based interpretation is to show that the target performance level is reasonable and appropriate, given the decision to be made, and that the cut score reflects the requirements in the target performance level.

For the standards-based interpretation, all test takers assigned to a performance category are taken to have achieved the performance level for that category, but not to have achieved the performance level for the next higher category. So, for example, for a licensure test on which increasing scores represent increasing competence in some domain, and setting a cut score (i.e., a passing score) adds a claim that scores above the cut score represent adequate (passing) performance and that scores below the cut score represent inadequate (failing) performance. In some cases, licensure agencies have chosen to report only on this 0/1 scale and to not report scores on the original score scale (or to report these scores only to failing candidates, who generally want to know how far below the cut score they scored). Once in place, the cut scores provide a clear, objective way of deciding whether each individual has passed or not.

In using the results of score-based decisions, we tend to talk and act as if we have a dosage-response relationship like that in Fig. 2.1, even though the relationship is more like that in Fig. 2.2, or more likely, like that in Fig. 2.3.

A test taker with a score at or just above the cut score is considered to be at the corresponding performance level, while a test taker with a score just below the cut score is assumed not to have achieved that level. In educational standard setting, we are imposing a sharp distinction where none exists to begin with (Shepard 1980). Wherever we set the cut score, there will not be much substantive difference between the test taker with a score one point above the cut score compared to the test taker with a score one point below the cut score. So some ambiguity is inevitable, but such ambiguity is a less serious problem than ambiguity in the general location of the cut score.

2.12 The Validity of Standards-Based Categorizations of Test Takers

Standard setting is concerned with how good is good enough; there is some goal to be achieved and some unintended side effects to be avoided, to the extent possible. The question of validity can be stated in terms of how well the goal is achieved and how well the side effects are avoided.

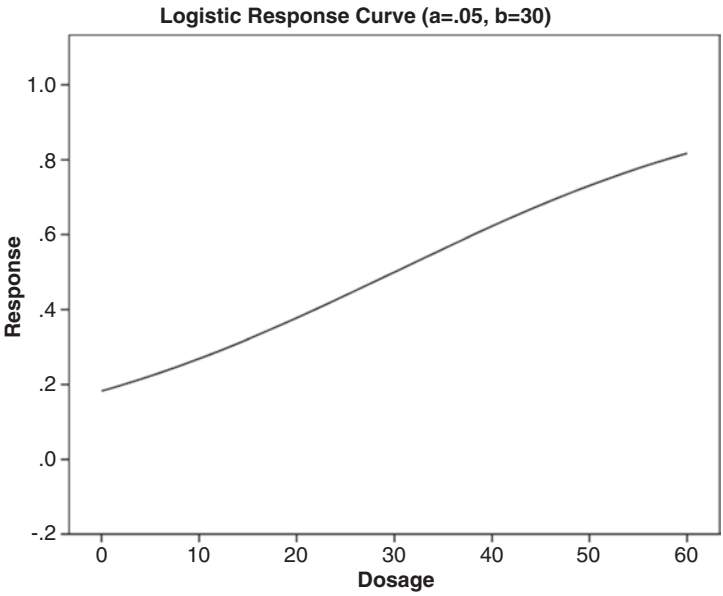

Fig. 2.3 Dose-response curve with a broad critical range

2.13 Standards-Based Inferences and Assumptions

The standards-based inference takes us from a scaled score to a conclusion about whether the test taker has achieved the performance level. In going from the scaled score to the categorical variable, the interpretation goes from a relatively fine-grained score scale to a much coarser-grained categorical scale. In making this shift, some information about performance differences is lost, but to the extent that the performance categories are well defined, overall interpretability may be improved.

There are at least two major assumptions needed to support this inference. First, the performance level specified in its label and in its descriptor is appropriate given the intended use of the categorical variable. Second, the cut scores are such that test takers assigned to a category have achieved the performance level defining the lower bound on that category, and have not achieved the performance level defining the lower bound of the next higher category. There are at least three kinds of evidence that can be used to provide evidence/backing for performance level warrants: procedural, internal consistency, and external relationships.

Procedural evidence for a performance level warrant would be derived from an evaluation of the methods used to define the performance levels and to set the corresponding cut scores; these procedures should be consistent with the intended use of the cut scores, should be thorough and transparent, and should be consistent with current standards of practice. The issues to be addressed in evaluating procedural evidence would include the relevance of test content and format to the intended use of the scores, the appropriateness of the standard-setting method given the test

design, the representativeness of the sampling of judges, the adequacy of the training of judges, the sampling of items or test-taker performances (where relevant), the appropriateness of the feedback to judges, and the confidence of the judges in the results. Procedural evidence can be especially decisive in undermining validity, but cannot, in itself, justify the performance level inference; it provides a limited but important check on the reasonableness of the standard setting.

Internal-consistency evidence uses internal relationships to check on the reasonableness of the standard setting. Analyses of the precision (reliability or generalizability) of the results over judges, panels, and occasions provide one important internal-consistency check on the plausibility of the results. For test-centered methods, like the Angoff method (Hambleton and Pitoniak 2006), agreement between item ratings and empirical item difficulties provides a check on how well the panelists understand how test takers are responding to the test tasks. Reasonableness of changes in ratings over rounds of the standard-setting process can provide an additional check on the ratings. Again, discrepancies undermine validity claims, but consistency is less decisive; internal consistency is a necessary condition for the acceptability of the standard-setting results, but it is not sufficient.

As discussed in more detail below, external validity evidence can take many forms. In some cases, it may be possible to compare the category assignment based on alternate measures (e.g., international benchmarks) to the categorizations based on the cut score, or to compare the cut scores to those obtained using other standard-setting methods. The value of such comparisons depend, in large part, on the suitability and quality of the external measures. For the performance level inference to be accepted, it needs to be backed by adequate evidence. An effective response to charges of arbitrariness is a demonstration of an appropriate relationship between the standards and the goals of the program.

2.14 Using Empirical Data to Evaluate Judgmental Standards

Empirical results can provide a particularly effective way to evaluate, or validate, performance standards, because they subject the proposed interpretation to serious challenges. As Cronbach (1980) suggested:

> The job of validation is not to support an interpretation, but to find out what might be wrong with it. A proposition deserves some degree of trust only when it has survived serious attempts to falsify it.

The cut scores in high-stakes testing programs should be able to withstand critical scrutiny.

As noted earlier, it tends to be easiest to set and validate defensible standards in cases where the standards are intended to achieve some well-defined goals, and some standard-setting efforts in education employ very precisely defined goals. In these cases, the performance standard is defined in terms of a specific observable

outcome, and the corresponding cut score can be set empirically by relating the test scores to the outcome variable. For example, "college readiness" as a standard of performance can be operationally defined in terms of some outcome variable (Beaton et al. 2012):

> Presumably this means earning at least passing grades. Others might suggest that the criterion should be higher – getting a B^- or better with a 50 percent probability, or a C^+ or better with a 75 percent probability, for example.

In these cases, college readiness is defined in terms of a particular level of performance on a particular scale (e.g., college grades), which is taken to define adequate college performance. The policy question of how good is good enough is addressed when the criterion is chosen (e.g., having a 50% chance of maintain a B or better in certain kinds of colleges or programs); finding a cut score on the test score scale corresponding to this criterion level of performance is an empirical, statistical issue of linking the cut score to the criterion performance. This kind of criterion-based analysis can be carried out without asking the basic standard-setting question of how good is good enough, and has more in common with criterion-related validity analyses than it does with standard setting as a policy making. The policy decision is made when the criterion value defining adequate college performance is specified.

McLarty et al. (2013) proposed Evidence-Based Standard Setting (EBSS) as a general framework for using criterion-related evidence to set and validate performance standards defined in terms of outcome variables like college readiness. They suggest developing multiple lines of evidence (empirical and judgmental) relevant to the proposed performance standard, which is then presented to panelists in a standard-setting meeting in which the panelists set the cut score. The judgments made by the panel focus on weighing and combining the different kinds of empirical data, rather than on judgments about expected performance of marginal test takers relative to a performance-level description. In the example presented by McLarty et al. (2013), the test was a high school algebra test, and the primary outcome of interest was preparedness for a 1st-year credit-bearing college mathematics course; in estimating the cut score, they considered criterion-based results for community colleges, typical 4-year colleges, and for more selective colleges, and then had a panel set the cut score based on all of these results.

The issue to be addressed in this section is the potentially more difficult problem of validating standards in cases where the performance level is defined in terms of performance-level descriptors (like that reported earlier for the NAEP proficient level). These performance levels are not defined in terms of a specific outcome variable, but they can suggest strong expectations about some outcomes, which can be used develop upper or lower bounds for the cut score. These expectations do not need to provide estimates of the cut score; rather, the upper or lower bounds provide empirical challenges to the reasonableness of the cut score.

The aim is to determine if the standard satisfies the Goldilocks Criteria, which require that the standard not be too low (i.e., below reasonable lower bounds) or too high (i.e., above reasonable upper bounds). The exact placement of the standard