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.

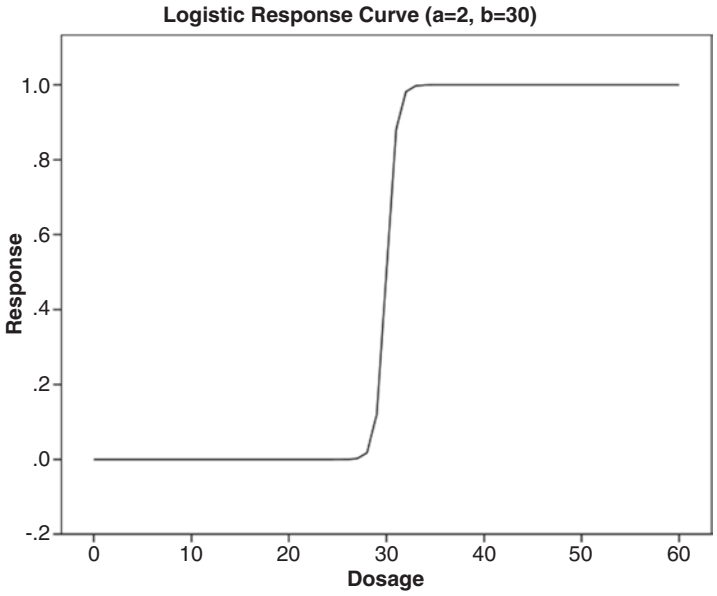


Fig. 2.1 Dose-response Curve with a Sharp Transition

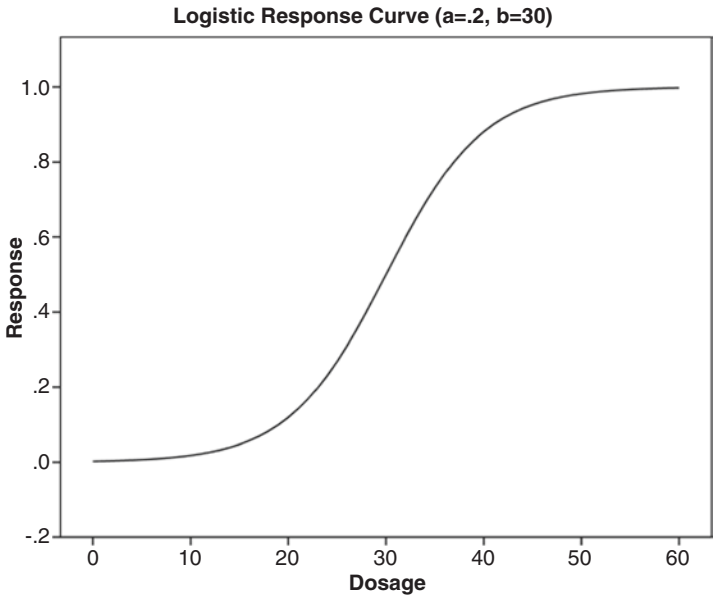


Fig. 2.2 Dose-response Curve with a well-defined Critical Range

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.

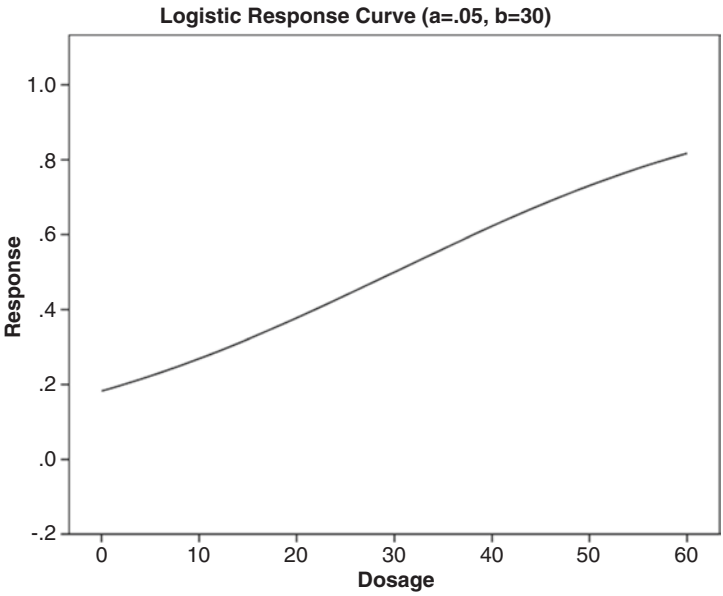


Fig. 2.3 Dose-response curve with a broad critical range

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

between the bounds would be a matter of judgment, and in that sense, arbitrary; however, to the extent that the bounds define a narrow interval, this residual arbitrariness would not necessarily be a problem. If the bounds are widely acceptable, they can provide serious empirical checks on the reasonableness of the standards. Standard setting is basically policy making, and policies are more-or-less reasonable rather than being right or wrong. The bounds provide a check on the reasonableness of the cut score by identifying the cut scores that would be considered unreasonable.

For this approach to work well, at least two conditions have to be met. First, the empirical relationship has to be well defined and fairly strong; the relationship does not have to consist of a precisely defined dosage-response curve, but the relationship between cut scores and outcomes should be clear and strong enough to support conclusions about bounds on cut scores from bounds on the outcomes. Second, the bounds on outcomes should be generally accepted by most or all stakeholders; for example, if we claim that cut scores above a proposed upper bound will produce serious negative outcomes, there needs to be agreement that the outcomes in question are negative and are serious enough to warrant adopting an upper bound to minimize the chance of their occurrence.

The intended goal to be achieved through standard setting tends to suggest lower bounds, and the potential side effects tend to suggest upper bounds on reasonable cut scores. In some cases, the needed empirical analyses may exist prior to the standard setting, and in some cases, it may be necessary to conduct empirical studies as part of the validation. It is certainly not necessary to evaluate all possible relationships; what is needed is enough strongly supported relationships to develop clear and persuasive bounds on the cut score.

First as noted above, some educational performance standards have been associated with preparedness for college. In particular, the proficient performance level on 12th grade NAEP has been associated with academic preparedness for college (Fields 2014). As indicated earlier, the NAEP performance levels have been defined in terms of performance-level descriptors, like the 12th-grade proficiency level descriptor presented earlier. Nevertheless, given the sophisticated academic content of the 12th-grade NAEP proficiency levels (see example given earlier), it certainly seems reasonable to associate this performance level with academic preparedness for college

Fields (2014) summarizes the results of more than 30 empirical studies of various kinds and uses these results to evaluate the reasonableness of the NAEP 12th-grade proficient level, as an indicator of college preparedness. A longitudinal study of students who had taken 12th-grade NAEP indicated that the average NAEP score of students who did not have to take any remedial courses and earned a B^- or better in 4-year colleges was just below the NAEP proficient level; this is a fairly high (but not unreasonable) bar for college preparedness, because the sample includes all students with averages over B^- (including those with A averages), and does not allow for any remediation.

Reasonable bounds depend on the intended use of the categorizations. If the intent is to identify test takers who are fully prepared for rigorous college work, the criterion of no remedial courses and B^- or better in a 4-year college could be taken

as a fairly high lower bound. If the intent is to identify test takers who should be awarded a high-school diploma, this criterion of no remedial courses and B⁻ or better in a 4-year college could be taken as an upper bound.

In testing programs involving multiple grade levels (e.g., NAEP, most state testing programs in the United States), it may be difficult to find external variables that can be used to suggest upper and lower bounds at some grade levels, but if such relationships can be identified at some levels (e.g., using indicators of college and work readiness at the12th-grade level), the methods described above for scaling across distinct but correlated score scales can be used to link bounds on one scale to create bounds at other scales (e.g., scales for other grade levels) or other testing programs (Beaton et al. 2012).

Potential negative side effects can be particularly relevant in setting upper bounds. For example, the location of a passing score can have a major impact on the incidence of adverse impact (Kane et al. 2006). If two groups of test takers have approximately normal score distributions with different means, a change in the passing score can increase or decrease adverse impact for the group with lower scores. If the passing score is near the middle of the score distribution for the lower scoring group (which is not uncommon), even a modest increase in the passing score can substantially increase the failure rate for this lower scoring group. Assuming that the passing score is in the lower end of the score distribution for the higher scoring group (where there are few scores), the increase in the passing score would have a relatively modest impact on the failure rate for this high-scoring group.

Discussions of cut scores in personnel selection have tended to give a lot of attention to adverse impact and have been concerned about the legal defensibility of standards. For example, in discussing physical ability tests for employment, Campion (1983) suggested that:

> The conceptual link between the job requirements and the cut-off scores chosen for the selection tests must be made explicit, and it must be documented and defensible. Physical abilities tests do have adverse impact against females; they probably will be legally challenged; and the cut-off scores determine the degree of adverse impact.

Given the impact of cut scores for high-stakes testing, and their potential for increasing or decreasing adverse impact, the rationale for the proposed cut score should be clear and persuasive. A cut score that is used to make consequential decisions needs to be justified.

A number of potentially negative outcomes of standards-based accountability programs in education (e.g., narrowing of the curriculum, teaching to the test, cheating) have been suggested, and data on the incidence of such behaviors could be useful in setting upper bounds. As the cut score goes up, such negative behaviors are likely to become more common. Establishing an empirical relationship between cut scores and undesirable systemic effects is likely to be difficult, but for purposes of developing upper bounds, the relationship does not need to be very precise (Tufts University 2015).

The choices made in setting upper and lower bounds will generally be complex and questionable; should a lower bound (or expectation) for college preparedness be associated with a first-year college average of B⁻ or C⁺ or C, and should an upper

bound be defined by a first-year average of B^+ or B? How much adverse impact is to be accepted in contexts where the stakes are high, and the proposed standard is highly subjective. Higher standards for licensure and certification can limit the availability of services and increase the cost of these services. Developing such Goldilocks-Criteria studies will not be easy, but they may give us a better idea of what we are doing from a policy point of view. In addition, the upper and lower bounds could provide the public with a better sense of what the performance standard means.

2.15 Summary

Standard setting occurs in contexts where it is necessary or useful to assign test takers to levels of performance. In education, it has typically involved the development of performance-level descriptors and corresponding cut scores by policy makers and standard-setting panels. This process is subjective and therefore not very transparent or precise. Reliance on empirical results in validating standards-based decisions can make the cut score less susceptible to charges of arbitrariness.

The choice of a performance standard and an associated cut score, like any policy decision, needs to be evaluated in terms of its outcomes, and the Goldilocks Criteria suggest that the standards need to be high enough to achieve the goals of the program, but not so high as to cause serious side effects. These criteria are consistent with those used to develop standards in health policy, and it provides a potentially useful model for using empirical results (e.g., dosage-response curves) in evaluating judgmental standard setting efforts. The question is not how accurate a policy is, but rather, how well it works, and how well standard setting works depends on its outcomes, positive and negative.

For the standards-based interpretation and uses to be valid, the inference from test scores to performance categories needs to be justified. The supporting evidence can be procedural, internal, and criterion-based. Standard setting tends to be least subjective/arbitrary when the standards are tied to a well-defined goal, and are most subjective/arbitrary when they are not tied to any specific external requirement. By identifying empirical bounds on the cut scores, which are generally accepted as being reasonable, we get a sense of how arbitrary the standards are, and by tightening the bounds, we can limit the degree of arbitrariness. Empirical evidence is especially important in high-stakes contexts, where the standards tend to be contentious.

The standards-based decisions are intended to promote certain outcomes, and are associated with potential side effects, and these relationships may be used to identify upper and lower bounds for an appropriate cut score. Lower bounds can be specified in terms of effective achievement of goals, and upper bounds can be specified so as to avoid or limit undesirable side effects. This approach does not eliminate subjectivity from standard-setting, but it can help to control it.

Standard setting involves the development of policy statements about how good is good enough. The results are arbitrary in the sense that they are judgmental, but they can be reasonable and well supported by data. In evaluating cut scores, the

question of validity is essentially the question of whether the decisions based on the cut scores are reasonable and broadly acceptable, are consistently applied, and have mostly positive consequences. Standard setting turns a general policy goal into a more explicit, detailed, and operational policy.

References

Beaton, A., Linn, R., & Bohrnstedt, G. (2012). *Alternative approaches to setting standards for the National Assessment of Educational Progress (NAEP)*. Sam Mateo: American Institutes for Research.

Campion, M. (1983). Personnel selection for physically demanding jobs: Review and recommendations. *Personnel Psychology, 36*, 527–550.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? New directions for testing and measurement: Measuring achievement over a decade, 5, 99–108.

Cronbach, L. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale: Erlbaum.

Fields, R. (2014). *Towards the National Assessment of Educational Progress (NAEP) as an indicator of academic preparedness for college and job training*. Washington, DC: National Assessment Governing Board.

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

Greenberg, D. (1998, June 11). Of Human Poundage. The Lancet, 352, n 9122, p. 158.

Hambleton, R., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport: Praeger Publishers.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73.

Kane, M., Crooks, T., & Cohen, A. (1997, March). *Justifying the passing scores for licensure and certification tests.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Kane, M., Mroch, A., Ripkey, D., & Case, S. (2006). *Impact of the increase in the passing score on the New York bar examination*. Madison: National Conference of Bar Examiners http://www.nybarexam.org/NCBEREP.htm.

Loomis, S. C., & Bourque, M. L. (2001). *National assessment of educational progress achievement levels, 1992–1998 for mathematics*. Washington, DC: National Assessment Governing Board.

McLarty, K., Way, W., Porter, A., Beimers, J., & Miles, J. (2013). Evidence-based standard setting: establishing a validity framework for cut scores. *Educational Researcher, 42*(2), 78–88.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan.

Porter, T. (1995). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton: Princeton University Press.

Shapiro, L. (1998, June 15). Fat, fatter: But who's counting? *Newsweek, 131*(24), 55.

Shepard, L. (1980). Standard setting, issues and methods. *Applied Psychological Measurement, 4*, 447–467.

Toulmin, S. (2001). *Return to reason*. Cambridge, MA: Harvard University Press.

Tufts University. (2015). How much exercise is enough? Two new studies seek the 'sweet spot' for activity and intensity. *Nutrition & Health Letter, 33*(5), 7.

Zieky, M. J., Perie, M., & Livingston, S. A. (2008). *Cut scores: A manual for setting standards of performance on educational and occupational tests*. Princeton: Educational Testing Service.

Chapter 3
Weaknesses of the Traditional View of Standard Setting and a Suggested Alternative

Mark Wilson and Maria Veronica Santelices

Abstract In this paper, we expand the traditional perspective on standard setting to include the necessary antecedents to a genuinely valid setting of standards, and use that conceptual framework to propose a new foundation for standard setting. These necessary antecedents include (a) the definition of an underlying variable on which the "standard" will be set in a way that is designed to be suitable for that standard setting, (b) the selection of a qualitatively definable point on that variable that corresponds to "enough" for the standard to be met, (c) the development of a suitable procedure ("test") and expression of its results in a suitable way to readily afford use in a standard setting procedure, and (d) the application of a suitable method for deciding the observable cut score that reflects attainment of the standard. From this new perspective, we critique two examples of the traditional approach, the "Modified Angoff" and the "Matrix method." We then describe an approach consistent with the more broadly-based foundation, centered on the Construct-Mapping line of thinking. We give an example of this in a unidimensional context. This approach is then generalized to address multidimensional constructs. We also illustrate a software application that has been developed to facilitate this process. We conclude by discussing some consequences of adopting the new approach, and survey needed next steps in research and development.

Keywords Standard setting • Construct mapping • Bear Assessment System (BAS)

M. Wilson (✉)
University of California, Berkeley, CA, USA
e-mail: markw@berkeley.edu

M.V. Santelices
Pontificia Universidad Católica de Chile, Santiago, Chile
e-mail: vsanteli@uc.cl

S. Blömeke, J.-E. Gustafsson (eds.), *Standard Setting in Education*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-50856-6_3

3.1 Introduction

The typical standard setting procedure, as it is currently implemented, takes as a starting point that a test has been developed and scaled, and then the standard-setting process begins. In a typical example of the standard setting process, a group of professionals (including, most likely, teachers in the relevant subject matter, policy-makers and/or administrators, and testing professionals) will engage in a discussion and decision-making process, as well as some analyses, that will result in the imposition of cut scores (one or more) on the test scale. Generally, the purpose is to divide the scale into segments according to a certain usage of the test, such as deciding between, for example, "Proficient" and "Basic." There are quite a few specific methods for doing so – see Cizek (2011) and Draney and Wilson (2011) for a survey of methods.

Perhaps the most well-established is the Modified Angoff Method (Angoff 1971; Livingston and Zieky 1982). This method is based on the idea of the "'borderline' test-taker...one whose knowledge and skills are on the borderline between the upper group and the lower group" (Livingston and Zieky 1982). Then "the judge considers each question as a whole and makes a judgment of the probability that a borderline test-taker would answer the question correctly" (Livingston and Zieky 1982). Then, the passing score is computed from the expected scores for the individual items. The logic of this procedure, which it shares to a greater and lesser degree with most of the others, is that (a) the test and the associated scale is accepted "as given," and (b) the standard setting proceeds from that point, relying on the expertise of the committee members in examining the individual items, to lead to appropriate and sound cut scores, via the nomination of probabilities of success on individual items.

It is our contention that this overall strategy is indeed, incomplete, and we will explicate the reasons for that view in the cases of the Modified Angoff and the Matrix method (Draney and Wilson 2011), in the sections below. In the immediate next segment of this section, we describe an alternative logic, one that we see as being more soundly based in the educational context of the standard setting. We see this as making the results of the test-development more useful to both the standard setters, and the teachers and other educational professionals who will use the results of the standard setting.

3.2 A New Approach to Standard Setting

The new explication of the logic that should be invoked when one carries out standard setting unfolds as follows, in four successive phases.

(1) In the first phase of the logic, one defines the outcome objectives for each discipline, which are commonly referred to as "standards."[1] Each standard reflects the tasks that students in a given grade should be able to achieve/perform.

[1] This label is confusing when it refers to just one aspect of "standard setting," but that is the typical usage.

(2) This initial set of standards then needs to be aggregated and in that aggregation it is important to decide qualitatively "what is enough." Do we expect success on *every* standard? Do we expect a *minimum* success, considered *enough*, on every standard? Do we expect enough success in a specific number of standards? Or on certain specific standards?
(3) The third phase involves the manifestation of student performance on the standards, therefore a test consisting of a certain number of items for each standard is constructed.
(4) Finally, the fourth phase deals with the question of which performances are acceptable. If students are scored on the test, the standard setters face an array of scores, with certain number of scored items per standard. Note that this is the same as the *typical* "Standard Setting" described above, which attempts to answer the question: how to decide what score represents "enough" of the subject-matter?

One might well ask: "Why should we consider these stages part of standard setting procedure and not part of, say, 'Test development'." The answer is that, indeed, what this logic is based on, is the observation that when one is developing a test that will be used for standard setting, then one should incorporate this planned usage at every stage of test development, right from the very start. This is consistent with the current view about validity (Kane 2013), under which assessments in general, and items in particular, are developed and validated with a particular use in mind. By following this approach one would be developing tests that can serve as basis for *both* large-scale assessments for use in standard setting as well as assessments that can be used for formative purposes within the classroom. The common framework of test construction applied in this case is an important safeguard that the use of standardized tests and standard-setting methods does not undercut classroom instruction.

Thus the standard setting procedure should include the definition of the outcome objectives and a way to define what is qualitatively "enough" of the standards through the definition of an acceptable performance exhibited in a manifest way, usually in the guise of a test (Wilson and Draney 2002). However, most "methods" of standard setting *start* at the very end of the process, phase 4 above, defining just the "acceptable" result on the scale. Traditional methods go on to develop in great detail the technical aspects of that phase, and to exhaustively test the methods of implementation to this phase. In our view, standard setting should be seen as much more than just defining the acceptable performance—it should explicitly integrate all of the four phases described above.

In this chapter we will next examine how the problems hinted at above are present in two traditional standard-setting methods: the Angoff Method (Livingston and Zieky 1982) and the Matrix Method (Draney and Wilson 2011). Subsequently, we will present an alternative method that is an attempt to respond to the problem raised above and integrate the setting of standards in the construction of the assessment. We then illustrate a software application that is aimed at facilitating a committee's work in carrying out the final phase of the standard setting, and we conclude by discussing issues and considering next phases.

3.3 Two Traditional Standard Setting Procedures

3.3.1 The Modified Angoff Method

A very brief description of the Modified Angoff method was given in the quotation above. A more complete description follows. The process begins by selecting a Committee of experts to serve as judges (Angoff 1971). This might be the examination committee for a professional certification exam, or it could be a group assembled specifically for this procedure. Members of the Committee who are not familiar with the details of the test might be asked to take it themselves. The Committee develops a (verbal) definition of a hypothetical "minimally competent examinee" (MCE) on that test. Then each Committee member is shown each item on the test, in turn, and asked to decide how many of a group of 100 MCEs are likely to answer that item correctly (i.e., there is an implicit assumption here that the items are all dichotomous).

These ratings are discussed by the Committee and the members are allowed to change their ratings, based on the discussion. This may be repeated several times. After all the items have been judged by all the Committee members, the numbers are averaged across all items and all judges are to determine the cut score. Additionally, a 95% confidence interval (CI) can be determined using the standard error of the ratings and the inter-rater reliability of the judges. The Committee members may be given additional information, where it is available, such as the actual observed frequency of correct responses on the items for a selected sample of students. Once the cut score is set, the standard then is maintained in other administrations and forms through statistical equating.

Now, let us compare this narrative with the logic described above. The Modified Angoff method assumes that the first phase of the process described in the preceding section, the definition of the standards, has been successfully completed, and is clear to the Committee members. It also assumes that the definition of what is (qualitatively) "enough" of the standards, has also been resolved and that the judges can apply that understanding to each item on the test. Both the first and the second phases, under this method, are assumed to have achieved a very high quality product. The same applies to the third phase, since it assumes that the test has been developed with the explicit goal of revealing how student performance compares to the standards. Finally, in the fourth stage, the Angoff method assumes (a) that the judges know the definition of what the standards are and (b) that "taking the average" of the probabilities (expectations) is the best way to summarize their judgments (this latter also constitutes an index of "what is enough").

In the development of this methodology, all of these assumptions were left unexamined, and equally, they are left unexamined in the instances of its application. Were they to be regularly examined, one might be convinced that a new application would not need such a deep examination, but given that it is never the case that they are examined, one must have skepticism that they are valid assumptions.

3.3.2 The "Matrix" Method

In order to discuss this method, we will refer to a particular examination program: The *Golden State Examination* (GSE) program (Draney and Wilson 2011). The Golden State Examination program in the state of California consisted of a set of high school honors examinations. These were end-of-course examinations in a number of subjects, including mathematics (Algebra, Geometry, High School Mathematics), language (Reading and Literature, Written Composition, Spanish Language), science (Physics, Chemistry, Biology, Coordinated Science), social science (US History, Government and Civics, Economics). Each GSE examination consisted of a set of multiple choice items and at least one written response item. Scores were categorized into one of six performance levels. The top three levels (4, 5, and 6) were considered "honors" levels (School Recognition, Honors, and High Honors, respectively). If a student achieves one of these honors levels on six exams (including US History, Reading and Literature or Written Composition, a mathematics exam, and a science exam), the student was also eligible for a State honors diploma and financial support for college.

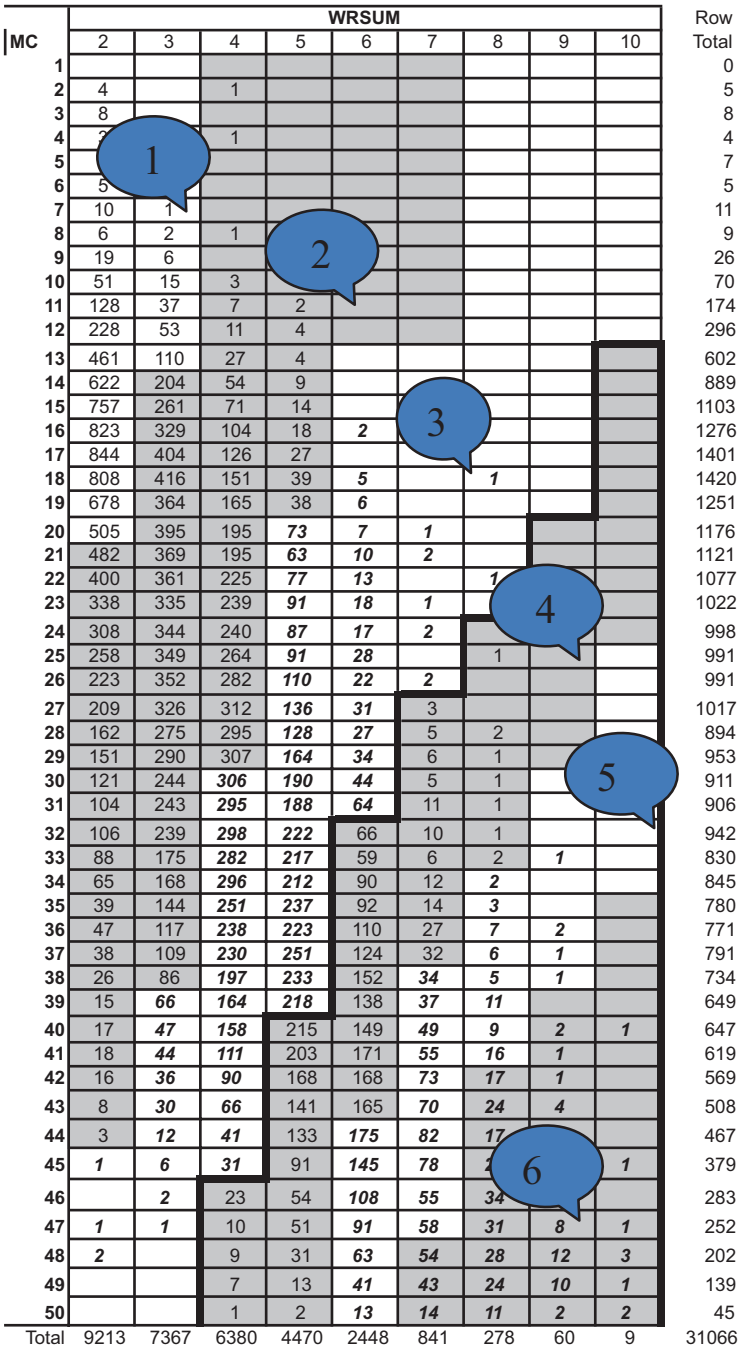
The Matrix Method is, as for the Modified Angoff method, based on the judgment of a Committee (usually seven to ten people), that might be composed of teachers and other professionals involved in the testing process. Correspondence between total scores on the two item types for each of the performance levels will be determined by the standard setting committee as follows. Committees meet and review all of the specific test materials, including performance level descriptions (PLDs; see Fig. 3.1). Once they are familiar with the material, they are then shown a blank version of a two-way matrix, known as a *mapping matrix*. See Fig. 3.2 for an example of such a matrix (although this example has been "filled in"). In this Figure, rows are defined by all possible scores on the multiple choice section and the columns are defined by all possible scores on the written response section of the examination. In the headings for the columns, "MC" indicates the total score on the multiple choice items, and "WRSUM" indicates the total score on the written response items. In addition, the numbers in "cartoon conversation" balloons indicate the final Committee decisions about the levels, and the numbers inside the matrix cells show the number of students who actually scored in each cell. The successive greyed-out and white areas indicate the extent of each score. The Committee is then asked, for each possible score combination (i.e., each cell of the matrix), to determine the most appropriate performance level. The Committee members may also be shown the counts in each cell of the matrix (as shown in Fig. 3.2). At the end of the discussion, if consensus has not been reached, the cutoffs are decided by majority vote.

Just as for the Modified Angoff Method, the Matrix Method assumes that the Standards have been developed and that they are related to the performance levels but, in fact, for the GSE, this was not usually the case—there was no linkage from specific items to the PLDs. In addition, the Matrix Method assumes that the judges have decided and know what is (qualitatively) enough of the standards for each

Level 6	Student work demonstrates evidence of rigorous and in-depth understanding of mathematical ideas; the work: • Is consistently correct and complete, and shows thorough understanding of mathematical content and concepts • Communicates clear and logical explanations of solutions to problems that are fully supported by mathematical evidence • Shows problem-solving skills that include appropriate generalizations, connections, and extensions of mathematical concepts • Includes effective use of mathematical language, diagrams, graphs, and/or pictures • Shows skillful and accurate use of mathematical tools and procedures, often with multiple and/or unique approaches
Level 5	Student work demonstrates evidence of solid and full understanding of mathematical ideas; the work: • Is essentially correct and complete, although it may contain minor flaws • Communicates explanations of solutions that are supported by mathematical evidence • Shows problem-solving skills that include connections and extensions of mathematical concepts • Shows appropriate use of mathematical language, diagrams, graphs, and/or pictures • Includes accurate use of mathematical tools and procedures
Level 4	Student work demonstrates evidence of substantial understanding of mathematical ideas; the work: • Is usually correct and complete, although it may contain flaws • Communicates explanations of solutions that are supported by mathematical evidence for most tasks • May contain evidence of problem solving without connecting or extending mathematical concepts • Includes frequent use of mathematical language, diagrams, graphs, and/or pictures • Usually shows evidence of appropriate use of mathematical tools and procedures
Level 3	Student work demonstrates evidence of a basic understanding of mathematical ideas; the work: • Is sometimes correct; however, it may lack either depth across the mathematical content areas or may show gaps in understanding of some concepts • Communicates explanations of solutions that are supported by mathematical evidence for some tasks, but explanations are very weak r missing for other tasks • May show ineffective or inconsistent problem solving • Shows some evidence of use of mathematical language, diagrams, graphs, and/or pictures • Shows some appropriate use of mathematical tools and/or procedures for some tasks
Level 2	Student work demonstrates evidence of limited understanding of mathematical ideas; the work: • Shows little evidence of correct solutions and is incomplete • Provides limited explanations of solutions that are not supported by mathematical evidence • Shows limited evidence of problem-solving, arithmetic computations may be correct but unrelated to the problem • Shows limited evidence of use of appropriate mathematical language, diagrams, graphs, and/or pictures • Includes limited or inappropriate use of mathematical tools and procedures
Level 1	Student work demonstrates little or no evidence of understanding of mathematical ideas; the work: • Is rarely correct and has major mathematical errors

Fig. 3.1 Example of performance level descriptions for the Golden State Examination program. High-school algebra

performance level. Now, this could actually be true if there had been a relationship established between the standards and the PLDs, however, this is not usually the case, and again, was not the case for the GSE. The Matrix Method also assumes that the test has been developed with the PLDs in mind, which, again, could be true of if there was a relationship between the standards and the performance level, and this had been used to develop the items, and yet again, this was not the case for the GSE. Finally, the Matrix Method assumes that judges can create or decide on the cut scores that define acceptable performance. Again, just as for the Modified Angoff Method, in the development of the Matrix methodology, all four assumptions were left unexamined, and equally, they are left unexamined in the instances of its application. Were they to be regularly examined, one might be convinced that a new application would not need such a deep examination, but given that it is not often the case that they are examined, one must have skepticism that they are valid assumptions.



MC	WRSUM 2	3	4	5	6	7	8	9	10	Row Total
1										0
2	4		1							5
3	8									8
4	3		1							4
5										7
6	5									5
7	10	1								11
8	6	2	1							9
9	19	6								26
10	51	15	3							70
11	128	37	7	2						174
12	228	53	11	4						296
13	461	110	27	4						602
14	622	204	54	9						889
15	757	261	71	14						1103
16	823	329	104	18	***2***					1276
17	844	404	126	27						1401
18	808	416	151	39	***5***		***1***			1420
19	678	364	165	38	***6***					1251
20	505	395	195	***73***	***7***	***1***				1176
21	482	369	195	***63***	***10***	***2***				1121
22	400	361	225	***77***	***13***		***1***			1077
23	338	335	239	***91***	***18***	***1***				1022
24	308	344	240	***87***	***17***	***2***				998
25	258	349	264	***91***	***28***		1			991
26	223	352	282	***110***	***22***	***2***				991
27	209	326	312	***136***	***31***	3				1017
28	162	275	295	***128***	***27***	5	2			894
29	151	290	307	***164***	***34***	6	1			953
30	121	244	***306***	***190***	***44***	5	1			911
31	104	243	***295***	***188***	***64***	11	1			906
32	106	239	***298***	***222***	66	10	1			942
33	88	175	***282***	***217***	59	6	2	***1***		830
34	65	168	***296***	***212***	90	12	***2***			845
35	39	144	***251***	***237***	92	14	***3***			780
36	47	117	***238***	***223***	110	27	***7***	***2***		771
37	38	109	***230***	***251***	124	32	***6***	***1***		791
38	26	86	***197***	***233***	152	***34***	***5***	***1***		734
39	15	***66***	***164***	***218***	138	***37***	***11***			649
40	17	***47***	***158***	215	149	***49***	***9***	***2***	***1***	647
41	18	***44***	***111***	203	171	***55***	***16***	***1***		619
42	16	***36***	***90***	168	168	***73***	***17***	***1***		569
43	8	***30***	***66***	141	165	***70***	***24***	***4***		508
44	3	***12***	***41***	133	***175***	***82***	***17***			467
45	***1***	***6***	***31***	91	***145***	***78***	***1***		***1***	379
46		***2***	23	54	***108***	***55***	***34***			283
47	***1***	***1***	10	51	***91***	***58***	***31***	***8***	***1***	252
48	***2***		9	31	***63***	***54***	***28***	***12***	***3***	202
49			7	13	***41***	***43***	***24***	***10***	***1***	139
50			1	2	***13***	***14***	***11***	***2***	***2***	45
Total	9213	7367	6380	4470	2448	841	278	60	9	31066

Fig. 3.2 An example mapping matrix (see text for explanation)

3.4 An Alternative: Standard Setting Using Construct Mapping

Following the logic above, and reading the critiques of the two exemplary typical standard setting methods, one must feel unsatisfied, and surely one would feel compelled to design a more complete and comprehensive approach. The following section describes one possible such approach, the Construct Mapping procedure (Wilson and Draney 2002) for standard setting which was developed as an extension of the Bear Assessment System (BAS, see Appendix: Wilson 2004). This method uses a key tool, the "Wright Map," which is described below, to making an explicit, rather than implicit or inferred, selection of cut scores on a latent trait scale. Note that there are other approaches that also integrate or could integrate the standard setting endeavor into the item and assessment development process, such as those presented by Mislevy et al. (2003), Wang (2003), and Reckase (1998). We are not making any claims here for advantage over these, but just using Construct Mapping as an example (albeit one we like). We will use the example of the Assessing Data Modeling and Statistical Reasoning (ADM; Lehrer et al. 2014) project to illustrate Construct Mapping.

As a reminder, the four phases of the process described in the introductory section of this chapter, are (briefly):

(1) A way to define the outcome objectives: The "Standards";
(2) A way to decide what is (qualitatively) "enough" of the standards;
(3) A way to make observable student performance that correspond to the standards; and
(4) A way to decide which performances are acceptable.

Brief Description of the ADM Project The ADM Project was led by Rich Lehrer and Leona Schauble, from Vanderbilt University, and Mark Wilson from University of California Berkeley. During the project, researchers developed an assessment system for a "Statistical Modeling" curriculum for middle school. The project was a multi-year, multidisciplinary collaboration of teachers, science and assessment experts. The goal was to design "a developmental perspective on learning," based on a learning progression with seven relational construct maps. ADM used a "reformed curriculum" approach, which centers on two linked aspects of classroom activities. The first is a whole-class conjecture approach to class-level instruction—in this approach, the teacher seeks to engage class-sized groups of students in discussions that are based on science conjectures based in their own experiences. For example, if students are exploring the life-forms associated with a small pond, then a conjecture might be that the recently-observed increase in a certain type of insect was associated with the increases or decreases in certain plants or other animals in or around the pond: students would be encouraged to make hypotheses about this, and defend them, in an ongoing class-discussion. The second element is small-group-based projects, where students might, for example, follow up on the class discussion by systematically observing the prevalence of the animals and plants that they hypothesized were linked, and report back to the class on what they found.

- **CoS4 -** Investigate and anticipate qualities of a sampling distribution.
- **CoS3 -** Consider statistics as measures of qualities of a sample distribution.
- **CoS2 -** Calculate statistics.
- **CoS1 -** Describe qualities of distribution informally.

Fig. 3.3 The construct map for *Conceptions of Statistics*, ADM project

- **CoS3F** Choose/Evaluate statistic by considering qualities of one or more samples.
- **CoS3E** Predict the effect on a statistic of a change in the process generating the sample.
- **CoS3D** Predict how a statistic is affected by changes in its components or otherwise demonstrate knowledge of relations among components.
- **CoS3C** Generalize the use of a statistic beyond its original context of application or invention.
- **CoS3B** Invent a sharable (replicable) measurement process to quantify a quality of the sample.
- **CoS3A** Invent an idiosyncratic measurement process to quantify a quality of the sample based on tacit knowledge that others may not share.

Fig. 3.4 Detailed view of **CoS3** (see Fig. 3.3), from the construct map for *Conceptions of Statistics*, ADM project

We will look at just one of the seven construct maps as an example: *Conception of Statistics* (CoS). Figure 3.3 shows the construct map for this construct, starting off at the lowest level (CoS1) with students who can attempt informal descriptions of distributions—"its lumpy", etc. At the next level (CoS2), students can calculate standard statistics that summarize features of a distribution (appropriate for 5–6 grade students) such as the mean, median and mode, but also measures of spread. The next level (CoS3) consists of students who can relate these statistics to the shape and nature of the distribution. The order of these two might be reversed in other curricula, but the situation in the schools where the assessments are to be used is that all students are taught to calculate these standard statistics via an algorithmic memorization approach, and so, the students had to be later taught what these statistics actually mean.

The highest level that was considered appropriate for the 5th and 6th grade students was CoS4 where the students are starting to learn about the "sampling distribution" of the statistics themselves, which one would expect, leads eventually to initial appreciation of the phenomenon of the reduction in the standard error of the mean as the sample size increases. This is way beyond the usual upper limit expected for 5th and 6th grade students, but was nonetheless reached by some students in the ADM program. Each of these levels can be considered a "standard" at a fairly coarse grain-size (and, in fact, a level that is too course for use as an instructional development guide—but, more on this in a minute). In fact, the construct maps do not stop at this grain size. In the ADM materials, each level is split up into a set of sub-categories that are at a more appropriate grain-size for the design of instruction, and the relevant sub-categories for CoS3 are shown in Fig. 3.4.

This way of expressing standards is quite different from the traditional approach, where individual standards are expressed in a stand-alone way. In this approach to standards definition, each standard is defined in terms of being a part of a single strand of a learning progression, and its place is defined in relation to the place of other standards that also are related to that single strand (or dimension) of the learning progression. For instance, consider the following standard shown in Fig. 3.4—CoS3E**:** "Predict the effect on a statistic of a change in the process generating the sample." This is a very specific standard that would guide a teacher to develop instructional materials and events where the student was learning about the characteristics of how a statistic would tend to change as the sample that it was being applied to was undergoing a change. It is at quite a detailed level, and hence could be considered an "instructional" standard. There are still many aspects of this standard that need to be specified in a given situation—what is the range of "statistics" that might be considered? how large might the sample be?, how complex might the change in the sample be?—these would need to be established as a part of the larger framing of the curriculum.

This illustrates the BAS approach to standards development (the first requirement above)—each standard gains in meaning and interpretability through its relationship with other standards in the learning progression. Thinking about assessment, with this finer degree of grain-size (such as CoS3E) would be most useful to daily and weekly design of instruction, for a larger grain-size tracking student progress across semester, the coarser grain-size shown in Fig. 3.3 would likely be more useful. This then provides the clue to how the BAS approach also addresses the second requirement—the decision about what is "enough" of the standards. This can be judged by using the construct map as the initial tool, using either the coarser levels (such as CoS3, etc.) or the finer levels (such as CoS3E, etc.). The construct map is the key element in connecting the construction of large-scale assessments, using the coarser levels, and formative assessments, using finer levels.

More material and information will be needed to complete the judgment of what is "enough" of the standards and that is what we go to now. Going further around the cycle of BAS (see Fig. 3.8), we can now see that the Construct map illustrated in Figs. 3.3 and 3.4 provides a way to set some design features for items that will be useful as assessments of *Conception of Statistics*. The Item Design is the BAS building block that items can generate student performances that relate to the standards. The principal design criterion is that such items will need to be ones that will generate student responses that will be likely to reveal their thinking in terms of the levels and sub-categories of the Construct map. For example, for CoS3E, one would try to develop items that engaged the students in thinking about how a changing sample would likely affect one of the statistics that they are learning about.

One example statistic would be the *sample mean*, and one possible change in the sample would be the addition of one extra case. A minute's reflection (for an adult) reveals that when the extra case is larger than the current mean, the mean would go up, and the reverse when the case was smaller than the mean. This leaves the interesting case where the extra case is in fact equal to the current mean. In fact, this is

the motivation behind the example item shown in Fig. 3.5. Looking at this item, you can see that, in order to efficiently solve this problem, you need to conceptualize the mean as the point that "balances" the effect of the case on the sample mean (as described above). Some students just don't get this at all, some simply grasp it intuitively. A third group, who are smart about math but just don't "get" this idea, actually experiment with different guesses, and record them, and some of those will actually figure it out, by using this "empirical" technique, right there in front of you! This illustrates the way that following the BAS can provide an approach to the third requirement above—a way to make manifest student performance on the standards—at least for each item, one at a time.

However, this is not the end of the story, as one needs to go beyond a one-item-at-a-time perspective to see how one can apply this logic to a test as a whole. For this, one needs to go to the next step around the BAS cycle (see Fig. 3.8)—to the Wright Map. The Wright Map is a tool that addresses the fourth requirement—the question of how to decide which performances are acceptable. The Wright Map does so in two sequential steps. First, we need to establish "bands" among the items that correspond to the levels of the Construct map—in this case the levels of the CoS Construct map shown in Fig. 3.3.

To do this, we first gather empirical data from having students take the test that has been developed according to the principles described above. Then we use that data to examine the order of difficulty of the items, and compare that to the expected order, given by the fact that the items were designed to reflect the order of levels in the construct map.

Students received their final grades in Science today. In addition to giving each student their grade, the teacher also told the class about the overall class average. When the teacher finished grading Mina's work and added her final grade into the overall class average, the overall class average stayed the same. What could Mina's final grade have been? (Show your work).

Student	Final grades
Robyn	10
Jake	9
Calvin	6
Sasha	7
Mike	8
Lori	8

Fig. 3.5 Items design: Open assessment prompt

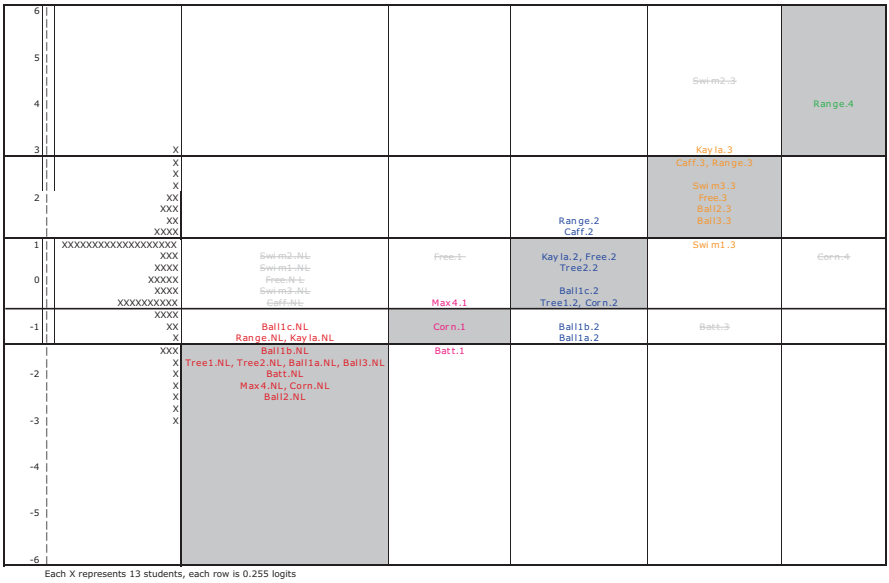


Fig. 3.6 An example of BAS banding

For example, the item in Fig. 3.5 was designed to match category 3E of CoS, and hence, to be one of the set of items matched to Level 3 of CoS.[2] Now, what happens whenever you take designed items and first investigate empirically whether they do or do not match the order expected, is that some do indeed match, some match fairly well, and some match quite poorly. This is illustrated in Fig. 3.6, which shows the initial results for the items that were developed for the CoS construct. In this Figure, note that the horizontal axis shows the expected level of each item (or item score in the case of a polytomous item), and the vertical axis shows the estimated locations of the students taking the test. The number of students at each estimated point is displayed on the left side as a "sideways" histogram of "x"s. The entries in the right-hand columns indicate the location of the item difficulties for each item, using a brief label for each item.[3] What we observe here immediately is that the locations of these difficulties do not occur in a neat step-like fashion, as one might expect if the item design was perfect. Now, we do expect some imperfections, but indeed we see that, in this initial data analysis, there are some item difficulties that are located way out of the expected range. For example, note that one item ("Corn-4") we had planned to be at level 5 has actually come up below most of the items we expected to be in level 4.

[2] Note that, if one had a polytomous item, say with 3 ordered response categories, one might expect that this would match, say two levels of a construct map, though they might not be consecutive.

[3] Note that for polytomous items, the labels actually show the difficulties of the respective scores—hence "Range-2" is the second threshold difficulty for the item "Range."

One needs to use this information to develop a more refined definition of the construct, the construct levels, and the specific design features of the items, in order to come up with a revised Wright Map including only the items that follow the revised specifications. This process should then be followed by a new data collection, to check that these findings were not affected by change, and, if one has the resources, to develop revised items that do indeed fit into the specifications. The results of this editing process are shown in Fig. 3.6. In this Figure, the bands are indicated by the dark grey boxes, and the items at each level that were excluded are the greyed out ones. As can be seen, in several cases, the items extend somewhat beyond the range of the bands. Then, with these bands, one can make criterion-referenced decisions (i.e., using the criteria of the construct map) to set cut scores for different levels. So, for example, if one wanted to create cut scores for each of the CoS levels, that can now be accomplished using the bands shown on Fig. 3.6.

3.5 Software to Help Construct Mapping

A software tool (*ConstructMap*; Hoskens and Wilson 1999) has been developed in order to dynamically display the item map, and provide the Committee with feedback about the consequences of setting the cut score at any point along the scale. The aim of this approach is to give committee members information that will help them balance information across different aspects of the test. The information could include dimensions or sub-components of content, or item-types. For example, we might use, instead of different aspects of content, different items types, such as multiple choice items and partial credit open-ended items. The relevant information is provided on a Wright map, which allows committee members to relate score levels to item locations and to indicate what the student's entire *response vector* tells us about what a student knows and can do.

We will illustrate this procedure using a somewhat simpler context than the seven-dimensional ADM context, described above. We will also show a somewhat different context, where the dimensions are not different content constructs (as they are for the ADM situation), but are different types of items, multiple choice and written response (i.e., having polytomous responses). We hope that this illustrates the flexibility of the approach.

For any chosen proficiency level, *ConstructMap* can show the probability of passing each multiple-choice item, the probability of attaining every level on the written response items, the expected total score on multiple-choice section or the expected score on each written response item. The information is presented using a Wright Map on which item types are scaled together to estimate the best-fitting composite, according to pre-determined item-weights (which is a substantive decision). The calibration is then used to create a map of all item/level locations. The Committee, through a consensus-building process, chooses cut points between performance levels on the map. The committee members are provided with copies of all exam materials, and a description of the performance levels, just as in the

Persons			GSE	Items			
FR	Percentile	Histogram	Scale	MC	WR 1	WR 2	
0	100.0		740		1.4		
0	100.0		735				
0	100.0		730				
42	100.0	<	725				
0	99.9		720			2.4	
0	99.9		715				6
0	99.9		710				
0	99.9		705				
0	99.9		700				
0	99.9		695				
0	99.9		690				
41	99.9	<	685				
63	99.7	<	680				
99	99.5	<	675				
0	99.2		670				
83	99.2	<	665		1.3		
23	98.9	<	660				
19	98.9	<	655				
192	98.8	X	650				
66	98.2	<	645			2.3	
89	98.0	<	640				
256	97.7	XX	635				
182	96.9	X	630				5
242	96.3	XX	625				
224	95.5	XX	620				
184	94.8	X	615				
203	94.2	XX	610				
266	93.5	XX	605				
276	92.7	XX	600				
308	91.8	XXX	595				
327	90.8	XXX	590				
413	89.7	XXXX	585		1.2		
433	88.4	XXXX	580				
301	87.0	XXX	575				
581	86.1	XXXXX	570	8			
475	84.2	XXXX	565				
598	82.7	XXXXX	560				4
461	80.7	XXXX	555				
661	79.2	XXXXXX	550				
596	77.1	XXXXX	545				
491	75.2	XXXX	540	21			
655	73.6	XXXXXX	535			2.2	
916	71.5	XXXXXXXXX	530	52			
459	68.6	XXXX	525	38			
744	67.1	XXXXXXX	520	43 46			
857	64.7	XXXXXXXX	515	35			
996	61.9	XXXXXXXXX	510	12 22 24 41 47			
899	58.7	XXXXXXXX	505	37			
661	55.8	XXXXXX	500	10 20 29 36 45 50	1.1		
853	53.7	XXXXXXXX	495	26 30 40 42			3
1035	51.0	XXXXXXXXX	490	5 14 25 48			
675	47.6	XXXXXX	485	23 33 34 44			
837	45.5	XXXXXXXX	480	39			
975	42.8	XXXXXXXXX	475	28			
967	39.6	XXXXXXXXX	470	51			
987	36.5	XXXXXXXXX	465			2.1	
1316	33.3	XXXXXXXXXXXX	460	27			
746	29.1	XXXXXXX	455				
936	26.7	XXXXXXXXX	450	7 11 49			
894	23.7	XXXXXXXX	445	15 16 19			
1032	20.8	XXXXXXXXXX	440				
755	17.5	XXXXXXX	435	6 9 13			
491	15.1	XXXX	430	17 31 32			2
656	13.5	XXXXXX	425				
701	11.4	XXXXXXX	420				
646	9.1	XXXXXX	415	4			
491	7.0	XXXX	410	3			

Fig. 3.7 A screenshot from *Construct Map*

traditional methods. They work with the software, which provides detailed information about all of the test items, based on the calibration.

A screen-shot from *ConstructMap* is provided in Fig. 3.7. In this Figure, the same sort of layout is used as for the other Wright Maps above, with the student distribution of the left, and the items on the right. There are more columns however. From the left, the columns are (a) the frequency of students at each location (FR), (b) the percentile that this represents, (c) the student histogram, (d) the scaled score (called "GSE" here), (e) the locations of the multiple choice levels, (f) the locations

of the item thresholds for the first open-ended item (WR1), (g) the locations of the item thresholds for the second open-ended item (WR2), and (h) the bands that were eventually developed for this example.

The committee members are given time to work with the program, and see what performance at various levels is like. In addition, they are able to choose various relative weights for the different item types, and see what effect that has on the map (i.e., the cut scores will move around, depending on the weights). They are led through a series of exercises to see the full effect of changing weights and choosing performance levels. Once they have worked extensively with the materials, they choose cut points between each of the performance levels, discussing their choices while led by a Committee Leader, and ideally coming to consensus about the location of each cut point. If consensus cannot be reached, the final cut points are chosen by majority vote, again as in the traditional method.

The software needs further development—in particular, to improve the link between what a student knows (in terms of the levels or bands) and can do (in terms of item responses) need to be developed. On occasion, committee members tend to justify their choice of cut scores in terms of the number of items that a student at a particular level would get right, rather than in terms of what that meant about their knowledge and skill in the subject at hand. In the future we intend to add a number of features to the *Construct Map* software, including actual (scanned in) examples of student work representing various levels, accessible by clicking, to improve this.

3.6 Summary and Conclusions

This chapter has shown how Standard-setting *must* be seen as more than a mere "technical exercise". Standard setting involves much prior work, both substantive and technical, including (a) How to develop standards that are "ready" for standard setting, (b) How to develop items that support that process, and (c) How to decide which student performances on the test are "enough". Because of these three conditions, standard setting requires an *overarching framework* that is based on coherence between the construct being measured, the items that are used to prompt student responses, the scheme used to score the responses, and the analysis method. Standard setting should not be left to the very end of the assessment development, but - on the contrary - should be part of the assessment development from the very beginning in order to ensure its validity.

We have discussed how some of the traditional standard setting methods fall short from this perspective and have offered an alternative approach that does attempt to address this issue. Standard setting is a complex problem, and hence one should not expect an easy solution. Solutions are simpler for single-dimension constructs, but more complex for higher-dimension constructs, or for tests that have different types of items. This alternative method, Construct Mapping, capitalizes on some of the advantages of item response modeling, without losing the essential element of human judgment. This approach also allows year-to-year equating through,

say multiple-choice items that are repeated from one year to the next (or even written response items, so long as they are not too memorable). It also allows a check on the quality of the year-to-year link. In addition, this method allows committee members to use the Wright Map as indicating a "probability profile" of what a student at a given level knows and can do.

The Construct Mapping procedure is a sound and valid general framework for standard setting because it incorporates the final use of the test (i.e., standard setting) from the very beginning of and all throughout test development. However, successful standard setting will always still require rich materials to help the Committee to understand the meaning embedded in the results, and generous training time, and meeting time, to accomplish the difficult task of setting sound standards. Year to year consistency for performance levels tends to be an issue, especially since written response items tend to vary quite a bit in overall difficulty. This can also be due to a variety of factors that influence the judgment of a particular committee in a particular year: for example, a particularly strong committee member may influence the others.

As mentioned earlier, we have found that committee members tend to focus on total scores when setting standards, as opposed to models of what a student knows and can do which should really be the focus of the discussion. Policy-makers and administrators need the results of standard-setting for decision making based on large-scale tests. Most teachers do not need them: Teachers need good *formative* assessments, and the positive effects of good formative assessment on student learning is well-documented (e.g., Black and Wiliam 1998). Thus, a major requirement of standardized tests and standard-setting methods is that they do not *undercut* classroom instruction. Thus the approach described above has the additional virtue that it provides a common basis for good *large-scale test construction* (i.e., for use in standard setting) as well as good *formative assessment*.

Appendix: The BEAR Assessment System

The BEAR Assessment System (Wilson 2004) consists of interrelated components (see Fig. 3.8), called *building blocks*, that are used to design measuring instruments and which are congruent with recent efforts to reform measurement in the domain of educational assessments (National Research Council 2001). The first building

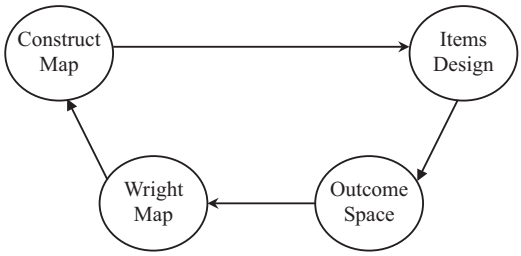


Fig. 3.8 The Bear Assessment System (BAS)

block is the *construct map*, which seeks to describe the variable being measured, from one extreme (say, low) to the other (say, high), and which is delineated by qualitatively-distinct levels. This is then used to develop an *items design*, which is the generic term for methods to generate responses from the respondents. These responses are then coded and valued using an *outcome space*. The resulting codes are analyzed using a *measurement model*, which is chosen to allow the analysis results to be related back to the construct map. In its development phase, these building blocks form a cycle of improvement for the measuring instrument. The building blocks enable and enhance the interpretation of the measures.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council of Education.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139–148.

Cizek, G. (2011). *Setting performance standards: Foundations, methods, and innovations*. New York: Routledge.

Draney, K., & Wilson, M. (2011). Selecting cut scores with a composite of item types: The construct mapping procedure. *Journal of Applied Measurement, 12*(3), 298–309.

Hoskens, M., & Wilson, M. (1999). *ConstructMap [Computer program]*. Berkeley: Berkeley Evaluation and Assessment Research Center.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Lehrer, R., Kim, M.-J., Ayers, E., & Wilson, M. (2014). Toward establishing a learning progression to support the development of statistical reasoning. In A. Maloney, J. Confrey, & K. Nguyen (Eds.), *Learning over time: Learning trajectories in mathematics education* (pp. 31–60). Charlotte: Information Age Publishers.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton: Educational Testing Service.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–67.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment* (Committee on the Foundations of Assessment. J. Pellegrino, N. Chudowsky, R. Glaser (Eds.), Division on Behavioral and Social Sciences and Education). Washington, DC: National Academy Press.

Reckase, M. D. (1998). *Analysis of methods for collecting test-based judgments*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Wang, N. (2003). Use of the Rasch IRT model in standard setting: An item-mapping approach. *Journal of Educational Measurement, 40*, 231–253.

Wilson, M. (2004). *Constructing measures: An item response modeling approach*. Mahwah/New York: Erlbaum/Taylor and Francis.

Wilson, M., & Draney, K. (2002). A technique for setting standards and maintaining them over time. In S. Nishisato, Y. Baba, H. Bozdogan, K. Kanefugi (Eds.), *Measurement and multivariate analysis* (Proceedings of the International Conference on Measurement and Multivariate Analysis, Banff, Canada, May 12–14, 2000) (pp. 325–332). Tokyo: Springer.

Chapter 4
Standard Setting: Bridging the Worlds of Policy Making and Research

Hans Anand Pant, Simon P. Tiffin-Richards, and Petra Stanat

Abstract Interpreting test scores in terms of whether examinees reach specific levels of achievement, provides a means to assess and communicate whether educational goals are being reached and expectations are being met. Whether these interpretations of test scores are informative, however, hinges on their validity. While validity plays an important role in educational assessment, it is rarely addressed in a systematic and comprehensive manner. The discussion focusses on the role of standard setting procedures which define minimum passing scores on test score scales. Our aim is to detail a theoretical framework in which validation is considered in the context of practical test development and large-scale assessment in Germany. To this end, we apply Kane's interpretive argument approach and Toulmin's inference model to the development of standards-based educational assessments and the interpretation of their results. A logical argument is presented to provide a theoretical framework for evaluating the rhetorical backing and empirical evidence supporting interpretations of educational assessment results.

Keywords Validity • Assessment • Standard setting • Cut scores • Function creep

H.A. Pant (✉)
Humboldt-Universität zu Berlin, Berlin, Germany
e-mail: hansanand.pant@hu-berlin.de

S.P. Tiffin-Richards
Max Planck Institute for Human Development, Berlin, Germany
e-mail: tiffin-richards@mpib-berlin.mpg.de

P. Stanat
Institute for Educational Quality Improvement (IQB), Berlin, Germany
e-mail: petra.stanat@iqb.hu-berlin.de

S. Blömeke, J.-E. Gustafsson (eds.), *Standard Setting in Education*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-50856-6_4

4.1 Introduction

The growth of standards-based accountability systems as a core element of evidence-based policy and practice in education has led to a recent upsurge in the scientific debate on the validity of educational assessments (e.g., Haertel 2013; Kane 2013; Koretz 2013, 2015; Lissitz 2009; Moss 2016). One of the most controversial aspects in this debate is the question to what degree the intended purpose and actual use of a test should be an issue of validity considerations (Messick 1995, 1998). Some validity researchers outright reject the idea that test impact evaluation is a legitimate element in the validation process and shift the larger part of this burden to the test users, e.g., political stakeholders or educational practitioners (cf. Lissitz and Samuelsen 2007). Others assign this burden of considering the consequential aspects of validity at least in part to the test developer (cf. Kane 2011). As Haertel (2013) states:

> Measurement professionals are obviously concerned with the question of how testing is supposed to improve schooling, although they might frame it more formally, asking first, with regard to a specific testing program, "What are the intended test score uses or interpretations?" Then, matters of test validity might be framed by asking, "What are the interpretive arguments and supporting validity arguments for those intended uses or interpretations?" These questions direct attention first to the intended use or purpose of the testing, then to the chain of reasoning from test scores to decisions or implications, and finally to the evidence supporting or challenging the propositions in that chain of reasoning.

In most cases, political stakeholders or educational practitioners are not expected to have scholarly expertise in reflecting and interpreting "raw" assessment results like test scores. It is believed that criterion-referenced interpretations of performance based on discrete proficiency levels (e.g., "basic", "advanced") can be communicated more easily to a wide variety of stakeholders than norm-referenced interpretations based on continuous proficiency scales. This communicative function makes standard setting procedures a critical gateway for validity concerns that pertain especially to consequential aspects of validity.

The purpose of this chapter is twofold. First, we want to elaborate on the chain of reasoning from test scores to decisions or implications. Our aim is to detail a theoretical framework in which validation is considered in the context of practical test development and large-scale assessment. We specifically discuss the role of standard setting within this validity framework.

From a practical perspective, the consequences of interpretations of test scores for specific purposes are difficult to separate from the general educational policy context, including the national strategy on accountability, curricular autonomy of schools, as well as other factors such as teachers' attitudes towards standardized large-scale assessment. We therefore, secondly, want to demonstrate how the application of theoretical validity considerations is useful in evaluating unintended consequences of standard setting using the case of large-scale assessment in Germany. We specifically show how the widening of the use of a test, which is called *function creep*, affects the validity of standard setting procedures.

We have divided the remainder of the chapter into four sections. In the following we briefly describe how educational accountability is designed in the case of Germany. In Sect. 4.3 we introduce standard setting procedures as a necessary step in interpreting test results, in terms of reaching educational goals and minimum or target levels of proficiency in specific domains. In the next Sect. 4.4 we integrate the process of standard setting in the theoretical frameworks of Kane (1992, 1994) and Toulmin (Kane 2011; Toulmin 1958) to construct a validity argument for test score interpretations. We then proceed to the core of this chapter and provide empirical evidence from studies conducted in Germany to examine the factors that may influence the validity of test score interpretations following standard setting procedures (Sect. 4.5). The discussion focusses on the role of standard setting in the evaluation of validity as the sequential weighing of rhetorical and empirical evidence supporting propositions of validity in interdisciplinary procedures, rather than a precise measurement procedure.

4.2 Educational Standards and Standardized Testing in Germany

As a reaction to the disappointing results for Germany in international large-scale assessments of student achievement, such as TIMSS (Trends in International Mathematics and Science Study) and PISA (Programme for International Student Assessment), a number of steps have been taken towards the development and monitoring of educational quality in Germany.

In 2003 and 2004, the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany (KMK)[1] introduced National Educational Standards for the primary level and for secondary level I detailing which competencies and skills students are expected to have developed by the time they reach certain points in their school career. In general, the key objective of introducing National Educational Standards was to shift attention to the learning outcomes of educational processes and to ensure greater comparability of educational requirements within the school system.

At the primary level, the focus of these standards was on the core subjects of German and mathematics. At secondary level I, the focus was on German, mathematics, the first foreign languages (English, French), and for the science subjects biology, chemistry, and physics. In accordance with the long-term strategy of the

[1] Under the German Basic Law and the constitutions of the federal states (Länder), the entire school system is under the supervision of the state. Supervision of the general school system is the responsibility of the Ministries of Education and Cultural Affairs in the 16 Länder. Hence, in Germany it is not the Federal Ministry of Education who is in charge with quality assurance of the general school system but the 16 Länder ministries. The Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany (KMK) is their coordinating body.

Standing Conference (KMK 2015) on educational monitoring in Germany, the 16 federal states (Länder) also decided to conduct regular comparative studies using standardized tests on three levels: First, Germany continues to participate in various international large-scale assessments of student achievement at the primary level (PIRLS, TIMSS) and secondary level I (PISA). Secondly, regular standards-based tests are used to assess the extent to which educational standards are being met at the state level. Thirdly, state-wide and standards-based comparison tests (*Vergleichsarbeiten*, VERA) were introduced in order to determine achievement levels of all students in grade 3 and grade 8, serving as a performance feedback to each and every school. Schools thus are provided with external data whether their students are "on track" one and two years, respectively, before the primary and secondary level standards are to be met.

However, international studies, the national state-comparison studies, and the state-wide tests feature important distinctions. These differences mainly concern the respective goals *(What purpose should the results serve?)* and the level of evaluation *(Who should be evaluated?)*. Whereas international and national assessments were designed to descriptively monitor the educational system at large, state-wide comparison tests aim at providing teachers with information on the strengths and weaknesses of their students with regard to the educational standards. Conjointly with aligned teaching material provided together with the test results, this feedback serves to support teachers in their efforts to optimize instructional quality and principals to evaluate their school's performance and, where necessary, to plan professional development of their staff.

Notably, in the national and the state-wide assessments the same type of standards-based tests and proficiency level models are being used for very different purposes. We will take on this point in Sect. 4.5 of this chapter. Table 4.1 summarizes the most relevant differences of the three pillars in Germany's test-based monitoring system.

The introduction of the educational standards and the long-term strategy on educational monitoring was accompanied by the foundation of the Institute for Educational Quality Improvement (IQB=*Institut zur Qualitätsentwicklung im Bildungswesen*). The IQB was in charge of developing proficiency level models as tools for interpreting test scores. These models provide a more concrete description of what students with a given test score are able to do ("can-do statements"). To ensure a clear presentation of the results obtained in the national performance assessments, student results in the IQB National Assessment Study reports are shown both as point scores that relate to a continuous proficiency scale and as levels that refer to established proficiency level models (Köller et al. 2010; Stanat et al. 2012; Pant et al. 2013).

To develop the proficiency models, *standard setting* comes into play. In standard setting procedures subject experts divide the continuous scale into multiple sections that can be usefully distinguished according to content and are then referred to as proficiency levels (Cizek and Bunch 2007). The experts work by systematically analyzing the cognitive requirements of items that students with a given test score are very likely to have completed. Each level description sets out the cognitive

Table 4.1 Types of standardized assessments as part of the long-term strategy on educational monitoring in Germany

	International tests	**National tests**	**State-wide tests (VERA)**
	PISA, PIRLS, TIMSS	*16-states comparisons*	*Within-state comparisons*
Standards-based tests?	No	Yes	Yes
Data base	Sample-based (approx. *N*=5000)	Sample-based (approx. *N*=50,000)	Population-based (approx. *N*=1.3 Mio.)
Periodicity	Every 5 years (PIRLS), 4 years (TIMSS), and 3 years (PISA)	Every 5 years (primary level) and every 6 years (secondary level)	Annually
Designated main purpose	System monitoring	System monitoring	School improvement, instructional improvement
Test administration and coding of responses	External test administrators	External test administrators	Teachers in their classes[a]
Data analysis	National project management	IQB	State evaluation agencies
Performance feedback level	Country	State	School, class, student[b]
Who is accountable?	Federal Ministry of Education; 16 State Ministries of Education	16 State Ministries of Education; supervisory school authorities	Principals, teachers

IQB Institute for Educational Quality Improvement, the national testing institute in Germany
[a]With the exception of the state of Hamburg where tests are administered and coded by the state evaluation agency
[b]Feedback on the individual student level was not officially encouraged

requirements that students can fulfil once they reach that proficiency level. This makes it possible to produce qualitative descriptions of the competencies that students have acquired, and can show what percentage of students are most likely to be able to fulfil specific requirements.

4.3 Standard Setting as a Key Validity Issue in Educational Assessment

The following sections focus on the setting of minimum test scores (henceforth referred to as *cut scores*) on educational assessments as a method of interpreting the results of such assessments. To ensure that standards-based interpretations of student test scores are sufficiently defensible as the basis for high- or low-stakes

decisions (Decker and Bolt 2008), inferences based on student test scores must be supported by evidence of their validity (Kane 1992, 2009; Messick 1995). The move towards standards-based assessment therefore makes the question of validity vitally important, as the results of standards-based testing can have consequences, both at the individual level of students and professionals, at the organizational level of schools and training programs, and at the system level of educational administration and policy. However, despite the central role of validity in interpreting test results, there is little consensus on the exact conceptualization of validity, how it is best evaluated, or indeed what constitutes *sufficient* validity (Lissitz 2009). Although some aspects of the current conceptualization of validity enjoy "fairly broad professional agreement" (Cizek 2011), there remain disputes concerning the importance of different sources of validity evidence. The view is supported here that different sources of validity evidence are necessary, although not individually sufficient, to support criterion-referenced test score interpretations.

4.3.1 The Place of Standard Setting in Educational Assessment

Standards-based interpretations of test scores are argued to facilitate the communication of assessment results to non-technical stakeholders, such as policy-makers, teachers, and parents (Cizek et al. 2004). From a policy perspective, setting educational standards and assessing achievement towards them allows the definition of educational goals and the implementation of teaching practices that facilitate achievement towards these goals.

Test-centered standard setting approaches focus on the items of an assessment and on how likely examinees are expected to be able to answer them correctly, depending on their level of proficiency. Popular methods include modifications of the Angoff (Angoff 1971) and Bookmark methods (Mitzel et al. 2001). Many of the methods currently employed combine expert judgment with psychometric analyses of item difficulty and examinee proficiency, allowing cut scores to be mapped directly onto item response theory (IRT)-derived metric scales (Cizek and Bunch 2007).

The significance attached to validating the inferences of test scores based on cut scores and performance standards is also emphasized in the *Standards for educational and psychological testing* (henceforth referred to as Standards) jointly published by the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (AERA et al. 2014):

> Such cut scores provide the basis for using and interpreting test results. Thus, in some situations, the validity of test score interpretations may hinge on the cut scores. 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.

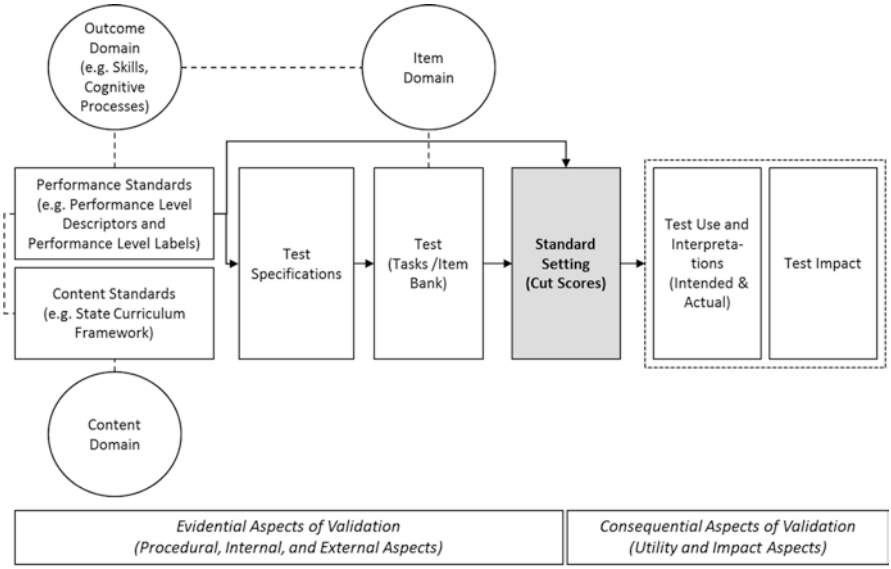


Fig. 4.1 The role of standard setting for evidentiary and consequential aspects of validation (Pant et al. 2009)

The guidelines provided by the Standards have two implications. First, cut scores must represent the characteristics and intended interpretations of a test—with regard to both content standards (construct representation) and performance standards (target levels of performance). Second, neither the method for placing cut scores, nor the procedure for evaluating the interpretations of test results in respect to cut scores is clearly specified; rather, each depends on the intended interpretations of test scores. A further important notion is that cut scores and associated performance standards are *concepts* employed to interpret test scores. Thus, there are no *true* cut scores that could be defined or applied, and therefore there is no method for establishing whether a cut score is correct or not. Rather than establishing whether a cut score is correct, the aim of validation procedures is to assess the degree to which there is "convincing evidence that the passing score does represent the intended performance standard and that this performance standard is appropriate, given the goals of the decision process" (Kane 1994).

As depicted in Fig. 4.1 standards-based assessment systems follow a typical schematic process. The initiator is typically an authorized *policy body*—such as a ministry of education or, as in the German case, the *Standing Conference* of the 16 state ministries (KMK). The policy body defines or commissions descriptions of the *content standards*, representing the characteristics of the knowledge, skills, and abilities which the assessment is to measure (e.g., reading comprehension), and *performance standards*, which represent specific levels of achievement on a scale of the measured competence. The operationalization of the content and performance standards is assigned by the policy body to a *technical body* (here: the IQB) which

designs the measurement instrument with which the achievement towards the learning goals can be described on a linear scale.

Test specifications detailing item characteristics, derived from the content standards, are developed and provide the blueprints used by item developers to design assessment tasks and items. These blueprints form an important basis for the content alignment, or construct representativeness of the measurement instrument. The definition of cut scores on the test score scale is thus key to the operationalization of corresponding performance standards (Kane 2001), allowing test-scores to be interpreted directly in terms of the performance levels descriptions.

It is beyond the scope of this paper to review all steps of this process with regard to validation aspects. However, the transformation of test scores on a continuous scale of ability to categorical levels of achievement is arguably a paradoxical part of the process. It aims to enhance the communication of inferences warranted from examinee test scores by reducing the information of test results from a continuous to a discrete variable reported on an ordinal scale. It is therefore important to provide a defensible case that the communicative and evaluative value of reporting of test scores on proficiency levels outweighs the loss of information.

4.4 The Argument Approach to Evaluating Validity

The process of validation can be conceptualized as formulating an *argument* for the validity of interpretations derived from test scores (Kane 1992), involving, first, the specification of the proposed interpretations and uses of test scores and, second, the evaluation of the plausibility of the proposed interpretations (Kane 2011). The first step is the construction of an *interpretive argument,* which builds the structure of the argumentative path from the observed performance to the inferred proficiency of an examinee in the domain of interest (e.g., proficiency in a foreign language). The interpretive argument hence "provides a framework for validation by outlining the inferences and assumptions to be evaluated" (Kane 2011). The second step involves the construction of the *validity argument,* which appraises the evidence supporting the inferences that lead to the test score interpretation and "provides an evaluation of the interpretive argument" (Kane 2011). The notion of establishing an evidence-based case for an interpretation was shared by Messick (1989) in his definition of validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores".

Validity is not considered a property of a test, but of the use of the test for a particular purpose. Importantly, Kane (1992) also proposes that "[t]he plausibility of the argument as a whole is limited by its weakest assumptions and inferences". It is, therefore, important to identify the assumptions being made and to provide supporting evidence for the most questionable of these assumptions. The greatest attention should be directed towards evaluating the availability of evidence in support of weak assumptions, to rule out alternative interpretations of test scores (Kane 1992).

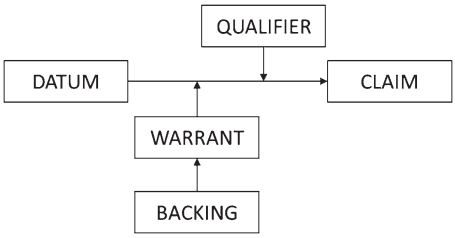


Fig. 4.2 Toulmin's model of inference

A formal representation of the interpretive argument is presented by Kane (2011) in Toulmin's model of inference (Toulmin 1958). Toulmin's model provides a framework with which an interpretive argument can be structured and the validity evidence supporting a test score interpretation can be evaluated. The basic structure of the model is described as the "*movement* from accepted *data*, through a *warrant*, to a *claim*" (Brockriede and Ehninger 1960), represented in Fig. 4.2. Kane (2011) summarizes three characteristics of such logical arguments. First, they represent disputable lines of argument that make substantive claims about the world and can be evaluated on the basis of empirical evidence, and in terms of how well they refute opposing arguments. Second, they provide arguments for probable or acceptable conclusions rather than certain facts, and may include an indication of their strength. Finally, informal arguments are *defeasible,* in the sense that they are subject to exceptions.

The movement from datum to claim is justified by a warrant, which is in turn supported by backing or evidence "designed to certify the assumption expressed in the warrant" (Brockriede and Ehninger 1960). The notion of a warrant in the context of validation could for instance refer to empirical evidence of the concurrent validity of a newly developed reading comprehension test, compared to an older established comprehension test. The claim being made in this case is the classification of an examinee on a proficiency level (e.g., pass or fail on the reading comprehension test) on the basis of their test score. A qualifier may be inserted if the claim is only valid under certain conditions. In the context of test score interpretations, the qualifier may relate to specific testing contexts (e.g., only suitable as a university entry requirement) and be designed to minimize unwanted consequences of testing.

4.4.1 A Structured Validity Argument

Criterion-referenced interpretations of test scores are based on the claim that examinee performance on an assessment can be interpreted in terms of levels of proficiency in the domain of interest, such as foreign language proficiency. The goal is to be able to generalize an examinee's performance on a sample of test items from the domain of interest to their estimated proficiency across the domain. Hence, the score on an English as a Foreign Language exam is extrapolated to infer the general communicative language ability of the examinee (see Chapelle et al. 2010 for an

application of the validity argument approach to the TOEFL, and Papageorgiou and Tannenbaum 2016 for applying it in language assessment in general). The argument that test scores can be interpreted in this way can be evaluated on the basis of whether the propositions or warrants for this claim are met, being supported by sufficient rhetorical backing and empirical evidence. In the example of a test of foreign language proficiency, these warrants could include evidence that the exam requires desired levels of vocabulary knowledge and text comprehension, that language skills can be applied in realistic communicative scenarios, and that the testing formats allow examinees to demonstrate their abilities adequately (e.g., productive as well as receptive language abilities).

The composition of the interpretive argument can be considered as a series of *propositions* of validity (Haertel 2002; Haertel and Lorié 2004), each of which must be supported by evidence of validity. Individually necessary—but not sufficient—propositions of validity can include (1) clear definition of the content standards detailing knowledge, skills and abilities relevant to the domain of interest, (2) representation of domain-relevant content in the assessment, (3) unambiguous definition of target performance standards, (4) reliable and unbiased test scores provided by the measurement instrument, (5) defensible cut score placements to differentiate proficiency levels with minimum test scores, and (6) alignment of intended test use to defensible interpretations of test scores provided by assessments.

In this context the datum is an examinee's observed performance in the content domain of interest, measured as a test score on an assessment. The claim that examinee performance can be interpreted in terms of whether they satisfy the criteria for reaching a specific level of proficiency can be supported by five warrants or propositions that indicate that such a claim is appropriate. Each warrant is backed by rhetorical and empirical evidence (Fig. 4.3). A qualifier is inserted to detail the conditions under which the proposed claims are valid for a specified population. The qualifier thus accounts for concerns of *consequential validity* (Kane 2011) and is a *justification of test use* for specific purposes (Cizek 2011, 2012). The interpretive argument is represented by the path from datum to claim with all inherent warrants of validity and the qualification under which conditions the claim holds. The validity argument in turn appraises the *backing* for the warrants supporting the proposed interpretation of the test scores. Again, it is beyond the scope of this chapter to expand in detail on all six warrants and backings, respectively, shown in Fig. 4.3. We will therefore focus on the weakest link in this chain of argument, the warrant of a defensible cut score placement.

4.4.2 Warrant of Defensible Cut Score Placements

The process of standard setting can be understood as a translation of policy decisions—such as the definition of educational standards or a passing criterion—through a process informed by expert judgment, stakeholder interests and technical

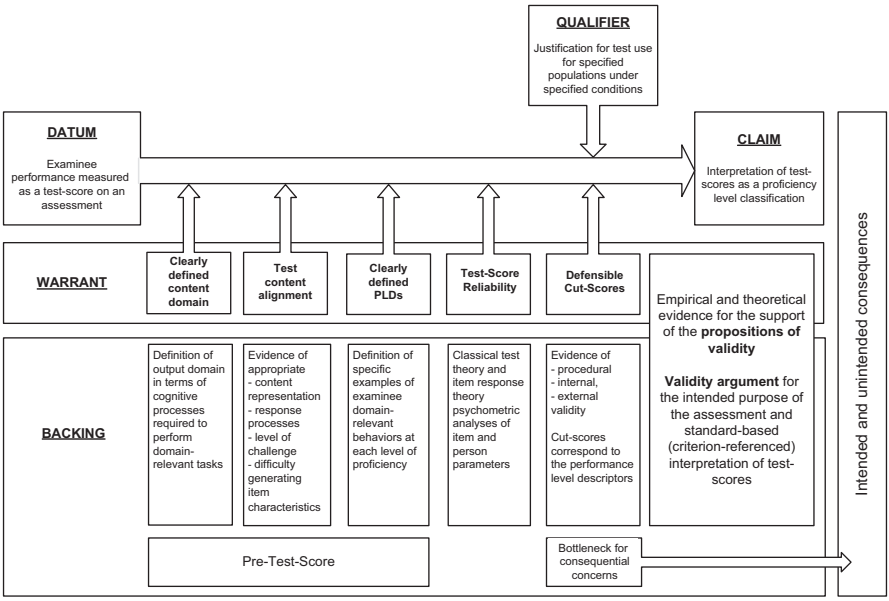


Fig. 4.3 Application of Toulmin's inference model and Kane's interpretive argument to the standards-based interpretation of test-scores presented as proficiency level classification

expertise (Cizek and Bunch 2007). This translation is achieved with the definition of cut scores by panels of experts, which differentiate discrete levels of proficiency on a continuous scale of examinee performance. The credibility of cut score recommendations has been the subject of strong criticism (Glass 1978), and may be considered a critical but weak element in the interpretive argument. Backing for the warrant of defensible cut scores can include sources of procedural, internal, external and consequential validity evidence (Cizek 1996; Kane 1994; Pant et al. 2009; Tiffin-Richards et al. 2013). *Procedural evidence* can include the selection and training of panelists, as well as their reported understanding of key concepts of the procedure (Raymond and Reid 2001), the psychometric calibration, selection, preparation, and presentation of materials, and the clear definition of performance standards (Cizek 1993). A central element of popular cut score placement procedures such as the Bookmark method is the ordered item booklet, in which test items are arranged by increasing difficulty (Karantonis and Sireci 2006). Panelists make their cut score recommendations by marking the boundaries between groups of items that represent the material an examinee is expected to have mastered at a specific level of proficiency. Tiffin-Richards et al. (2013) demonstrated that the selection of the items which are included in these item booklets can influence how expert panelists set their cut scores. In particular, items with a high mismatch between their content descriptions and their empirical difficulty presented panelists with difficulties in the standard setting procedure, and in many cases resulted in more extreme cut scores for the highest and lowest proficiency levels (Tiffin-Richards et al. 2013). This

indicates that the materials used in standard setting procedures may have a significant influence on cut score recommendations.

Internal evidence can be evaluated by assessing the inter- and intra-panelist consistency of cut scores across cut score placement rounds, while *external evidence* can evaluate the consistency of cut scores across different standard setting procedures or between parallel expert panels. A factor that may impact external evidence of validity was demonstrated by Pant et al. (2010), who showed that cut score placement recommendations appeared to differ between expert panels with different constellations of participants. Expert panels, which included both educational practitioners and participants representing educational policy makers, set stricter minimum pass scores than did panels solely made up of educational practitioners. It appeared therefore that the experience panelists had of the examinee population, influenced their perception of what examinees could and should be expected to achieve. Of course, the nature of standard setting studies, in which expert panels of highly qualified professionals are necessary, makes controlling for panelist experience and qualifications exceedingly difficult. Nevertheless, the constellation of expert panels may be critical in establishing both the appropriateness and the defensibility of cut score recommendations.

However, the warrant of appropriate cut score recommendations for the operationalization of educational standards on assessments critically depends on the prior propositions of validity: well-defined content domain and performance standards, well-aligned assessment items, and reliable measurements of examinee performance. Without these preconditions, the cut score placement procedure does not offer the basis to make appropriate recommendations for minimum test scores on competence-based assessments.

In the following section we will illustrate the relationship between the reliability of a measurement instrument, as described in validity proposition (4), and the potential consequences of different uses of test scores, as described in proposition (6).

4.5 Function Creep of Test Use: A Threat to Cut Score Validity

As mentioned in Sect. 4.2, the national and the state-wide assessments in Germany apply the same standards-based tests and proficiency level models for very different purposes, namely system monitoring and improvement of instructional quality, respectively. Although participation in state-wide assessments (called *VERA*) is mandatory for all schools and classes in grades 3 and 8, no formal consequences or sanctions were attached to the results of the tests. Hence, state-wide tests in Germany can be described as “no-stakes” for the individual student and as “low-stakes” to “no-stakes” for teachers and principals. The notion of teacher/ principal *accountability* given in Table 4.1 refers to the vague expectation that schools will use class

level and school level performance feedback from state-wide assessments as a formative evaluation tool in order to adjust their teaching routines.

Moreover, in its formal and binding *Agreement on the Further Development of VERA*, the Standing Conference of the Ministries (KMK) in 2012 stressed that the main function of the state-wide comparison tests was to improve teaching and school development. Grading of test results was not allowed. The agreement also states that aggregated results from individual schools must not be published as league tables. In addition, giving supervisory school authorities access to *VERA* results must follow strict rules that are aligned with *VERA*'s main task of school and instructional improvement.

However, none of the 16 states implemented a coherent support system for teachers and principals that enabled them how to understand performance feedback reports, in the first place, and how to transform feedback results into appropriate instructional and organizational interventions. Hence, as study results show, a substantial proportion of teachers and principals in Germany has a very reserved attitude towards the usefulness of the state-wide tests (Maier et al. 2012). Many teachers do not use them for instructional improvement (Richter et al. 2014) or organizational development (Bach et al. 2014; Wurster et al. 2013) but rather for the—much more familiar—diagnostic purpose of evaluating individual students' achievement. Such a gradual widening of the use of a test (or any other system) beyond the purpose for which it was originally intended is referred to as *function creep*.

A number of state agencies, who are in charge of the content and design of the performance feedback reports, "gave in" to teachers' asking for more detailed and fine-grained "diagnostic information" about their students. Figure 4.4 gives an example of how individual feedback of *VERA*-results is provided to teachers and parents. As can be seen, imprecision of the individual test score leads to a 95-% confidence interval that spans three to four of the five proficiency levels. Moreover, at the *class-level* the uncertainty attached to the estimated percentages of students reaching a certain proficiency level is not even communicated in this report (cf. Fig. 4.4). However, the uncertainty can be substantial, as we will demonstrate with a case example.

Case Example: State-Wide Testing of English as a Foreign Language in Grade 8 As measurement instruments are never error-free, classification errors are unavoidable and establishing the *accuracy of classifications* therefore becomes a critical consideration (Zhang 2010). In the present example, the German *Educational Standards* for secondary language education and the *European Framework of Reference for Languages* (CEFR, Council of Europe 2001), on which the Educational Standards are based, provided the performance standards to define the five levels A1 (lowest), A2, B1, B2 and C1 (highest) on an assessment of English as a Foreign Language (EFL) reading comprehension. The positions of the cut scores defining the proficiency levels on the present reading comprehension assessment were set by panels of expert judges—including educators, as well as psychometric, linguistic and didactic experts (Harsch et al. 2010). Whether or not the ambiguity of recommended cut scores is taken into account, the allocation of examinees to proficiency

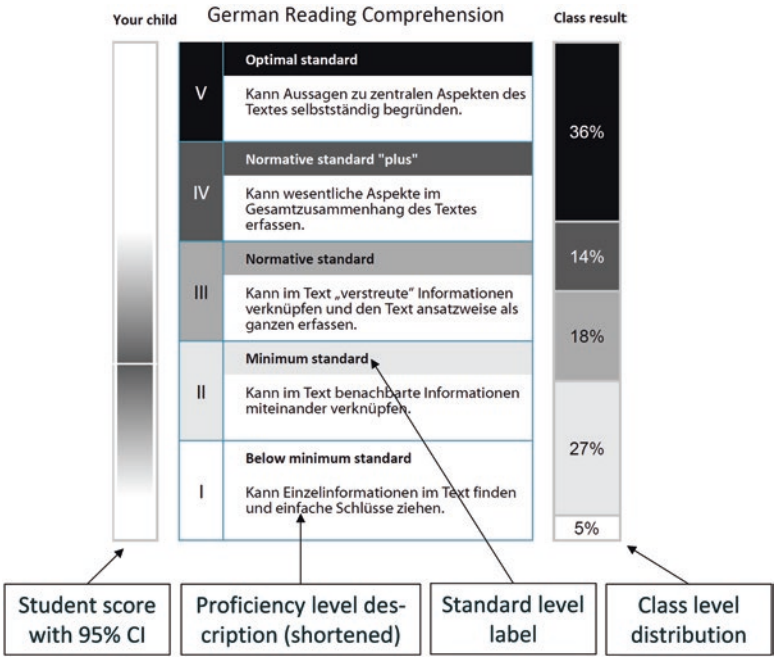


Fig. 4.4 Sample performance feedback in a state-wide test on student and class level

levels cannot be considered error-free, as the estimation of student ability is prone to both *sampling error* and *measurement error*. Betebenner and colleagues clearly illustrate that measurement error evident at the individual level should also be considered when reporting aggregated percentages of examinees on mutually-exclusive proficiency levels (Betebenner et al. 2008).

The present study employs the method proposed by Rudner (2001, 2005) for dichotomously scored response data to estimate the accuracy of classifications of examinee proficiency, based on a single test administration. As summarized in Wyse (2010), several factors have been documented which systematically affect classification accuracy. These include the underlying distribution of examinee ability and number of tested examinees (Lee 2010), the number of proficiency levels set on an ability scale and the reliability of examinee ability estimates (Ercikan and Julian 2002), the distance on the ability scale between cut scores (Zhang 2010), the number of examinees with ability estimates close to a cut score (Emons et al. 2007), and the measurement error associated with the examinee ability estimates (Betebenner et al. 2008).

In our example the response data of N=33,520 grade eight students on a test of EFL reading comprehension as part of the state-wide *VERA* assessment was analyzed using Rudner's (2001) approach to calculating expected classification accuracy. The examinee sample (N=33,520) came from N=1706 classes with an

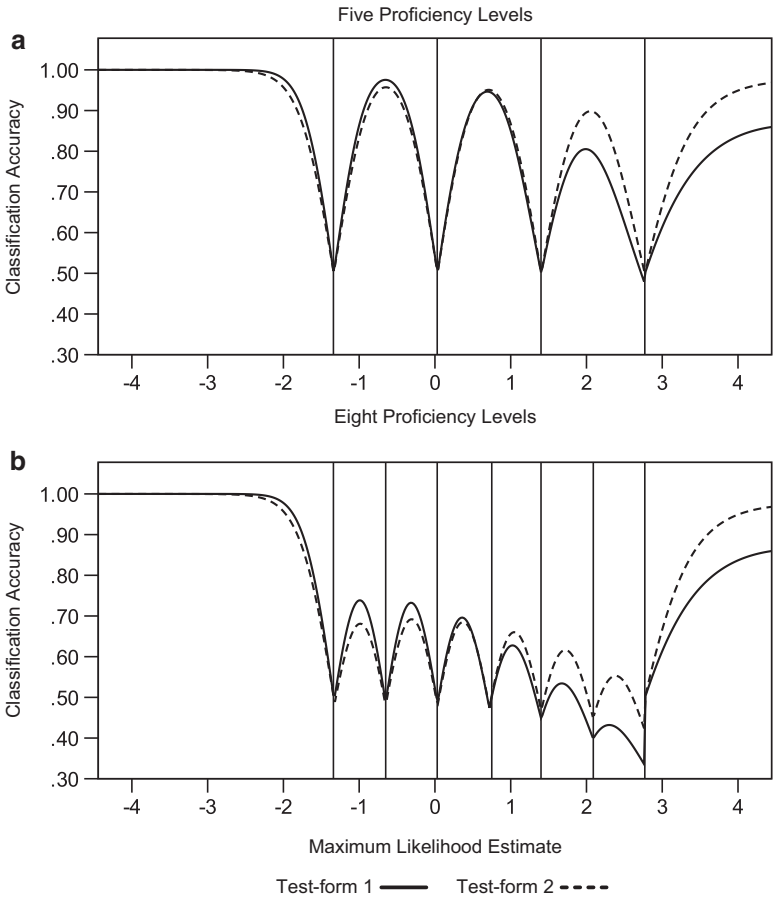


Fig. 4.5 Classification accuracy for two alternate test-forms for (**a**) five and (**b**) eight proficiency levels

average size of 19 students. (For more technical details of the study, cf. Tiffin-Richards 2011.)

The classification accuracy with which examinees with different ability estimates would be assigned to proficiency levels on two parallel test-forms[2] is illustrated in Fig. 4.5. As can be seen, classification accuracy is lowest directly at the cut scores, highest at the peripheries and the further the ability estimates are from the nearest cut score, and decreases from the lower to the higher proficiency levels. *Test-form 1* ("the easier one") is shown to have higher classification accuracy for examinees with lower ability estimates and lower accuracy at the higher ability levels compared to *test-form 2*. From a statistical perspective this may appear trivial,

[2] Most states administer two test-forms in *VERA* with differing mean difficulties to better match the different ability levels in a two-tiered school system (Gymnasium vs. all other school types).

however, cut scores are in part set to facilitate communication of test score results to non-technical recipients with specific interests and expectations.

More importantly however, some states applied a feedback model of eight (instead of five) proficiency levels to accommodate for teachers' desire for more fine-grained information. In this case the proficiency levels A2, B1 and B2 were subdivided into the levels A2.1, A2.2, B1.1, B1.2, B2.1 and B2.2. The comparison of the upper (a) and lower (b) graph in Fig. 4.5 reveals the substantial decrease in expected classification accuracy from an average 81% for the five-proficiency-levels model to 65% for the eight-proficiency-levels model. The expected classification accuracy using the "fine-grained" feedback model drops to 43% in classes with a high level of "true" ability and a mismatched ("too easy") test-form. Misclassification may play a negligible role when the purpose of the assessment is system monitoring at state level. However, it is arguably a threat to the validity of inferences drawn at smaller aggregate levels like schools or classes where trend information about changes in proficiency level distributions may become the focus of interest for teachers and principals. In any case, misclassification of the demonstrated magnitude is a massive validity concern for diagnostic inferences at the individual student level.

The case example of the paradox relationship between the increase in perceived detail of reports and the accuracy with which examinees can be assigned to these levels leads to the recommendation to set only as many cut scores to define proficiency levels as is necessary for the purpose of the assessment. And it strengthens the case to avoid function creep of test-use and the use of proficiency level models by all means.

Cut score placement procedures are arguably the weakest link in the validity argument for criterion-referenced interpretations of test scores, and thus present a bottleneck for validity concerns. Cut score placement procedures therefore require particular attention in the validation process and are particularly dependent on the quality of earlier stages of the validation process.

4.6 Conclusions

The argument approach to validation illustrates the complexity of the operationalization of competence-based testing programs, as well as the consequent complexity of the interpretive and validity arguments that can be constructed to provide a convincing case for the interpretation of test scores. The perspective of considering validation as an argumentative case supporting a proposed interpretation of an examinee's test score as an indication of his or her level of proficiency in the domain of interest, leads to two general conclusions.

First, it is evident from the sequential structure of the argument approach to validation that each element of the validity argument relies, at least in part, on preceding propositions of validity being met. Poor definition of the content domain and content standards will pose difficulties in the definition of clear proficiency level

descriptors (PLDs), ambiguously defined PLDs provide a poor basis for cut score placement procedures to differentiate proficiency levels, and so on. Deficits in rhetorical and empirical backing for a warrant supporting the proposed interpretation of test scores can thus lead to weaker support for subsequent warrants, as well as weakening the overall support for the validity argument's claims. Being aware of the interdependence of the evidentiary support for each warrant of the argument's validity is therefore critical. This is particularly important for any institution or program responsible for the development and operationalization of educational standards, as validity evidence may need to be drawn from different groups of professionals at different stages (e.g., content domain experts and practitioners for construct definition, item developers for content alignment, psychometric experts for test reliability, etc.).

Second, cut score placement procedures not only rely on the quality of prior propositions of validity, but also reflect expert opinions rather than exact measurement procedures. Cut score recommendations represent expert judgments on how best to operationalize educational standards, based on the content of an assessment and proficiency level descriptions. Under ideal circumstances, content domains would be perfectly defined in terms of the cognitive processes required to complete domain-related tasks, test items would cover the entirety of the relevant content domain or represent a random sample of all its elements, and PLDs would perfectly describe examinee behaviors relevant to the content domain at each proficiency level. However, content domains and PLDs are usually described in general terms, item samples are limited and possibly not representative of the entire content domain, due to practical limitations. Cut score recommendations are at best approximations of appropriate and defensible numerical criteria for reaching proficiency levels on assessments where the content domain and proficiency level descriptors are usually defined in generalized terms and there is a limited availability of assessment items and testing time.

What the validity argument clearly demonstrates is that the validity of criteria-referenced test score interpretation depends on a sequence of warrants of validity being met. A stronger focus on the definition of the constructs of interest (e.g., reading, writing, mathematics, natural science) in terms of underlying cognitive processes (e.g., word decoding, text comprehension, number sense, abstract reasoning) is the necessary basis for making standard setting and cut score placement procedures meaningful.

The argument approach to validity in general provides a suitable framework for the challenging task of combining theoretical concepts and measurement procedures with practical considerations and policy aims, to develop and operationalize theoretically and psychometrically sound, practical and socially acceptable standards-based assessments. The evaluation of the degree to which inferences are valid and resulting actions are justifiable is, in the end, necessarily embedded in a social discourse whose participants typically bring to the table diverse frameworks, assumptions, beliefs, and values about what constitutes credible evidence. For the future, it will therefore be crucial to implement a strategically oriented dialogue between research and policy involving educational researchers, educational policy

makers, educational administrators, and educators themselves in order to arrive at a coordinated and coherent system of setting validity priorities.

Acknowledgments This research was supported in part by a grant (PA-1532/2-1) from the German Research Foundation (DFG) as part of the Priority Program "Competence Models for Assessing Individual Learning Outcomes and Evaluating Educational Processes" (SPP 1293).

References

AERA, APA, & NCME (American Educational Research Association, American Psychological Association, National Council on Measurement in Education). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Bach, A., Wurster, S., Thillmann, K., Pant, H. A., & Thiel, F. (2014). Vergleichsarbeiten und schulische Personalentwicklung – Ausmaß und Voraussetzungen der Datennutzung. *Zeitschrift für Erziehungswissenschaft, 17*, 61–84.

Betebenner, D. W., Shang, Y., Xiang, Y., Zhao, Y., & Yue, X. (2008). The impact of performance level misclassification on the accuracy and precision of percent at performance level measures. *Journal of Educational Measurement, 45*, 119–137.

Brockriede, W., & Ehninger, D. (1960). Toulmin on argument: An interpretation and application. *The Quarterly Journal of Speech, 46*, 44–53. doi:10.1080/00335636009382390.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice., 29*(1), 3–13. doi:10.1111/j.1745-3992.2009.00165.x.

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

Cizek, G. J. (1996). Standard setting guidelines. *Educational Measurement: Issues and Practice, 15*(1), 13–21. doi:10.1111/j.1745-3992.1996.tb00802.x.

Cizek, G. J. (2011, April). *Reconceptualizing validity and the place of consequences*. Paper presented at the meeting of the NCME, New Orleans.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods, 17*, 31–43. doi:10.1037/a0026975.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on test*. Thousand Oaks: Sage.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: contemporary methods. *Educational Measurement: Issues and Practice, 23*(4), 31–50. doi:10.1111/j.1745-3992.2004.tb00166.x.

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

Decker, D. M., & Bolt, S. E. (2008). Challenges and opportunities for promoting student achievement through large-scale assessment results: Research, reflections, and future directions. *Assessment for Effective Intervention, 34*, 43–51. doi:10.1177/1534508408314173.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2007). On the consistency of individual classification using short scales. *Psychological Methods, 12*, 105–120.

Ercikan, K., & Julian, M. (2002). Classification accuracy of assigning student performance to proficiency levels: Guidelines for assessment design. *Applied Measurement in Education, 15*, 269–294.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement, 15*, 237–261. doi:10.1111/j.1745-3984.1978.tb00072.x.

Haertel, E. H. (2002). Standard setting as a participatory process: Implications for validation of standards-based accountability programs. *Educational Measurement: Issues and Practice, 21*(1), 16–22. doi:10.1111/j.1745-3992.2002.tb00081.x.

Haertel, E. (2013). How is testing supposed to improve schooling? *Measurement: Interdisciplinary Research and Perspectives, 11*(1–2), 1–18.

Haertel, E. H., & Lorié, W. A. (2004). Validating standards-based test score interpretations. *Measurement: Interdisciplinary Research and Perspectives, 2*, 61–103. doi:10.1207/s15366359mea0202_1.

Harsch, C., Pant, H. A., & Köller, O. (Eds.). (2010). *Calibrating standards-based assessment tasks for English as a first foreign language. Standard-setting procedures in Germany*. Münster: Waxmann.

Kane, M. T. (1992). Quantitative methods in psychology: An argument-based approach to validity. *Psychological Bulletin, 112*, 527–535. doi:10.1037/0033-2909.112.3.527.

Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*, 425–461. doi:10.3102/00346543064003425.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319–342.

Kane, M. T. (2009). Validating the interpretations and uses of test scores. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 39–64). Charlotte: Information Age.

Kane, M. T. (2011). Validating score interpretations and uses. *Language Testing, 29*, 3–17. doi:10.1177/0265532211417210.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73.

Karantonis, A., & Sireci, S. G. (2006). The bookmark standard-setting-method: A literature review. *Educational Measurement: Issues and Practice, 25*(1), 4–12. doi:10.1111/j.1745-3992.2006.00047.x.

KMK (2015) = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2015). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. München: Wolters Kluwer.

Köller, O., Knigge, M., & Tesch, B. (Eds.). (2010). *Sprachliche Kompetenzen im Ländervergleich*. Münster: Waxmann.

Koretz, D. (2013). Commentary on E. Haertel "How is testing supposed to improve schooling?". *Measurement: Interdisciplinary Research and Perspectives, 11*(1–2), 40–43.

Koretz, D. (2015). Adapting educational measurement to the demands of test-based accountability. *Measurement: Interdisciplinary Research & Perspectives, 13*, 1–25.

Lee, W.-C. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement, 47*, 1–17.

Lissitz, R. W. (2009). *The concept of validity: Revisions, new directions, and applications*. Charlotte: Information Age.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*, 437–448.

Maier, U., Bohl, T., Kleinknecht, M., & Metz, K. (2012). Impact of mandatory testing system and school context factors on teachers' acceptance and usage of school performance feedback data. *Journal for Educational Research Online/Journal für Bildungsforschung Online, 3*, 62–93.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons` responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749. doi:10.1037/0003-066X.50.9.741.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research, 45*(35), 44. doi:10.1023/A:1006964925094.

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: Erlbaum.

Moss, P. A. (2016). Shifting the focus of validity for test use. *Assessment in Education: Principles, Policy & Practice, 23*, 236–251.

Pant, H. A., Rupp, A. A., Tiffin-Richards, S. P., & Köller, O. (2009). Validity issues in standard setting studies. *Studies in Educational Evaluation, 35*, 95–101. doi:10.1016/j.stueduc.2009.10.008.

Pant, H. A., Tiffin-Richards, S. P., & Köller, O. (2010). Standard-Setting für Kompetenztests im Large-Scale-Assessment [Standard setting in large-scale assessment]. *Zeitschrift für Pädagogik,* (Beiheft 56), 175–188.

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (Eds.). (2013). *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I*. Münster: Waxmann.

Papageorgiou, S., & Tannenbaum, R. J. (2016). Situating standard setting within argument-based validity. *Language Assessment Quarterly, 13*(2), 109–123.

Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 119–157). Mahwah: Lawrence Erlbaum.

Richter, D., Böhme, K., Becker, M., Pant, H. A., & Stanat, P. (2014). Überzeugungen von Lehrkräften zu den Funktionen von Vergleichsarbeiten: Zusammenhänge zu Veränderungen im Unterricht und den Kompetenzen von Schülerinnen und Schülern. *Zeitschrift für Pädagogik, 60*, 225–244.

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment Research & Evaluation, 7*(14). http://PAREonline.net/getvn.asp?v=7&n=14. Accessed 10 May 2016.

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment Research & Evaluation*, *10*(13). http://PAREonline.net/getvn.asp?v=10&n=13. Accessed 10 May 2016.

Stanat, P., Pant, H. A., Böhme, K., & Richter, D. (Eds.). (2012). *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011*. Münster: Waxmann.

Tiffin-Richards, S. P. (2011). *Setting standards for the assessment of English as a foreign language: Establishing validity evidence for criterion-referenced interpretations of test-scores* (Unpublished doctoral dissertation). Freie Universität Berlin, Berlin.

Tiffin-Richards, S. P., Pant, H. A., & Köller, O. (2013). Setting standards for English foreign language assessment: Methodology, validation and a degree of arbitrariness. *Educational Measurement: Issues and Practice, 32*(2), 15–25. doi:10.1111/emip.12008.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press.

Wurster, S., Richter, D., Schliesing, A. & Pant, H. A. (2013). Nutzung unterschiedlicher Evaluationsdaten an Berliner und Brandenburger Schulen. Rezeption und Nutzung von Ergebnissen aus Schulinspektion, Vergleichsarbeiten und interner Evaluation im Vergleich. Die Deutsche Schule*,* 12 (Suppl.), 19–50.

Wyse, A. E. (2010). The potential impact of not being able to create parallel tests on expected classification accuracy. *Applied Psychological Measurement, 35*, 110–126.

Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing, 27*, 119–140.

Chapter 5
Standard Setting in PISA and TIMSS and How These Procedures Can Be Used Nationally

Rolf Vegar Olsen and Trude Nilsen

Abstract In this chapter, we compare and discuss similarities and differences in the way the two large-scale international studies, PISA and TIMSS, formulate and set descriptions of standards. Although the studies use similar methods, different decisions have been made regarding the nature and properties of the final descriptions of student achievement. In addition to this overview, we treat PISA and TIMSS as case studies in order to illustrate an under-researched area in standard setting: the nature of and empirical basis for the development of performance level descriptors (PLDs). We conclude by discussing how these procedures may be relevant for formulating useful standards in tests and assessments in the Norwegian context.

Keywords Performance level descriptors • Standard setting • Large-scale assessment • PISA • TIMSS

5.1 Introduction

One of the more powerful ways to report the PISA 2000 scores was to use performance level descriptions (PLDs). The state of shock communicated by policymakers in several countries following the presentation of the PISA results may have been caused in part by the power of these descriptions. For instance, policymakers were warned that "…[E]ducation systems with large proportions of students performing below, or even at, Level 1 should be concerned that significant numbers of their students may not be acquiring the literacy knowledge and skills to benefit sufficiently from their educational opportunities" (OECD 2001, p. 48).

R.V. Olsen (✉)
Center for Educational Measurement, University of Oslo, Oslo, Norway
e-mail: r.v.olsen@cemo.uio.no

T. Nilsen
Department of Teacher Education and School Research, University of Oslo, Oslo, Norway
e-mail: trude.nilsen@ils.uio.no

S. Blömeke, J.-E. Gustafsson (eds.), *Standard Setting in Education*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-50856-6_5

In short, the scales were partitioned into a finite number of intervals, and information about students' relative success on test items was used to develop verbal descriptions characterizing students' performance as they progressed on the scale. PISA was not the first large-scale assessment to develop and implement these types of descriptions—similar procedures were developed and applied successfully in both the National Assessment for Educational Progress (NAEP) (Beaton and Zwick 1992) and TIMSS (Kelly 1999).

With some exceptions concerning national assessments[1], standard setting rarely occurs in Norway. However, teachers and exam judges are given the task of grading students, and at least at a superficial level, the end product of grading resembles the end product of standard setting procedures, because they both consist of a limited number of levels or cut scores that are intended to represent a coarse measure of the student achievement. In other words, some rules or procedures that are applied result in grades; however, there is very little understanding of what the grades actually represent or of teachers' reasoning when making grading decisions.

In this chapter, we discuss the similarities and differences in the way the two large-scale international studies, PISA and TIMSS, formulate and set their descriptions of standards. In doing so, we also briefly relate these procedures to the wider literature on standard setting (e.g., Cizek 2012; Cizek and Bunch 2007; Smith and Stone 2009). Previous studies emphasized various aspects of how to use expert judgments to identify substantially meaningful cut scores along a scale representing the measurement of achievement in a specified domain. Here, we investigate how decisions are operationalized in large-scale international studies and extend the discussion to a less researched area: the nature of and basis for the development of the descriptions of student achievement along the scale. These are potentially powerful tools for communicating the results of the studies. In concluding, we discuss how these procedures may be relevant to conceiving and operationalizing useful standards in assessments in the Norwegian context.

5.2 PISA and TIMSS: Differences and Similarities

Before describing how PISA and TIMSS produce their descriptions of students' proficiency at different points along the scales, it is necessary to give a short account of how the two studies differ. To some extent, the nature of the final descriptions of students' proficiencies could be regarded as reflecting the somewhat different perspectives and aims guiding the two studies. We use science as the example domain in this chapter; hence, we refer to some specific aspects of how the two studies have

[1] The national assessment is comprised of compulsory reading, English and numeracy tests conducted at the start of the school year as students enter upper primary school (5th grade) and lower secondary school (8th grade). They are low-stakes assessments meant to be used formatively for students, but they are also used for accountability purposes for schools.

defined and operationalized this domain. Similar statements could, however, be made for mathematics.

Both TIMSS and PISA include measures of students' competencies in science and mathematics. In addition, PISA includes a measure of reading competency and one so-called innovative domain varying from cycle to cycle (e.g., collaborative problem solving). The major important difference between the assessments is that while the TIMSS framework and design is firmly based on a model of school curriculum (Mullis et al. 2009), PISA is based on a more future-oriented perspective that seeks to identify knowledge and competencies needed for further studies, careers, and citizenship in general, emphasizing what could be termed a *systems perspective* (OECD 2006). Consequently, TIMSS samples intact classes in order to study instructional processes, whereas PISA samples students across classes within schools. Also, TIMSS includes grades in the middle of primary school (4th grade) and the beginning of lower secondary school (8th grade), whereas PISA samples an age cohort toward the end of compulsory schooling (students turning 15 in a specific year).

The tests are constructed quite differently in the two assessments. PISA uses clusters of items with a common stimulus material, often in the form of an extended text, while TIMSS mainly contains stand-alone items, including "pure or context-free" items. While TIMSS places equal emphasis on science and mathematics in each survey, PISA has a system in which one of the three core domains is allocated more time every third cycle. One consequence of these differences in how the tests are constructed is that TIMSS includes a far greater number of total items in each domain. For instance, when science was the major domain in PISA 2006, a total of 109 items covered the domain, while TIMSS always has more than 200 items in each of the two domains.

There are other similarities and differences between the two assessments, but the aforementioned are the most relevant differences in terms of factors that directly or indirectly affect the standards they have developed (for a more detailed comparison of the two assessments, see Olsen 2005). In oversimplified terms, while PISA has a more future-oriented goal closely related to monitoring the sustainability and development of society, TIMSS aspires to study the educational effectiveness of factors proximal to what happens within classrooms.

5.3 Standard Setting Procedures

In education, the term *standard* refers to a range of different phenomena. First, standards are often used to refer to formulations of expectations. In official curriculum documents, the intended aims of the education system are described through content or competency standards. In some countries or jurisdictions, schools have to meet expectations of average performances to be achieved, and at the system level, expectations of future performance may also be defined in relation to international surveys. These expectations are often referred to as *standards* or *benchmarks*. Another use of the term *standard* refers to agreed-upon quality criteria for certain objects, such as in standards for teaching, standards for assessments, etc.

Here, we refer to a family of meanings that are related to both standards as expectations and standards as quality descriptions. Both types of standards are related to measuring performance, proficiency, or achievement within some domain of relevance for education (such as science). For our purposes, standard setting may in its widest sense be defined 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" (Cizek 1993, p. 100). In other words, standard setting refers to procedures that are implemented in order to identify points or intervals along a scale designed to measure student achievement within a specified domain. In what follows, the associated verbal descriptions of these points or discrete levels along the scale are regarded as parts of the standard setting procedure.

In the literature, these are often referred to as *achievement* or *performance level descriptors* or *PLDs* (Egan et al. 2012; Perie 2008). Over the last decades, standard setting has emerged as a response to several types of questions or purposes. First, standard-setting procedures have been initiated in order to provide a more rational and judicial basis for pass/fail decisions. This could, for instance, be for certification purposes aiming to ascertain that persons entering into a profession meet a standard regarded as appropriate. In this case, the procedure would involve identifying a specific cut score on an assessment. Second, particularly in the US context, standard setting serves to promote and develop criterion-based assessments, as opposed to a simple reference to a norm or a distribution. Numbers in the form of percentage correct, percentile ranks, etc. alone do not communicate what students know or are able to do—they simply state that a student is relatively more or less able as compared to a distribution of items or students. Third, in systems with several exam providers, such as the UK, regulatory processes have been installed to ensure that the exams are comparable across the different providers; this process is also referred to as *standard setting*.

Many education systems seek to ensure that standards are maintained over time (i.e., that the numbers used to report student performance in one year have a rational basis for comparison with the apparently similar numbers used in the past and in the future). Hence, standard setting is intimately related to purposes of linking and equating scores.

Standard setting procedures usually rely on judgments by panels of content or subject matter experts (e.g., teachers). These experts are tasked with deciding where along the scale they find it meaningful (based on theory and/or tacit expert knowledge) to create a cut score. The great number of specific procedures used to organize the work of such panels may be grouped into two distinct approaches*, item-* or *person-centered*, depending on whether the procedure primarily involves judging items or test takers (for details see, for instance, Cizek and Bunch 2007; Zieky et al. 2008).

Although most of these methods were originally developed within the framework of classical test theory, they are now increasingly implemented using item response theory (IRT). In particular, for several of the item-centered methods, the advantage of using IRT is that students' proficiency is placed on the same scale as the difficulty estimates of the items, often referred to as *item maps* or *Wright maps* (see Fig. 5.1). This enables development of verbal and probabilistic descriptions of

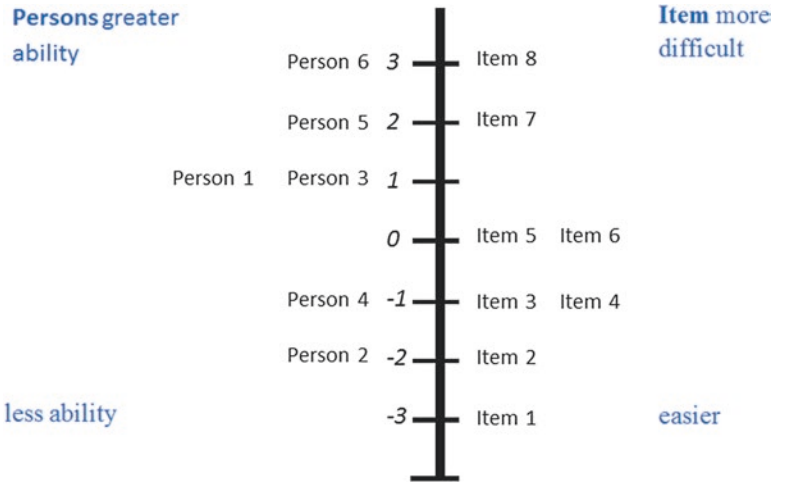


Fig. 5.1 A generic example of a person–item map

the proficiencies demonstrated by a typical student at different points or within different intervals along the scale. These methods are therefore often referred to as *item mapping* (Huynh 2009).

Figure 5.1 contains a generic and simple example of an item map. As stated above, with the help of IRT, the difficulty of the items and the ability of the students are placed along the same continuous underlying scale. The scale ranges from easier items and students with less ability at the bottom to more difficult items and students with more ability at the top of the scale. A default option in most IRT applications is to locate the items on the point of the scale where students have a 50/50 chance of succeeding. This default option is also referred to as *RP50*[2]. For example, in Fig. 5.1, Person 4 has a 50% probability of providing a correct response to Items 3 and 4 and an even higher probability of success on Items 1 and 2. However, not everyone would agree that a 50% chance of responding correctly to an item represents mastery. A somewhat stricter criterion of at least an 80% chance (RP80) could be perceived as more useful in some contexts. This adjustment is easily accommodated and would simply result in items being shifted upward in the person–item map.

5.4 Standard Setting Procedures in TIMSS and PISA

The methods used in PISA and TIMSS are based on the interpretation of person–item maps. To a large extent, the procedures applied in both studies, particularly in TIMSS, are based on those first implemented as part of the National Assessment of

[2] *RP* refers to *response probability*.

Educational Progress (NAEP) called *scale anchoring* (Beaton and Allen 1992). Both PISA and TIMSS use the following procedures:

1. Expert groups write frameworks explicating to some degree the construct being measured, including a generic hypothesized notion of the characteristics of performance from low to high on the scale.
2. Items are developed and implemented according to the specifications in the framework.
3. Item writers and expert groups develop item descriptors (IDs), which may include coding the items according to the categories used in the framework and open-ended statements with specific descriptions of the knowledge and skills involved in solving the item
4. Data are analyzed, parameters for students and items are extracted (using IRT), and graphics and tables with data (as in Fig. 5.1) are produced.
5. A (pragmatic and empirically based) decision is made about the number and location of cut scores to be used.
6. Items are identified as markers of the performance levels to be reported.
7. Performance level descriptors (PLDs) are developed based on detailed descriptions of the clusters of items identified (see Step 3) and the general description of the construct included in the framework (see Step 1)

The major difference between the procedures used in PISA and TIMSS as compared to the standard setting procedures described in the literature is that identifying cut scores is not based on a process involving a panel of judges. Instead, the practice is rooted in the premise that it is not possible or meaningful to derive substantial qualitative descriptions of thresholds along the scale from explicit or implicit theory alone. The scales are continuous and unimodal, suggesting that any cut score is equally meaningful in a qualitative sense. Decisions about the number and location of cut scores are therefore solely based on a combination of pragmatic criteria regarding usefulness for communication and empirical criteria. However, expert judgments are still vital to the process, particularly in Steps 1, 2, 3, and 7. Although PISA and TIMSS are very similar in their approach to standard setting, there are important differences between how their cut scores are developed and communicated.

Figure 5.2 provides a more detailed description of the nature of the standard setting in TIMSS and PISA. The details of the procedures are described in the technical reports (see, for instance, the latest versions: Mullis 2012 ; OECD 2014). The figure illustrates that PISA identifies more cut scores than TIMSS: six[3] and four, respectively. Another major difference is that PISA uses the cut scores to define intervals of proficiencies, while TIMSS defines what is referred to as *anchors* along the scale; these anchors are interpreted as fuzzy points. Another striking difference is that TIMSS has placed the cut scores along some preselected and well-rounded

[3] This number is typical, but five and seven cut scores have also been used in PISA.

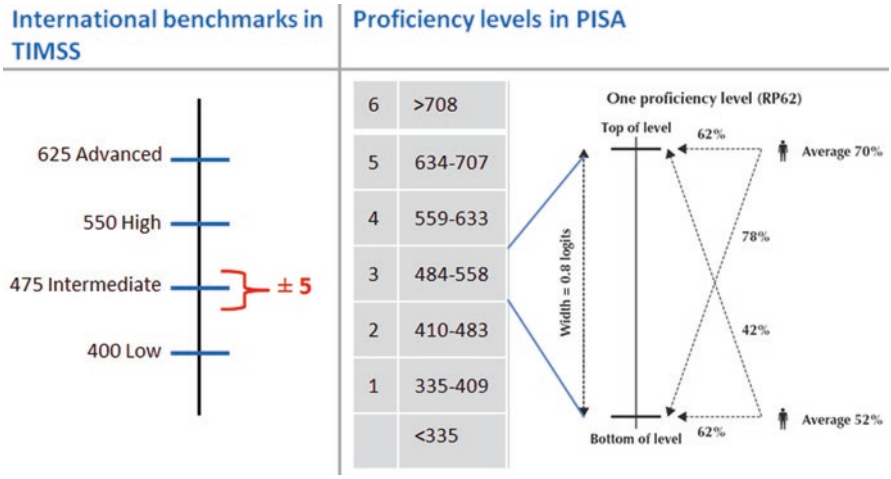


Fig. 5.2 Principles for deciding cut scores in TIMSS and PISA. The numbers for PISA refer to the science scale. Note: The right-hand figure is copied from OECD (2014, p. 293)

values on the scales, while PISA has applied another criterion for placing the cut scores, resulting in somewhat irregular values. In addition, the exact location of the cut scores is generally not equal across the domains in PISA, while TIMSS operates with the same cut scores across mathematics and science. It is interesting to note that the distance between two adjacent cut scores are rather similar in the two studies, representing about 75 points (corresponding to 75% of one standard deviation unit in the internationally pooled sample).

5.4.1 Defining the International Benchmarks in TIMSS

Standard setting is called scale anchoring in TIMSS, referring back to procedures first developed for the NAEP (Beaton and Allen 1992) and implemented for the first time in TIMSS 1999 (Gregory and Mullis 2000; Kelly 1999). Initially, these anchors (or international benchmarks) were placed on a percentile scale. However, for the 2003 assessment, the test centre realized that in order to report trends, they needed to reference defined points on the underlying scale (Gonzalez et al. 2004). The values in Fig. 5.2 have been in use ever since.

The anchoring process in TIMSS begins by identifying students who score within five scale-score points of each cut score. For these students, the percentages correct are computed for all items. Several criteria are then used to identify item anchoring at the different benchmarks. First, for a multiple-choice item to anchor at a specific benchmark, at least 65% of the students in the benchmark interval must answer it correctly. Additionally, less than 50% of the students belonging to the next

lower benchmark must respond correctly[4]. For open-ended response items, the criterion is to place the item in the lowest of the benchmarks with at least 50% correct responses[5].

5.4.2 *Defining Proficiency Levels in PISA*

The procedure used in PISA identifies a set of equally spaced intervals along the scale. The starting point for defining these levels is the idea that students at a particular level will be more likely to solve tasks at that level than to fail them. Students are therefore assigned to the highest level in which they are expected to correctly answer the majority of the assessment items. Then, a pragmatic choice is made for the width of the equally spaced intervals[6]. The last procedural step in setting up the proficiency levels in PISA is to place the lower end of Level 1 at the lowest score point possible given the requirements above. In practice, using a response probability of 62% (RP62) produces intervals with these properties. As illustrated in Fig. 5.2, by using RP62, a student at the lower end of any proficiency level is expected to give correct responses to more than 50% of the items belonging to this interval. A student at the very top of a level is expected to respond correctly to approximately 70% of the same item set.

Specific arguments about the width of the intervals applied in PISA and the distance between adjacent benchmarks in TIMSS are, to our knowledge, not explicitly documented. However, it may be reasonably assumed that these choices are affected by what is perceived to be a useful number of categories for reporting combined with the limitations given by the total number of items. The latter is important to consider because the end products of the standard-setting process in TIMSS and PISA are not a set of cut scores. Having identified these, the next step is the development of verbal descriptions of what students know and are able to do at different levels of the construct. Hence, a fair number of items are needed at each benchmark or within the proficiency levels in order to develop meaningful descriptions (Step 7 in the list above).

5.5 From Items to PLDs

As identified in the list of steps involved in the standard-setting procedures in PISA and TIMSS, items are described by IDs that reflect both generic categories used to define the construct in the framework and the specific content and cognitive demand of each item. This is the raw material used to generate PLDs. In the following

[4] This discrimination criterion for the low international benchmark could not be applied for obvious reasons.

[5] In addition, a less strict criterion was used to identify items labeled “almost anchored.”

[6] The exception is the categories at each of the ends, which are unbounded.

sections, we use the domain of science in both assessments to exemplify this process. However, the points we make are generalizable to any domain in the assessments, and they serve as examples for our discussion on the choices made when generating PLDs from item maps.

5.5.1 IDs and PLDs in PISA and TIMSS

Figure 5.3 contains examples of one item from PISA and one item from TIMSS 8th grade. Both items belong somewhere above the middle of the scale in the item map, with percentages correct at 43% internationally. The PISA item is one of the items that define Level 4, while the quite similar TIMSS item anchors at the high international benchmark. The specific statement used to describe the TIMSS item is "Recognizes the major cause of tides" (Martin and Mullis 2012). In the international report from PISA, the item is presented as follows: "This is a multiple-choice item that requires students to be able to relate the rotation of the earth on its axis to the phenomenon of day and night and to distinguish this from the phenomenon of the seasons, which arises from the tilt of the axis of the earth as it revolves around the sun. All four alternatives given are scientifically correct" (OECD 2004, p. 289). In addition, the listing identifies the item as belonging to certain categories in the content and the procedural dimensions defined in the frameworks for the assessments[7].

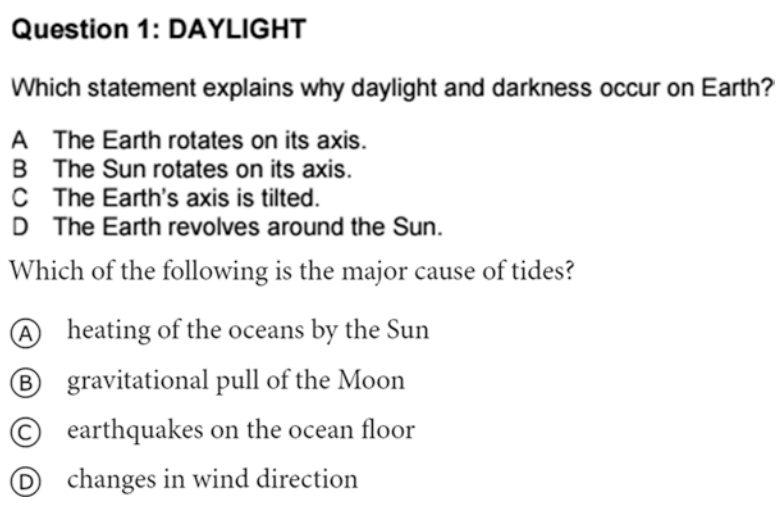


Fig. 5.3 Examples of items from PISA and TIMSS (the upper question is from PISA 2003, the lower question is from TIMSS 2011)

[7] In PISA, this aspect of the construct is defined by three competencies, and in TIMSS, by three cognitive domains.

These specific texts in the form of items are transformed into content-specific claims about what students with success on the items are able to do, or *IDs*, as we have coined them.

This stage involves some degree of generalization and removal from the original item-specific information. Finally, the full set of statements are reviewed, reduced, and synthesized into more overarching statements, PLDs, which express students' capabilities at discrete levels through a process involving consensus among subject matter experts.

There are some obvious differences in the two assessments' PLDs (see Table 5.1 for examples). The PLDs developed from TIMSS are longer and more detailed; furthermore, they refer more specifically to the content covered by the items. Also, the PLDs in TIMSS, given the more item-dependent language used, are not identical from one assessment to the next, while the statements used in PISA are almost identical over time. Although the PLDs in TIMSS also refer to what students are able to do with their knowledge ("compare," "contrast," etc.), the PLDs in PISA have a unique focus on such procedural aspects and include more generic competencies such as "reflect" and "communicate."

These differences reflect the divergent definitions and operationalizations of the domain of science in the two assessments. TIMSS has a framework with a high degree of content specification that is based on analyses of curricula in the participating countries. PISA is concerned with what students at the age of 15 are able to do in situations where an understanding of and about science (as a knowledge-generating process) is needed. The stimulus and items are therefore crafted to be less dependent upon very specific content knowledge.

Table 5.1 Examples of PLDs in PISA and TIMSS

PISA Level 4 (559–663)	TIMSS High (550)
"At Level 4, students can work effectively with situations and issues that may involve explicit phenomena requiring them to make inferences about the role of science or technology. They can select and integrate explanations from different disciplines of science or technology and link those explanations directly to aspects of life situations. Students at this level can reflect on their actions, and they can communicate decisions using scientific knowledge and evidence" (OECD 2007, p. 43)	"Students apply their knowledge and understanding of the sciences to explain phenomena in everyday and abstract contexts. Students demonstrate some understanding of plant and animal structure, life processes, life cycles, and reproduction. They also demonstrate some understanding of ecosystems and organisms' interactions with their environment, including understanding of human responses to outside conditions and activities. Students demonstrate understanding of some properties of matter, electricity and energy, and magnetic and gravitational forces and motion. They show some knowledge of the solar system, and of Earth's physical characteristics, processes, and resources. Students demonstrate elementary knowledge and skills related to scientific inquiry. They compare, contrast, and make simple inferences, and they provide brief descriptive responses combining knowledge of science concepts with information from both everyday and abstract contexts" (Martin et al. 2012, p. 83)

5.5.2 The Number and Nature of PLDs

The aim of assessments like PISA and TIMSS is to develop solid measures of students' proficiency in a few defined domains. At the outset, the items selected for the assessments are standalone, single observations of situations in which we can reasonably assume that students' overall ability on the measured trait is involved. However, single items are very unreliable observations of these abilities. The items involve unique content, make use of idiosyncratic language and representations, and various response formats. For a typical test, simple isolated right/wrong items have a point biserial correlation with the overall test score in the order of 0.3–0.4. In psychometric terms, this means that only 9–16% of the variance for an item can be seen as "true" variance related to the common trait being measured, whereas the major portion of the variance is residual variance (Olsen 2005). The obvious question regarding the PLDs developed from tests as part of a standard-setting process is to what degree we should include and rely on item-dependent information in the proficiency level descriptors. After all, the proficiencies we seek to describe are regarded as independent of the actual items included in the assessment.

Another important decision to be made is how many cut scores to identify and how to use them to assemble performance levels. In addition to reflecting the purpose of the PLDs, this decision is contingent upon the number of items available and how they are distributed across the scale. For many reasons, TIMSS has almost twice as many items in each domain as even the major domain in each PISA cycle. In this respect, TIMSS has a more favorable starting point for the process because more items mean more information to potentially include in the item maps. TIMSS has taken advantage of this by describing students' proficiency at or close to a few points or benchmarks on the scale. In this process, items with very similar difficulties are identified and clustered, which enables the development of PLDs by aggregating and synthesizing information across a set of data points with shared properties. This also allows for the development of PLDs that are well separated along the continuous scale. As a result, PLDs with a clear progression from one level to the next are produced. However, this method also leads to excluding items that do not meet the strict anchoring criteria; in fact, only half of the items fully satisfy the criteria. But, by including items that are almost anchoring, the number of items used to develop most of the PLDs in TIMSS is relatively large and should constitute robust data in the process. PISA, given its more limited amount of items, opts to describe intervals along the whole scale; thus, it includes all of the items in its process of extracting PLDs. Given this relatively low number of items, the decision to develop PLDs for six distinct levels on the scale seems rather ambitious.

We have not seen an overview of the number of items in the different levels in PISA, but assuming that the item difficulties resemble a normal distribution, we suggest that the number of items in the top and bottom levels is rather low. For the reading assessment developed for PISA 2009 and the mathematics assessment for PISA 2012, efforts were made to include a larger number of easy items. This was a well-reasoned improvement, given that the cut score between Level 1 and Level 2

receives policymakers' attention. Another issue related to the high number of PLDs in PISA is the risk that the progress in proficiency involved in advancing from one level to the next may not be that easy to grasp when reading the PLDs from low to high. For the same reason, Perie (2008) recommends using no more than four levels.

Thus far, we have established that the standard setting processes in PISA and TIMSS more or less follow the same principles: item maps are produced, IDs capturing both highly item-specific information and the more generic aspects involved in the construct are formed, and PLDs are extracted through an expert consensus process. Figure 5.4 illustrates this step in the standard setting process as a continuous scale ranging from completely item-specific statements to descriptions of generalized proficiencies. TIMSS has developed PLDs with a closer reference to the content of the items than PISA. PISA has developed PLDs with generic statements more closely resembling a theory of what constitutes progress in scientific literacy.

We argue that depending on the number of items at hand, the purpose of the assessment, and the intended use of the reported results, a decision should be made regarding where on this continuum it is possible and useful to target the PLDs. To the far left of this spectrum (see Fig. 5.4) are extremely item-specific PLDs, for instance in the form of a listing of all of the IDs. To the far right are very generic PLDs, for instance, in the form of simple labels such as low, intermediate, and advanced with only short and unspecific descriptions. The very item-specific information available in PLDs toward the left-hand side of the figure, could provide teachers with relevant subject matter information to be used in their formative practices. However, PLDs at this end are less robust in that they are more contingent on the actual items included in the test. Descriptors belonging to the right-hand side in Fig. 5.4 are less dependent upon the actual items included in the test and could for instance be used to communicate more generalized understandings of what constitutes performance on different levels in the construct being measured. Such PLDs could serve grading purposes and they could also potentially be used in assessments where learning progressions over longer time-spans are monitored.

Given the very high number of items available, we suggest that TIMSS should develop PLDs that are more generic and stable over time. After all, the assessment aims to report measures that are linked from one assessment to the next. For PISA, we suggest that more items are needed in order to develop PLDs in their current generic form. One possibility to remedy this situation is to create a new standard-setting process in which all available assessment material used since the first assessment in 2000 is assembled into one item map. With the 2015 assessment, all three domains in PISA have served as major domains twice, and pooling all of the item

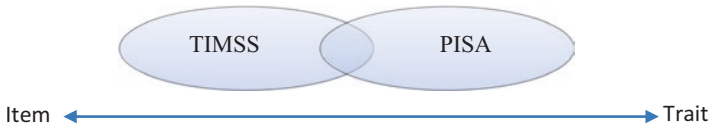


Fig. 5.4 Item versus trait specificity in TIMSS and PISA

information across assessments would significantly increase the number of IDs available for generating PLDs.

5.6 Possible Implications for the Norwegian Context

Other chapters in this book include more explicit descriptions of the current use of standard setting in the Nordic countries. Thus, here, we focus on suggestions for why and how standard setting where PLDs are developed for reporting purposes should be considered in the Norwegian context. With our partial knowledge of the situation in the other Nordic countries, we assume that this discussion is relevant for other Nordic countries. We first discuss some issues related to the national assessments before returning to issues related to the interpretation of grades in exams. Both types of assessments are reported using a limited number of reporting categories organized from lower levels to higher levels of achievement. In this sense, both assessments use standards established in an empirical setting.

5.6.1 PLDs in National Assessments

Explicitly formulated PLDs have already been created for the Norwegian national assessments[8]. The assessments are low stakes and are conducted at the beginning of 5th and 8th grade in the cross-curricular domains of reading, numeracy, and English. A description of how these PLDs were developed is, as far as we know, not publicly available. Without going into details of the nature of these PLDs, some similarities and some differences exist across the three domains. They all describe three levels for 5th grade and five levels for 8th grade. Originally, these cut scores were determined via specific percentiles. The PLDs in the reading domain resemble those used in PISA, while those developed for numeracy and English include more content-specific statements. Numeracy operates with overarching and generic PLDs in addition to a list of very content-specific statements.

We suggest that the methods applied for standard setting in the international assessments could be helpful in revising and document a transparent basis for the current PLDs. After initially using classical test theory and percentiles as the basis for reporting, all of the national assessments in Norway are now being developed within an IRT framework. Moreover, the assessments are now linked over years to support interpretations of trends. With these changes, new challenges and possibilities for standard setting have emerged. In the process of revising the standards we suggest that two issues should be emphasized. First, with the new scales developed to link performance over time, there is a need for robust descriptions at a more

[8]The actual descriptions are available in Norwegian from http://www.udir.no/Vurdering/Nasjonale-prover/

general level without reference to specific item content (also suggested for TIMSS). Furthermore, it is now possible and potentially very helpful to develop a joint item map for each assessment domain and each grade level in the national assessments, including the complete item material from several years of testing. This would help producing even more IDs, which would be particularly helpful for developing more robust PLDs for low- and high-performing students (in line with what we recommend for PISA). Furthermore, the progress from 5th to 8th grade, possibly extending to include 11th grade[9], should be explicitly modeled in the new PLDs. Standards with such a vertical scaling perspective are more challenging to develop because aligning PLDs across grade levels must be taken into account.

5.6.2 New Standard-Based Exams?

Given that the exams have multiple purposes, are high stakes for students, and are laborious and resource-intensive processes, it is unfortunate that the grading system appears to be unfixed and allows for inconsistencies and arbitrariness. A few examples supporting this claim can be found from official statistics reported in yearly national publications, e.g., The Norwegian Directorate for Education and Training (2014):

- Half of all pupils achieve lower written exam results than they do coursework grades in the same subject.
- The difference between coursework grades and exam grades varies systematically across schools.
- Even though the general descriptions of grades are the same for all subjects, the variation in average grades across subjects is large.
- Average grades, particularly for exams, vary over years

These observations illustrate that not all aspects of current grading practices are well understood. Assigning grades to students is defined as a judicial act, and these examples indicate a lack of transparency in current grading practices. Establishing more robust standards could be one helpful way to improve the situation.

However, standard setting in this situation is far more complex than setting cut scores and extracting PLDs. First, grading coursework typically includes evaluating products, not just assessments in the form of standard tests. Second, grades are formally defined to represent the degree to which the students have demonstrated mastery of the intended curriculum. In reading the curricular aims for a subject, it is quite evident that they are not formulated to reflect a unidimensional trait that lends itself to measurement on a single scale. Instead, most appear piecemeal with non-related descriptions of knowledge and processes that students should master.

A recent committee touched upon this issue in their series of white papers discussing the future of the Norwegian education system (NOU 2014:7, 2015:8). In

[9] Similar assessments are available for 11th grade, but they are not compulsory.

these reports, they recommend developing systems that support deep learning and learning progression. They do not explicitly state how learning progressions should be formulated or achieved, but in order to support progression, the formal curriculum needs revision. Care should be taken in reformulating curricular aims with a clear conceptual progress across grades. It is unreasonable to expect that subject matter expert groups working in isolation could formulate curriculum standards with such properties. Standard setting procedures, including collecting and analyzing empirical data in some form or another, are needed to support this process.

We do not claim that all of the issues related to the complexity of grading students in school may be fixed by simply performing one or several standard-setting procedures. However, as exams are already very systematic and large-scale logistic operations, it is possible to collect data and develop item maps as described above. This could constitute the first small step toward a more robust foundation for grading in the Norwegian school system.

References

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics, 17*(2), 191–204.

Beaton, A. E., & Zwick, R. (1992). Overview of the national assessment of educational progress. *Journal of Educational and Behavioral Statistics, 17*(2),95–109.doi:10.3102/10769986017002095.

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement, 30*(2), 93–106. doi:10.1111/j.17453984.1993.tb01068.x.

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

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting : A guide to establishing and evaluating performance standards on tests*. Thousand Oaks: Sage Publications.

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

Gonzalez, E. J., Galia, J., Arora, A., Erberber, E., & Diaconu, D. (2004). Reporting student achievement in mathematics and science. In M. O. Martin, I. V. S. Mullis, & S. J. Chrostowski (Eds.), *TIMSS 2003 technical report* (pp. 274–307). Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College.

Gregory, K. D., & Mullis, I. V. S. (2000). Describing international benchmarks of student achievement. In M. O. Martin, K. D. Gregory, & S. E. Stemler (Eds.), *TIMSS 1999 technical report* (pp. 265–276). Chestnut Hill: International Study Center, Boston College.

Huynh, H. (2009). Psychometric aspects of item mapping for criterion referenced interpretation and bookmark standard setting. In E. V. Smith Jr. & G. E. Stone (Eds.), *Criterion referenced testing: Practice analysis to score reporting using Rasch measurement models* (pp. 148–159). Maple Grove: JAM Press.

Kelly, D. L. (1999). *Interpreting the Third International Mathematics and Science Study (TIMSS) achievement scales using scale anchoring.* (PhD), Boston College, Boston.

Martin, M. O., & Mullis, I. V. S. (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., Mullis, I. V. S., Foy, P., & Stanco, G. M. (2012). *TIMSS 2011 International results in science*. Chestnut Hill: TIMSS & PIRLS International Study Center.

Mullis, I. V. S. (2012). Using scale anchoring to interpret the TIMSS and PIRLS 2011 achievement scales. In M. O. Martin & I. V. S. Mullis (Eds.), *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschof, C. (2009). *TIMSS 2011 assessment frameworks*. Chestnut Hill: TIMSS & PIRLS International Study Center Lynch School of Education, Boston College.

NOU. (2014:7). *Elevenes læring i fremtidens skole. Et kunnskapsgrunnlag [Pupils' learning in the school for the future. A knowledgebase]*. https://nettsteder.regjeringen.no/fremtidensskole/

NOU. (2015:8). *Fremtidens skole. Fornyelse av fag og kompetanser [A school for the future: Renewing school subjects and competencies]*. https://nettsteder.regjeringen.no/fremtidensskole/

OECD. (2001). *Knowledge and skills for life. First results from PISA 2000*. Paris: OECD Publications.

OECD. (2004). *Learning for tomorrow's world. First results from PISA 2003*. Paris: OECD Publications.

OECD. (2006). *Assessing scientific, reading and mathematical literacy. A framework for PISA 2006*. Paris: OECD Publishing.

OECD. (2007). *PISA 2006: Science competencies for tomorrow's world*. Paris: OECD Publishing.

OECD. (2014). *PISA 2012 Technical report*. Paris: OECD Publishing.

Olsen, R. V. (2005). *Achievement tests from an item perspective. An exploration of single item data from the PISA and TIMSS studies, and how such data can inform us about students' knowledge and thinking in science*. Oslo: Unipub forlag.

Perie, M. (2008). A guide to understanding and developing performance level descriptors. *Educational Measurement: Issues and Practice, 27*(4), 15–29. doi:10.1111/j.1745-3992.2008.00135.x.

Smith Jr., E. V., & Stone, G. E. (Eds.). (2009). *Criterion referenced testing: Practice analysis to score reporting using Rasch measurement models*. Maple Grove: JAM Press.

The Norwegian Directorate for Education and Training. (2014). *The education mirror 2014: Facts and analysis of kindergarten, primary and secondary education in Norway*. Oslo: The Norwegian Directorate for Education and Training http://www.udir.no/globalassets/upload/rapporter/educationmirror/the-educationmirror_english.pdf.

Zieky, M. J., Perie, M., & Livingstone, S. A. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests*. Princeton: Educational Testing Service.

Chapter 6
In the Science and Practice of Standard Setting: Where Is the Science??

Barbara Sterrett Plake

Abstract Standard setting is a complex process that involves the application of social science, psychometrics, content expertise, politics, and economics. Over the last 60 years of the practice of standard setting, many methods have been proposed, and many implementation decisions have been made that affect the practice of standard setting. Some of these decisions have been made based on scientific studies about their impact on the standard setting results, but many have been made purely on factors of human judgment or for streamlining the process without the benefit of research to support these decisions. The purpose of this chapter is to focus on where additional research is needed to support many of the practical decisions that are found in many standard setting applications.

Keywords Standard setting • Performance level setting • Research

6.1 Introduction

Standard setting is a complex process that involves the application of social science, psychometrics, content expertise, politics, and economics. Over the last 60 years of the practice of standard setting, many methods have been proposed, and many implementation decisions have been made that affect the practice of standard setting. Some of these decisions have been made based on scientific studies about their impact on the standard setting results, but many have been made purely on factors of human judgment or for streamlining the process without the benefit of research to support these decisions. The purpose of this chapter is to focus on where additional research is needed to support many of the practical decisions that are found in many standard setting applications.

B.S. Plake (✉)
University of Nebraska-Lincoln, Lincoln, NE, USA
e-mail: bplake@unl.edu

S. Blömeke, J.-E. Gustafsson (eds.), *Standard Setting in Education*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-50856-6_6

This chapter will follow the sequence of practices within a judgmental standard setting study. The chapter expands on and updates the Plake 2008 article that articulated a research agenda for standard setting (Plake 2008). This chapter will address research in these practices for (a) selection of panelists, (b) panelists' training, (c) panelists' operational ratings, (d) feedback provided to panelists' during their operational ratings, (e) cross panel/vertical articulation of panelists' cuts scores, (f) policy smoothing of cut scores, and (g) the uses of cut scores. It will also address challenges to conducting standard setting research and propose some solutions to these challenges.

6.2 Selection of Panelists

Central to the validity of cut scores from a judgmental standard setting study is the ratings provided by panelists, and central to the validity of these ratings is the composition of the panel(s) who provide these ratings. In most standard setting applications, the composition of the panel is designed to reflect the stakeholders who will use the cut scores for policy and classification purposes. However, in some standard setting practices, policy makers have instead insisted that the panel be reflective of the population for whom the cut scores will be applied. For example, in standard settings in educational settings, some applications of standard setting studies have required that the panel be representative of the intended student population. This sometimes creates a disjunction between the characteristics of the population of potential panelists and the characteristics of the student population. This is also the case in some professions when the membership within the candidate population is not congruent with the available population of panelists. In the extreme, this creates some blatantly inappropriate characteristics for the panelists. For example, in educational settings where the intent is to set cut scores for alternative assessments designed for students with significant cognitive disabilities, it would be absurd to compose the panel with members of that population. Policy makers should be mindful of the implications of policy statements regarding the composition of the panels to ensure such incongruent statements about the relationship between the characteristics of the panels and the candidate populations are not presented.

Another issue regarding the composition of the panel is whether, and how, all relevant stakeholders should be involved in the standard setting process (Kane 2001). In some standard setting applications (NAEP is an example) the membership of the panel, by policy, must include members of the public (Loomis and Bourque 2001). In other applications, especially in educational settings, members of the business community, post-secondary educators, and parents are included in the standard setting panels. Although the intent of such diverse standard setting panels is often to engender buy-in from these groups, and/or to ensure these groups have a voice in the decisions about the cut scores, the inclusion of these groups can be detrimental to the successful implementation of a judgmental standard setting study. This is in part because it is critical in judgmental standard setting studies that the panel

members have a clear understanding both of the candidates for which the cut scores are to be applied and the curriculum/educational experience of these candidates (Hambleton and Pitoniak 2006). Further, the panelists need to be able to determine the intricacies of the individual test questions. If the panelists do not have a firm foundation in the content of the test material, it will be nearly impossible for them to estimate the performance of target candidates on the questions that comprise the test. Further, these non-discipline panelists may have hidden (or maybe not-so-hidden) agendas regarding the placement on the cut scores.

Business members may feel pressure to provide rigorous cut scores in order to reduce their costs for training recruited personnel, and parents may want to impose less rigorous cut scores in order to be protective of their children. Often policies that add these non-discipline stakeholders to the standard setting panels are frequently created when the decision stakes for the candidates or agency are especially high (such as the highly public results from NAEP assessments or high school graduation tests). Unfortunately, by including these non-discipline panelists, the resulting cut scores may suffer from invalidity when the intent is typically to enhance validity, but in many cases that validity is often just face validity, not construct-related validity.

I have served on technical advisory committees for several states and consortia. As part of that experience, I have reviewed multiple standard setting studies and proposals. The examples below are from this experience. To protect confidentiality, I have not indicated the states in which these examples are derived.

Some researchers have tried to implement standard setting procedures with the full population raters, or with untrained raters. In one instance, in an attempt to use teachers of students in a state-wide standard setting study for an Algebra I end-of-course test, teachers state-wide were asked to participate in the standard setting study. In this case, a contrasting groups type approach was used. Teachers were asked, prior to their students taking the test, to classify their students into "proficient" and "not proficient" categories. Following the administration of the test, distributions for the students classified by their teachers as proficient or not proficient were compared. These distributions were fully overlapping, indicating, based on the teachers' categorizations, there was no difference in overall test performance between the students classified by their teachers as proficient and not proficient. However, there were fundamental flaws in the study that affect interpretations of the results. For example, teachers were not given the performance level descriptions for "proficient" and "not proficient," did not discuss the characteristics of students at the borderline for the proficient category, and were not provided any feedback on their ratings.

Another attempt to use a full population occurred recently with the standard setting study for Smarter-balanced consortium assessments. In this instance a "crowd sourcing" model was used as one component of the standard setting process. A website was set up for the general population to use to make item ratings. There were several problems with the implementation, including issues with the website developed to let people access the site. More problematic from a validity perspective is that these "raters" were not given any training, therefore the validity and interpretability of their ratings is highly questionable. To my knowledge this is the first time such a crowd-sourcing full-population panel was attempted for a standard

Box 6.1 Research Questions: Selection of Panelists

1. What is the effect on standard setting results by including non-stakeholders and members of the public on the panel?
2. How does using full population, crowd-sourcing techniques with untrained raters impact results from a standard setting procedure?
3. How does having a panel set standards for adjacent grades affect results from a standard setting procedure?

setting study. Research is needed to study the utility of ratings derived from untrained raters.

Along with the selection of panelists, another decision, often in educational settings, is to have panelists consider more than one grade level for setting multiple cut scores across multiple grades. With the federal requirement that all students in grades 3–8, for example, be tested in reading and mathematics, many states and consortia are required to set these multiple cut scores. Some standard setting efforts have one panel (sometimes split into two parallel panels) set standards for two adjacent grade levels. This is seen as advantageous because the total number of panelists is reduced, and there can some continuity in the standards set for these two adjacent grades (even though there will typically be a vertical articulation following the setting of the individual grades' cut scores). However, there is also the risk of having a problem panel or a dominant panelist influence the standards set for two grades. There has not been any research into whether the standards set by a single panel across adjacent grades are similar to those set by independent panels. Because the standards set by these panelists in educational settings have high stakes implications, research is needed to support procedures that result in valid cut score interpretations and use (Box 6.1).

6.3 Panelists' Training

Because, as stated previously, the validity of the resultant cut scores relies directly on the ratings provided by the panelists, the preparation that the panelists receive is critical to the validity of the cut scores from a judgmental standard setting study. There are several critical incidences where training should be provided to the panelists. The first opportunity for training of the panelists occurs at the time of recruitment. Unfortunately, this is sometimes turned over to a third party or a staff assistant who is not well-versed in the purposes of the recruitment. In these cases, panelists are sometimes asked to participate in an effort designed to improve the tests or related to assessment practices. These panelists then have a mistaken understanding of their task and sometimes are confused during the initial phases of the standard setting study because there is a disjuncture between their expectations and actual

tasks to be completed. It is very important, as well, that the panelists have a full understanding of the time and effort expectations.

In addition to fully explaining to the panelists the purpose of the project and the expectations, in some cases panelists have been provided with pre-meeting materials. These materials are sometimes made available on websites and often contain documents with performance level descriptors, orientation materials, etc. Even when there are clear instructions to the panelists about how they should process these materials in advance of the meeting (in some cases the materials that the panelists interact with is noted by the system), often not all panelists follow through reading these materials. Further, even if they do read these materials, many times their understanding is incomplete, incorrect, or confused. Typically, there is not an opportunity for the panelists to ask questions or benefit from the questions posed by other panelists. Because the coverage of these advance materials in uneven, and because it is essential that all panelists have a full understanding of these materials, it is typically necessary to repeat these materials, with full discussion and deliberations, at the standard setting meeting. Research is needed on the efficacy of providing advance materials to panelists.

Training is a critically important part of the standard setting activity. Current training practices consists of multiple components: presenting to the panelists a general overview of the standard setting process, having panelists take the test, leading a discussion of the general performance level descriptors (PLDs) and the development of performance level descriptors of candidates at the borderline for each category (called *borderline performance level descriptors, BPLDs*), providing panelists with a practice task to experience the process of making their ratings. In some cases, panelists also receive during training indications of the kinds of feedback that they will receive across multiple rounds of standard setting. Although there is clearly some prescribed order in these components (it makes sense that the orientation comes first, for example, and that the practice activity comes last), research is needed on whether the order of these components makes a difference in the panelists' understanding of the tasks they are to undertake in a standard setting study (Box 6.2).

Box 6.2 Research Questions: Panelists' Training

1. How do panelists' preliminary ideas about what they are going to be doing affect their ratings and evaluations of the standard setting procedure?
2. What is the impact of pre-meeting materials on panelists' ratings and evaluations of the standard setting procedure?
3. How does the sequence of training affect the panelists' ability to provide appropriate ratings during the standard setting procedure?

6.4 Panelists' Operational Ratings

Because there are many different standard-setting methods, each with their own procedures, there are a variety of research questions that are relevant regarding panelists' ratings. When the standard setting method focuses on the test questions themselves, and focuses the panelists' ratings on how candidates at the borderline of each performance category will likely answer the test questions correctly (e.g., Angoff 1971; Bookmark, Lewis et al. 2012), research could focus on several features of these ratings. For example, in some situations where there are multiple performance categories, for policy reasons some standard setting studies have instructed panelists to begin their rating process by considering the most important category (in educational settings, this is often the cut score between Proficient and Not-proficient). Research is needed to support this policy decision, considering whether panelists will systematically differ in their ratings depending on which category they rate first. In addition, research is needed on whether, in this situation, a panelist should rate all items in the test, focusing on a single performance category (e.g., between proficient and not-proficient for all items in the test) or whether it is better for the panelist to keep the item in focus and make ratings across the performance categories.

Another factor in the validity of panelists' rating has to do with the criterion they are instructed to apply when making their item level judgments. In some cases the panelists are not required to use a pre-articulated criterion, instead they are asked to provide the probability that a randomly selected, borderline candidate would be able to answer the item correctly (Angoff Standard Setting Method). In other standard setting methods (most notably the Yes/No Method and the Bookmark Standard Setting Method), panelists are asked to use a pre-specified criterion on whether they believe the item will be answered correctly by candidates at the borderline of each performance category (Impara and Plake 1997). With the Bookmark method, decisions need to be made about how to order the test questions in the ordered item booklet. Unlike with classical test theory, where the proportion correct value can be used to order the test questions by difficulty, when item response theory is used, different IRT models can yield different item orderings. Research is needed to examine whether providing differing item orderings for the panelists to use actually has a meaningful impact on their ratings. This can be especially important in situations where different decisions for ordering the items are used across content areas. It is not uncommon for one approach to ordering the items is used for one content area (say, mathematics) across all grade levels and a different approach used for another content area (say, reading language arts). Then when the results from the standard setting process are completed, these cut scores from mathematics and reading language arts are considered to be comparable. Research is needed to address this assumed comparability.

Another research question pertains to how the test questions are presented to the candidates and/or the panelists. When the test questions are presented to the candidates in an adaptive fashion, the questions posed to the candidates vary in part on

how the candidates respond to the test questions. If candidates are answering the items correctly, the test becomes systematically harder, "tailoring" the test to the ability level of the candidate. Conversely, candidates who are answering the items incorrectly will be given easier items, again tailoring the test to the ability level of the candidate. Therefore, there is not a "set" test that is given to the candidates. In some standard setting studies with adaptive tests, panelists are given a fixed form and they use that to set the cut scores (Way and McLarty 2012). These cut scores, on the theta metric, are then applied to the results from an adaptively administered test. Research is needed to examine whether the cut score on the theta metric, obtained from a fixed form test, is appropriately generalizable to tests administered in an adaptive fashion.

Research has addressed whether a shortened, but representative, set of questions could be substituted for a full-length test in a standard setting study. In this research, after gathering panelists' ratings on a full-length test form, the researchers performed secondary analyses to form subsets of items, of varying lengths and degrees of representation, to ascertain the generalizability of cut scores derived on the shortened tests to the full-length test form (Ferdous and Plake 2007). The results were very promising, indicating that theta level cut scores were virtually identical when 50% of the items were used in the reduced form test, and when content and statistical properties of the shortened test matched this information from the full-length test. However, there could be differences in panelists' ratings if they were given the reduced length test instead of the full-length version. Further research is needed to follow up on the generalizability of these results in actual operational settings.

In operational programs with historical trend data, it is often desirable, with the introduction of new assessment blueprints, to validate that the current cut scores are still appropriate. In such instances, the process is referred to as *standards validation* instead of *standard setting* (Mattar et al. 2012). The goal is to keep trends valid, even though new blueprints or even new content standards have been introduced. It is critical that the performance level descriptors (PLDs) remain the same; otherwise a new standard setting would need to be implemented. Often a standards validation procedure uses a Bookmark Standard Setting method. Using equipercentile equating, the locations of the current standards are identified in the Ordered Item Booklet. Then panelists are asked to review the items in the vicinity of these cut scores. Panelists may decide to move the location of the cut score bookmarks, but often within a narrow range. Multiple rounds are frequently used, with traditional feedback similar to a regular Bookmark Standard Setting method. In many cases, fewer panelists are used for a standards validation effort. Dwyer (2016) examined three approaches for maintaining equivalent cut scores when new forms are introduced (common-item equating, resetting the standard, and rescaling the standard). This current research provides an excellent example of how decision-making can be informed by research. Research is needed to examine how the location of these cut score bookmarks influences panelists decisions and how range restrictions impact panelists' confidence in their final bookmark placements (Box 6.3).

Box 6.3 ResearchQuestions: Panelists' Operational Ratings

1. What is the impact of having panelists, when making multiple ratings, start at different decision points on the continuum of cut scores (e.g., starting with Proficient and moving to Advanced and then to Basic as opposed to starting with Basic, then Proficient, and finally Advanced)?
2. Should panelists focus on a single item and make multiple ratings (e.g., Basic, Proficient, Advanced) or should panelists focus on a cut score (e.g., Basic) and rate all items with that cut point in mind?
3. What is the impact of using different item ordering approaches across different content areas on the comparability of cut scores across the content areas?
4. How should panelists change their rating process if items are presented in an adaptive fashion to candidates?
5. What is the impact of using a shortened, but representative, sampling of test questions instead of the full test on the results from a standard setting procedure?
6. How do results from a standards validation approach differ from that from a standard setting procedure?

6.5 Feedback to Panelists

It is generally accepted practice that feedback be provided to panelists between rounds of ratings. Typically, this feedback consists of two distinct forms, either "panelists-based" or "candidate-based" (Reckase 2001). Panelists-based feedback often consists of information provided to the panel regarding the location of individual panelist's cut scores based on the most recent round of ratings. Candidate-based feedback often consists of information about how candidates performed on the individual items, and impact data, showing what percentage of candidates would fall into each performance category based on the panel's most recent round of ratings. Research is needed to investigate the impact on panelists' ratings of these differing kinds of feedback, and the impact of these different kinds of feedback at different points in the standard setting process.

In some cases, panelists are provided with information about how the population performed on other assessments. This is sometimes referred to as *benchmarking* (Phillips 2012). This information may come from national assessments, such as NAEP. In recent efforts to set cut scores for college and career readiness on state or consortia high school assessments, panelists were given information about performance of students who took the national SAT and ACT examinations. Because the population taking these assessments can vary substantially from the target popula-

Box 6.4 Research Questions: Feedback to Panelists

1. How does the sequencing of different kinds of feedback to panelists affect panelists' ratings?
2. What is the impact of including Benchmark data on panelists' ratings?

tion of high school students taking the high school tests (because, for the most part except for states that offer these college readiness tests to all students in their state, only college bound students take these college readiness tests). Panelists have difficulty processing all of this additional information, especially if the information is not directly comparable to the target population for setting cut scores. Research is needed to help understand how to best present Benchmark information to panelists, including the sequencing, timing, and communicating strategies (Box 6.4).

6.6 Cross-Group Articulation of Cut Scores

In some applications, most often in educational settings, there is a desire to ensure that cut scores across adjacent grades in the same content area are reasonably smooth in terms of impact (Cizek and Agger 2012). Typically, cut scores are set using within grade level panels. Even when panels are assigned adjacent grades within a content area, substantive differences in percentages of students within performance categories across grades can result. It is common in these situations that a post standard setting process, called *vertical articulation*, is undertaken. Typically, representatives from each grade panel convene in a single group and consider the consistency of percentages of students that fall into performance level categories sequentially across grades. It is the task of this cross-grade panel to still maintain the perspective of the borderline performance level descriptors and the test questions to reconsider the location(s) of the cut scores on the score continuum. This is an easier task to accomplish using a Bookmark standard setting method as the cut scores in this case are directly tied to item locations in the Ordered Item Booklet. Research is needed to consider how cross-grade articulation can be meaning applied in situations with other judgmental standard setting methods, such as the Angoff Standard Setting Method (Box 6.5).

Box 6.5 Research Questions: Cross-Group Articulation of Cut scores

1. What procedures work best when doing cross-group articulation with other standard setting approaches other than a Bookmark Standard Setting Method?
2. How can a content focus be maintained when doing a cross-group articulation with an Angoff Standard Setting Method?

Box 6.6 Research Questions: Policy Smoothing of Cut Scores

1. How can policy makers implement a policy-smoothing methodology without introducing concerns about arbitrary decision-making?
2. What criteria should be applied from when policy makers implement a policy smoothing procedure?

6.7 Policy Smoothing of Cut Scores

In addition to cross-grade vertical articulation by representatives from the respective standard setting panels, a policy level articulation process is sometimes used, again often in an educational setting. In this case, policy makers (state board of education representatives, for example) are given the opportunity to "smooth" the cut scores to achieve a desired consistency in percentage of students falling in each performance category. This is sometimes viewed as being very arbitrary as it is not clear what criteria should be used by these policy makers in making these adjustments. These policy makers are not necessarily content experts and have not received the needed training to make item based decisions about performance of students at performance level borderlines. It would be very helpful if policy research studies could be conducted to provide guidance for policy makers when put in the position of making such adjustments to the cut scores that have already undergone in depth scrutiny by grade and content experts and already been considered for consistency by the cross-grade panels (Box 6.6).

6.8 Uses of Scores

In addition to research to inform the practices in setting cut scores, research is also needed in the uses test scores derived from administration of these tests (Hambleton and Slater 1997). In some settings, sometimes in educational settings where there are high-stakes consequences to test performance or in credentialing examinations,

Box 6.7 Research Questions: Uses of Cut Scores

1. What policies should be put into place when retakes are permitted on an assessment that has cut scores in place?
2. How do policies regarding "banking" of scores affect the validity of scores interpretations when cut scores are used with a multi-component examination?
3. How do compensatory and conjunctive decision policies impact the validity of performance interpretations?

policies need to be implemented about retakes. When a retake is allowed, questions need to be addressed regarding whether documented educational remediation is required prior to the retake or whether there is a required interval of time that must have passed prior to retake. Because there are the concerns about regression to the mean and standard error of measurement, in some cases an adjustment of the cut score to counteract these measurement artifacts is considered. In other cases, no attention is paid to the ability of candidates to capitalize on chance in gaining passing level scores (Millman 1989). Further, in some applications where there are multiple components to the assessment, decisions need to be made regarding whether the overall passing decision is conjunctive (candidates have to meet the passing score on each component in order to pass overall) or compensatory (candidates have to reach an overall score across the respective components). These different decision rules can have profound impact on the overall passing rate (Clauser and Wainer 2016). Policy makers should be informed, and consider the long term impact on these differing decision rules. Policy level research could help inform these decisions (Box 6.7).

6.9 Challenges to Conducting Research in Standard Setting

Although there are many topics ripe for research in standard setting, there has been limited research to support the practice of standard setting. Notable exceptions to this statement is the research being conducted at the National Board of Medical Examiners by Brian Clauser and colleagues (see Clauser et al. 2009, 2014; Margolis et al. 2016). All of these studies address practical issues in implementing an Angoff standard setting study in the context of medical licensure, such as whether it is desirable to show the panelists the scoring key when they are rating the items. However, the ability of such an organization to be able to fund and conduct standard setting studies may be unique.

There are several reasons why there is limited research in standard setting. Most importantly, standard setting impacts the lives of candidates, especially in high-stakes testing programs. It is not appropriate to "try-out" new methods or new procedures during an operational standard setting when the results of the standard setting process could be impacted by the research process. Further, it is very costly to conduct a standard setting research studies. In other social science settings, proxy studies are conducted using volunteers or paid participants, often using college students as participants. These "pseudo panelists" would be asked to take on the role of actual standard setters. This does not have the same validity as the use of actual stakeholders and the utility and generalizability of their ratings are questionable. Further, it is challenging to simulate panelists' ratings using simulation studies. This has been tried, using models to simulate specific kinds of biases in ratings (Plake and Kane 1991), but again the utility and generalizability of these results have been called into question.

There are some promising strategies for conducting research in standard setting, though. One example is to negotiate a research component with a client agency that permits conducting multiple panels, one to do operational ratings and the second to perform the research study with the opportunity then to compare the results across the two panels. Another opportunity to conduct research is to insert in an operational standard setting a step, prior to the operational ratings, that includes a research component. Although there is still the potential that inserting such a step could upset or distort the operational ratings, careful planning can help to overcome such concerns. With funding from grants or contracts, it has been possible to explore new standard setting approaches. Such funding sources, however, are few and far between. Finally, another source of research is to conduct secondary analyses of the data from a standard setting study. For example, rating by subsets of the panelists, or subsets of the items comprising the test, could be analyzed to determine how stable the final results are to various conditions of panelists or test construction features.

6.10 Conclusion

The purpose of this chapter was to address multiple research areas where the practices in standard setting could benefit from additional research to support the validity of score interpretations from standard setting studies. The components of standard setting that were considered in this study included the selection of panelists to participate in the standard setting study, the training of the panelists, the operational ratings made by these panelists, the feedback provided to panelists during their operational ratings, the use of cross-grade articulation panels in adjusting cut scores from the standard setting panels, the role of policy makers in smoothing the cut scores from the standard setting process, and issues surrounding the use and applications of cut scores from standard setting studies.

Standard setting has served an important role in assessment practices. The results from the standard setting procedure provide meaning to test scores and these test

scores (and their classifications into performance categories) have been used in multiple ways that affect the lives of test takers and others who use these test results. The validity of these interpretations of test scores relies directly on the quality of the standard setting process. This chapter indicates several opportunities where research could yield improved validity of the interpretations of test scores.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Cizek, G. J., & Agger, C. A. (2012). Vertically moderated standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 467–484). New York: Routledge.

Clauser, A. L., & Wainer, H. (2016). A tale of two tests (and of two examinees). *Educational Measurement: Issues and Practice, 35*, 19–28.

Clauser, B. E., Mee, J., Baldwin, S. G., Margolis, M. J., & Dillon, G. F. (2009). Judges' use of examinee performance data in an Angoff standard-setting exercise for a medical licensing examination: An experimental study. *Journal of Educational Measurement, 46*, 390–407.

Clauser, J. C., Margolis, M., & Clauser, B. (2014). An examination of the replicability of Angoff standard setting results within a generalizability theory framework. *Journal of Educational Measurement, 51*, 127–140.

Dwyer, A. C. (2016). Maintaining equivalent cut scores for small sample test forms. *Journal of Educational Measurement, 53*, 3–22.

Ferdous, A. A., & Plake, B. S. (2007). Item selection strategy for reducing the number of items rated in an Angoff standard setting study. *Educational and Psychological Measurement, 65*(2), 185–201.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Washington, DC: American Council on Education.

Hambleton, R. K., & Slater, S. C. (1997). Reliability of credentialing examinations and the impact of scoring models and standard-setting policies. *Applied Measurement in Education, 10*, 19–38.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*, 353–366.

Kane, M. T. (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. 52–88). Mahwah: Lawrence Erlbaum Associates.

Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Schultz, M. (2012). The bookmark standard setting procedure. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 225–254). New York: Routledge.

Loomis, S. C., & Bourque, M. L. (2001). From tradition to innovations: Standard setting on the National Assessment of Educational Progress. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 172–217). Mahwah: Lawrence Erlbaum Associates.

Margolis, M., Mee, J., Clauser, B. E., & Winward, M. (2016). Effect of content knowledge on Angoff-style standard setting judgments. *Educational Measurements: Issues and Practice, 35*, 29–37.

Mattar, J., Hambleton, R., Copella, J., & Finger, M. (2012). Reviewing or revalidating performance standards on credentialing examinations. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 399–412). New York: Routledge.

Millman, J. (1989). If at first you don't succeed: setting passing scores when more than one attempt is permitted. *Educational Researcher, 18*, 5–9.

Phillips, G. W. (2012). The benchmark method of standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 323–346). New York: Routledge.

Plake, B. S. (2008). Standard setters: Stand up and take a stand. *Educational Measurement: Issues and Practices, 27*(1), 3–9.

Plake, B. S., & Kane, M. T. (1991). Comparison of methods for combining minimum passing levels for individual items into a passing score for a test. *Journal of Educational Measurement, 28*, 248–256.

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: Lawrence Erlbaum Associates.

Way, W. D., & McLarty, K. L. (2012). Standard setting for computer-based assessments: A summary of mode comparability research and considerations. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 451–466). New York: Routledge.

Part II
Standard-Setting in the Nordic Countries

Chapter 7
Standard Setting in Denmark: Challenges Through Computer-Based Adaptive Testing

Peter Allerup and Christian Christrup Kjeldsen

Abstract The objective of this chapter is twofold: (1) to provide an overview of standard settings in the Danish school system and (2) to guide the reader through the history of standard setting in a Danish context. This includes periods when Denmark provided education without using the formal standard setting approaches we use today, where standard setting involves more than creating limits or cut points in various distributions measuring student achievement. The Danish school system has recently started making actual use of several traditional methods of determining the minimum pass mark for an exam, which opens the possibility for students to move on to a higher educational level, similar to the transfer from compulsory lower secondary to general upper secondary education. This chapter will introduce the former practice, in a time when there was no formal standard setting, up to the current situation, focussing in particular on the introduction of adaptive computerized systems based on the psychometrics used by PISA and IEA. This new form of endeavour forms a sophisticated use of standard setting. This highly complex use of assessment systems takes advantage of the Rasch model terminology that has been applied as part of Danish national tests (NTs). This assessment system secures scalability of students both horizontally, across students within the same grade level, and vertically, across students at different grade levels. To find this system being implemented across an entire educational system, is unique. The chapter explains how this system developed with a particular focus on the Danish NT.

Keywords National tests • Rasch model testing • Psychometrics • Computerized adaptive tests • Longitudinal testing

P. Allerup (✉) • C.C. Kjeldsen
The Danish School of Education, Aarhus University, Copenhagen, Denmark
e-mail: nimmo@edu.au.dk; kjeldsen@edu.au.dk

S. Blömeke, J.-E. Gustafsson (eds.), *Standard Setting in Education*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-50856-6_7

7.1 Introduction

This chapter addresses several aspects of standard setting considered here as relevant when focusing on the Danish situation, using the historical development as a foundation for understanding the innovative use of adaptive testing applied to the current so-called *national tests* (NTs). However, the contemporary meaning of *standard setting* is closely related to another reference for the word *standard*. In fact, *standard* is referring to the psychometric properties of the measuring instruments by which the proficiency of students are measured. By this, the focus is directed towards the elements of the test – the *items* of the instruments and how they interact, rather than on the output, such as scores produced by the instruments. The traditional setting of, for example, cut points in score distributions are, at best, derived from psychometric properties.

This chapter zooms in on and finally focuses in particular on the Danish national tests – the *National Tests* (NTs) – In 2006, the Danish Parliament decided to make national tests compulsory in the *Folkeskole* (UVM 2013). One reason for having a broader look than is provided by a narrow discussion on score cut points or centrally defined goals is that a large group of tests in Denmark has been classified as *standard* only after specific rigorous statistical analyses of the items that comprise the test. Another group of tests applied over time in the Danish school system has not been subjected to such item analyses prior to their implementation. Therefore, both aspects will be discussed in order to obtain a comprehensive description of the situation in Denmark.

7.2 Brief History of Prior School Assessments in Denmark

The Danish combined pre-primary and lower secondary school in a uniform compulsory school named *Folkeskole*. The contemporary *Folkeskole* caters to students from the age of approximately 6 – grade 0 – to approximately 15 – 9th grade. Grade 0 (zero) has, until recently, only to a limited extent included formal teaching activities; therefore, the main compulsory learning within the *Folkeskole* takes place from grade 1 to grade 9. When moving from one grade to the next, the students are not subjected to any high- or low-stake testing. In other words, no testing or examination could eventually result in a student either failing or passing a school year. Therefore, tests are not part of the standard procedures for students progressing automatically from one grade to the next.

7.2.1 The Early Twentieth Century Until Post-WWII

In 1903, a school reform was enacted that defined a new level, referred to as *Mellemskole*, intended to serve as a bridge between the compulsory *Folkeskole* and upper-secondary school (high school or *Gymnasium*). The students were tested in

grade 5 and, according to the test results they achieved, their school life followed either a transition into a 4-year academic track or a 3-year vocational track. The tests were in reading and mathematics, and the decision regarding which of the two tracks the student would follow was based upon consultations between teachers and parents in light of the test results. It is, however, difficult to find any justification for the use of specific cut points in these tests for the separation of students into one of the two tracks.

In 1958, the Danish Education Act underwent a major change, eliminating the two tracks and, subsequently, eradicating the need to separate eleven-year-old students into different groups. The 1958 law meant that students up to grade 9 were no longer subject to high-stake evaluations by means of standardized tests. However, that does not mean that the students were not subjected to standardized testing from the early 1960s up until today.

7.2.2 The Period After 1960 with a Focus on Rasch Models

Around 1960, the Danish statistician Georg Rasch introduced what became known as the Rasch models in the field of psychometrics, creating a philosophical background for measuring and comparing student achievement using standardized tests. This was an important shift in the way assessments were validated in comparison with the time prior to the introduction of these models. The Danish Institute for Educational Research (DPI) was established in 1955, with Rasch as a selected member of the board of the institute.

Due to Rasch being responsible for the development of modern IRT (Item Response Theory) tests in the main topics of reading and mathematics, and through his personal relationship with one of the central people behind the publishing company Dansk Psykologisk Forlag (DPF)[1], most of the tests used in the *Folkeskole* in subsequent years became standardized in the sense of being Rasch model-approved before being applied in practice. These tests were not accepted as valid until explicit data had been collected and analyzed under the Rasch model. Test items were eventually modified during this process and the final standardized tests could be altered according to the prior analysis.

The standardized tests published by DPI and DPF generated standard questions in terms of how to validate them, use them, score them and interpret the results. If a test was intended to form part of standard setting, all these areas had to be properly analyzed. Validation has become mainly a question of undertaking statistical analyses by means of applying Rasch models to pilot data collected specifically for this purpose. Standardized test scores are subsequently either *norm-referenced* (e.g., students' performance is determined by how well they performed in comparison to

[1] A publishing company producing psychological tests, intelligence tests. And subject-oriented tests mainly within reading and mathematics. It was originally started in 1949 by a group of psychologists.

their peers), or they can be *criterion-referenced*. In relation to the earliest classification of standard setting approaches, one would find a mixture of standardized tests with interpretations of test results as either norm-referenced or criterion-referenced (Cizek 2012; Hays 2015).

It is somewhat ironic that while the Danish statistician Georg Rasch (1960) formulated the principles for *objective* comparisons, which among other things made it possible to put student achievements on common scales, even if the students had not responded to the same set of items, the National Assessment of Educational Progress (NAEP) in the United States had already in the 1980s introduced test designs based on the validation of items by means of Rasch models, while in Denmark only the Danish Psychological Publishing Company (DPF) was producing tests fulfilling these requirements.

To some extent, the same irony may be found in relation to the methodological refinement of defining the performance standards for the cut scores of NTs; whereas methodological development of standard setting procedures related to cut scores have been used and developed for decades in the US (Zieky 2012; Cizek et al. 2004), the setting of cut scores in the information to teachers and parents on a five-point Likert scale in the NT is mainly simple and to some extent arbitrary. However, right up until the end of the twentieth century, the official system in Denmark continued as usual; it was not until the late 1990s that the focus of the Ministry of Education on Denmark's strategy for education, learning, and IT involved developing plans and ideas in future education enhanced by IT.

Before providing details on the adaptive test system introduced as a result of this new focus, an overview is provided of the basics of the Rasch model that form the back-bone for making it possible to implement adaptive testing procedures in the NT.

7.3 The Rasch Model and Its Implications for Adaptive Testing in Denmark

The simple Rasch model for two response categories employs two sets of parameters: one set for the item difficulties and another for the student achievements. It is impressive that Rasch (1968) proved a mathematical equivalence among the following three statements (i)⇔(ii)⇔(iii):

(i) Student scores (and item scores) exhaust all knowledge of "severity" and "skill" (sufficiency).
(ii) It is possible to calculate and compare student achievement using any subgroup of items.
(iii) The Rasch model is a valid statistical description of data (item response level "true"/"false").

The first characteristic (i) can be called a *validation* of the practical use of student scores. Standard setting for student achievement must, consequently, involve estimates

of student achievement under the Rasch model. The second property (ii) is called *specific objectivity* and is extremely useful in test situations where not all students are responding to the same items, like the NT, PISA and IEA's TIMSS and PIRLS studies. Especially, in adaptive test-designs, the ability to use any subgroup of items is of major importance. The third point (iii) deals with the fit of the Rasch model to data collected using the test; that is, an evaluation that can be handled properly by professional statisticians who have access to methods and computer software for this task (Allerup 2007).

The student achievement scale consists of values from minus infinity to plus infinity; in practice, however, values vary from about −3.00 to about 3.00 (the possible range has been limited technically within the NT from −7 to 7 on the logit scale), with the "neutral" student measured in the middle with the value 0.

The Rasch Model for two response categories assigns the following probability to a correct response: $\mu = 1$,σ_v and θ_i referring to respectively student achievement for student No v and item difficulty for item No i:

$$P\left(X_{vi} = \mu\right) = \frac{e^{\theta_i + \sigma_v}}{1 + e^{\theta_i + \sigma_v}} \tag{7.1}$$

Rasch provided a proof specifically for the necessary and sufficient condition in the special case $M = 2$ response categories (Rasch, not dated, but approx. from 1965), later presented (Allerup 1994) using his original mathematics notation.

The Rasch model has been selected by the international PISA test and the IEA's TIMSS and PIRLS studies as the basis for evaluation of test validity and for the calibration of international scales. However, since negative values as measures of student achievement are not popular, the values have been shifted parallel to the original scale to a new scale of around 500, with a standard deviation ± 100 instead of the original 0.00 ± 3.00.

This is a purely mathematical exercise that does not affect the interpretation; rather, it makes it possible to avoid the issue of negative values sending the wrong signals to the receiver; an example is the discussion that has taken place in Denmark in relation to the introduction in 2007 of a new marking system in education – the so-called '7-point grading scale'[2] – where the failing category has minus three (−3) as a mark, which is reworded "For an unacceptable performance" on the 7-step CTS scale. This scale is used for evaluations in the daily school routine of the *Folkeskole* in grade 8 and 9, where pupils receive biannual continuous assessment marks in all exam subjects.

It should be noted that although the principles of "specific objectivity" look like a purely theoretical concept, the items of a test, (e.g., in the Danish national test), cannot work as a proper scale of items unless a pilot of the sample of items has been tested for item homogeneity by means of the Rasch Model. Thus, it is impossible to

[2] The scale consists of five marks designating a passing level (12, 10, 7, 4 and 02) as well as two marks designating a non-passing level (00 and −3). The 12, 10, 7, 4 and 02 are equivalent to the ECTS marks A, B, C, D and E, whereas 00 and −3 are equivalent to Fx and F.

develop items solely on a desk without also arranging practical testing. During these test stages (Allerup 1997), up to 50% of the original items are usually (e.g., PISA) eliminated because of deficient psychometric properties. Under an adaptive regime, this could be a challenge for the development of new items because the items are never presented to the students in a simple linear test run.

Therefore, new items to be tested for homogeneity have to be included as "hidden" items without direct consequences for the calculation of student achievement. Evaluation of item homogeneity in these cases must follow other techniques known from the simple linear paper and pencil tests. If not carried out carefully, this may harm the validity of the standards successively set by the items when presented to students (Allerup 2005).

7.4 Differentiation Between Horizontal and Vertical Testing

When comparing students with respect to their level of achievement, this will usually be undertaken by means of traditional tests consisting of a fixed number of items: so-called *linear testing*. Horizontal testing is carried out every year in Denmark for students in grade 9, by the end of compulsory schooling. This test is called *Folkeskolens Afgangsprøve*, which consists of the final exams of the compulsory schooling stage and has nothing to do with the national tests (NTs).

The students' results are compared using normative techniques, with total distributions of score values for all students in the country as a reference. The subjects are Danish, Mathematics, English, Chemistry and Physics. The subjects are chosen at random from the subjects, written English, German and French are added, as well as History and Geography. Sport may be selected as well. The performance is evaluated by means of the aforementioned 7-point grading scale.

The new marking scale has been developed in part out of the desire to simplify the compatibility and comparability between Danish and foreign grading scales, and, pivotally, to provide a clear correlation between the descriptions for the individual marks and the academic objectives.

In tests and examinations that according to the rules of the individual study programs etc. require documentation in the form of tests, examinations or leaving certificates, students are to be given an assessment according to the following grading scale (7-point grading scale):

12: "For an excellent performance"
10: "For a very good performance"
7: "For a good performance"
4: "For a fair performance"
02: "For an adequate performance"
00: "For an inadequate performance"
−3: "For an unacceptable performance"

The marks from written examinations are adapted to a distribution, with 10% of the students receiving "12," 25% receiving "10," 30% receiving "7" and, symmetrically, 25% receiving "4" and 10% of the students receiving "02." No prefixed levels for the marks "00" and "−3" are determined.

Students taking *Folkeskolens Afgangsprøve* (although compulsory, approximately 19% of students in the Copenhagen areas did not take part in the examination) cannot be compared across years, because new linear tests are developed each year and because of the distributional constrains on the marks from the 7-point scale. Horizontal testing can, therefore, be undertaken by means of the *Folkeskole* final examination. The exam tests cannot, however, be used for vertical testing or evaluating standard setting over continuous years. It is interesting, however, that vertical testing has been carried out since 1955 in another domain of the national state: namely, as part of the recruitment system of the Danish armed forces.

Based on the Rasch model, Børge Prien developed a series of mental tests for screening purposes which aspirant soldiers have to undergo before entering the military. These "intelligence tests" have been active ever since and are applied even today as a screening instrument, proving to be an efficient tool for vertical testing (Kousgård 2003). For civilians no official standard settings based on the intelligence scores from BPP have been developed. However, comparing tests results across the years between 1950 and 2003 gave rise to debate concerning possible change in the general level of intelligence.

Returning to the Danish national tests: these are based on the same psychometric principles as the BPP intelligence tests and, consequently, have the same properties regarding the possibility to estimate *a trend* across several years of testing (vertical). As will be clear from the section describing the technical details of the Danish NTs, this capability to estimate a trend is even applied at an individual level across several grades. If sufficiently tested and validated, therefore, the tests are appropriate means for performing reasonable vertical testing.

However, in order to be able to properly interpret a sequence of successive testing (e.g., with regard to a measured trend), one will have to look at how an established curriculum has defined the content domain, as the test items are representative of it.

For the subject reading, some reading experts favour the view that the development in reading from the lowest grade to grade nine is just a matter of presenting items to the students with increasing levels of difficulty. The curricula reference remains the same. This is, obviously, not the case in a subject like mathematics, where the curriculum changes drastically across the school years. No easy solutions have been found as to the impact of shifting curricula on the interpretation of successive results from testing in, for example, mathematics.

The Danish national tests are, moreover, implemented in an adaptive framework that allows the testing to take place using solely test items with an adequate range of difficulties for the individual students. As such, the adaptation has no direct influence on the possibility of performing vertical testing. However, the system in charge of the adaptive testing has to make use of an item bank in order to be able to offer the adaptive testing possibility. As will be shown in the section describing the technical details behind the Danish national tests, this fact, together with the

psychometric restrains of homogeneity on items in the bank, leads automatically to a situation where vertical testing is possible.

It will be easier to conduct vertical testing in the future, because a general shift is taking place from curricula-oriented evaluations in Danish evaluation research to more literacy-oriented evaluations. Hence, a change of curriculum between successive testing's will have less effect. This fact will influence the very construction of test items in the item bank of the Danish national test by making the concept of necessary ability in order to provide a right answer to an item more independent of a specific curriculum. Consequently, emphasis will be placed on *student competencies* rather than "old fashioned" *curriculum-referenced knowledge*, making it easier to construct Rasch model homogeneous achievement scales across several grade levels.

7.5 The Former Set of School-Leaving Tests at Grade 9

Norm referenced standardized testing compares a student with all the other students who take the same test. This is what has happened for most of the tests produced by DPI and Dansk Psykologisk Forlag (DPF), which allows teachers to use the tests in everyday evaluations with their students to provide a general picture of the level of the class. Some of the standard tests provide means for ranking students, classes and even schools. However, this had not been a visible part of the evaluation programmes in the Danish *Folkeskole* until 2000 when PISA introduced another way of looking at the test results.

Criterion-referenced standardized tests determine how well students meet specific requirements or fulfil previously established standards set within the subjects of mathematics, reading, language etc. It could be argued that standard setting in Denmark is a compulsory exercise for the *Folkeskole* at grade 9, when students sit the final exams. Within this endeavour, a mixture of criterion and norm-based standards are applied by specially trained groups of teachers who define required cut points for each topic in order to create the basis for a mark on the named (CTS) seven-point scale. This has been the practice to date; however, following a decision of the NT in 2006, a new standard setting based on electronic testing has been devised. In the following, the focus is on the latest developments of national adaptive tests.

7.6 A New Set of Danish National Computerized Adaptive Tests at All Grades

The key words behind the development of the new national tests were:

1. Living up to modern psychometric requirements (in practice, Rasch modelling)
2. IT-based (electronic)

3. Adaptive
4. Valid
    a. Content validity
    b. Construct validity
    c. Predictive validity
5. Reliable

In the requirements listed above, the need for adaptive testing implies that an item bank is constructed. The term *reliability* refers to the usual definition whereby groups taking the test over time, within reasonable margins of error, will get the same results. The first point and the term *validity* hint at evaluations taking place under the Rasch Model. The full name of the new Danish national test is the "National IT-based Adaptive Testing." Unlike the abbreviation (NT), the full name highlights the more important aspects of these tests in comparison with many other types of tests. The tests and their electronic implementations are the response to weak PISA results in 2003 combined with political requirements to monitor students across several years (vertical testing), instead of the former test system described above (horizontal testing), which allows only one year at a time to be used to set the standards for comparing student achievements (Fig. 7.1).

Ever since the company DPF managed to compare student achievements from their tests on common latent scales, it was also deemed desirable that such tests be a part of the public *Folkeskole* testing system. An important feature of the international PISA tests and of the tests developed under IEA's TIMMS and PIRLS[3] is that

Subject of the test	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9
Danish, reading		X		X		X		X	
Mathematics			X			X			
English							X		
Geography								X	
Physics/Chemistry								X	
Biology								X	
Danish as second language					X		X		

Fig. 7.1 Grades and subjects in the present NT system of testing (http://uvm.dk/Uddannelser/Folkeskolen/Elevplanernationale-test-og-trivselsmaaling/Nationale-test/Fag-og-klassetrin)

[3] The International Association for the Evaluation of Educational Achievement (IEA) is an independent, international cooperative of national research institutions and governmental research agencies. It conducts large-scale comparative studies of educational achievement and other aspects of education. TIMSS 2015 is the sixth cycle in IEA's series of assessments of maths and science achievement at fourth and eighth grades. TIMSS Numeracy is at the primary school level. PIRLS 2016 is the fourth in a 5-year cycle of assessments monitoring trends in reading literacy in primary school. PIRLS Literacy, a study of fundamental reading skills, and the e-PIRLS assessment of online reading offer further opportunities to investigate children's experiences in learning to read.

these latent scales can be constructed from booklets consisting of items that have only little overlap of items. This is possible because all the items in each latent scale have been tested for validity in relation to the Rasch model. In other words, each scale is developed on the basis of pilot testing data followed by statistical analyses eventually resulting in modifications of the scale items.

It was a representative of the Ministry of Education, who, after traveling to Norway in the early 2000s, conveyed the first ideas for the design of IT-based adaptive tests developed under modern psychometric conditions like the Rasch model specifies. Right from the start, they wanted to develop a test for many grade levels and in as many subjects as possible. They strongly wished to be able to compare results from year to year, so that the effect of key initiatives, such as school reforms, could be measured and evaluated dynamically over a number of years.

As mentioned previously, this was not possible with the existing final grade 9 exams, whereas the standard setting procedures applied could be characterized as *relative methods* of standard setting (Cizek 2012). There are currently approximately ten mandatory tests in school subjects and two optional tests in the subject of Danish as a second language; however, it is the intention to expand the list across both subjects and grade levels within NT.

A few key elements from a new school reform can illuminate the important properties of comparability that the NT is intended to satisfy: for example, if students from different grades are compared or the same student's measurements over several years are compared. The reform has the following objectives and will have to be followed empirically:

- All students must be presented with academic challenges so that they become as proficient as possible.
- The number of the most talented students in Danish and mathematics must increase year by year.
- The number of students with poor results in the national tests for reading and mathematics must diminish year by year.
- Students should, in the long run, perform on the same level in the 8th grade as they do now in the 9th grade.

It is already clear from this shortened list that a standard setting based on the NT must make possible comparisons over years and across groups of students so that the results of the NT for a specific student can be compared with the same student's results at a later date and with other students' results. However, the new school reform was launched after the introduction of the NT and, therefore, cannot fully take credit for the requirements presented. From a psychometric angle, the current NT is one of the few of its kind in the world.

The PISA and IEA scales centered on 500 have become familiar to the users of those international test results; that is, to the public through the media, and, more importantly, to researchers and decision makers at the Ministry for Children, Education and Gender Equality. They fully understand that, for example, a Danish reading result on the PISA scale of 492 is less than the average OECD level on a scale of 500. But do these eight points demonstrate a significant deviation from the

value 500? Does it mean that Danish students fail to pass relevant criteria for standard settings? These questions cannot be answered properly based solely on the construction of the scale. However, the official achievement categories proposed by the PISA consortium put students in up to seven groups with descriptions according to their performance, which allows the user to judge if a student result is "acceptable" or not.

It is even postulated from the descriptions of these boxes whether, based on the actual category, it will be possible for the student in the category to complete an educational programme after finishing the compulsory *Folkeskole*. Although this is not true (Allerup et al. 2013), these classifications have achieved the status of standard setting among politicians in Denmark who are responsible for initiatives that, like the last *Folkeskole* reform, have been partly based upon such interpretations of the classifications.

Another view on standard setting, which is of a solely mathematical nature, comes from the fact that secondary analyses are frequently desired after the primary presentation of test results. It is, for instance, clear that after an initial presentation of the international PISA results, there might be a need for more detailed analyses concerning achievement by students who do not have Danish as their mother tongue. In many cases of ordinary test results, the presentation of test results is achieved by means of the percentage of correctly solved items ("percentage correct").

In continuation of the tables and analyses presented, it might be tempting to do the secondary analyses using the same reference to the percentage correct. However, in most cases, this would be an inappropriate method because, mathematically speaking, the variable percentage correct does not keep the necessary properties in order for the calculations to hold valid statements. In this case, the standard setting by means of the 500 ± 100 scale offers simple possibilities to do secondary analyses using standard statistical software, thus avoiding a situation with invalid statements.

As in the case of the NT, standard setting in connection with the dissemination of test results has two facets. One for the close users of results, like the students, teachers and parents, and, for the statistics, the school. However, there is another one for researchers who wish to perform secondary analyses, like evaluation programs with respect to the effect of changes in school reforms. The researchers want access to the Rasch scale achievement measures while the first group gets the results on a completely different scale: namely, a percentile scale.

This scale attempts to take advantage of the familiarity with the 7-point marking scale – with the inherent feeling about knowing whether the performance is "good" or "weak" – with a normative message about where on the latent scale the student belongs. It is a 0–100 point scale that is cut down in a five-point Likert scale:

- Extensively above average [91:100]
- Above average [66:90]
- Average [36:65]
- Below average [11:35]
- Extensively below average [1:10]

A student profile consisting of three values of the subject oriented sub-domains is built on these "translations" from the Rasch scale to the used 100-point percentile scale.

From a statistical point of view, it is slightly strange that the official dissemination of NT results takes place using 0–100 point values for the three profiles individually. For the aforementioned group of persons to receive the results in this "percentile language," it is clear that evaluations and comparisons can be carried out solely on the grouped values: that is, the five-point scale. It would have been much more accurate to compare and carry out evaluations by means of the Rasch scale scores.

7.7 Scaling and Adaptivity of the Danish National Tests

The scaling of student ability under the National Adaptive Tests takes place during iteration. The student receives an item randomly selected from the pool (bank) of items possessing average level of difficulty. For a test to be adaptive, it has to be carried out in such a way that the student is continuously exposed to tasks that are appropriate to their skill level. Specifically, this means that one strives to have both weak and strong students presented with tasks in the test run, which they have about 50% chance of answering correctly. At the beginning of the test, it is impossible to know whether it is a weak or a strong student in front of the computer screen; therefore, the start item has a "mid"-level difficulty: a difficulty level in the middle of the items in the item bank. Depending on whether the student can answer the item correctly, the next item selected is either more difficult or easier than the first item.

All obtained answers are used to estimate the latent student achievement and it can be decided whether the student behind the computer is performing well or not. Student achievement within three specialized sub-domains are estimated within each subject. In mathematics, for example, these are algebra, geometry and maths application. Taken together, the results from the three domains constitute a profile of the student, which is why some call the national test a *profile test*. The student is presented continuously with new items within each of the three sub-domains until the student can answer yet another item with the same level of difficulty with a probability of 50%.

The statistical uncertainty on the final achievement estimate depends on the number of items the student has been through and the actual process of changing to more difficult / easier items. As a rule, the student will find that the statistical requirement of low statistical variance on achievement estimate is fulfilled by about 20 items in each profile area. The adaptive feature was originally included in the original construction of the national testing for several reasons. First, there already existed a great many adaptive tests on the market, eventually presented as the so-called CAT test (Computerized Adaptive Testing) on the web. The NT aimed to offer especially low performing students an opportunity to experience the feeling of obtaining correct answers to several items, and definitely more items than they were

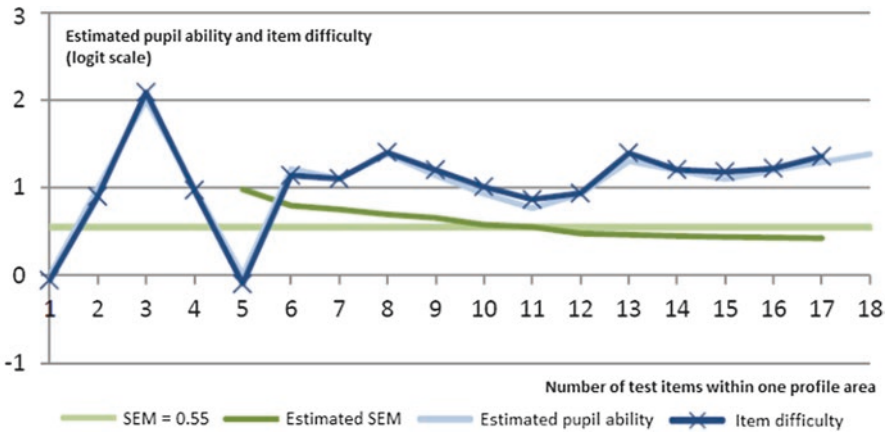


Fig. 7.2 The convergence of successive student ability estimates towards a point (ordinate) with increasing number of items given to the student (abscissa) according to the adaptive system in three specific subject sub domains

used to during routine testing, which do not adapt but have a pre-defined difficulty that meets the average pupil's skills and competences. (Fig. 7.2).

Furthermore, while being in the process of upgrading the importance of IT in the classroom and in testing situations, it is the right moment to depart from the so-called *linear test system* that uses paper and pencil with a fixed number of items, that are the same for all students.

The items are chosen from the relevant subject group, within one of the three profile areas[4], constituting the total test subject. Depending on answering correctly or incorrectly, the process is repeated with a random selection of items in every step, matching the successively updated estimate of student ability according to the Rasch Model. Based on the assumption that student ability is calculated from a situation where the student has 50% chance of a correct answer, it can be seen from the mathematical formula of the Rasch model that this situation is met with $\sigma_\nu = -\theta_i$, in which case the total probability becomes 50%.

The iterative process continues until standard error (SEM) falls below a certain limit. In the case of the national tests, therefore, the pupils are continuously presented with new items until the process leads to a situation where the student has about a 50% chance of providing a correct answer. It is slightly technical to describe how the computer balances a statistical requirement of precision (SEM error of measurement) on successively updated estimates of student achievement with the aim of achieving a value of item difficulty in a way that ensures an approximately 50% chance of responding correctly. The implementation of the Rasch model in the NT takes care of this process. The adaptive principle, however, is a technique for selecting the next item in a student's current responses of test items in NT.

[4] In mathematics the profile areas are "algebra," "geometry" and "application."

Nearly all items are in the multiple choice (MC) format (i.e., items where the student can choose between several possible answers that are specified in advance). It is due to the adaptive principle that contributes to the selection and presentation of the next item in such a way that past responses form the basis for a mathematical calculation of the student achievement. This calculation or successive estimation is continued, item for item, targeting a situation in which the next item has a level of difficulty matching the – so far – estimated student achievement, such that the probability of a correct response is 50%.

The Rasch model, implemented as the basis for all successive calculations and estimations, is accountable for the number of times a student is faced with new items. Then, the presentation of new items stops when the statistical standard error of measurement (SEM) becomes smaller than the pre-fixed level. The adaptive principle affects the precision by which the determination of student achievement takes place.

It is assumed that the adaptive process is continued until the statistical standard error of measurement (SEM) is below the pre-fixed limit of 0.55. This limit has practically been set to a value which allows the student to complete all items in the three profiles within approximately 45 min. A definition of an appropriate limit is not simple but can be evaluated in light of the Rasch model. This is used as a stop criterion and, thereby, a matter to take into account in the interpretation of the standard setting of the cut scores (the aforementioned ordinal scale for dissemination to teachers, parents etc.).

Having provided some insights into the more technical side of the adaptive testing in Denmark by means of the national tests, the focus is now on how this system is perceived by the pupils.

7.8 How Pupils Feel About Standard Testing

Pupils were previously encouraged to prepare for common test sessions, which were announced the day before the testing, carried out with paper and pencil, and for which there were known standards, enabling them almost immediately to feel whether they did well or not on the test(s). The introduction of the national tests has changed the way they feel about the tests significantly. Now, students are brought to the school computer room, where each student sits behind a computer.

After a short introduction, the students start responding to the items, which are continuously presented to them in the adaptive system online on the web, where "the next item" is selected from an item bank according to the outcomes ("correct" or "non-correct") from earlier responses. Every student runs a sequence of items that is different from other students' sequences and the only visible aspect shared by all students in the room is how quickly the computer screen turns green, which is the sign from the system that the computer has calculated the student achievement with satisfactory precision[5].

[5] Stop criteria is SEM standard error of measurement calculated under the Rasch Model.

From interviews (Kousholt 2013) it is furthermore clear that the students are very conscious about the total number of items presented during the test session. Combined with the fact that all students get a "green screen," i.e., are stopped, when the probability of responding correctly to the next item is 50%, all students in retrospect have the feeling that they have managed to answer about half of the items correctly, irrespective of the student's academic level. This happens for each of the three profile areas of the test subject. The adaptive principle affects the precision by which the determination of student achievement takes place.

Despite adequate student instruction, it does not appear that students understand that it is a "machine" and not they themselves that determines the number of tasks they are set to answer; it is considered prestigious to be presented with as few items as possible. It is also slightly "embarrassing" to be the last student that is tested. Everyone else has left the room or has stopped while the last student remains sitting alone. Here, too, there is a lack of understanding of the technical terms behind the test. From qualitative analysis of the test process, it became clear that some students were very interested in the actual number of items, and they were very eager to do as many tasks as possible and as quickly as possible (Kousholt 2012).

Another issue is that it is not easy for capable and weak students to understand that when they meet after the test session and exchange opinions about the testing, both students will have resolved about 50% of the items correctly. Is this in accordance with their mutual, fairly accurate general sense of each other's daily academic level? Because the allocation of "next item" is done randomly from the item bank, it is difficult, or practically impossible, to create a comprehensive picture of the test results like the one a teacher can collect in the case of, for example, a standard IEA reading test where the student reads several sentences forming a narrative and subsequently responds to a series of questions (items).

In all forms of evaluation, mediation of the test results is very important. This is also the case for the NT, because a proper presentation requires the use of various technical concepts in order to understand the meaning of the test results. For many public school performance tests, it has been common practice to count the number of correctly answered items and use this number when the student is informed whether the results are "good" or "less good." Specific numbers of correctly solved items may also lead to a certain mark on the 7-point mark scale[6] (CTS). The relation between the number of correctly solved items and a specific mark applies only to the particular sample and to specific students who have taken the test. However, this is not suitable for the NT, due to its adaptive structure (CAT). This raises the following question: How is standard setting influenced by this adaptive method?

It is an important part of new standard setting carried out by means of the national tests that as soon as the test has been completed, the system immediately provides the three-dimensional profile of the student built up from responses to each profile area. It should be regarded as a great improvement that the results of the national tests, unlike many old-fashioned linear tests, can now provide standard setting in

[6] The scale contains the 7 points -3, 00, 02, 4, 7, 10 and 12.

terms of comparisons over time, making it possible to follow the students and make comparisons across different groups of students.

The student sitting in front of the computer screen, never sees the genuine Rasch estimates or other technical aspects, like the number of items or the time for testing. Only the final evaluation in categories are available to the pupil/parents and the 100-point scale for the teacher. As its name suggests, this percentile scale reflects purely normative information. However, in contrast to those grade 9 exams, which are performed with paper and pencil, the normative reference is not this year's students' performance in comparison (relational).

The reference distribution for the national tests was created together with the construction of the item bank and the tests. In principle, this distribution is a fixed reference which ensures that comparisons between students in later years can be interpreted in a fixed framework. In contrast, the paper-and-pencil tests developed at grade 9 provide a normative reference that is based solely on the student performances "this year." This difference means the standard setting using the new national tests allows for comparisons across different years, which is not possible in the case of the old-fashioned paper- and-pencil tests.

7.9 Discussion

In retrospect[7], it should be recognized that there are both positive and negative aspects to the way the adaptive capacity is implemented in this national testing system. One positive aspect is that, for the first time, it has now become possible to test the students with responses given to completely different (adaptive customized) items and still be able to compare the test results. The national test utilizes the Rasch property of "objectivity," or the item validity in harmony with the adaptive bulk of items in the item bank to conduct valid comparisons and measurements that could not be carried out before. As mentioned previously, PISA and IEA's TIMSS and PIRLS all operate with items placed in many separate booklets. In this way, these studies enjoy the same objectivity principles as are built into the item bank for the national tests, making it possible to compare students who have not responded to the same items. A few negative aspects of the way the adaptive principle works can be highlighted (see also Allerup 2013); here, only a few facets are included.

7.9.1 Technical Aspects of Implementation of Adaptivity

The way the adaptive principle is implemented in the Danish national tests represents only one of several possible methods. In this respect, the actual procedure sets the standard for Denmark. A criterion-referenced assessment means that the student

[7] See also the list of references with specific listing of papers etc. concerning adaptive testing.

evaluation results are linked to the learning objectives behind the task with a didactic reference.

From a technical point of view, it is easier to create elements for the standard setting in the case of a *normative framework,* compared to standard settings with *goal-oriented evaluations*. As mentioned previously, a desire to measure student performance in terms of simple *percentage-correct solved items* encounters the problem that students do not solve the same items because of the adaptive principle. There are various attempts to create an *information basis* for goal-oriented services of the test results from adaptive test systems. The wish to gain access to evaluation results, which relate better to the learning objectives that lie behind it, (i.e., enhance their *criteria oriented goals* better than is now the case), is not fulfilled by the NT.

7.9.2 Educational Suitability of the Adaptive Principle

An important part of the assessment of the educational suitability of the adaptive principle is how quickly and accurately it can achieve reliable estimates of the student's achievement level. The suitability of the adaptive principle should be seen in light of the influence of the adaptive principle, the speed with which it can calculate the student's "true" achievement level, and the precision with which the estimate is determined.

The level of speed for the single student during testing depends on management of the above described stop criterion and on the pupil's behaviour when responding to items in the national tests. Seen through a theoretical statistical lens, it is important to distinguish between two types of students: (1) students who are challenged purely by responding to an item and (2) students who are inspired by everything else other than the actual item they need to answer.

Regarding standard setting, the first type allows for a *functional measure of the speed*, and it seems that the adaptive principle leads to total test times, which is lower than the time used by usual traditional linear testing (TLT – paper based). Only in extreme cases, for example, where a student on a usual linear test submits a "blank" (i.e., failure of all questions), or is so skilful that he/she immediately answers all items correctly, can it occur that the adaptive principle possibly will extend the total test time.

The second group of students represents a major problem in the adaptive tests and challenges the efforts of standard setting with the national tests. This group of students may delay the test time in principle indefinitely because the student is more concerned with "teasing" the system than with solving the tasks. The total test time for these pupils can therefore easily exceed the time that would otherwise be used, (e.g., in the linear test).

It could be considered negative for the standard settings, by means of the national tests, that the student, through the adaptive process, may be presented only with items that are special with respect to the content. An argument for that could be that the adaptive procedure in the "next choice of item" focuses much more on the

difficulty of the item than on the *content*. This leaves almost no possibility for content informed evaluation on a class level for the teacher, whereas all pupils have had their particular test run on items from the large pool of items; the NT may, therefore, on a continuum from content to performance oriented assessment be found clearly towards the measurement of performance.

A further related issue could be the fear, for the standard setting, that this takes place without any consideration of aspects of content validity; in other words, the reference for the standard setting would, in such an event, be biased. However, this issue is not a real concern since the items originally chosen for the item bank – within subject-defined limits – are all "equal" in the sense that they can replace each other in a test. This is a consequence of the items being approved by the Rasch model prior to their inclusion in the item bank. Therefore, in this case, there is no benefit of controlling content validity in linear testing through a rigorous choice of items compared to the random selection undertaken by the adaptive procedure.

The reasoning that favors linear testing is faulty, because of the structure of the items in the data bank. All items in each profile area differ mainly by their difficulty and not (ideally seen) by their content (or reference to learning objectives). It can be concluded that the association of the adaptive principle with the construction technique behind the item bank ensures that academic precision in the sense described is the same as that found by the corresponding linear test.

Currently, there is an ongoing discussion concerning how to implement new principles for the evaluation of student works and presentations at exams. This discussion is about a shift from goal-oriented evaluations, executed relative to a set of didactic elements behind the test item, to an evaluation of competencies. It is thought that standard setting of evaluations based on evaluation competencies will encounter problems of inaccuracy, and they will be difficult to manage compared to the goal-oriented procedures currently applied.

7.9.3 Further Advantages and Disadvantages of the Adaptive Principle

Clearly, linear tests are easier to fit into a standard setting because of the replicability of the test. In order to evaluate the advantages and disadvantages of the adaptive principle (vs. linear test) means for testing, it is necessary to clarify that such tradeoffs depend on the psychometric method (model) that is used in the actual implementation. In this test, the Rasch model has been used in the construction.

Advantages and disadvantages of the adaptive principle should also include an evaluation of the item bank, with or without the property of holding items that are homogeneous in the Rasch model sense. In any event, implementation of any adaptive test system requires the presence of a relatively large item bank. It can be considered a disadvantage compared to usual linear testing that, for example, maintenance of the item bank calls for extensive work and analysis resources.

Unlike the traditional linear test (TLT paper based), the elaboration of test items to the bank goes through several stages of work. In a development phase, members of the committee, usually experienced teachers, propose a series of items. This process is monitored by Ministry of Education officials and will be common for national tests and ordinary paper and pencil linear tests. In the pilot phase, testing of items is carried out on students from appropriate grade levels.

In the case of the national tests, it is done among approximately 700 students, which is the minimum number needed to ensure that the trial takes place on a valid level. This process is common to both test types. In the pre-testing phase, internationally known as the *Field Trial* (from the OECD's PISA and IEA's TIMSS and PIRLS studies), all practical elements are tested for the later main testing – practical aspects of test execution and theoretical aspects concerning the psychometric properties of the items belong to this phase. All materials from the main testing will undergo various types of analysis.

In the case of the ordinary linear paper and pencil tests applied in grade 9, the CTS scale, under the fixed distributions of percentages (10%, 25%, 30%, 25%, 10%) on the five upper marks (2, 4, 7, 10 and 12), will create the current year's distribution of results. As mentioned previously, this is a purely normative process. In the case of the national tests, the material will be added to the data already available for running the national tests. It is obviously better to establish a standard setting under the new national tests compared to the situation with ordinary linear paper and pencil tests. In other words, it is by no means an easy task to build an appropriate item bank and keep it up to date.

The most significant way the adaptive principle differs from TLT is that all students will find that they have solved about 50% of the items correctly. This fact is a product of the adaptive principle; technically, it can be easily controlled to be at a different level from just 50%, if so desired. There seems to be a consensus that students do best in test situations when their chance to solve items lies somewhere between 20% and 80%. Outside this range, students suffer from either frustration or boredom. The talented students can usually solve most of the items and might become discouraged with respect to their self-understanding if presented continuously with easy items. The adaptive procedure takes care of this efficiently.

There is a marked difference between the practical test circumstances of ordinary linear tests and the national tests, because with the ordinary paper-and-pencil tests, the student can "go back" and correct an already given response. This is not possible with adaptive national tests because the path by which a student is led through a series of items is determined by the responses already given. Hence, it must be considered a shortcoming in the NT that the student has no opportunity to correct earlier answers.

The stop criterion and the understanding of the ongoing allocation (adaptiveness) of new items make this impracticable. However, it raises no new fundamental problems concerning the mathematical calculation of the student's skill level. It can be said that solving the problem, in retrospect, can be regarded as a version of a traditional linear test, still consisting of Rasch homogenous items.

Part of the way the adaptive principle works is that a student is presented with "the next item" by selecting, at random, an appropriate item with a certain level of difficulty. Therefore, two students with the same estimated level of achievement face different interpretations of the "next item" depending on the outcome of the random selection. There is a risk that this side of the adaptive principle can cause a sense of confusion among students, as they experience the implementation of NT as a "bouncing around" from issue to issue – although the items actually adopted are within the same profile area. It is normal practice in the design of TLTs to avoid a "mixed" content in successive items.

This causes the operationalization of the adaptive principle to stop a basic principle, which is otherwise used for the construction of tests: where students are presented successively to a number of items, there must be some kind of internal coherence among the items. In conventional linear tests (final examinations, etc.), standard settings carried out by means of items designed as in the national tests are, therefore, very different from the kind of items presented in "bundles" like OECD, PISA and IEA PIRLS. Similarly, a certain consistency is stressed by a series of mathematics items where, for example, particular geometric items often use the same graphic basic structure as a common frame for the single items.

References

Allerup, P. (1994). *Rasch measurement, theory of the international encyclopedia of education* (2nd ed.). London: Pergamon Press.

Allerup, P. (1997). Statistical analysis of data from the IEA reading literacy study. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences*. New York: Waxmann Verlag.

Allerup, P. (2005). *Statistics and test – Some conditions and opportunities* (p. 140). Vejle: Kroghs Forlag.

Allerup, P. (2007). Identification of group differences using PISA scales. In S. T. Hopmann & G. Brinek (Eds.), *PISA according to PISA – Does PISA keep what it promises?* Vienna: University of Vienna.

Allerup, P. (2013). Evaluation of national tests – Expert assessment 2. Ministry of Education (Quality and Supervision Agency). Annex to the evaluation of the national tests in primary schools.

Allerup, P., Klewe, L., Torre, A. (2013). Young people's choice and opt-out in secondary education; quantitative perspective. Aarhus University, Department of Education (DPU).

Cizek, G. (2012). An Introduction to contemporary standard setting. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed.). London: Routledge.

Cizek, G., Bunch, M., Koons, H. (2004). *Setting performance standards: Contemporary methods*. NCME.

Hays, R. (2015). Standard setting: The clinical teachers toolbox. *The Clinical Teacher, 12*, 226–230.

Kousgård, E. (2003). *50 år med intelligensprøven BPP, Militærpsykologisk Tjeneste*. Copenhagen: Defence Command Denmark.

Kousholt, K. (2012). The national tests and their meanings – From the children's perspective. *Educational Psychology Journal, 49*(4), 273–290.

Kousholt, K. (2013), Children as participants in social practice test. *Educational Psychological Journal*, special issue.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Munksgaard; (1980) Chicago: University of Chicago Press. ISBN 0-941938-05-. LC # 80-16546.
Rasch, G. (1968). *A mathematical theory of objectivity and its consequences for model construction*. Report from the European Meeting on Statistics, Econometrics and Management Sciences, Amsterdam.
UVM. (2013). *Evaluering af de nationale tests i folkeskolen: Rapport*. Copenhagen: Rambøll.
Zieky, M. (2012). So much has changed: An historical overview of setting cut scores. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed.). London: Routledge.

Specific Publications Concerning Adaptive Testing

Babcock, B., & Weiss, D. (2012). Termination criteria in computerized adaptive tests: Do cariable-length CATs provide efficient and effective measurement. *Journal of Computerized Adaptive Testing, 1*(1), 1–18. doi:10.7333/1212-0101001.
Eggen, T. J. H. M. (2009). Three-category adaptive classification testing. In W. J. Van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 373–387). New York: Springer.
Eggen, T. J. H. M. (2010). Three-category adaptive classification testing. In W. Linden & C. Glas (Eds.), *Elements of adaptive testing*. New York: Springer.
Glas, C. A. W., & Vos, H. J. (2010). Adaptive mastering testing using a multidimensional IRT model. In W. Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing*. New York: Springer.
Makransky G. (2012). Computerized adaptive testing in industrial and organizational psychology. (Ph.D. thesis), University of Twente, The Netherlands. 2012 ISBN: 978-90-365-3316-4.
Rapport fra REVIEW-panelet. (2007). *De nationale it-baserede test i folkeskolen*. Devo Team Consulting.
Scheuermann, F., & Bjørnsson, J. (Eds.). (2009). *The transition to computer-based assessment new approaches to skills assessment and implications for large-scale testing*. Luxembourg: European Communities.
Tao, Y. H., Wu, Y.-L., & Chang, H.-Y. (2008). A practical computer adaptive testing model for small-scale scenarios. *Educational Technology & Society, 11*(3), 259–274.
Triantafillou, E., Georgiadou, E., & Anastasios, A. (2006). The design and evaluation of a computerized adaptive test on mobile devices. *Computers & Education, 50*(4), 1319–1330.
Vos, H. J., & Glas, C. A. W. (2000). Testlet-based adaptive mastery testing. In W. Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice*. Dordrecht: Springer.
Walter, O. (2010). Adaptive tests for measuring anxiety and depression. In W. Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing*. New York: Springer.
Ware Jr., J. E., Bjorner, J. B., & Kosinski, M. (2000). Practical implications of item response theory and computerized adaptive testing: A brief summary of ongoing studies of widely used headache impact scales. *Medical Care, 38*(9), II73–II82.
Weiss, D. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development, 37*(2), 70–84.
Weiss, D., & von Minden, S. (2011). Measuring individual growth with conventional and adaptive tests. *Journal of Methods and Measurement in the Social Sciences, 2*(1), 80–101.

Chapter 8
Experiences with Standards and Criteria in Sweden

Gudrun Erickson

Abstract After decades of norm referencing, a criterion-referenced grading system was introduced in Sweden in the mid 1990s. The shift brought about a number of changes at different levels of the educational system. In this, the national tests, with their long tradition and high degree of acceptance, were seen as one way of implementing the new system. Hence, the tests were given several explicit aims, from clarification of subject syllabuses and criteria and active, positive impact on learning, to advisory tools to enhance fair grading at the individual level, as well as stability over time. After a brief introduction of the system and its various developments and challenges, the chapter focuses on the current nature and status of the national tests, including the issue of standard setting. For a number of years, this has been the responsibility of the different universities developing the tests, which has brought about certain differences regarding methods as well as outcomes. Lately, however, attempts have been made to develop a common framework for the tests, including procedures for setting standards. In the chapter, a brief report will be given of this on-going work, which is part of a general analysis of the Swedish national assessment system at large.

Keywords Swedish school-system • Criterion referencing • National tests • Multiple aims • Standard setting • Common framework

8.1 National Assessment in Sweden: A Brief Overview

The Swedish educational system, including national assessment, has undergone major changes during the last few decades, which has brought about discussions of a number of crucial issues. However, it needs to be pointed out that not everything

G. Erickson (✉)
Department of Education and Special Education, University of Gothenburg, Gothenburg, Sweden
e-mail: gudrun.erickson@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_8

has changed. For example, national curricula and subject syllabuses have been kept, albeit in partly new forms, and teachers still have the responsibility for awarding individual grades. Grading is supported by nationally provided tests, marked by the students' own teachers. The tests, as well as additional materials aimed to help teachers in their evaluating role, are developed by different universities in the country, appointed by the National Agency for Education (NAE). To facilitate the understanding of the system as well as the reading of this chapter, where different aspects are elaborated on, a brief overview is given in Table 8.1.

Table 8.1 National assessment in Sweden: an overview

Time	Situation/Action	Issues
Before 1994	Norm-referenced system Grades from school-year 8 Five-point grading scale Standardized tests (En, Ma, Sw) to establish group mean	The fairness and impact of norm-referencing The role of the three standardized tests in grading for all subjects
1994–2011	Introduction of a goal- and criterion referenced system; No specific content or methods stipulated in the national curricula and subject syllabuses; schools and teachers to decide Four-point grading scale; initially no criteria for the highest level (introduced in the 2000 revision of the subject syllabuses) National tests (En, Ma, Sw) with several aims – to enhance learning, implement curriculum, support fair grading, contribute to studies of development As compared to the earlier system, a large proportion of performance assessment tasks, requiring teachers' qualitative judgement Gradual development of a wide array of support materials for teachers' continuous assessments and for clarification of national standards Gradual awareness of increasing inequity across schools and considerable variability in teachers' marking and handling of national test results in relation to individual grading Towards the end of the period, gradual introduction of national tests for younger students and for new subjects	Issues related to the interpretation of verbal descriptors for levels of competence Effects on fairness and equity due to local decisions about subject content and criteria for the highest grade level Effects on reliability and equity due to the different aims of the national tests Variability in marking of national tests, especially regarding performance assessment tasks The advisory role of the national tests for teachers' final grading: the balance between too much and too little support

(continued)

Table 8.1 (continued)

Time	Situation/Action	Issues
2011	New curricula and subject syllabuses introduced A section labelled "Central Content" introduced in all subject syllabuses Criteria/Performance standards with generic "value words" describing progression (similar across subjects) Grades from school year 6 Six-point grading scale (A-F), with three levels verbalized (A, C, E); teachers to decide on B, D and F "Threshold rule" for final grades, allowing no compensation for uneven profiles; all aspects of the intended standards have to be met Mandatory national tests of Natural and Social Sciences for grade 6 and 9; the mandatory grade 6 tests abolished in 2015 (turned into assessment support materials) According to an NAE study (2016), severe problems with the generic performance standards and the threshold rule; major investigation and possible revision suggested Framework for the national testing system underway	Issues of workload in relation to the national tests The interpretation of the performance standards, in particular the generic 'value words' The role of "Central Content" in relation to the national tests The role of grade levels D and B in national tests The effect of the threshold rule in national tests The structure and function of the future framework for the national tests

8.2 Background

In the Swedish educational system, there is a long tradition of trust in teachers' continuous assessments of students' knowledge and skills, and in teachers' ability to aggregate their various observations into single subject grades. These grades are used, to a large extent, for admission to higher education, which makes them extremely important for individual students' educational choices, or, put differently, for young people's life chances and self-image. As compared to most countries, grades are introduced fairly late; at present from school year six, when students are typically twelve years old. In the earlier school years, students' achievements are summarized and reported in individual development plans, including written reports.

Major changes to the Swedish school system were introduced in the mid 1990s, the most noticeable being a shift from a highly centralized system to the opposite, namely a system in which local municipalities and independent schools were given the freedom to make most decisions about their own schools. However, binding national curricula, including subject syllabuses, were kept, as was a national grad-

ing system. In parallel, and after many years of norm-referencing, a "goal and knowledge-related" system for grading was launched (Gustafsson 2006). The latter implied that teachers were required, quite suddenly, to award individual grades to students based on verbal criteria, or performance standards, that were not always perceived as clear enough and that initially did not cover all grade levels. Furthermore, the degree of novelty introduced in the content standards varied considerably among subjects, the characterization of mathematics, for example, being distinctly different from the previous one, whereas, English to a large extent resembled what had been defined in the preceding syllabus.

National curricula, albeit to some extent varying in type, have a long tradition in the Swedish educational system as part of a centralized definition of goals for schooling in general, as well as for individual subjects. In addition, national standardized tests, or assessment systems, have been provided since the middle of the twentieth century (Gustafsson et al. 2014; Marklund 1987). Both these phenomena – curricula and national tests – have traditionally been well accepted by teachers, which may, to some extent, be due to the fact that teachers' autonomy has remained at a high level. A clear manifestation of this is that the national tests have been – and still are – advisory in their function and are meant to be combined with teachers' continuous observations; consequently, there are not exams of the traditional, decisive kind. Furthermore, teachers usually mark their own students' national tests, with recommended but not mandatory co-rating, something that has been widely criticized, both nationally and internationally (Nusche et al. 2011; Skolinspektionen 2013). Types of test materials have varied, as have subjects, however with a common core of Swedish (and later Swedish L2, Swedish as a second language), Mathematics and English. After the shift to criterion/standards related grading, the role of the national tests has been expanded to the individual level. Furthermore, issues of format in relation to standards have become focused upon, for example generating questions like "does the demand for reflection in social sciences require constructed response items, or can this ability be assessed using selected response formats?"

In the early 1980s, a decision was made that the national tests were to be developed by different universities in the country with proven, strong research within the subjects in focus and in educational assessment. The reason for this was related to quality as well as to the legitimacy of the system; it was considered less appropriate that the authority responsible for national curricula also took charge of the development of instruments used to check students' attainment of these curricula. The delegation is still the case, with the National Agency for Education (NAE) as the coordinating and responsible authority for the national assessment system. However, since the establishment in the early 2000s of the Swedish Research Council ("Vetenskapsrådet"), the NAE does no longer fund research in connection with test development, hence, funding for these activities has to be sought from other sources.

8.3 Early Discussions and Dilemmas

The introduction of the national tests and diagnostic materials accompanying the new, criterion-referenced grading system in the middle of the 1990s followed a period of intense discussions of various issues of assessment, not least questions of aims and functions. It became clear that there was a strong ambition from the national educational authorities to distance the national tests, not only from the preceding norm-referenced system but from the traditional concept of measurement as well, and instead strengthen their role in the implementation of the new syllabuses and the enhancement of learning, for students as well as for teachers. In this, aspects of exemplarity and washback were strongly focused upon. This widening of the assessment concept was by no means an isolated, Swedish phenomenon but coincided with developments in many countries in the western world. For example, Caroline Gipps' *Beyond testing,* published in 1994, and the work by Paul Black and Dylan Wiliam and other members of the Assessment Reform Group in the UK (see Gardner 2012), played an influential role in the discussions. In this, aspects of validity were emphasized and – unfortunately – sometimes depicted and perceived as the opposite of reliability. A manifestation of this in the national tests was the reluctance of the NAE and some test developing universities to use points as indicators of quality – and when points were actually used, cut-offs were in some cases recommended rather than stipulated.

Another sign of the ambition to change was that different forms of performance testing, not seldom generating extensive written responses, were strongly preferred to selected response formats. Furthermore, confidentiality was toned down to enable the use of the national tests as didactic tools in the classroom (thereby effectively preventing the use of anchor items). Another aspect of the wish to tone down the test character of the materials was the decision not to develop standardized specifications for the national materials, but rather to encourage specificity, and thereby differences, among the different subjects, regarding development processes as well as products.

As already mentioned above, the issue of aims of the national tests was one of the initial concerns of the national authorities. For a number of years after the introduction of the new system in the 1990s, it was publically announced that the Swedish national tests aimed to

- enhance students' educational achievement,
- clarify goals and indicate strengths and weaknesses in individual learner profiles,
- concretize goals and criteria,
- enhance equity and fairness in assessment and grading,
- provide data for local and national analyzes of educational achievement.

Following criticism from different experts, pointing out that a single test cannot, with maintained quality, cater for a number of different aims, and in particular

expressed in a government investigation (SOU 2007), the five aims were eventually reduced to "only" two, namely to

- enhance equity in assessment and grading and to
- provide evidence for local and national analyzes of educational achievement.

The pedagogical aims were kept to some extent, however, stating that "the national tests may also contribute to the concretization of goals and criteria and enhanced student achievement." Consequently, the change to the description of aims was quite small, with little visible effect on test development and instruments.

During the first decade of the new system, roughly between 1995 and 2005, the educational authorities were clearly cautious about interfering too much at the local level – fearing that the national tests would provide too much support and be perceived, and treated, as exams rather than advisory assessment materials. Consequently, individual schools and teachers were given very much responsibility. However, a growing number of observations and studies indicated considerable differences in handling national tests and in awarding grades at the school as well as the municipality level. Following this, issues of fairness and equity were raised and given considerable attention in the general debate. In 2007, a report was published by the National Agency for Education focusing on the relationship between national test results and final grades, questioning the fairness of the system (Skolverket 2007). This may be seen as a kind of turning point, leading to more analyzes of issues related to equity, an increased demand for clarification of rules and criteria, and more documentation at the classroom and school levels. In a way, it can be claimed that the situation changed quite quickly from distinct worry about the national tests influencing *too much* at the local level to the opposite, namely concern about *too little* influence, with the balance between the two seemingly very difficult to find.

Underpinning recent developments and the current situation regarding the national assessment system, in particular two external types of criticism need to be mentioned, namely from the Swedish Schools Inspectorate and from the OECD.

Between 2009 and 2013, the Swedish Schools Inspectorate (SSI), commissioned by the Government, conducted annual re-rating studies of randomly sampled national tests. The results indicated considerable variability in marking, and also some evidence of teachers' more positive ratings as compared to external raters, especially of performance-based tests. The latter was especially noticeable in essays of Swedish and, to some extent, English. The conclusion of the Inspectorate was that the inclusion of essay tasks in national tests should be investigated further, with the aim of determining whether this type of test should at all be included in large-scale, standardized tests (Skolinspektionen 2013). The results of the SSI studies were heavily and promptly publicized and contributed to a debate about teachers' professionalism and responsibility for rating and grading. However, there are also studies to some extent contradicting the SSI result (Erickson 2009), and raising serious methodological concerns, warning about the conclusions drawn on the basis of the SSI studies (Gustafsson and Erickson 2013).

Nusche et al. (2011) identified several positive features, when, on behalf of the OECD, they evaluated the Swedish national assessment program. The involvement of many categories of stakeholders is mentioned, as well as the transparency and availability of data. Furthermore, the trust in teachers in marking the tests is seen as a way of enhancing professionalism. However, they also criticize the strong emphasis on performance based testing, which, in their opinion, contributes to inconsistency in ratings, and suggest the addition of items and tasks with more closed formats. One of their main conclusions is that there is a definite need for a coherent framework for the whole assessment program. We will return to this issue later in this chapter.

8.4 The Current System

In 2011, new national curricula and subject syllabuses were introduced, which brought about additional changes with clear impact on the national assessment system. Criterion-referencing remained, but a distinct novelty was the six-point grading scale (A-F) replacing the four-point scale that had been in existence since 1994. Five of the six levels (A-E) in the new scale imply a Pass or above, however only three of them (A, C, E) with verbalized performance standards; levels D and B are meant to illustrate a level where the lower grade is completely attained, as are the majority of the standards for the higher grade. Teachers are required to make these interpretations and judgements.

Emanating from decisions at the political level, a number of new national tests have been introduced, for new subjects and for new age groups (for an overview, see Table 8.2), and the point in time when subject grades are to be awarded for the first time has been lowered from, typically, 14 to 12 years of age (from school year 8 to 6). This also means that a new group of teachers with no previous experience of highly summative evaluation of this kind now has to award grades, supported by the national curriculum and the subject syllabuses, and by the national tests. Moreover, it should be mentioned that the national assessment system comprises a wide array of support materials for teachers, either in the form of diagnostic or formative materials, or as explicit "training kits" for internal use at, or, among schools.

The subject syllabuses accompanying the 2011 curricula comprised a new section labeled *Central Content*. Reflecting the extensive decentralization in the 1990s, the former syllabuses laid down the goals of each subject and the criteria for grading but gave no indications of concrete content or methods; individual schools and teachers were expected to make the necessary decisions. The growing criticism of this, especially from the point of view of transparency and equity, brought about several changes aimed at increased clarity and standardization, the Central Content section in the syllabuses being one, the growing number of commentary and exemplifying materials from the NAE previously mentioned another.

As before, the content standards for individual subjects were more or less changed as compared to previous descriptions. The performance standards, referred

Table 8.2 Mandatory national tests in Sweden (2016)

School year (Lower sec.)	Subject/Subject course	Comments
3	Mathematics; Swedish/ Swedish L2[a]	*Summative function + support for educational planning*
6	English, Mathematics; Swedish/Swedish L2	*One test of Natural Science + one of Social Science mandatory 2013–2014*
9	English; Mathematics; Swedish/Swedish L2	
	One test of Natural Sciences	
	One test of Social Science[b]	
Courses (Upper sec. and municipal adult ed.) Swedish for adult immigrants	English 5; English 6	*Course specific tests*
	Mathematics 1–4	
	Swedish/Swedish 2–1 & 3	
	Course B, C, D	*Course specific tests*

[a]From 1 July 2016, assessment materials – not regular tests – for school year 1, focusing on the development of literacy and numeracy, are mandatory for teachers to use; the materials are accompanied by performance standards for reading comprehension for school year 1

[b]There are three subjects in Natural Science (biology, chemistry, physics) and four in Social Science (geography, history, religious studies, social studies). National tests are developed for all subjects; schools are randomly assigned one test from each category

to as *knowledge requirements*, however, were distinctly different with generic descriptions of quality used across subjects. These descriptions are quite general and abstract, comprising generic so called *value expressions* to illustrate differences and progression among levels of knowing. Strong doubts have been expressed from the beginning regarding the degree of support they actually provide for equal assessment and grading, and for the use in defining and enhancing knowledge in an adequate way (Carlgren 2015; Gustafsson et al. 2014).

A recent report from the National Agency, following a government initiated study, confirms this criticism, showing, for example, that more than half of teachers in lower and upper secondary school find the performance standards unclear (Skolverket 2016). Further, it has been shown that the number of teachers and students who feel that the national standards serve a clarifying function is significantly lower than before the reform in 2011. One problem frequently highlighted is that the expressions used to describe qualitative differences need to be interpreted – by individual teachers and for individual subjects. Also, there are considerable differences between the structure and principles of different subject syllabuses, which is perceived as further complicating the use of the documents.

Another factor which has been criticized already from the introduction of the 2011 curricula and syllabuses is a rule against compensation in grading, i.e., a requirement that all aspects of the performance standards have to be met for a student to be awarded a certain grade, referred to by the NAE as *the threshold rule* (Gustafsson et al. 2014; Vlachos 2013). Thus, unevenness in knowledge profiles – or, expressed differently, multidimensionality within the ability demonstrated – is

penalized. Critics have focused on theoretical objections as well as pedagogical implications, the latter regarding a number of effects perceived as negative; teachers have to focus, to a large extent, on detecting what their students cannot do, and many students express that they give up from the beginning, if they consider their knowledge profiles uneven.

In the recent report from the NAE (Skolverket 2016), the non-compensatory rule for grading in combination with the unclarity of the performance standards are identified as important factors affecting fairness and equity in a negative way. However, it is pointed out that there are additional factors as well, and that different remedial actions are being planned to change the situation. Following this, the NAE have taken immediate action based on the results of the study, opening up for a somewhat more generous interpretation and use of the compensatory rules, and, importantly, suggesting a major investigation of the system at large.

Traditionally, the national tests in Sweden have been very well received by teachers, as shown, for example, in a recent report by the National Agency (Skolverket 2014). Positive features have concerned perceived alignment to the national syllabuses and thereby a clarifying and implementing function; correspondence with teachers' own judgements; appreciation from students, etc. Since the materials are quite extensive, many teachers have also described the workload in administering and marking the tests as considerable, but it has generally been felt "to be worth it" in relation to the strong support provided.

However, due to the rapid introduction of a number of new tests in new subjects and for new age groups – and thereby also new groups of teachers – around 2012, the attitudes to the national assessment system started to change, rapidly and quite considerably. Especially for teachers in school year 6, who had no experience in awarding grades, and who sometimes taught all subjects to their class, the system was very demanding, with a maximum of six national tests to handle during the spring term, all of them including several subtests. In a hearing organized by the National Agency in August 2014, the changing attitudes became very clear; a large number of stakeholders met for a day to share their views on the national assessment system. Although the tests as such were generally given positive comments, criticism of workload, lack of digitalized tests, lack of external rating, and vagueness regarding the aims and weight of the tests became quite evident.

8.5 Test Results – Characteristics, Stability and Use

The national assessment system generates a large amount of data, forming the basis for a number of different analyses within and between subjects. However, due to the lack of designated research funding from the NAE, these studies cannot be carried out on a regular basis but have to be accommodated for within the test development budget, unless funding from other sources is granted. In the following, a brief account will be given of some of the observations made of test results, focusing on those related to issues of stability over time.

Table 8.3 Aggregated national test results for school year 9, 2013–2015

Subject	Aggregated nat. Test score 2013	Aggregated nat. Test score 2014	Aggregated nat. Test score 2015	Range of variability 2013–2015
English	14.9	15.0	15.1	*0.2*
Mathematics	12.2	11.4	10.7	*1.5*
Swedish	13.3	13.6	13.5	*0.3*
Swedish L2[a]	9.2	9.4	9.2	*0.2*
Biology	10.9	12.5	11.9	*1.6*
Chemistry	11.2	11.4	13.3	*2.1*
Physics	11.9	11.4	11.8	*0.5*
Geography	13.0	13.1	12.6	*0.5*
History	11.2	12.4	10.7	*1.7*
Religious studies	11.3	12.8	13.8	*2.5*
Social studies	13.8	12.6	12.9	*1.2*

[a]The results from the test of Swedish as a second language will not be commented on, since the number of test-takers is distinctly lower than for the rest of the subjects (approx. 1/10 of English, Mathematics and Swedish). What can be noted, however, is that the results are consistently at a very low level, below 10.0 which is the pass level.

First of all, it is clear that uneven profiles are very common, both at individual and aggregated levels, and that they can be traced in all subjects. This may be seen as an internal, subject-related phenomenon but is actually of general interest in relation to the non-compensatory national rule for grading described earlier.

Another observation, which needs to be considered from different angles, is the consistent difference in results among subjects, both regarding level of difficulty as manifested in the national test scores, and the variability among years. In Table 8.3, the aggregated national test results from school year 9 (end of compulsory school) from 2013 to 2015 are shown, hence the first 3 years with the new, six-point grading scale. (All students' results are collected, and the maximum test score equals 20.0.)[1] In spite of certain weaknesses, for example, different sample sizes and partly different degrees of data loss, a number of interesting observations can be made. One concerns the clear, and consistent, difference regarding the level of results among subjects (that has been relatively stable since the introduction of the criterion-referenced system in the 1990s). Looking at the core subjects, there is an obvious discrepancy between mathematics and English, with distinctly lower results in mathematics and higher in English.

This can be further exemplified by the grades from the 2015 national tests, where 19% of the students had an F in mathematics, whereas the corresponding number for English was 4%. Looking at the other end of the scale, 6% of the students gained an A in mathematics, i.e. the highest grade, as compared to 20% in English. These results can be interpreted in different ways and should be discussed from different

[1] The number of students per sample varies from > 80,000 for Swedish, Mathematics and English (the full cohort), between 25,000 and 30,000 for Natural Sciences (a third of the cohort, due to the random distribution of the three subject tests), around 20,000 for Social Sciences (a fourth of the cohort), and c. 8000 for Swedish as a second language (a relatively small group of students).

angles, epistemological, pedagogical, methodological, etc. Questions that can be raised are, for example, whether mathematics is more cognitively demanding than languages, or in this case English in the Swedish context; if the syllabuses differ in essential ways regarding structure, level of standards, number of standards etc.; whether the test of mathematics is too difficult and/or the test of English too easy in relation to their respective constructs; if teachers of mathematics are less professionally skilful/successful; if teaching materials of mathematics are not sufficiently aligned to the national standards; whether extramural exposure to the subjects differs in significant ways; if traditions in the two subjects vary considerably, etc. A single answer to these question is not likely to be found, but rather a combination of contextually sensitive aspects of the issue at large.

The question of variability over time is complex and involves aspects of curriculum as well as test development (including specifications), analyses and, not least, standard setting. Examples of subjects with distinctly varying results across years, as demonstrated in Table 8.3, are Religious studies and Chemistry. One aspect to be noted here is that the national tests of Natural Sciences and Social Sciences are relatively new in the system, introduced in 2009 for Natural Sciences and 2013 for Social Sciences. In spite of this, however, two subjects – Geography and Physics – demonstrate much lower variability than the others in the same group of subjects. Looking at the core subjects, with a very long tradition, Mathematics emerges as the subject with the most noticeable variability across years. Again, it needs to be remembered, that there are many ways of analyzing the issue of stability and its function, not least in a system like the Swedish, with advisory national tests.

To further consider the characteristics of the core subjects – English, Mathematics and Swedish – the following graphs illustrate the difference between the year 9 aggregated national test grades (online version: green, paper version: gray line) and teachers' final grades (black line) per subject between the years 1998 and 2012. (It should be mentioned that there was a major confidentiality problem, a leakage of the tests of English and mathematics in 2003, which to a large extent invalidated the results.) (Figs. 8.1, 8.2 and 8.3)

Several observations can be made on the basis of the graphs. First of all, teachers' grades seem more stable over time than the national test grades. This is particularly noticeable for mathematics. Further, the levels of Swedish and English, in particular the latter, seem to increase gradually, both in national test scores and teachers' grades, something that can be discussed from societal as well as methodological points of view. Another observation is that teachers' grades are consistently higher than test grades for Mathematics and Swedish, whereas for English the two types of grades coincide to a large extent. What needs to be remembered, is that there are no clear regulations concerning this relationship; the only information given is that the national tests results are intended to "support teachers' grading" and that, when awarding final grades, teachers need to take "all available information" about individual students' knowledge and competences into account. Consequently, the national tests are not to be used as single examinations, overruling the continuous observations made by teachers.

In spite of the relative vagueness in the descriptions of the weight of the national test results in relation to teachers' final grades, the National Agency publishes

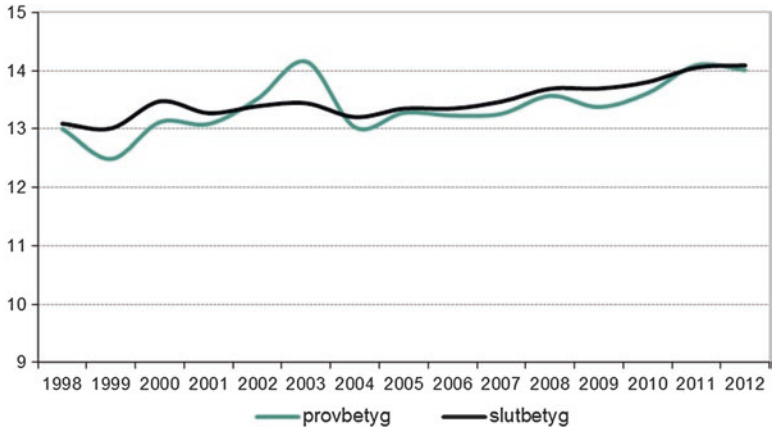


Fig. 8.1 Aggregated test scores (more variable line) and teachers' final subject grades (more stable line); school year 9, 1998–2012, ENGLISH

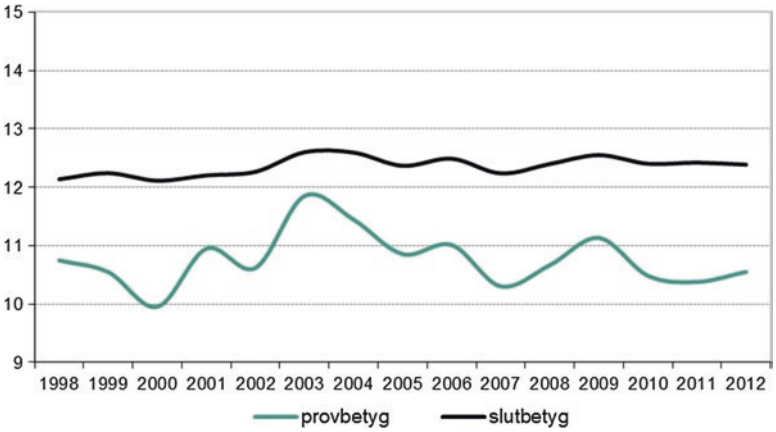


Fig. 8.2 Aggregated test scores (lower line) and teachers' final subject grades (upper line); school year 9, 1998–2012, MATHEMATICS

annual reports focusing on this issue. Also, the Schools Inspectorate frequently uses the information in their inspections of individual schools. This is sometimes criticized but is also regarded as an important indication of the functionality of the system, and as one aspect of the complex issue of fairness and equity.

In Tables 8.4 and 8.5, the relationship between teachers' final grades (TFG) and the aggregated national test grades (ATG) is shown for school year 9 and some courses in Upper Secondary School, spring 2015. The middle column shows the percentage of cases where the two grades coincide; the left column cases where the

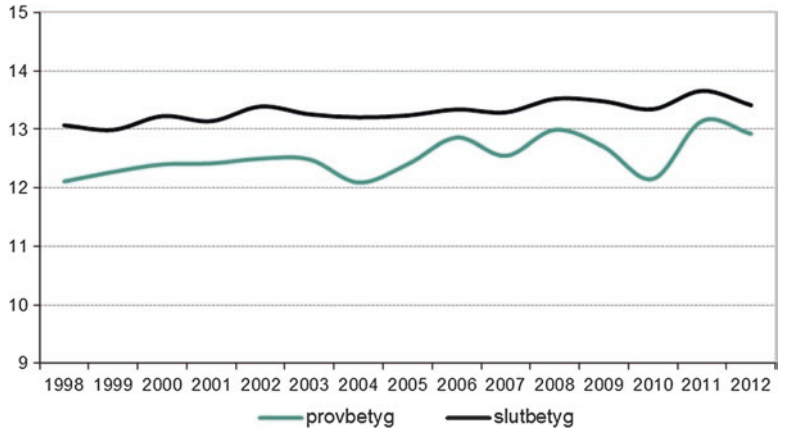


Fig. 8.3 Aggregated test scores (lower line) and teachers' final subject grades (upper line); school year 9, 1998–2012, SWEDISH

Table 8.4 Aggregated test grades (ATG) in relation to teachers' final grades (TFG); year 9, spring 2015 (n≈85,000 En, Ma, Sw; NSc 1/3 sample; SocSc 1/4 sample)

SUBJECT YEAR 9	TFG<ATG (%)	TFG=ATG (%)	TFG>ATG (%)
English	15	74	11
Swedish	9	64	27
Swedish L2	7	65	28
Mathematics	2	60	38
Chemistry	15	65	20
Biology	6	58	37
Physics	5	60	34
Geography	5	64	31
History	2	40	58
Religious studies	9	71	20
Social studies	6	65	29

Table 8.5 Aggregated test grades (ATG) in relation to teachers' final course grades (TFG); upper secondary school, spring 2015 (n > 20,000)

SUBJECT/COURSE UPPER SEC. SCHOOL	TFG<ATG (%)	TFG=ATG (%)	TFG>ATG (%)
English 5	14	73	13
Swedish 1	10	61	30
Mathematics 1a	2	69	29
Mathematics 1b	2	75	24
English 6	11	69	20
Swedish 3	10	51	39
Mathematics 2b	1	48	51

final grade awarded by the teacher is lower than the test grade, and the right column the opposite, where teachers' grades are higher than the aggregated test scores.

A number of observations emerge from the tables, the first regarding the similarity between the patterns for lower and upper secondary school; in both cases, with the exception of English, teachers' grades are higher than the aggregated test results, often quite significantly so. Teachers apparently either find the test standards too demanding or evaluate what the students do in the classroom more positively as compared to the national tests. Further, the number of cases where the two grades – teacher- and test generated – coincide range from 40 to 75% (History in year 9 and Mathematics 1b, respectively).

The case of English is interesting, since it actually shows that a number of teachers in lower secondary school consider the national test too generous. This is a complex issue well worth analyzing and discussing, however not within the scope of this chapter.

8.6 Test Development and Standard Setting

The quality of assessment – the degree of validity and reliability of inferences, decisions and actions (Moss et al. 2006) that can be claimed – is due to a process of test development, in which each step is of vital importance. This process obviously includes decisions about cut-scores and benchmarks, commonly referred to as *standard setting*.

Test development, as well as standard setting, can be done in many different ways but are clearly inter-related activities, where actions in the one obviously impact on the other.

When looking at the gradual development of the Swedish national assessment system, it becomes quite obvious that there is no real standard setting tradition in the country. This is partly due to the long period of norm-referencing at the national level (more than 30 years), when cut-off points were statistically generated. However, it is also related to the sovereign role of teachers in awarding grades, without binding exams having the final, summative and decisive function. It also needs to be borne in mind that standard setting is strongly associated with measurement and psychometrics, areas which for quite a long time have been regarded with certain suspicion at the national educational level in Sweden (cf. what has been said previously about the introduction of the new assessment system in the 1990s).

In addition, it can be assumed that the vagueness regarding the weight of the national tests in teachers' decision making further contributes to the relatively weak focus on, and experience in, standard setting as such. However, dropping school results – discovered through international surveys (since the Swedish national tests, due to their change and variability over time, cannot be used for trend studies) – in combination with an increase of between-school variation, have brought about discussions regarding education and equity. This has led to more attention being paid to issues of test quality, including standard setting, which is a highly complex area

with significant effects on test results and their use, that is on validity at large (Koretz 2008).

Test Development Processes at the different universities involved in the national assessment system show both similarities and differences, the latter being somewhat more conspicuous. However, there is a firm, common basis in alignment to the national subject syllabuses, defining the constructs for the different tests. Furthermore, all test developing groups collaborate with different stakeholders, among which teachers and researchers are the strongest, but students also play a role, however, clearly more so for some of the groups. Also, piloting of materials is a self-evident element in the test development process, albeit varying to a considerable extent from one subject to the other, some with quite small groups, sometimes convenience sampled, others using iterative, small-scale piloting to prepare for large-scale pretesting rounds (n=>400) in randomly selected groups in the country. On the whole, the different institutions report very different experiences when it comes to finding schools and teachers willing to take part in pre-testing activities; for some subjects this seems very difficult, whereas others routinely have to say no due to too many willing candidates. It also needs to be mentioned that there are differences regarding changes allowed to items after final, large-scale pre-testing, ranging from generosity to strict restrictiveness.

As for clear differences across the test developing projects, quite a number can be mentioned, one of the most distinct probably being the existence of explicitly level differentiated items or tasks in some tests, but not in others. In tests with this differentiation, there are specific, non-compensatory demands at each grade level regarding the types of mastered items required, whereas in non-differentiated tests, all items are counted equal and eventually aggregated into a total sum, based on which the standards are set. In other cases, however, differences are more a matter of degree than of actual existence. One example is test specifications and internal frameworks that most often exist but show considerable variation regarding level of detail and transparency. Other differences concern the use of – and feelings about – different test formats, or put more explicitly, the proportion of selected and constructed response items and tasks.

The degree of, and reliance on, performance assessment, most often in written form, is one example of this, with aspects of validity involved, regarding possible construct irrelevant variance as well as construct under-representation (Messick 1989). In addition, linking procedures and use of anchor items vary to a considerable extent, from no empirical linking to anchor items in pretesting as well as in sharp tests. Further, both similarities and differences can be found regarding analyses performed, with classical methods being generally used, whereas Item response theory (IRT) is more common in some subjects than in other ones. Finally, and as part of test development, standard setting procedures vary to a considerable extent.

Standard Setting is undertaken by all test development groups, albeit in different ways. However, in this case as well, there are similarities at a more general level. First of all, this concerns strict focusing on the content and performance standards

stipulated in the national subject syllabus, which requires interpretation of verbal descriptions and transformation into benchmarks and cut scores. Consequently, a judgemental approach is generally used, in which teachers have an active, often central, role, and where the procedures followed are strongly influenced by Angoff-related methods in a wide sense (Cizek and Bunch 2007). Finally, and importantly, the standards identified are meant to define minimal levels for the different grades, not the kind of "average" or "normal" levels that can sometimes be seen and that inevitably leave room for interpretation of how much less can be accepted within a certain level.

As for differences in the standard setting procedures, it needs to be borne in mind that, preceding the decision what is required for a certain level, is the need for reliable and stable differentiation in some subjects concerning level specific items and tasks, i.e., what distinguishes an E task from a C task etc. And once this is done, decisions have to be made regarding what should be required in terms of number of mastered items, and combinations of items and tasks, for each level. Another considerable difference across the different subject groups is the active use – or none-use – of empirical, pre-testing data in the standard setting rounds: are data presented to the participants, and if so, what data are shown and discussed, and how are they intended to be used? Other differences across subject groups concern the number of participants and the proportion across categories of participants, preparatory training procedures for participants, and the organization of the standard setting sessions. Furthermore, variation among procedures also concern the role of the suggestions made by the panellists, the analyses of these suggestions in relation to other sources of evidence, and the final decision making procedure – when this is done, where it is done, and by whom.

The different subject groups developing national tests work fairly independently and with their own, internal routines, which are also described in documentation provided for the NAE. These documents are often quite extensive and describe the whole test development process. Emphasis is on construct related issues, including the performance standards provided in the national subject syllabuses, and on process related factors. It is quite clear that not very much is said about standard setting, more than in quite general terms. This may be seen as an indirect example of the lack of tradition regarding this link in the test development chain.

In a recent analysis, the National Agency (Skolverket 2015) analyzed the relative importance of different sources of error in the national tests, aiming both to understand the unreliability of test results at the student level, and the lack of stability of the aggregated test results. A distinction was made among three main sources of errors: unreliability due to random sources such as item selection and guessing, unreliability due to inconsistency of teacher ratings of open-ended items, and unreliability due to lack of test stability, primarily caused by variation in standards over time. More precisely, lack of test stability was defined as the random variation in test means across time, after the long-term trend in test score had been removed. The main conclusion was that random errors and rating errors are important sources of error, which account for the major part of misclassifications of students. Lack of

test stability was found to account for only a few per cent of the student misclassifications and it was concluded that this source of error only has marginal effects.

But even though the stability error exerts only little influence at the student level, the error certainly is easy to identify in the quite dramatic differences in mean test scores over time and particularly so for some subjects. At this aggregate level, the random errors can safely be assumed to cancel out, so variation in standards (or test difficulty) is the only factor influencing differences in test means over time, apart from true change. The NAE (Skolverket 2015) did not explicitly evaluate the consequences of the observed stability errors, and they did not consider the possibility that the long-term trend in test scores could also be due to successive changes in standards over time. Nevertheless, it seems that the issue of stability errors due to imperfections in standard setting does not need to be turned into an empirical question, but that efforts should rather be made to eradicate the stability errors as completely as possible.

There is obviously not a single, unambiguous recipe for successful standard setting, applicable in all contexts and for all aims. On the contrary, there is a considerable number of different methods, often based on either a test centered or an examinee centered approach. However, some basic parameters can be identified as essential prerequisites for standard setting processes that contribute to fairness and stability. An analysis of the different methods used by the test development groups in Sweden shows clear variability but emphasizes some factors as crucial, the most essential of course being valid and reliable items and tasks. This requires clear specifications and solid piloting and pretesting with a sufficiently large number of candidates. Not altering items to be included in a test in any substantial way after the final, large pretesting round is another principle that can be identified. Anchor items, preferably both in pretesting and in sharp tests, obviously facilitate arriving at stable standards.

Further, a robust analytical approach is needed. Here IRT plays an important role, although classical methods may also be used as part of the analytic approach. Having enough, well-prepared/trained panelists is of course a crucial aspect, given the strong emphasis on expert judgement in the system. Combining the qualitative data provided by the standard setting groups with solid pre-testing data enhances quality. Finally, when possible, a final trial of the whole test (conducted under rigorous conditions) is a beneficial step in the test development process, including standard setting.

It also needs to be borne in mind that what has been said about standard setting here, refers mainly to tasks where a points system is used. With performance related tasks, for example essays and oral tests, other procedures are used, in which teachers individually analyze, mark and comment on student samples (for example, 12 teachers work on 50 essays). The results are then collected, collated and analyzed, and a number of benchmarks are chosen, commented on and included in the teacher guidelines.

8.7 Ongoing Activities

As discussed above, both test development as such and standard setting have crucial roles in creating valid and reliable national tests and assessment materials, contributing to fairness and equity. In this, the issue of standardization of processes and products has to be addressed. Suggestions about this were made, and small-scale preparatory work was undertaken, already some 10 years ago, but it is only quite recently that the issue was taken up in a more official way, following different internal and external observations and recommendations, for example from the OECD (Nusche et al. 2011). A small working group was given the task to develop, in collaboration with different experts, a suggestion for a common framework for all national tests. In the agreement made, it is stated that the framework should contribute to theoretically founded quality assurance, consist of a theoretical and a more practical part, and be mainly aimed at the National Agency as the responsible coordinator of the system, and for the test developing university departments in charge of test development. One of the functions of the framework is to form the basis for the different test development groups to develop subject specific specifications, however, clearly aligned with the general framework, in which standard setting is likely to be an essential issue. Throughout the agreement, clarity and transparency are emphasized. Preliminary reporting to the NAE was agreed for the autumn of 2015. However, this date was postponed due to a politically initiated, extensive inquiry ("Statens offentliga utredning"/SOU) on the national assessment system, which was to report in the spring of 2016.

In the instructions for the Inquiry, it was stated that the main tasks of the investigator were to

- analyze the aims, function and scope of the national tests,
- propose a system for continuous national evaluation for trend measurement over time,
- propose how the rating of student responses and performances should be designed to ensure equal procedures,
- draw up a proposal aimed to increase the proportion of external marking of national tests in a cost effective way,
- analyze the possibilities for digitalization of national tests and suggest how, to what extent, and at what pace this could be done.

The inquiry was reported on 31 March 2016 in two volumes, comprising some 800 pages (SOU 2016:25). The title of the final report summarizes the conclusions and suggestions of the investigation: "Equivalent, fair and efficient – a new system for national assessment." A very brief summary of the most important proposals shows the following: An assessment system with three components is suggested: mandatory national tests, national assessment support materials (with one section providing support for grading, the other for teaching and learning; both of them optional for schools to use), and a national evaluation system. Each of the components will have a distinct aim, and the ones intended for grading purposes should be supported by a common framework, defining quality measures. Furthermore, it is proposed that the national tests should have one aim only, namely to enhance fair-

ness and equity at the individual level. According to the suggestion, the number of mandatory national tests should be reduced and comprise only the core subjects, English, mathematics and Swedish/Swedish L2, and a National Evaluation system should be introduced to cater for trend studies. A gradual digitalization of the assessment system is proposed for the next 5 years, and pilot studies of external marking and co-rating are to be conducted. A clearer and somewhat stronger relationship between national test results and grades is suggested, and a tentative model for criteria at the group level is presented; however, it is emphasized that the issue will have to be analyzed further by the NAE.

The proposals made have been considered by a large number of stakeholders in the country, their responses required in the late summer of 2016, and a political decision is foreseen for the late spring of 2017. Whatever that will incur, it seems reasonably clear that a common framework for the national tests, and maybe the whole assessment program, will be required. Hence, the work initiated by the NAE has been taken up again, with a tentative outline scheduled for the late autumn of 2016, and a full proposal by the summer of 2017.

8.8 Looking Ahead

The system of national assessment in Sweden, with a long tradition and a generally high degree of acceptance, is currently in a dynamic phase, with a number of changes foreseen. In this, it seems essential both to look back and to look forward. This means building on, maintaining and further develop the positive aspects of the system, not least the tradition of viewing assessment as an integrated aspect of the pedagogical process, the positive attitudes of teachers and students, and the sometimes clearly beneficial impact of the materials – and at the same time take the opportunity to change and/or strengthen what has obviously been weak and leaves ample room for improvement. The latter undoubtedly means increasing stability, thereby contributing to validity and reliability – or expressed differently, enhancing fairness and equity. A balanced, gradual introduction of digital assessment forms seems an inevitable and positive aspect of this. Finally, further development and elaboration of methods of collaboration with broad groups of stakeholders seems essential, as well as strengthening the relationship and respect among the policy, practice and research levels.

References

Carlgren, I. (2015). *Tänk om… (den ‘nya kunskapsskolan’ inte är en kunskapsskola)*. [What if… (the ‘new knowledge school’ is not a knowledge school)]. Skola & Samhälle S.O.S. (Nov.12.2015). http://www.skolaochsamhalle.se/om/. Retrieved 29 Apr 2016.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting. A guide to establishing and evaluating performance standards on tests*. Thousand Oaks: Sage.

Erickson, G. (2009). *Nationella prov i engelska – en studie av bedömarsamstämmighet*. [National tests of English – a study of inter-rater consistency.] http://nafs.gu.se/publikationer. Retrieved 29 Apr 2016.

Gardner, J. (Ed.). (2012). *Assessment and learning* (2nd ed.). London: Sage.

Gipps, C. (1994). *Beyond testing: Towards a theory of educational assessment*. London: The Falmer Press.

Gustafsson, J.-E. (2006). *Barns utbildningssituation. Bidrag till ett kommunalt barnindex* [The educational situation of children. A contribution to a municipal children's index.] Stockholm: Rädda Barnen.

Gustafsson, J.-E., & Erickson, G. (2013). To trust or not to trust? – Teacher marking versus external marking of national tests. *Educational Assessment, Evaluation and Accountability, 25*(1), 69–87.

Gustafsson, J-E., Cliffordson, C. and Erickson, G. (2014). *Likvärdig kunskapsbedömning i och av den svenska skolan – problem och möjligheter*. [Equal assessment of knowledge in and of the Swedish school system – problems and possibilities.] Stockholm: SNS förlag. http://www.sns.se/sites/default/files/likvardig_kunskapsbedomning_web.pdf. Retrieved 29 Apr 2016.

Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Marklund, S. (1987). *Skolsverige. Del 5: Läroplaner.* [Educational Sweden. Part 5: Curricula.]. Stockholm: Liber.

Messick, S. A. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education/Macmillan.

Moss, P., Girard, B. & Haniford, L. (2006). Validity in educational assessment. Review of Research in Education*,* 30(1*)*, 109–162. London: Sage Publications.

Nusche, D., et al. (2011). *OECD reviews of evaluation and assessment in education Sweden*. Paris: OECD. http://www.oecd ilibrary.org/education/oecd-reviews-of-evaluation-and-assessment-in-education_2223. Retrieved 29 Apr 2016.

Skolinspektionen. (2013). *Olikheterna är för stora*. [Differences are too large.]. Stockholm: Skolinspektionen. http://www.skolinspektionen.se/sv/Beslutoch-rapporter/Publikationer/. Retrieved 29 Apr 2016.

Skolverket. (2007). *Provbetyg-slutbetyg-likvärdig bedömning? En statistisk analys av sambandet mellan nationella prov och slutbetyg i grundskolan, 1998–2006*. [Test grade-final grade-equal assessment? A statistical analysis of the correlation between national tests and final grades in compulsory school.] (Rapport nr 300). Stockholm: Skolverket.

Skolverket. (2014). *Så tycker lärarna om de nationella proven*. [Teachers' opinions about the national tests.] *2013*. Stockholm: Skolverket; Rapport 401. http://www.skolverket.se/publikationer?id=3218. Retrieved 26 May 2016.

Skolverket. (2015). *Provbetygens stabilitet*. [The stability of test grades]. Aktuella analyser Stockholm: Skolverket. http://www.skolverket.se/publikationer?id=3488. Retrieved 29 Apr 2016.

Skolverket. (2016). *Utvärdering av den nya betygsskalan samt kunskapskravens utformning.* [An evaluation of the new grading scale and the design of the knowledge requirements.] Stockholm: Skolverket; Dnr 2014:892. http://www.skolverket.se/publikationer?id=3652. Retrieved 26 May 2016.

SOU. 2007:28. *Tydliga mål och kunskapskrav i grundskolan - Förslag till nytt mål- och uppföljningssystem*. [Clear goals and knowledge requirements in compulsory school.] Stockholm: Utbildningsdepartementet. http://www.regeringen.se/rattsdokument/statens-offentliga-utredningar/2007/05/sou-200728/. Retrieved 29 Apr 2016.

SOU. 2016:25. *Likvärdigt, rättssäkert och effektivt – ett nytt nationellt system för kunskapsbedömning*. [Equivalent, fair and efficient – a new national system for knowledge assessment.] Stockholm: Utbildningsdepartementet. http://www.regeringen.se/rattsdokument/statens-offentliga-utredningar/2016/03/sou-201625/. Retrieved 29 Apr 2016

Vlachos, J. (2013). *Betygssystem som bäddar för misslyckanden*. [A grading system heading for failure.] Ekonomistas, June 2013. https://ekonomistas.se/2013/06/14/. Retrieved 29 Apr 2016.

Chapter 9
Validating Standard Setting: Comparing Judgmental and Statistical Linking

Anna Lind Pantzare

Abstract This study presents a validation of the proposed cut scores for two test forms in mathematics that were developed from the same syllabus and blueprint. The external validity was analyzed by comparing the cut scores set by an Angoff procedure with the results provided by mean and linear observed score equating procedures. A non-equivalent group anchor test (NEAT) design was also used. The results provide evidence that the cut scores obtained through both judgmental and statistical linking are equivalent. However, the equating procedure revealed several methodological and practical challenges.

Keywords Standard setting • Equating • National testing • Equivalence • Fairness

9.1 Introduction

A cornerstone of high-quality test development is that results from different test forms used for the same purposes are comparable (American Educational Research Association et al. 1999, 2014). Hence, what test form the test taker receives should not affect the result. Even though extensive theory exists for how to develop valid and parallel test forms (e.g., Crocker and Algina 1986; Downing 2006; Gronlund and Waugh 2009), strict parallelism is difficult to accomplish in practice. Furthermore, even if test forms have been developed to be as parallel as possible, i.e., with the ambition to measure the same construct with the same difficulty (American Educational Research Association et al. 1999, 2014), unsubstantiated cut scores, e.g., the score a test taker needs to get a certain grade on the test, can undermine the trustworthiness of the whole assessment system. Standard setting, the process of establishing reliable and valid cut scores, has been a concern for the last 50 years, which is as long as criterion-referenced tests have been common in educational contexts

A.L. Pantzare (✉)
Department of Applied Educational Science, Umeå University, Umeå, Sweden
e-mail: anna.lind.pantzare@umu.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_9

(Berk 1986; Cizek and Bunch 2007; Zieky et al. 2008). Most of the research on this topic has focused on providing general guidelines for how to organize high-quality standard setting practices, what methods to use, and how to evaluate the cut scores. This might be due to the fact that standard setting has been – and still is – seen as a rather complicated process (e.g., Cizek 2001), that usually has to rely on subjective judgments regarding test items and/or test takers. The need to involve panelists, which may lead to not knowing the basis for the established cut scores, is often stated as the reason to conduct research on standard setting. Another problem that has been highlighted in standard setting studies is the difficulty of defining performance level descriptions and then getting the panelists to make common interpretations of these descriptions (e.g., Hambleton and Pitoniak 2006). In addition, some have argued that it is challenging for the panelists to evaluate the difficulty of the test and suggest accurate cut scores (Shepard et al. 1993). However, other studies have provided evidence that panelists are able to accurately estimate item difficulty (e.g., Brandon 2004; Ferdous and Buckendahl 2013; Giraud et al. 2005).

Research about what panelists in standard setting panels are thinking when deciding which cut scores to recommend also exists. This research has shown that the judgments are influenced by various factors, such as the types of students the panelists are teaching, the types of items that are being judged, how the panelists are trained, and how the item or test data is presented to the panel (Clauser et al. 2013; Ferdous and Plake 2005; Impara and Plake 1998; Plake and Impara 2001; Skorupski and Hambleton 2005). These findings are important to consider when validating a standard setting process.

Even though there is a lot of research on standard setting, studies that have investigated the equality of cut scores for parallel test forms are scarce. One of the few articles found focuses on maintaining equivalent cut scores for small sample test forms (Dwyer 2016). If the results from parallel test forms are evaluated against the same performance standards but the cut scores are decided individually for each test form then there should be a control ensuring that the cut scores are equally demanding. This can be challenging, especially from a perspective of fairness, when several test forms are administered. Furthermore, when test results are used in school evaluations or to monitor change over time, it is important that variations in the results between test forms stem from changes in the test takers' knowledge levels rather than from differences in the difficulties or cut scores of the tests. The following chapter will use data from Swedish national tests in mathematics to investigate how the standard setting of parallel test forms can be validated.

9.2 Standard Setting – A Note on Terminology

The term standard setting is frequently used in literature, but with somewhat different meanings. Usually, standard setting reflects the whole process, from defining the performance standards or level descriptions to establishing cut scores (e.g., Cizek and Bunch 2007; Hambleton and Pitoniak 2006). However, standard setting

sometimes describes only the first part of the activity, i.e., the definition and description of the performance standards (e.g., Baird et al. 2007). A third definition limits standard setting to only include the operational part of establishing cut scores (e.g., Giraud et al. 2005). The first definition of standard setting, (i.e., the whole process from defining the performance standards or level descriptions to establishing the cut scores), has been adopted in this chapter.

9.3 The Swedish School System and National Tests

Since 1994, the Swedish school system is criterion-referenced. The most recent curriculum, which included new syllabi and a revised grading scale, was introduced in 2011 (The Swedish National Agency for Education 2012). Even though the steering documents are national, Sweden has one of the most decentralized school organizations in the world, entrusting teachers to teach, assess, and grade their own students (Dufaux 2012).

The current grading scale has six grade levels, with A–E reflecting passing grades (A as the highest grade) and F reflecting a fail. A national syllabus that includes aims, goals, and content, as well as the knowledge requirements for the grades E, C and A, exists for each subject. The grades D and B are awarded to students who have met all the requirements for the lower grade (i.e., E or C) and a majority of the requirements for the higher grade (i.e., C or A) (The Swedish National Agency for Education 2012).

Swedish upper secondary education includes national tests in Swedish, English, and Mathematics that support teachers in the grading of their students (The Swedish National Agency for Education 2005). Upper secondary school courses can last either one or two semesters, and a new test form is provided every semester to minimize the disturbance to school planning. Before the test is administered, the test takers must be informed of the cut scores, i.e., the score required for each grade level. This may seem rather unorthodox because it will prevent any adjustments after the test has been administered, but this requirement has existed ever since the transition from the norm-referenced system to the current criterion-referenced system in 1994. The main argument that has been given for this model is that it will prevent teachers from interpreting the test scores in a relative manner, i.e., to grade on a curve. To make this possible, the cut scores for each new test form are determined through a standard setting procedure before the test is administered. A relevant question is if the model really works as intended, or rather, if it is naïve to assume that the cut scores are equally demanding. From an international perspective, it is rather uncommon to handle standard setting in this manner. In educational assessment systems the cut scores are generally set *after* test administration, and include a thorough analysis of how students solved the items (e.g., Massachusetts department of education 2007; Newton 2005). This type of model makes it possible to correct for unexpected results.

9.4 The Swedish National Tests in Mathematics

The various national test forms used for measuring upper secondary school mathematics courses are based on the same blueprint. This blueprint is quite elaborate, defining how many items are necessary to measure each goal and knowledge requirement. The overarching goal of the test development is, of course, that each test form should be parallel to the test forms that have been previously administered. To achieve a balanced test form, field test information, together with classifications of the items and their possible answers, are taken into consideration as indicators of parallelism. The objective is that this test development procedure will result in test forms that have similar cut scores.

The Swedish national tests in mathematics usually consist of 30–40 items. A majority of the items are polytomous, constructed-response items, in which the student is asked to present a short or extended answer. The multiple-choice format is only used for a few items in each test form. The choice of format is based on the type of knowledge and competence measured, as defined in the knowledge requirements. The final answer is not only of interest, but also the process the student followed to reach the answer.

The national tests are scored by school teachers, who generally score their own students. There is no system of external control, such as scorer training or moderation. There are only the items and an item-specific scoring guide, which includes evaluated examples of student work for some of the items. The lack of control mechanisms may be a problem, but a previous study shows that upper secondary school mathematics teachers in Sweden are quite good at scoring national tests in an equivalent way (Lind Pantzare 2015). The scoring guide also includes additional information that defines what aim and which knowledge requirement each item is intended to assess. This item information guides the teachers when they are evaluating students' tests. Additionally, the item information guides panelists in standard setting sessions used for recommending cut scores.

9.5 Standard Setting of Swedish National Tests in Mathematics – Judgmental Linking

The research is in agreement with the conclusion that a standard setting procedure must be implemented in a sound way to yield valid cut scores. Also, in the research the standard setting procedure is often described similarly. Generally, the procedure follows the subsequent steps: selection of a representative panel, the choice of a suitable method, preparation of performance level descriptions, training participants to use the selected method, collection of the first round of ratings, discussion of the ratings and providing panelists with supplementary information (e.g., empirical item data), collection of one, possibly two, round(s) of reviewed ratings, evaluation of the standard setting process, and documentation of the process (Cizek and Bunch 2007; Hambleton and Pitoniak 2006).

The standard setting method that is most commonly used worldwide is the Angoff (1971) method, as well as all of its variations (Cizek and Bunch 2007). While the original method was designed for dichotomously scored items, Hambleton and Plake (1995) extended the method to also include polytomously scored items. This modified Angoff method is used in establishing the cut scores for the Swedish national tests because: (1) it has the capability to handle both dichotomously and polytomously scored items, and (2) it offers the possibility to establish cut scores before test administration. The Angoff method is one of the few methods that have both of these attributes.

The Swedish method for standard setting follows the approach recommended in literature except for one alteration: it does not include a separate step for the determination of performance level descriptors. This is because the syllabus defines the knowledge requirements, and these are used as the performance level descriptors. Since the teachers regularly work with these knowledge requirements when they teach, assess, and grade their students, they are supposed to be well acquainted with them. The teachers' grading experiences allow them to identify the group of borderline examinees at each grade level, which is essential in the standard setting procedure. In addition, mathematics is perceived to be a hierarchical subject with some sort of consensus about which concepts are difficult and which concepts are easy, possibly to a greater extent than other subjects. Therefore, it has been seen as logical to use mathematics teachers as panelists in the standard setting panels, and their capability to establish equivalent cut scores between test forms has been trusted.

However, over time the aims of the Swedish national tests have changed, and the stakes of the tests have increased. The national tests not only aim to support teachers in assessing and grading their students, but they are now also expected to provide information that can be used to evaluate a school's performance. Due to rather poor Swedish results in international comparative studies (e.g., TIMSS and PISA) there has been a discussion about the possibility to use national tests in regular evaluations of the achievement level in schools, both on a municipality and national level. One issue that has been highlighted as a major drawback is the quality of cut scores, even if previous research (Näsström and Nyström 2008) has reported that the cut scores appear to be trustworthy, at least when it comes to how the panelists interpret the knowledge requirements that serve as performance standards. However, this study did not provide any information about the stability of the test difficulty or the cut scores over time. Although the knowledge requirements can remain unchanged and should serve as the basis for equally demanding cut scores for different test forms, a change in the knowledge levels of the student population might affect the panelists and their visualizations of the borderline examinee. Hence, several uncertainties exist in the judgmental linking procedure as there is no solid evidence other than that the procedure is implemented in a sound way.

Since it is unclear how valid cut scores set by a judgmental procedure really are, it is important to investigate their comparability and the consequences of variation between cut scores for different test forms. This study will attempt to validate the proposed cut scores for two test forms in mathematics developed from the same syllabus and blueprint.

9.6 Validation

Messick (1989, 1995) has stated that validity is not a property of the test or the items, but it is about the interpretation of the test scores. By this definition, the objective of validating standard setting is to find evidence that the proposed cut scores are reasonable in relation to the performance standards and the aim of the test. Kane (1994) also stated "The aim of the validation effort is to provide convincing evidence that the passing score does represent the intended performance standard and that this performance standard is appropriate, given the goals of the decision process." Since no single method for validation exists and no specific value is reported, Kane argues for a systematic review and proposes a framework to use when validating performance standards and cut scores. This framework uses a three-part, systematic investigation of validity evidence. The first and second parts are concerned with the actual standard setting and include evidence of procedural validity and internal consistency. Kane (2001) argues that these analyses should be given more attention since the required data are rather easy to collect and handle. In Sweden, the standard setting procedure follows the steps recommended in literature for the evaluation of procedural validity as closely as possible. The internal consistency can be investigated by calculating the standard deviation of the cut scores suggested by panelists. A smaller deviation represents better consistency.

The third part of the framework compares the standard setting with some external criterion. Kane (1994) suggests several alternatives for this part of the validation and one is to conduct another standard setting study for the same test form by using another method. However, a more objective procedure that compares test forms and provides cut scores must be implemented if the goal is to achieve comparable cut scores without relying on judgmental procedures. In many testing systems, the cut scores for a new test form are determined through statistical linking, for example, equating (Kolen and Brennan 2004). Equating is a statistical process that determines the comparability of scores from different, but parallel, test forms and makes it possible to establish equivalent cut scores for a new test form in relation to those for the old ones.

Kolen and Brennan (2004) distinguish three different concepts: equating, scaling, and linking. Equating is the strictest, as it requires that the test forms were developed from the same blueprint and measure the same skills at approximately the same level of difficulty. Scaling is used to make comparisons between tests, but the scaled scores cannot be used interchangeably. Linking is a broad term that is identical to equating when the test forms fulfil the requirements for equating. These requirements are generally met by the Swedish national tests in mathematics and the concepts of equating and linking are used interchangeably in this chapter. Studies on equating have not investigated whether equating should be implemented, but has rather assessed the different kinds of equating designs, the necessary conditions for equation, and how to evaluate the equating results (Crocker and Algina 1986; Kolen and Brennan 2004; Livingston 2014; Lord 1980). Moreover, a require-

ment of many large-scale standardized testing programs has been that different test forms should be linked or equated (Kolen and Brennan 2004).

There must be a link between the old and the new tests for equating to be feasible. The link is created either by common items, so-called anchor items, in the two tests, common test takers, or randomly equivalent groups (Crocker and Algina 1986; Dorans et al. 2010; Kolen and Brennan 2004). The use of randomly equivalent groups is popular since it only requires the administration of a test form to one group of test takers without a need for common items. A random selection of test takers from a larger group is performed to obtain a group that is statistically comparable with a group from another test form. If this is manageable, then it is possible to directly compare the two different test forms. However, this method requires large groups of test takers. Therefore, many large-scale assessment programs include common items within a test or administer them together with the regular tests. One crucial feature of equating is that the result is dependent on the quality of data connected to the anchor items. If the test takers do not answer the anchor items in the same manner as they answer the items in the regular test, then conclusions drawn from the data can be spurious. The risk of problematic data sets increases when the test takers can identify the anchor items. Research has shown that students in low-stakes testing can lack motivation, which can affect the item parameters (Eignor and Stocking 1986; Kolen and Harris 1990). This lack of motivation might not be a problem if all items are affected equally, as it will then be possible to take the difference between the field trial and the regular test into account. However, the changes in item parameters might also differ and it may not be possible to take these differences into account in the equating procedure.

9.7 Equating the Swedish National Tests in Mathematics

The Swedish national tests in mathematics are not regularly equated due to several factors. First, some of the upper secondary school courses can only provide a small number of students for the field trials, which is a major problem. Second, it has not been possible to administer field trials of new items during the regular administration of the national tests. This concern is mainly related to test security issues, since test takers have been known to memorize certain items or take photos with hidden cameras, after which the items have been released on websites. Such item exposure would make the development of upcoming test forms very difficult.

Even though the implementation of an equating procedure causes several problems, there is also a certain degree of uncertainty involved with cut scores from a judgmental approach. Standard setting can be rather easy to implement, but it is not possible to know whether the cut scores for different test forms are comparable since the estimates are subjective. At least it is necessary to find ways to investigate the validity of the cut scores.

In this study, the validation will focus on external criteria, comparing *judgmental linking* (using the regular Angoff standard setting procedure) with *statistical linking*

(that takes an equating procedure). The rest of the chapter is organized as follows: the next section will describe the methods and data, with a description of the national tests that were investigated, the results will then be discussed, and finally, the conclusions will be presented.

9.8 Study Design – Statistical Linking

Two national test forms constitute the basis of this analysis. The test forms were developed for the second course of five in the upper secondary school. The main focus of this course is solving exponential and second-degree equations, as well as linear equation systems. The course also includes properties of quadratic functions and fundamental theorems in geometry concerning similarity, congruence, and angles.

One of the test forms is the "old" form and will be denoted as the regular administration. This test form is denoted Y, following the notation used in Kolen and Brennan (2004). The second test form, administered to group 2, is the new test. This new test should be parallel to the old test form in both content and difficulty. The new test form will be denoted X. Both of the test forms consist of three parts. The first part contains short answer items, while the second and third parts contain items that require extended answers. The use of a calculator is not allowed in the first or second parts, while all kinds of digital tools can be used in the third part. The distribution of content areas and the sum of the items (57 points) are the same in both test forms.

In the study there is also an anchor test consisting of nine tasks, which comprise thirteen items that add up to a total of 18 points. This anchor test is used as the link between the two test forms. The distribution and difficulty of item types in the anchor test are similar to those of the regular test forms. The items differ in mathematical content, complexity, and the type of answer required. One of the items is multiple-choice, eight are short answer, and four are constructed-response items that demand extended answers. These thirteen items were used as an external anchor test, i.e., they were administered separately to the new test form.

The same 13 items from the anchor test were included in the first part of the old test form and therefore, served as an internal anchor, as shown in Fig. 9.1. Even though the anchor test was internal in one test form and external in the other, the stakes in this study can be assumed to be similar since the students who took the external anchor knew that it was not a part of the national test but they also received information that the result of the anchor would affect the grading at the end of the study course.

9.9 Participants

Group 1 represents the students who received the old test form. This is the data from the regular reporting of results connected to the national test. In this type of reporting, the teachers who have administered the national test are supposed to report item

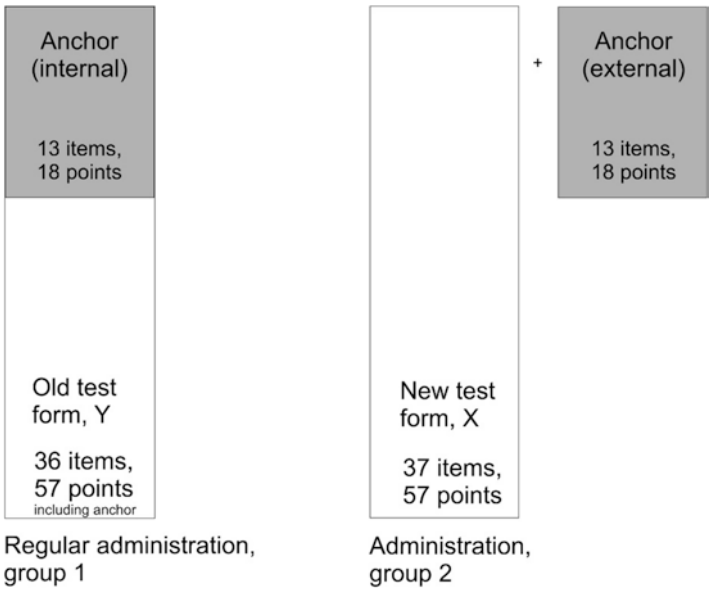


Fig. 9.1 Schematic figure of the old and new forms and how the anchor test is related to them. The old test form was administered to group 1 and the new test form was administered to group 2

data for students born at four given dates in each month. This should yield a rather representative sample of results. For the old test form, results from 3330 examinees were reported and used in this study. Group 2 received the new test form, which included a separately administered anchor test, and comprised 191 students. All of the students completed both the anchor test and the new test form. Both test forms were administered to their respective groups on the same date. The external anchor test was administered to group 2 during the next math class, which was within a couple of days of the national test. The students that participated in group 2 were selected from among the classes that had to carry out the regular national test during the current semester. Also, these selected classes were from schools that had reported results representative of the entire population in previous years.

The teachers in these classes were asked to administer the new test form and the anchor test instead of the regular test form. They were also asked to report the results from both of these tests. The difference in group sizes posed a problem, especially if one group performed very well and the other performed very poorly. However, large differences between the two groups were not expected because group 2 had previously been representative of the entire population in terms of the national tests and an anchor test, which would provide information about the performance levels, was included in both test forms.

9.10 Standard Setting of the New Test Form

The cut scores for the new test form were estimated with a modified Angoff procedure. Two panels were used; panel 1 included twelve panelists and panel 2 included eleven panelists. The panelists' task was to estimate the item difficulties for the grades E, C, and A. The two panels included panelists who came from two different geographical areas in Sweden and the participating panelists represented ten different schools. Twelve of the panelists were female and all of the panelists had more than 5 years of teaching experience. All but five had participated in a standard setting meeting at least four times.

The standard setting meetings followed a strict agenda. Before the meeting, all of the panelists received a copy of the test form and the scoring guide. The panelists were instructed to thoroughly work through the items in the material and independently make a holistic estimation of the cut scores before attending the meeting. When the panelists in each panel had gathered, they started the meeting by discussing their holistic estimations and the differences in their estimations. In addition, specific items, as well as demands for the scoring guide in relation to the knowledge requirements, were discussed.

Next, the chair introduced the Angoff method. Since all of the panelists had participated in previous standard setting meetings, this introduction was viewed more as a recapitulation of the method. Thereafter, a first round of individual item estimations for the grade E was carried out. These estimations served as a basis for discussions regarding the interpretations of the knowledge requirements for each item. There was a special focus on items with large variation in the estimated item difficulties. After this discussion, a second, and final, round of estimations for the grades E, C, and A were collected.

9.11 Statistical Analysis Methods

The judgmental standard setting procedure of each panel was analyzed separately by calculating the mean and standard deviation of the judgments, which will serve as information about internal consistency. As the number of test takers in group 2 was rather small, only a few methods could be used for equating the test forms. The simplest, and least statistically demanding, method is mean equating. In this method, only the mean values for the anchor test and the test form administered to the two groups of test takers need to be taken into account. The mean anchor test results for the two test taker groups are used to define the difference in proficiency. This information is then used to adjust the scores on the new test form. The advantage of this method is that it is easy to implement and requires rather few test taker results. However, a major weakness is that the score adjustments are the same along the whole score scale, even if the standard deviation might differ.

A slightly more robust alternative is linear equating. In this method both the mean values and standard deviations of the observed scores are used to make adjust-

ments between the score scales of the two test forms. A thorough derivation of the mathematical relationships underlying both mean equating and different linear equating methods can be found in Kolen and Brennan (2004).

This study implemented observed score equating. The statistical equating procedure was carried out using Common Item Program for Equating (CIPE) (Kolen 2004). The calculations of this statistical package follow the mathematical relationships described above and it is possible to analyze the quality of the anchor and the results from equating with Tucker and Levine mean and linear methods. These two methods differ based on the assumptions that are made about how group 1 will perform on form Y and how group 2 will perform on form X. In the calculations the anchor items were used as internal items in relation to the old test form, Y, and as external items in relation to the new test form, X, in line with how they were administered.

9.12 Angoff Cut Scores and Internal Consistency

Since there are two groups of panelists it is possible to compare the cut scores and standard errors of the new test form. The results show that the cut scores that were set using the modified Angoff procedure are quite consistent. The standard errors are rather small, which indicates that the panelists shared a common view of how a borderline examinee will perform on the test (Table 9.1).

9.12.1 External Validity Evidence – Results from the Equating Procedure

In the equating procedure, the regular test form, Y, is considered to be the baseline to which the new test form will be linked. The regular test form, Y, had cut scores of 14, 30, and 45 for the grades E, C, and A, respectively. A univariate analysis of the anchor item results from the two groups showed that group 1, in which the anchor was internal, had a mean value of 5.4 (SD 3.5) and group 2, in which the anchor was

Table 9.1 Mean cut scores and standard errors (in brackets) for the three grades E, C, and A for the new test form. Results from the two Angoff standard setting panels and the final cut score

	Grade		
	E	C	A
Angoff, panel 1	15.0 (0.3)	30.1 (1.3)	45.1 (1.2)
Angoff, panel 2	14.7 (0.4)	29.9 (0.7)	44.8 (0.9)
Final cut score	14	30	44

external, had a mean value of 5.9 (SD 3.3). The mean value for the regular test form (group 1) was 16.7 (SD 9.9) and 17.2 (SD 7.9) for the new test form (group 2). These results indicate that the two groups are rather comparable despite their considerable size difference.

Since the anchor items are internal in the regular test form and external in the new test form, one can assume that the correlation between the anchor and the regular test form should be higher than between the external anchor and the new test form because the scores from the internal anchor comprise approximately a third of the total test score. This proved to be the case, as the correlation between anchor and regular test form was 0.9, compared to a correlation of 0.8 when the anchor is external. Figure 9.2 shows the equated score curves for all of the implemented methods. Figure 9.3 shows the standard errors of equating for the two linear methods.

The equated cut scores for the new test form are reported in Table 9.2 The standard errors of the cut scores for E, C and A determined through linear equating are 0.5, 0.8, and 1.4, respectively. The practical impacts of the two methods (statistical linking and judgmental linking) can be illustrated by comparing the classifications of the equating in a confusion matrix (Tables 9.3 and 9.4).

CIPE allows the user to vary how the results from each of the two groups should be weighted in the equating procedure. However, calculations with different group weights did not result in significant differences in the equated scores.

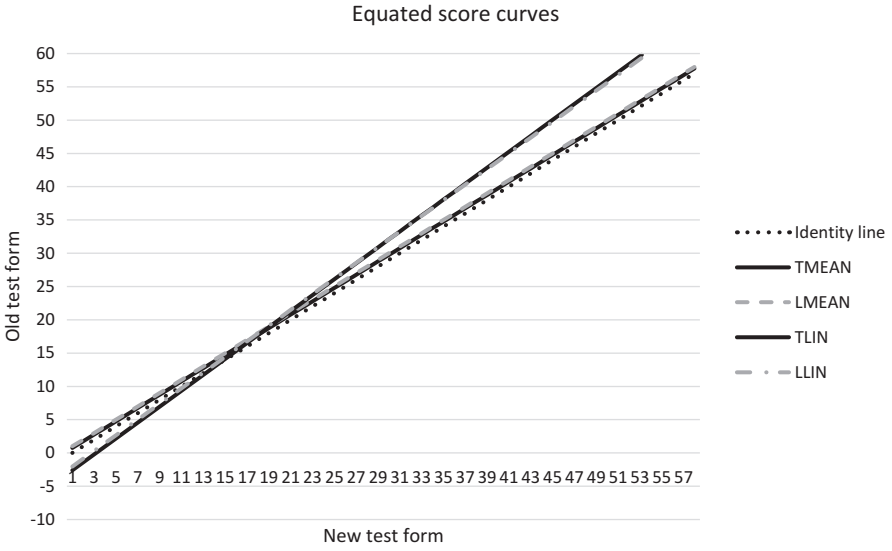


Fig. 9.2 The equated score *curves* and the *identity line*. Both the mean and linear equating are made with the Tucker method (TMEAN and TLIN) and Levine observed score method (LMEAN and LLIN)

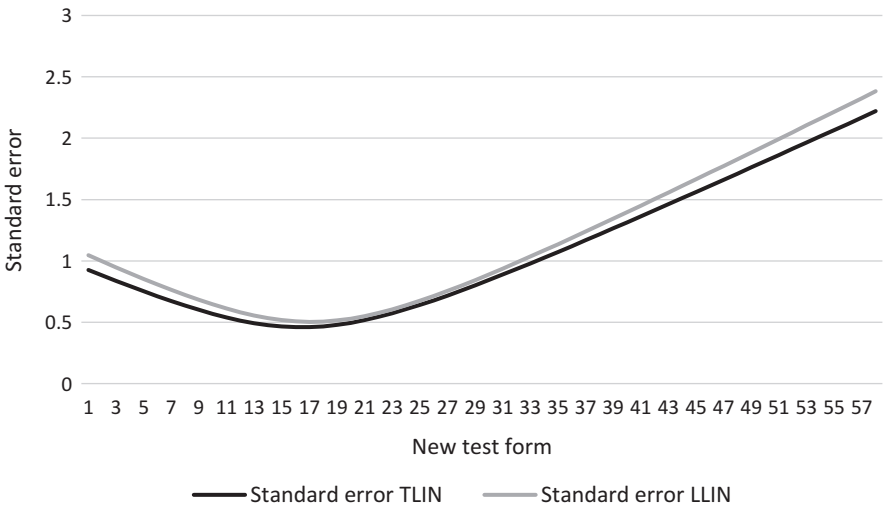


Fig. 9.3 Standard errors of equating for the linear Tucker method (TLIN) and the linear Levine observed score method (LLIN)

Table 9.2 Mean and linear equated cut scores for the grades E, C, and A for the new test form

	E	C	A
Mean equating	13	29	44
Linear equating	14	27	40

Table 9.3 Confusion matrix comparing the grades for the students in group 2 depending on the cut scores from the standard setting, judgmental linking, or the mean equated cut scores, statistical linking

			Angoff standard setting, judgmental linking			
			Grade			
			F (<14)	**E (14)**	**C (30)**	**A (44)**
Equated score mean equating, statistical linking	**Grade**	**F (<13)**	30.4%	–	–	–
		E (13)	5.2%	55.6%	–	–
		C (29)	–	1.0%	7.7%	–
		A (44)	–	–	–	–

The cut scores for the grades E, C, and A are given in the brackets. A is the highest grade and F is the grade fail

9.12.2 Are the Cut Scores Valid?

In the beginning of this chapter, the question "Is it naïve to assume that the cut scores for different test forms are equally demanding?" was posed. It is a relevant question, since the Swedish model, and possibly other testing programs, do not have

Table 9.4 Confusion matrix comparing the grades for the students in group 2 depending on the cut scores from the standard setting, judgmental linking, or the linear equated cut scores, statistical linking

			Angoff standard setting, judgmental linking			
			Grade			
			F (<14)	**E (14)**	**C (30)**	**A (44)**
Equated score mean equating, statistical linking	**Grade**	**F (<14)**	35.6%	–	–	–
		E (14)	–	50.9%	–	–
		C (27)	–	5.7%	7.2%	–
		A (40)	–	–	0.5%	–

The cut scores for the grades E, C, and A are given in brackets. A is the highest grade and F is the grade fail

the necessary information to regularly equate test forms and therefore, must rely on cut scores set by a standard setting procedure. In order to investigate the question, it was necessary to set up a special study to collect the required data.

The results from the standard setting and the equating procedures show that the equated cut scores are one or two scores below the judgmental standard setting for the grade E. Prior experience from standard setting sessions for other test forms in Swedish national mathematics tests has shown that the standard setting estimates for the grade E often are slightly higher than those from the field trial data. Discussions among panelists in standard setting sessions have revealed that it is a rather common opinion that, at least for the lowest passing grade, it is necessary to maintain high standards to ensure that students do meet the required skills. This common opinion might influence the estimates from judgmental linking, which normally tend to be higher than expected based on the field test data. In contrast, mean equating gives the same values for cut scores as the judgmental procedure for grades A and C. However, linear equating yields cut scores for the grades C and A that are lower than those from the judgmental procedure. These results were positively surprising because they provide evidence for the hypothesis that the cut scores are valid. Another result that is important to note is that the standard errors of the standard setting and equating procedures are similar.

As Kane (1994, 2001) argues, no such thing as a "true" cut score exists; it is rather a question how valid the cut scores are. The results from mean equating show that judgmental linking through the modified Angoff standard setting procedure yields almost identical cut scores as statistical linking through mean equating, which could be taken as an indicator of validity. Hence, a conclusion could be that judgmental standard setting seems to work rather well in a Swedish mathematics context.

The next question is: can these results be generalized? There are, as was stated earlier, a lot of special circumstances that probably influenced the results. This study concerns a mathematics test, a subject with a rather common opinion of what is difficult and what is easy, which may facilitate the ability to set equivalent cut

scores. In addition, the test forms in the study have a scoring guide that includes information that can contribute to the interpretation of the items. Also, in school systems like those in Sweden, where teachers are well aware of the knowledge requirements and have a general consensus about what it takes to reach a certain grade, teachers can be quite proficient at evaluating the difficulty of test items.

9.13 Methodological Challenges with the Equating Procedure

As mentioned before, one of the main reasons that the Swedish national tests in mathematics are not regularly statistically linked, for example, through equating, is the problem of including anchor items in field trials or in the regular tests. The main objections for this practice are, like in many other educational systems, the time available, the possibility to include a representative selection of items, and the possibility to maintain confidentiality. Also, some courses have too few students. In addition, since the Swedish teachers score the national tests themselves, it would be even more problematic to handle anchor items in the regular test forms. It would be difficult, perhaps even impossible, to ask teachers to neglect students' results on the anchor items, especially if the student has solved the anchor item but not the regular item that measures the same skill. Furthermore, over time the teachers will learn to recognize the anchor items and then the problem of confidentiality becomes relevant. The equating procedures should be developed, also for settings like Sweden, as the demand for test quality is the same as in other countries.

When test forms are equated with the help of an anchor, it is necessary to rely on the results from the anchor items. As mentioned earlier, the mathematical assumption for equating methods states that anchor items function the same way irrespective of whether they are an internal or an external anchor. However, normally the anchor is either internal or external in both groups, not as in this study, where one group included an internal anchor and the other included an external anchor. In this study, the scores from the anchor items were a bit higher when the anchor was external. This was unexpected, as generally the scores are higher when the anchor items are included in a test so that the test takers cannot know which items are anchor items and which are regular items. The problem in this study is that there is no way of knowing if the results from the anchor test in group 2 would have been even higher if the anchor had been internal. However, as argued earlier, the test takers in group 2 were informed and requested to do their best on the anchor items, and the results indicate that the students took the anchor test seriously. In this way, the application of two different types of anchor tests to the two groups may not have been a problem, and the assumption is that the data from the two groups are comparable.

One problematic feature of the equating procedure is that students who received the anchor items separately, as in group 2, could have exerted less effort towards answering these items, which could lead to a lower score on the anchor. This, in

turn, would result in the statistical model assuming that students in group 2 are less qualified than students from the regular group, which would raise the equated cut scores. Also, if the anchor items are too easy in comparison to the regular items, then the statistical equating model will lower the cut score, especially for higher grades. This is due to ceiling effects and the fact that the anchor does not fulfill the requirement of being a representative sample for the total test. However, based on the results from this study, none of these problems seemed to be apparent.

9.14 Conclusion and Suggestion for the Future

The main conclusion from this study is that the information available supports that the cut scores in Swedish national tests in mathematics tests are valid. However, since the statistical linking contains some uncertainties, stronger evidence could have been obtained if the anchor was applied to both groups in the same way. This was a limitation of the study. One approach that has been discussed as a possibility for future validations is to administer the regular test to all test takers and use that test as an anchor, after which the new test would be split into parts and administered as field trials. In this approach, the anchor would be administered in the same way, which is a strength, but the results for all of the new items will come from field trials in which the students might not have done their best. Also, this type of procedure is even more demanding when the required number of participating test takers is considered, since every part of the new test has to be trialed on a sufficiently large group.

Another future possibility is the digitalization of the tests. Digitalization is on the political agenda in Sweden, as in many other countries. If a test was digitally administered, then it would be easier to include anchor items among the regular items without the students being able to differentiate between the two. Also, digitalization would allow item information to be registered digitally, which could be used to link the new items to the old.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–601). Washington, DC: American Council of Education.

Baird, J., Newton, P., Goldstein, H., Patrick, H., & Tymms, P. (2007). *Alternative conceptions of comparability. Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion referenced tests. *Review of Educational Research, 56*(1), 137–172.

Brandon, P. E. (2004). Conclusions about frequently studied modified Angoff standard setting topics. *Applied Measurement in Education, 17*(1), 59–88.

Cizek, G. J. (Ed.). (2001). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah: Lawrence Erlbaum.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks: SAGE Publications.

Clauser, B. E., Mee, J., & Margolis, M. J. (2013). The effect of data format on integration of performance data into Angoff judgments. *International Journal of Testing, 13*(1), 65–85.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont: Wadsworth Group.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). *Principles and practices of test score equating*, *Research report. ETS RR-10-29*. Princeton: Educational Testing Service.

Downing, S. M. (2006). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3–25). Mahwah: Lawrence Erlbaum.

Dufaux, S. (2012). *Assessment for qualification and certification in upper secondary education: A review of country practices and research evidence*, *OECD education working papers* (Vol. 83). Paris: OECD publishing.

Dwyer, A. C. (2016). Maintaining equivalent cut scores for small sample test forms. *Journal of Educational Measurement, 53*(1), 3–22.

Eignor, D. R., & Stocking, M. L. (1986). *An investigation of possible causes for the inadequacy of IRT pre-equating*, *ETS research report series, 1, i-46*. Princeton: Educational Testing Service.

Ferdous, A. A., & Buckendahl, C. W. (2013). Evaluating panelists' standard setting perceptions in a developing nation. *International Journal of Testing, 13*(1), 4–18.

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*(3), 257–267.

Giraud, G., Impara, J. C., & Plake, B. S. (2005). Teachers' conceptions of the target examinee in Angoff standard setting. *Applied Measurement in Education, 18*(3), 223–232.

Gronlund, N. E., & Waugh, C. (2009). *Assessment of student achievement* (9th ed.). Upper Saddle River: Pearson.

Hambleton, R. K., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (Vol. 4, pp. 433–470). Westport: American Council on Education.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8*(1), 41–55.

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.

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

Kane, M. T. (2001). So much remains the same: conception and status of validation in standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 53–89). Mahwah: Lawrence Erlbaum.

Kolen, M. J. (2004). *Common item program for equating (CIPE), Windows GUI* (2.0 ed.). Iowa: Iowa Testing Programs, The University of Iowa.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. New York: Springer Publishing.

Kolen, M. J., & Harris, D. J. (1990). Comparison of item preequating and random groups equating using IRT and equipercentile methods. *Journal of Educational Measurement, 27*(1), 27–39.

Lind Pantzare, A. (2015). Interrater reliability in large-scale assessments – Can teachers score national tests reliably without external controls? *Practical Assessment, Research & Evaluation, 20*(9), 1–14.

Livingston, S. A. (2014). *Equating test scores* (2nd ed.). Princeton: Educational Testing Service.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: Lawrence Erlbaum.

Massachusetts Department of Education. (2007). 2007 MCAS standard setting report. Massachusetts.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education & Macmillan.

Messick, S. (1995). Validity of psychological assessment. Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749.

Näsström, G., & Nyström, P. (2008). A comparison of two different methods for setting performance standards for a test with constructed-response items. *Practical Assessment, Research & Evaluation, 13*(9). http://pareonline.net/getvn.asp?v=13&n=9. Accessed 7 Apr 2016.

Newton, P. E. (2005). Examination standards and the limits of linking. *Assessment in Education, 12*(2), 105–123.

Plake, B. S., & Impara, J. C. (2001). Ability of panelists to estimate item performance for a target group of candidates: An issue in judgmental standard setting. *Educational Assessment, 7*(2), 87–97.

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

Skorupski, W. P., & Hambleton, R. K. (2005). What are panelists thinking when they participate in standard-setting studies? *Applied Measurement in Education, 18*(3), 233–256.

The Swedish National Agency for Education. (2005). *National assessment and grading in the Swedish school system* (p. 32). Stockholm: The Swedish National Agency for Education.

The Swedish National Agency for Education. (2012). *Upper secondary school 2011*. Stockholm: Fritzes.

Zieky, M. J., Perie, M., & Livingston, S. A. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests*. Princeton: Educational Testing Service.

Chapter 10
National Tests in Norway: An Undeclared Standard in Education? Practical and Political Implications of Norm-Referenced Standards

Idunn Seland and Elisabeth Hovdhaugen

Abstract Educational standards were not the official object of national tests, when they were introduced as a tool for quality assessment in Norwegian schools in 2004. As the national curriculum relies on teachers' professional judgement for setting criteria for student learning, there are no direct links between the standardised tests and the managerial and pedagogical employment of the norm- referenced test results. In this study, we investigate how municipalities and teachers conceptualise and utilise results from national tests. We find that whereas school owners simply set future results from national tests to be above the national mean, many teachers either disregard or do not seem to comprehend the relationship between the norm-referenced test results and the national curriculum. Consequently, teachers seem to under-exploit test results for student learning development, while school owners seem to over-exploit the same results, as the national norm-based mean demonstrates that there is little variance at a local level, nor does it provide explanatory power. Results and teaching have never been linked to through authors' explanation: teaching in Norway has of course been linked to curriculum aims. Our point is that results from national tests have not been "fed back" to teaching linked to these curricular aims. Results and teaching have never been linked through set curricular aims, which is partly a political process. Instead, national tests emerge as an undeclared standard in Norwegian education, causing ambiguous political demands and signs of professional frustration.

Keywords National tests • Standards in education • Norm-referenced standards • Quality assessment • Competence aims

I. Seland (✉) • E. Hovdhaugen
NIFU (Nordic Institute for Studies in Innovation, Research and Education), Oslo, Norway
e-mail: idunn.seland@nifu.no

S. Blömeke, J.-E. Gustafsson (eds.), *Standard Setting in Education*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-50856-6_10

10.1 Introduction

Since the early 1980s, policies for standards in education have emerged in Organisation for Economic Co-operation and Development (OECD) member states, furthered by a broad and growing interest in the organisational and structural reform of public administration. Effective schools and accountability issues were put on the international agenda, and the need for demonstrating and measuring the actual results of educational input became more prominent. In order to improve and secure quality in schools, educational standards and standard setting have become a new concern for national authorities, and this process has political as well as technical, methodological and pedagogical aspects (Lowe 1995; Dowson et al. 2007; Snyder 2010).

In Norway, from 1970s onwards, the national curriculum had been distinctively input-oriented, emphasising the content of education rather than levels of student achievement. The focus of educational policy then shifted incrementally from educational inputs to educational outcomes over the late 1980s into the 1990s. The *Knowledge Promotion* reform, coming into effect in 2006, then represented distinct curricular changes followed by structural and legal adjustments of responsibility between state and municipal levels of government. After the reform, the guiding principle for school activity is achievement understood as students' competencies, whereas administration and management are left to the local schools supported by their municipal school owner (Møller et al. 2013).[1] However, there are no specific goals or thresholds set in the national curriculum. Student achievement is formulated in the *Knowledge Promotion* curriculum as broad competence aims to be further defined and operationalised into indicators for learning by teachers locally (Tveit 2013; Skedsmo 2011).

Norwegian authorities introduced tools and procedures for assessing the quality of the national educational system in 2004, preceding the *Knowledge Promotion* reform. Among the main assessment components were national tests of core academic skills. The white paper presenting national tests as elements of this national quality assessment system mentioned the need for educational standards, but not in relation to the new test system. Despite some debate, the subject of standards was left unresolved (Utdannings- og forskningsdepartementet 2004; Tveit 2013). Four years later, a white paper instructed municipal school owners to present annual reports describing the current situation and to set goals for further development, using students' results on national tests as an indicator of quality in education (Kunnskapsdepartementet 2008).

In this study, we investigate the use of national tests in order to discuss how these results are conceptualised and employed as standards in primary education by municipal school owners, teachers and principals. In Norway, the results from

[1] In 2013, 95.3 per cent of Norwegian schools at the primary and lower secondary level were public, and 96.9 per cent of the pupils attended public schools (Statistics Norway 2015). The municipal political body is designated owner of public schools within the municipality, and holds the overarching responsibility for financing and for employment of teachers. School principals are responsible for the school's budget, and report to the municipal administrative head.

national tests are presented on a norm-referenced scale and group means are used to describe the results for a specific school, a municipality or the country as a whole. These means are then used by school owners when formulating propositions for high-standard, future student achievement stated as results 'above national average,' sometimes defined by exact percentage points. These ambitions are then imposed as standards on local schools, where teachers struggle to come to terms with the pedagogical implications of the same test results.

Previous analysis of data used in this article (Seland et al. 2015) showed that many teachers feel that they are burdened with the responsibility for preparing students for national tests and to use the results to develop students' future learning outcomes. At the same time, a majority of teachers and principals are pleased about having access to more information on student performance in general. In this study we argue that the principals and teachers do not regard this information in relation to national standards, which could be viewed as the key managerial component of the tests. This under-exploitation of information seems to make the test results more confusing to the teachers and less useful for setting future local aims and strengthened pedagogical effort for both municipal school owners and principals. As a norm-referenced mean, the test results show little variance (nor explanatory power) at a local level, and the relationship between the national tests and the curriculum is not evident to the teachers.

10.2 Background

The psychometric aspects of the Norwegian national tests relate to the definition of standard setting in its strictest form: to establish cut-off scores on examinations (Bunch and Cizek 2007). However, when the OECD first investigated the use of standards in a sample of member countries, one insight was that the concept *standard* was given a number of different meanings in the national implementation of quality in education (Lowe 1995). This is also evident in the literature, as standard setting can be understood both as establishing cut-off scores and as upholding high standards in education. Bunch and Cizek (2007) expand their definition of standard setting to include 'a procedure that enables participants using a specific method to bring to bear their judgements in such a way as to translate the policy positions of authorising entities into locations as a score scale.' In the following we concentrate on how policy positions on quality assessment of the Norwegian educational system have been translated into standards as what Dowson et al. (2007) call 'expectations of what individuals are expected to know and do,' in order to raise the level of all students' educational capability. This way of defining standard setting is closer to current practice in Norwegian schools. A related conceptualisation of standard setting is presented by Snyder (2010), stating that standards are determined through the political process that decides what the focus of schooling should be. This focus can be related to the students' individual development and preparation for future employment and life-long learning.

In Norway, national tests in reading (in Norwegian), numeracy and English are conducted annually. Hence, the tests focus on core academic skills and knowledge judged to be essential in the national education system (Dowson et al. 2007; Utdannings- og forskningsdepartementet 2002, 2004; Snyder 2010). The tests build on content standards that rely on a set of curricular outcomes or objectives (Bunch and Cizek 2007), defined at the end of the 4th and the 7th years of compulsory schooling, tested in the first semester of 5th and 8th grades respectively. These outcomes are constructed exclusively for the tests and do not appear in the national curriculum as such. The standards are then established as norms, that is level of mastery with respect to relative standing or performance within the entire group of students (Bunch and Cizek 2007). When carried out on 5th grade students, the results are presented on a three-level norm-referenced scale, whereas in 8th grade there are five levels of mastery. The mean score test results at national, municipal and school level are made public by the Norwegian Directorate for Education and Training (Utdanningsdirektoratet). Hence, the tests can be said to provide transparency and accountability measures, albeit in a low-stake fashion (Seland et al. 2015; Snyder 2010).

Even though the national tests fall within the definitions of standard and standard setting as described above, this is not made explicit in communication to the public. According to the Directorate for Education and Training, the aim of the national tests is twofold; firstly, to provide information about overall student achievement, and, secondly, to utilise this information to improve students' learning results, as a form of pedagogical tool. Operationalised, the tests are thus presented as a management tool for schools, municipal school owners and national educational authorities, as well as to serve as a formative assessment tool for teachers and students (Utdanningsdirektoratet 2010).

National tests had a rough start in Norway, stirring a vivid and heated debate in which teachers, principals, educational researchers and politicians alike protested against the new order in 2004. After having only been in place for 2 years, the tests were put on hold and then re-introduced in their current shape in 2007 (Utdanningsdirektoratet 2010; Tveit 2013). Skedsmo (2011), analysing material from the implementation of national tests in 2005, refers to the widespread boycotts that took place when the tests were first introduced. Seven years later, the tests were not only a compulsory and regular part of the schools' evaluation activities, studies also indicated that the tests were being used for both managerial and pedagogical purposes within a majority of schools (Seland et al. 2013; Mausethagen 2013).

To get a broader picture of the institutional and professional landscape that met the tests with such pronounced resistance in 2004, it is helpful to look at the recent history of assessment in Norwegian primary and secondary education. In 1988, Norwegian authorities asked the OECD for an evaluation of the national educational system. Preparing for the evaluation, the Norwegian report to the OECD (Kulturdepartementet and Kirke- og undervisningsdepartementet 1988) precedes and foretells the administrative changes that were completed with the introduction of the *Knowledge Promotion* reform almost two decades later. The transfer of responsibility for public services from the government to local political bodies had

just started in the late 1980s. A total of 450 municipalities (today there are 428) became equally responsible for public welfare. A key concern for Norwegian authorities in this situation was upholding the equal quality of education regardless of municipal financial means.

The OECD report in turn remarked on the apparent lack of information about student results and indicators of quality in Norwegian education, and commented on what their experts perceived as 'anxiety for standards' in schools where they visited (OECD 1990). According to Tveit (2013), this 'anxiety' was grounded in a state committee's proposition to abolish formal marks in primary and secondary education entirely in the early 1970s. One central reason behind this proposal was the growing unease about rankings and the sorting of children based on their academic achievements. Instead, a holistic purpose for the national education system was emphasised. Marks had then been abolished in lower primary schools, whereas they were upheld in lower secondary and upper secondary education. As the OECD did not want to challenge the apparent professional unease about the individual ranking of students, they recommended *sample tests* to measure students' learning outcomes at group level and to monitor quality development at the institutional and national level. Tests of this kind were then discussed on a national political level through the 1990s, although inconclusively (Tveit 2013).

How can we understand this 'anxiety for standards'? In their request to the OECD, Norwegian authorities pointed to the manifold purpose of the Norwegian schools, fearing that stricter control measures would disturb the balance among scientific, professional, social and regional aims of national educational policy (Kulturdepartementet and Kirke- og undervisningsdepartementet 1988). Standards in education hold the potential for affecting teachers' practices in the classroom for instrumental purposes, stressing certain criteria and leaving others out in a 'teaching to the test' manner (Tveit 2013; Ball 2011). Moreover, a more result-oriented school was believed to constitute a threat to a safe learning environment (Kulturdepartementet and Kirke- og undervisnings-departementet 1988), a condition for developing the collaborative, deliberative and social skills that traditionally and currently dominate the Norwegian national curriculum. Moreover, the teacher union's opposition to national standards has been persistent (Tveit 2013).

In an evaluation of the implementation of the *Knowledge Promotion* reform, Møller et al. (2013) found that by no means all school owners found that by no means all school owners were engaged in school development processes, while, at the same time, teachers' assessment practices seemed to have undergone a remarkable change. National programmes, headed by the Norwegian Directorate for Education and Training to define, develop and improve educational assessment, have gained considerable attention and support from schools and teachers after the implementation of the reform, above all the *Assessment for learning*-programme (Hopfenbeck et al. 2013). In accordance with the national curriculum, these assessment programmes avoid level-specific criteria when marking student results. An OECD (2011) report on evaluation and assessment for improving school outcomes in Norway labelled this feature of teachers' assessment practices problematic (Nusche et al. 2011):

> Many teachers find it difficult to translate [the curriculum] competence aims into concrete lesson plans, objectives and assessment activities. The broad competence goals have the advantage of giving teachers ownership in establishing their teaching programme, but there seems to be a need for more structure for a substantial number of teachers.

The Norwegian Ministry of Education and Research subsequently ordered the Directorate for Education and Training to evaluate the need for standards understood as national criteria for assessment. In 2015, the Directorate asked a reference group consisting of teachers, principals, school owners, union representatives and educational researchers, as well as the state county governors (*fylkesmannsembetene*), to partake in this evaluation process. A majority of these representatives stated that the current guiding principles for competence goal attainment in the national curriculum, which are of a voluntary nature, should be upheld (Utdanningsdirektoratet 2015). Hence, teachers can choose to use these guiding principles as a standard by proxy, but they can also choose not to use them. The current situation shows a general increase in professional interest in assessment in Norwegian schools, but also by a professional opposition to formal and definite national standards in education.

10.3 Data and Research Methods

Our analysis builds on data collected for a large evaluation of how the national tests work to meet the diverging aims set for them (Seland et al. 2013). The data for the evaluation project were collected in the autumn of 2012. In this study we combine quantitative and qualitative data from this project, expanded with a new qualitative document analysis of a sample of municipal school owners' annual reports on primary and lower secondary education.

The quantitative data from the evaluation consist of three surveys, one of municipal school owners, one of principals and one of teachers. The surveys of school-owners and principals are part of an omnibus covering a representative sample of Norwegian municipalities, and schools in the same municipalities.[2] The responses to the omnibus used in the analysis derives from 118 municipalities and 612 schools within these 118 municipalities. The response rate among school owners was 78%, while 65% of the principals answered the survey. A separate teacher survey was also collected at 97 schools in the omnibus sample of 612 schools, and a total of 469 teachers that had experience of national tests within the last 5 years participated in

[2] The omnibus is a project initiated by the Norwegian Directorate for Education and Training, in order to reduce the burden of research requests on the school sector. In order to do this, the population of municipalities in Norway has been divided into three comparative samples, also covering the schools in these municipalities. However, ten large municipalities are included in every sample and the schools in these municipalities are divided into three equal samples, drawn randomly. Hence, municipalities are not drawn randomly, but the samples are composed to be equal, using criteria such as size, geography, type of municipality (rural/urban) and type of school (primary, lower secondary). For further information on the design on the samples in the omnibus, see Vibe et al. (2009), and for this specific sample, see Vibe and Hovdhaugen (2012).

that survey (response rate 72%). The sample of schools from which the teachers come is representative for the population of schools in most respects, although small schools are a little under-represented, compared with larger schools.

The qualitative interviews were conducted at six case schools in three municipalities, and in each municipality we undertook site visits to one primary school (covering 5th graders) and one lower secondary school (covering 8th graders). This sample covered three schools whose students performed above the national average and three schools whose students performed below the national average on national tests, and was drawn from the sample of schools from which principals had answered the survey. The interview with the principal was a semi-structured face-to-face interview.

In preparing for the school visit, we also asked the principal to arrange interviews with teachers who had experience with national tests within the last 5 years. The eligible teachers would then be limited to teachers of Norwegian, mathematics and English, the subjects that are normally associated with national tests in Norway. We had no control over the teachers' real motives to participate in the interview, but as it turned out, many of them were recruited by the principal for the primary reason that they did not teach classes at a certain time on the day scheduled for our visit. The particular day for visiting the lower primary schools were in accordance with an actual national test taking place, which we observed in the classroom (observational data not included in this study). Following this procedure, we interviewed a total of 16 teachers, mainly in groups. In all group interviews, teachers responsible for different subjects were present. About half of the teachers interviewed were teachers of Norwegian, but many of them (especially, in lower primary schools) had experience with more than one of the three subjects, Norwegian, mathematics and English.

For all the interviews we used semi-structured guides. We asked the principals for step-by-step preparation, actual arrangement of the tests and the use of test results. The teachers were asked more or less similar questions with formulations intended to let them discuss the pedagogical and formative assessment aspects of the test results among them. In order to clarify the intentions of the tests and the rules and regulations surrounding the test we have also studied documents provided by the Norwegian Directorate of Education and Training.

The tapes from the qualitative interviews were transcribed in full and anonymised. Then each transcription was analysed, seeking to establish a bottom-up understanding of individual statements that could be grouped together to form certain categories of usage, problems and attitudes relating to teachers' work with national tests. The categories were partly derived from the survey questions, as we wanted a better understanding of what teachers do (i.e., to prepare their students for the tests), or how they work with the results. Many principals and teachers volunteered their opinions freely and associatively in the interviews, which allowed us to generate new analytical categories based on the transcriptions, i.e., on students struggling with the tests or media displays of test results. In this study, our original clusters of individual statements have been used to extract qualitative examples that can be read as illustrations and tentative explanation of survey results.

Survey data from the school owner level have been expanded with a recent qualitative document analysis. We sampled ten reports from the administrative school owner

body in ten Norwegian municipalities of varying size. These reports are statutory, produced for local political authorities, local schools and the public. The sample was extracted by putting the key words *kvalitet i grunnskolen* ('quality in primary education'), *nasjonale prøver* ('national tests'), and *målsetting* ('goal') into the Norwegian search engine Kvasir. We downloaded and read the first 10 reports that resulted from this random search. This sample method was chosen to uphold the restrictions of anonymity on the survey data, to make sure that any connections among municipalities sampled for the survey and the document study that had to be thoroughly referenced in this study, were purely incidental. The majority of municipal reports are from the school year 2014–2015, and their goals are set for a limited period of time within the span of 2014–2018. In our qualitative analysis of school owners' situation reports, we read each document in its entirety to find out how national test results were conveyed, whether these results were employed as indicators of further student academic development and if so, how these indicators were presented.

10.4 Analyses

In order to answer the question about how national tests are conceptualised and employed as a standard in Norwegian schools, we start by looking at how school owners and teachers use the test results. We then move on to discussing the implications of different types of usage of the results.

10.4.1 School Owners' Use of National Test Results

The national tests are supposed to equip the municipal school owners with data to plan and support school development actively (Utdannings- og forskningsdepartementet 2004). In general, school owners do think that the results are important for the further development of primary education; 85% agree to this statement (Fig. 10.1).

Discrepancies among municipalities based on size are clearly discernible from Fig. 10.1. The same difference is systematically upheld when we asked the school owner representatives what kind of contact or information they provided for local schools. All school owners in the largest municipalities communicate directly with principals in local schools through administrative meetings held on a regular basis, but this applies to only 83% of the school owners in the smallest municipalities. The school owners in large municipalities also utilise more channels of information, as they use circular letters, web pages and seminars to inform principals about national tests to a greater extent than is done in smaller municipalities. We will return to the significance of the number of inhabitants at the end of the study, as difference in size also applies to answers from schools with many vs few students.

The three most important topics of information were i) the intentions with national tests (79% of all school owners); ii) advice on how schools should employ

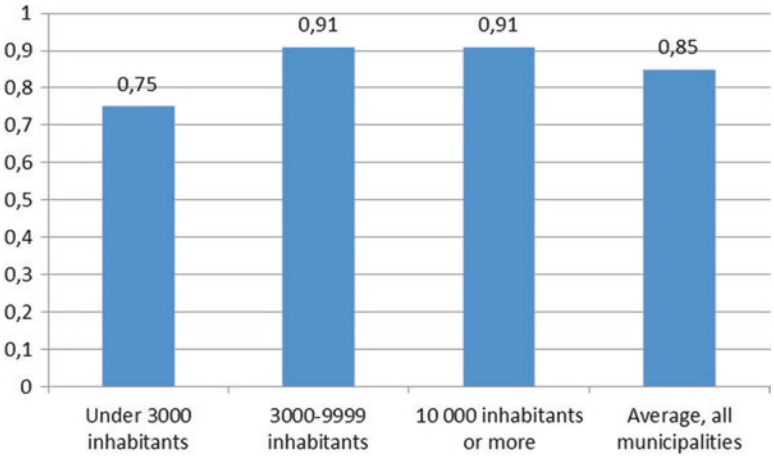


Fig. 10.1 Proportion of school owners who agree with the statement 'Results from national tests are important for school owners in the further development of primary education', by size of municipality

test results (73% of all school owners); and iii) regulations of exemption of individual students from national tests (68% of all school owners). These topics were considered important in municipalities of all types.

The qualitative analysis of school owners' situation reports shows that, as an introductory statement, several reports state the vision for primary and lower secondary education within the municipal jurisdiction. Here are a few examples (authors' translation):

> **Larvik** will become the best municipality in the county of Vestfold regarding results on national tests. (Larvik 2014)
> **Skaun** will offer the best environment for learning and upbringing among the municipalities in the region of Mid-Norway. (Skaun 2015)
> The quality of primary education in **Molde** will be among the top ten in the country. (Molde 2013)

It should be noted here that national tests are not the only indicators of quality in primary education that are considered in the municipal statuary reports. However, as instructed, all ten school owners in our sample present the results from national tests, in most cases accompanied by descriptions of the intentions and regulations of the test system copied from the Directorate of Education and Training website. Out of the ten municipalities, nine school owners state future goals anchored in national test results. Only one of the municipalities, Stange, gives no statement of future goals for national tests. Instead, a formulation about the limitations of the test results is provided: 'Small schools with a small cohort may distort the mean result' (Stange 2014).

Stange, a municipality that had a population of a little over 20,000 at the end of 2015, is, of course, right in this reflection on cohort size; especially, given the fact that they have nine primary schools with around 20 students in 5th grade on each of these schools (*skoleporten.no*). Interestingly, such reflections do not exist in the

other reports, regardless of the number of inhabitants in the municipality writing the report. The Ullensaker report states that 'the strategy plan for 2014–2018 aims for our students to score above the national average on national tests' (Ullensaker 2014). The same goes for the municipality of Kristiansund (2014). The municipality of Larvik simply aims for 'fewer students on the lowest performance level and more students on the highest performance level for every national test in 2016 compared to 2013' (Larvik 2014). Other school owners define it more elaborately, stating the exact increase/decrease in percentage points on the norm-referenced scale. Importantly, all future aims set in these reports are higher than the national average (Kristiansand 2015; Molde 2013; Gran 2014; Bergen 2014).

The report from the municipality of Smøla stands out among the reports, as Smøla has a relatively low average score, while all the other municipalities show reasonably good scores on national tests. The results from national tests in Smøla in 2012 showed that 66% of 5th-graders scored at the lowest performance level and about 14% scored at the highest performance level on the national test in numeracy. In the report, the school owner states that the aim for Smøla was to have fewer than 20% of students on the lowest performance level, and more than 30% of students on the highest performance level in 2013. Obviously, since the results are meagre, the goal has been adjusted to 'more than 20% of students on the highest performance level' (Smøla 2012).

However, Smøla is a municipality of a little over 2100 inhabitants, the smallest in this qualitative sample. The actual interpretation of results at school level could therefore be even more erroneous than in the municipality of Stange (see above). The number of students in the cohort was not stated in the report, but the website *skoleporten.no* shows that there were 22 students in 5th grade in 2012. This means that only three students were at the highest level and an increase to 20% would mean that another student reaches the highest level. The school owner representative in Smøla underlines in the report what we also have stated as the main opposing fronts in the Norwegian educational system, which is the conflict between measuring learning results and catering for students' personal growth. 'These two mandates should never be put against one another,' the report says (Smøla 2012).

10.4.2 Teachers' Use of National Test Results

In order to investigate how national tests are used in practice by teachers, teachers answering the survey were presented with a range of statements on the use of the national test with which they could agree or disagree. These statements and teacher responses are presented in Table 10.1. Strongest agreement, by over half of all respondents, is found in statements related to dissemination of results to pupils and parents. This is not surprising as it is in line with governmental guidelines, as teachers are supposed to use the tests to provide feed-back to pupils and parents. Common for the statements in Table 10.1 are that agreement indicates engagement with the test, and an interest in trying to use them for school development and pedagogic

Table 10.1 Statements on how the school work with national tests (NT)

	Disagree	Neither agree nor disagree	Agree
At my school, teachers are interested in NT results	21.0	39.9	45.1
At my school, we spend time discussing NT results	30.5	33.5	36.0
At my school, teachers and the principal work together to analyse NT results	30.1	26.3	43.6
At my school, teachers work together to analyse NT results	29.7	29.9	40.4
At my school, the principal uses NT results strategically in educational development	32.0	31.1	36.9
At my school, teachers use NT results strategically in educational development	29.6	33.0	37.4
I emphasise students' results from NT when I plan lessons	21.4	32.3	46.3
I emphasise students' results from NT when I tailor teaching to students' needs	15.9	32.4	51.7
It is important to inform parents about pupils' NT results	12.8	15.8	71.5
It is important to involve parents in improvement of pupils' NT results	28.9	28.7	42.4
There is great interest from the local community in NT results	45.0	33.9	21.1

development. The results show no clear resistance by teachers, but a leaning towards wanting to put the test results to use. However, a hesitant attitude is also indicated by about a third of the respondents, as they neither agree nor disagree with the statements.

Teachers' interest in the tests is also evident in other research on national tests in Norway. In her study of teachers' discourse on national tests, Mausethagen (2013) finds that teachers seem confident and at ease talking about test results with colleagues. She perceives this to be different from the initial opposition to the tests formed by teachers in the initial implementation of the Norwegian quality assessment system back in 2004. In our survey, teachers were asked if they view national tests as a good tool, and the answers were mixed at best: 41% agree, about a third are indifferent and a little under a third (29%) disagree. There could be a range of reasons why teachers are reluctant to support the tests, but both the survey and the interviews indicate that this may be linked to how teachers experience the usefulness of test results and how the test results are presented in the media.

Further analysis of the correlation of how teachers make use of the test indicates that teachers who utilise test results in the actual planning of lessons are more inclined to view the tests as a good tool for their work.[3] Test results employed in a strategic way or at a managerial level do not comply as well with teachers' view of

[3]This statement is in part based on a regression done in the evaluations report, which indicated a clearly significant effect, that teachers who are using test results in their lesson planning also see

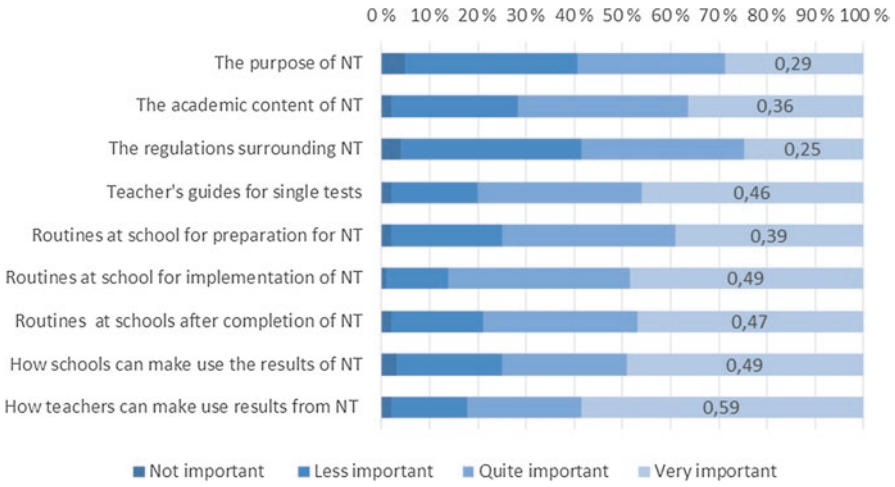


Fig. 10.2 Responses to the question 'Concerning national tests (NT): how important is the following information for you, in your job as a teacher?'

whether the tests are a good tool for them. Teachers want to use the tests as a pedagogic tool, but results used in a strategic/managerial way do not resonate well with this. To an even lesser degree is there a positive attitude towards national tests associated with community or public interest in the test results.

The type of information concerning the national tests that teachers request also indicate that they expect to be able to use the tests as a pedagogic tool. Figure 10.2 shows that 59% of the teachers state that information on how they can make use of results is very important. So far, we have seen that teachers are quite eager to know how they can utilise test results, and that the teachers who succeed in utilising results when they plan lessons, are most content with national tests. Still, Table 10.1 also shows that less than half (46%) of the teachers utilise test results for these purposes, and 40% work with colleagues in order to analyse test results. However, 30% disagree that they work with colleagues in this way. Only 37% of the teachers answered that colleagues collaborate to use test results when they plan for pedagogical development.

Why do not more teachers use test results for planning lessons and pedagogical development? Comments in the survey and interviews with teachers indicate that one reason lies in the teachers' own comprehension of the tests as well as the test results. As mentioned before, the results from national tests are presented on a norm-referenced scale with levels of mastery with respect to relative performance within the entire group of students (Bunch and Cizek 2007). Here are three illustrative statements from our teacher respondents (authors' translation):

---

the tests as a good. Tool (Seland et al. 2013: 129). The bivariate correlation between the two items has also been checked and is significant (Pearson's r =0,545, Cronbach's alpha=0,705).

> I have little knowledge of how to utilise test results in my teaching. To be presented with a leaflet and be ordered to just go on with it [i.e., use the test results] is less inspiring. Courses on how and why we should make use of the tests should be mandatory.
>
> I felt bad in that parent-student-teacher conference, when I said, 'well, your daughter is at level 2, that is somewhat critical and means she has to practise.' Luckily there were no follow-up questions, like 'why am I at level 2?', because that I could not explain.
>
> [I] could not find any explanation or guidance on the internet on what my students really were tested for, so I could not use the results for anything. At what point did they fail?

Teachers' apprehension about the tests seems related to the level of cooperation within the local school on how test results should be understood and what measures could be taken. Our teacher survey shows that 41% of the teachers' experience that the teaching staff work together to analyse national test results, while 43% report that the principal works along with teachers to analyse results. It should be noted that in the survey, these answers are not mutually exclusive. A proportion of between a quarter and a third of the teachers replied 'both yes and no' to these questions, which gives the impression that there are vague routines for cooperation on result analysis in a great number of schools. 'The tests and their results concern the whole teaching staff', says one teacher whose school had experienced disappointing results over the years, but were now at a turning point:

> …It provides a reflection on how well we succeed. We cooperate, (…) we try to track every student's individual progress. That is why we have to know what the students learn in the first, second and third year. We have to understand all possible explanations [for results on national tests].

'Teachers in 1st to 4th grade find the tests exciting,' says a teacher whose school has good test results, 'because in reality, the tests measure *their* success.' There are also teachers' comments and interview statements that give indications of lack of cooperation among colleagues, which causes confusion: 'We have just had the test results,' one teacher says, 'and the principal is interested, the students are eager and the teacher who had them last year is excited. But what do we do now?' 'This is a lonely job,' says one teacher at a school that has the municipality's poorest test results. 'It seems hard for the other teachers to realise the importance of making the students practise reading, for instance, when it's not in their syllabus.'

What these statements illustrate is that some teachers or entire teaching staffs either do not understand or are unwilling to realise what the national tests measure and indicate as standards in education. Teachers, who report on collegial cooperation for analysis and utilisation of test results, convey an understanding that the results show actual student achievement compared to 4th and 7th grade curriculum criteria respectively. Hence, at these schools, teachers responsible for the first 4 years of schooling are involved and made responsible. But the opposite is also evident in the qualitative data: one group of 5th grade teachers said that their school had had 'terrible results' over several years, and the principal just told them to 'fix' it. These teachers, however, had little or no knowledge of their students before they reached 5th grade, and had even less contact with their previous teachers. Not surprisingly, these 5th grade teachers were perplexed as how to deal with the situation and were only able to try to patch up their students' skills for testing in 8th grade. Here, the

lack of correspondence between the national tests' scoring criteria and curriculum poses a significant problem to the teachers when it comes to using the test as a pedagogic tool. A teacher expresses this as:

> Using the national tests, I can find out that a student's score is low, but not *why* the student's score is low. In order to find out why, I have to use [a diagnostic] test instead.

Teachers who have a negative attitude towards the tests express the view that the results do not show anything they did not already know, and that the tests are too narrow in scope to be able to measure pedagogical quality. However, there are also expressions of praise for the national tests, as the tests 'stimulate schools, since what is tested in the national tests are important skills students should master,' as one teacher puts it.

An aspect of the tests that teachers frequently mentioned in the interviews was the publication of results. As information about test results is a natural starting point for making comparisons, they are inevitably used for the benchmarking of schools by the Norwegian media (Elstad 2009). As mentioned earlier, small municipalities and small schools may experience great variations in results from year to year, simply because of cohort variations, rather than variations in the quality of teaching. Teachers are fully aware of this, and are in general quite negative towards the publication of results. This comes across as comments in the survey:

> Results of the national test depend on many different variables, and are therefore uncertain. They have to be coupled with the teachers' observations of students.
>
> The composition of a cohort may vary from one year to another, and this is not addressed in the publication of results.

There are also teachers who state that results should be published, as schools are a public service provider. However, the proportion of teachers showing negative attitudes toward publication of results significantly outnumber those who are positive about it. Several of the principals interviewed expressed resignation and resentment against publication of the test results in the media. One principal said that media has made the statistical national mean test results into set equivalents for school quality, which he knows to be wrong. Other principals agree:

> When the newspapers start on such rankings, (…) as if they can tell if one school is better than the other. I for one know the reality, but the public does not know. It is a shame they are left with such impressions.

'A mean score of 1.7 could actually prove to be a good result,' says another principal, pointing out that every score has to be evaluated according to students' and parents' socioeconomic background. 'But this knowledge never reaches the public,' he adds.

Here we touch on the ambivalence that the principals show, when they talk about national tests and their use. In interviews, the principals talk about how working to improve results actually implies comparing results, either to former years or to other schools. This is a challenge not only for schools that are scoring under par, but also for schools that have good scores, that have to try to use the results of the test for school development. Both schools that score well, and schools that score less well,

may use external factors such as the composition of the student body, to explain their results. However, one principal states that all schools should work to develop student learning outcomes further by comparing themselves to schools where similar conditions exist.

In the survey, principals showed good comprehension of the tests in general, both as an educational standard and as a pedagogic tool for further development of student achievement. A majority (60%) agreed to national tests being a good tool for school development. Again, there are differences based on size, both the size of the municipality and the size of the school. Principals of larger schools and in larger municipalities tend to agree more with the statement that the tests are a good tool for school development than those in smaller schools and smaller municipalities do. In interviews, some principals argue that there has been a development towards more serious and systematic work with test results over the last years. 'We may have taken it too lightly,' one principal says, 'viewing these tests as rather trivial matters. Over recent years there seems to have been an "awakening" – the tests are a good tool, and they show us where to place our efforts.'

10.5 Conclusions

Educational standards were not the official object of the national tests, when they were first introduced in Norwegian schools in 2004. The central idea of the national tests is presented in a downplayed manner, as sources of management information for national authorities, municipal school owners and principals. Test results are also intended to provide principals, teachers, students and their parents with pedagogical information to improve further learning.

In this study, we have investigated the use of national test results to see how these results are conceptualised and employed. School owners appear to have a very high interest in results of national tests. While representatives of school owners seem to regard the national statistical mean result of the national test as the 'acceptable' level of academic achievement, their vision for future high standard achievement is commonly projected above this statistical average. School owners consequently envision and formulate new statistical aims to be fulfilled by local schools.

Principals function as the link between school owners and the teachers, and our survey shows that the majority of principals see national tests as a good tool for school development. However, survey results as well as interviews with principals and teachers reveal that by no means all principals engage in the analysis of results, but leave this for the 5th and 8th grade teachers to sort out. In these circumstances, national tests are utilised as formative assessment in order to strengthen and develop students' skills for the future. While this is not wrong, future student achievement will not improve unless the teaching staff responsible for the early years of schooling also strengthen their efforts. This is a management responsibility, and serves as an explanation to why only about 40% of the teachers confirm that national tests are a good tool for working with students' learning.

The most striking feature of teachers' comments in the survey and in interviews are lack of coherence among curricular aims, teachers' practices and national test results. Teachers appear to be inadequately informed about what national test scores actually measure and what kind of information about student achievements can be drawn from the results. This results in either a struggle to grasp the true meaning of the tests or indifference bordering on disregard for the tests altogether. We interpret this as teachers not being aware of the key managerial element of the tests constructed as a summative, standard-based measure of student learning after the initial 4 years and the end of 7 years of primary education, respectively. With this piece of information missing, teachers seem to under-exploit test results for student learning development, as they try to use the test results for formative purposes. Conversely, school owners' reports seem to over-exploit the same results.

The main problem with national tests not being openly presented and discussed as standards in education is the lack of coherence between curriculum competence aims and test results. As our study points out, this problem stems from a long history of professional unease and political scepticism towards having standards understood as individual ranking of students. Consequently, there are no set aims for measuring student achievement in the national curriculum, only broad competence aims to be further operationalised into indicators of learning by individual teachers. In turn, this does not contribute to creating a common understanding of what the standards are and should be. The descriptions of the levels of mastery in the national tests, made available to the teachers and the public by the Norwegian Directorate of Education and Training, is an effort to bridge this gap. However, the actual standards and the consequences of not fulfilling them are not spelt out in these documents. Somehow the 'anxiety for standards' that the OECD experts remarked on in 1988 seems to prevail at national as well as school level – but, interestingly, not at the municipal school owner level.

A second problem is posed by the choice of a norm-referenced scale, and this is evident even in situations where national tests are accepted as a standard by school owners and principals. We propose that this mode of presentation in itself leads to under- as well as over-exploitation of results by teachers and school owners respectively. The lack of detail in what the students do not know or master, makes the tests a less valuable tool for the teachers and forms a basis for school owners' excessive ambition on behalf of local schools and teachers. As a norm-referenced mean, the test results provide little variance and explanatory power at a local level, and when applied to small cohorts, the value of this information decreases even more. The data clearly show that school owners in large municipalities are more active and more engaged with national tests than school owners in small municipalities. The same tendency is visible regarding school size in data from the survey on principals. Norway is a sparsely populated country with about 38% of the municipalities housing 3000 or fewer inhabitants, and at the national level, about 50% of the schools have fewer than 30 students in the 5th grade. Here, national test results can possibly provide a map for future tailored education at the individual level, but they are of less value for school development.

The situation of national tests as educational standards in Norway seems inconclusive. The tests fulfil the definitions of standards both in a psychometric methodological sense (Bunch and Cizek 2007) and in the meaning of 'expectations of what individuals are expected to know and do' (Dowson et al. 2007). However, the political process to determine what the focus of schooling should be for individual development (Snyder 2010), has been inconclusive. Applied as standards by school owners, future aims are result-driven. At school level, aims for future development are assessment-driven, but these two phenomena are never linked with set curricular aims, which is partly a political process. Instead, national tests emerge as an *undeclared* standard in Norwegian education, causing ambiguous political demands and signs of professional frustration.

References

Ball, S. J. (2011). *Education debate: Policy and politics in the twenty-first century*. Chicago: The Polity Press.

Bunch, M., & Cizek, G. J. (2007). *Standard setting. A guide to establishing and evaluating performance standards on tests*. Thousand Oaks: Sage Publications.

Elstad, E. (2009). Schools which are named, shamed and blamed by the media: School accountability in Norway. *Educational Assessment, Evaluation and Accountability, 21*(2), 173–189.

Dowson, M., McInverny, D. M., & Van Etten, S. (2007). The state of play in standards and standard reform. In D. M. McInverny, S. Van Etten, & M. Dowson (Eds.), *Standards in education* (pp. 3–11). Charlotte: Information Age Publishing.

Hopfenbeck, T., Tolo, A., Florez, T., El Masri, Y. (2013). *Balancing trust and accountability? The assessment for learning programme in Norway*. A governing complex education systems case study: OECD. http://www.oecd.org/edu/ceri/Norwegian%20GCES%20case%20study%20OECD.pdf. Accessed 12 Oct 2016.

Kulturdepartementet og Kirke- og undervisningsdepartementet. (1988). *Reviews of national policies for education: Norway*. Oslo: Kulturdepartementet og Kirke- og undervisningsdepartementet.

Kunnskapsdepartementet. (2008). *Kvalitet i skolen. (St.meld. nr. 31 2007–08)*. Oslo: Kunnskapsdepartementet.

Lowe, J. (1995). Overview. In: *OECD*, *Performance standards in education: In search of quality* (pp. 7–26). Paris: OECD.

Mausethagen, S. (2013). Talking about the test. Boundary work in primary school teachers' interactions around national testing of student performance. *Teaching and Teacher Education, 36*, 132–142.

Møller, J., Prøitz, T. S., Rye, E., & Aasen, P. (2013). Kunnskapsløftet som styringsreform. In B. Karseth, J. Møller, & P. Aasen (Eds.), *Reformtakter. om fornyelse og stabilitet i grunnopplæringen* (pp. 23–41). Oslo: Universitetsforlaget.

Nusche, D., Earl, L., Maxwell, W., & Shewbridge, C. (2011). *OECD reviews of evaluation and assessment in education: Norway*. Paris: OECD.

OECD. (1990). *Reviews of national policies for education: Norway*. Paris: OECD.

Seland, I., Hovdhaugen, E., & Vibe, N. (2013). *Evaluering av nasjonale prøver som system. NIFU-rapport 4/2014*. Oslo: NIFU.

Seland, I., Hovdhaugen, E., & Vibe, N. (2015). Mellom resultatstyring og profesjonsverdier. *Nordisk Administrativt Tidsskrift, 3*(92), 44–59.

Skedsmo, G. (2011). Formulation and realisation of evaluation policy: Inconsistencies and problematic issues. *Educational Assessment, Evaluation and Accountability, 23*(1), 5–20.
Snyder, C. (2010). Standards in assessment in education. *Development, 53*(4), 540–546.
Statistics Norway. (2015). *Fakta om utdanning 2015 – Nøkkeltall fra 2013*. Oslo: Statistisk Sentralbyrå.
Tveit, S. (2013). Profiles of education assessment systems worldwide: Educational assessment in Norway. *Assessment in. Education, 21*(2), 221–237.
Utdanningsdirektoratet. (2010). *Rammeverk for nasjonale prøver*. http://www.udir.no/Upload/Nasjonale_prover/2010/5/Rammeverk_NP_22122010.pdf?epslanguage=no. Accessed 12 Oct 2016.
Utdanningsdirektoratet. (2015). *Svar på oppdragsbrev 18–14 punkt 3, kartlegge og vuerdere nytteverdien av kjennetegn på måloppnåelse*. Letter from the Norwegian Directorate of Education and Training to the Norwegian Ministry of Education and Research.
Utdannings- og forskningsdepartementet. (2002). *Førsteklasses fra første klasse. Forslag til rammeverk for et nasjonalt kvalitetsvurderingssystem for norsk grunnopplæring. (NOU 2002: 10). Delinnstilling fra Kvalitetsutvalget*. Oslo: Utdannings- og forskningsdepartementet.
Utdannings- og forskningsdepartementet. (2004). *Kultur for læring.* (St. meld. nr. 30 2003–04). Oslo: Utdannings- og forskningsdepartementet.
Vibe, N. & E. Hovdhaugen. (2012). *Spørsmål til Skole-Norge høsten 2012*. NIFU-rapport 47/2012. Oslo: NIFU. Accessed 12 Oct 2016.
Vibe, N., M. Evensen, & Hovdhaugen, E. (2009). *Spørsmål til Skole-Norge: Tabellrapport fra Utdanningsdirektoratets spørreundersøkelser blant skoler og skoleeiere våren 2009*. NIFU-rapport 33/2009. Oslo: NIFU. Accessed 12 Oct 2016.

Overview of Documents Used in the Document Analysis of School Owner Reports

Bergen kommune. (2014). *Kvalitetsmelding for bergensskolen 2014*. http://www3.bergen.kommune.no/BKSAK_filer/bksak%5C2015%5CBR1%5C2015245634-5209033.pdf. Accessed 12 Oct 2016.
Gran kommune. (2014). *Tilstandsrapport for grunnskolen i Gran 2014–2015*. http://www.gran.kommune.no/Documents/Grunnskole/Tilstendsrapport%20for%20grunnskolen%20i%20Gran%202014-2015.pdf. Accessed 12 Oct 2016.
Kristiansund kommune. (2014). SPOR. Styringsdokument for grunnskolen i Kristiansund 2014–2017. http://www.kristiansund.kommune.no/Handlers/fh.ashx?MId1=897&FilId=3523. Accessed 12 Oct 2016.
Kristiansand kommune. (2015). *Kvalitets- og utviklingsmelding 2015*. https://www.kristiansand.kommune.no/globalassets/barnehage-og-skole/skole-og-sfo/pedagogisk-senter/kvalitets--og-utviklingsmelding-2015.pdf. Accessed 12 Oct 2016.
Molde kommune. (2013). Kvalitetsplan for grunnopplæringen i Molde kommune. Del 3: Aktuelle indikatorer for kvalitet i skolen for perioden 2013–2016. https://img5.custompublish.com/getfile.php/2761915.2125.asyewfqwvy/Kvalitetsplan.p?return=www.molde.kommune.no. Accessed 12 Oct 2016.
Larvik kommune. (2014). *Kvalitet i skolen 2014–2016*. Læring først og alt annet etterpå. http://www.larvik.kommune.no/Global/Skole/Dokumenter/KvalitetISkolen%202014-2016.pdf. Accessed 12 Oct 2016.
Skaun kommune. (2015). *Kvalitet i skolene i Skaun. Med blikk for framtida. Sektorplan for grunnskolen 2015–2018*. https://img0.custompublish.com/getfile.php/3004239.1720.qbdyfrwtrp/Sektorplan+grun?return=www.skaun.kommune.no. Accessed 12 Oct 2016.

Smøla kommune. (2012). *Tilstandsrapport for grunnskolen – Smøla kommune*. http://www.smola.kommune.no/Handlers/fh.ashx?FilId=935. Accessed 12 Oct 2016.

Stange kommune. (2014). *Tilstandsrapport for grunnskolen i Stange kommune 2014*. http://www.stange.kommune.no/getfile.php/Filer/Stange/PDF/Tilstandsrapport_skole.pdf. Accessed 12 Oct 2016.

Ullensaker kommune. (2014). *Kvalitet i Ullensakerskolen – tilstandsrapport for grunnskolen 2014–2015*. https://www.ullensaker.kommune.no/Documents/Ullensaker%20dokumenter/05%20Ull%20Skole%20og%20barnehage/Skole/Tilstandsrapport_UllensakerKommune2015.pdf. Accessed 12 Oct 2016.

Chapter 11
Setting Standards for Multistage Tests of Norwegian for Adult Immigrants

Eli Moe and Norman Verhelst

Abstract This chapter reports on the procedures applied when setting standards for a multistage test in Norwegian for adult immigrants. Cut scores were set for listening and reading tests between the levels (1) A1 and A2, and (2) A2 and B1 on the Common European Framework of Reference for Languages (CEFR). In addition to documenting the quality of the procedures, the question of whether and how to take the multistage design into account is discussed. Nineteen judges took part in the standard setting. The method used was a generalisation of item-mapping strategies of Cizek and Bunch, which leans heavily on a graphical representation of item and task difficulty. The judges did two rounds of standard setting for each test. The average standards for listening and reading, A1/A2 and A2/B1, were remarkably stable across all rounds, including a round of cross validation. While the natural choice would be to use a conditional test response curve as a basis for setting standards, an unconditional curve was used in the end due to empirical and ethical reasoning. Until a convincing rational argument is found, empirical evidence will have to support the standards chosen.

Keywords Tests for adult immigrants • Reading test • Listening test • Item Response Theory • Multistage testing

E. Moe (✉)
Research Group for Language Testing and Assessment,
University of Bergen, Bergen, Norway

Kompetanse Norge (Skills Norway), Bergen, Norway
e-mail: Eli.Moe@uib.no; Eli.Moe@kompetansenorge.no

N. Verhelst
Eurometrics, Tiel, The Netherlands
e-mail: norman.verhelst@gmail.com

S. Blömeke, J.-E. Gustafsson (eds.), *Standard Setting in Education*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-50856-6_11

11.1 Introduction

The main aim of this study is to document the quality of the standard-setting procedures applied when linking results from a Norwegian language test for adult immigrants, *Norskprøven*, to the Common European Framework of Reference for Languages, the CEFR, (Council of Europe 2001). Insights gained from the study should inform judgement of the effectiveness of the standard setting procedures applied, and what operational procedures need to be implemented in order to ensure a valid and reliable link to the CEFR in the future.

An important additional aim is to discuss a problem, hitherto not discussed in the standard setting literature. Standards have been set using a test-centred procedure that results in a standard on a latent variable. To find the standard in the score domain, one uses commonly the test response curve, finding the expected score corresponding to the latent standard. This is a straightforward procedure: since the latent standard is, by definition, the competence of a person on the border between two categories (pass or fail, or between two neighbouring categories of the CEFR descriptive scale), the expected score associated with this competence is the average score of such a borderline person. But it is reasoning that applies unequivocally to the case of linear tests where all students answer the same items in the same order.

As the population of immigrants taking the tests of Norwegian is very heterogeneous – the CEFR level varying from below A1 to B2 or higher – a single linear test for all candidates is not realistic. Targeted testing – using extra background information from the candidates to determine their approximate level – was not realistic either, since a substantial percentage of the candidates do not follow any course in the Norwegian language. Therefore, it has been decided to use multistage testing, whereby candidates take one or two routing tests, and depending on their score on this routing test, they take a main test at a low (A1/A2), an intermediate (A2/B1) or an advanced level (B1/B2). See Sect. 11.3 for more about the CEFR and the different levels. As the routing test takes substantial testing time, the performance on the routing test (s) will be taken into account when making e a decision on the candidate's language proficiency level. The important question then is to know whether – and how – the multistage design should be taken into account in this decision.

11.2 Background

Vox is the Norwegian Agency for Lifelong Learning and belongs mainly to the Norwegian Ministry of Education and Research. Among other things, Vox is in charge of curricular and pedagogical issues regarding the teaching of Norwegian and social studies to adult immigrants. The agency monitors the implementation of the national curricula and the national tests for this group, initiates research and development and disseminates information to stakeholders in the field (Vox 2016a). January 1st 2017, Vox changed its name to Kompetanse Norge. The new English

name is Skills Norway. Since this chapter was written in 2015/2016, and since everything described in this chapter took place before the name change, the former name Vox is used throughout this chapter.

Until 2014, Folkeuniversitetet developed tests in Norwegian for adult immigrants at the CEFR levels A2 and B1, *Norskprøve 2* (A2) and *Norskprøve 3* (B1) on behalf of Vox and the Ministry of Labour and Social Inclusion. The results of these language tests were reported as 'pass' or 'fail'. In 2011, the ministry commissioned Vox to be in charge of the development of new tests for the same group of immigrants with a graded reporting system.[1] The result was *Norskprøven*, a multistage language test measuring language proficiency at the levels A1, A2 and B1.

Norskprøven includes tests of listening, reading, speaking and writing. All parts of the test, except the speaking test, are computerised. The focus of this chapter is the listening and reading tests since these tests are constructed as a series of items, where the answers are scored as correct or incorrect. The speaking and writing performances of the candidates are assessed by two raters on the basis of rating criteria mirroring the competence required at the levels A1–B1. The speaking and writing tests are therefore not subjected to standard setting procedures. *Norskprøven* is administered twice a year during a two-week period, in May/June and in November/December. The first test administration took place in May/June 2014. This chapter reports on the standard setting event which took place prior to the first test administration. This standard setting event aimed to set listening and reading standards for the CEFR levels A1/A2 and A2/B1.

11.3 The Common European Framework of Reference (CEFR)

The Common European Framework of Reference of Languages: learning, teaching and assessment was published in 2001. It assigns language learners into three main groups according to their language competence. Basic users (A1 and A2) focus on learning the most important, everyday language in order to be able to communicate in everyday situations. Independent users (B1 and B2) can cope independently in social and educational settings, while advanced users (C1 and C2) are able to use the language effortlessly, coherently and effectively in professional settings.

In the CEFR there are 56 scales of language descriptors covering (1) five different skills (listening, reading, spoken production, spoken interaction and writing) and (2) the six levels A1–C2. On one hand, the CEFR levels should be firmly set. If not, they would lose their function as common reference points. On the other hand, the CEFR is not meant to be dogmatic, prescriptive or absolute. The document is a framework of reference, which allows different interest groups to adapt the levels and basic principles to national or more local situations.

[1] Folkeuniversitetet's latest test administration was in February 2014, while the first administration of *Norskprøven* (Vox) was in May/June 2014.

Table 11.1 Reading for information and argument (CEFR 70)

C2	As C1
C1	Can understand in detail a wide range of lengthy, complex texts to be encountered in social, professional or academic life, identifying finer points of detail including attitudes and implied as well as stated opinions.
B2	Can obtain information, ideas and opinions from highly specialised sources within his/her field.
	Can understand specialised articles outside his/her field provided he/she can use a dictionary occasionally to confirm his/her interpretation of terminology.
	Can understand articles and reports concerned with contemporary problems in which the writers adopt particular stances and viewpoints
B1	Can recognise significant points in straightforward newspaper articles on familiar subjects.
A2	Can identify specific information in simple written material he/she encounters such as letters, brochures and short newspaper articles describing events.
A1	Can get an idea of the content of simpler informational material and short, simple descriptions, especially if there is visual support.

Originally, the CEFR descriptors were developed with adult foreign language learners in mind; for instance, tourists and teenage or adult students. Later, the levels and descriptors were adapted and used, for instance, in second language contexts (Vox 2012) and for children (Hasselgreen 2003; Hasselgreen et al. 2011).

Table 11.1 gives an example of original CEFR descriptors for the scale "Reading for information and argument" A1–C2. While the A1 descriptor points to very basic reading competence the C1 descriptors describes advanced reading competence. The list below shows A2 listening descriptors adapted for the curriculum aims of VOX for immigrants studying Norwegian (Vox 2012). The descriptors are translated into English by the authors.

- Can understand sentences and frequently used expressions, when speech is slowly and clearly articulated.
- Can identify the topic of conversations provided speech is slowly and clearly articulated.
- Can catch the main point in short, simple messages and announcements provided speech is slowly and clearly articulated.
- Can follow short and simple directions.
- Can identify some main points in the news on TV.
- Can understand some frequent dialect expressions.

The main aim of the standard setting procedures applied was to set cut scores for listening tests and reading tests between the levels (1) A1 and A2 and (2) A2 and B1. Listening and reading descriptors were used to prepare the standard setting judges for the standard setting.

11.4 The Test Candidates

The persons sitting for Norskprøven form a heterogeneous group with respect to age, immigration status in Norway, language and school background. All immigrants are offered courses in Norwegian and social studies. For asylum seekers and refugees these courses are free of charge. Other participants have to pay a fee. According to Statistics Norway, the age range in the group attending Norwegian courses was from 16 to 56 and above in 2014 and 2016; however, close to 80% of the students were between 16 and 35 years of age in 2013 and 2014. Approximately 85% of them were refugees, asylum seekers or persons who had been reunited with their families. The largest immigration groups were from Eritrea, Somalia, Thailand, Syria, the Philippines, Afghanistan, Ethiopia, Sudan and Iran. The educational background of the participants varies a great deal; some have a university background, others have to learn to read and write when they start their Norwegian courses.

In the schools and educational centres, the participants attend classes according to their educational background. According to *The National Curriculum in Norwegian and Social Studies for Adult Immigrants* (Vox 2012), there are three teaching 'streams', stream 1, 2 and 3. Teaching and training in the different streams are based on the educational background of the participants. This means that pace and progression of the training will be different for the three streams. In addition, work methods, teaching materials and group size may vary. Stream 1 is tailored to persons who had little or no schooling, and therefore are not able to use written language as a means for learning. Some of the students attending this stream have not learned to read and write. Stream 2 is for those with some formal education who are able to use the written language in the language learning process. Stream 3 is tailored to participants who had a good general education. Some of the students in this stream have also started or completed education at college or university level in their home countries. Table 11.2 gives an overview of the number of participants in the three streams in 2013 and 2014. The table shows that most participants attend stream 2 classes, while around 20% attend stream 1 and stream 3 classes.

It is also possible for candidates who have not attended the courses to sit for *Norskprøven*. In 2013, 36% of the test candidates sitting for *Norskprøve 2* and *3*

Table 11.2 Participants in Norwegian and social studies classes divided into streams

	Number of participants					Percentage of participants			
	Stream 1	Stream 2	Stream 3	No stream[a]	Total	Stream 1	Stream 2	Stream 3	No stream
2014	7949	21,176	7517	2033	38,675	20.6	54.8	19.4	5.3
2013	7502	22,620	8683	3846	42,651	17.6	53.0	20.4	9.0

[a]'No stream' refers to students who are attending Norwegian courses, but who are not assigned to any specific stream. This may happen in rural areas, in small schools, where there are only a few immigrants learning Norwegian

Vox (2016b)

were private candidates. The fact that the Ministry of Labour and Social Inclusion asked Vox to develop one Norwegian language test for this diverse group made constructing such a test challenging.

11.5 The Listening and Reading Tests – Overview

In 2014, tests were developed for the CEFR levels A1 to B1 (in 2015, tests for B2 were added). Since Norskprøven is aimed at a heterogeneous group of test takers, a decision was made to develop multistage listening and reading tests. Candidates with a higher level of Norwegian language proficiency would sit more challenging tests than those with a lower level of proficiency. In this way, low-level candidates would not feel over-challenged, as they could work with items that they were able to answer correctly. Candidates would end up taking one of two main tests, an A1–A2 test or an A2–B1 test. The level of the main test would be determined by the outcome of one or two routing tests, R1 and R2. Figure 11.1 gives an overview of the routing system of the listening and reading tests.

All candidates start a listening or reading test by taking Routing test 1, R1. This is a short test of six easy items developed for A1 and A2 language learners. Candidates who answer three or less items correctly, go directly to an A1/A2 main test. Those answering four or more items correctly proceed to Routing test 2, R2, with six items aimed at A2 and B1 learners. If the outcome of R2 is three or less correctly answered items, candidates are routed back to the A1/A2 main test. Candidates with four or more correct answers go on to the A2/B1 main test. The A1/A2 main test consists of 20 items for both listening and reading proficiencies; the A2/B1 main test consists of 23 listening items and 25 reading items. This means that all candidates who are taking the A2/B1 listening main test will take two routing tests. All in all, they will answer 35 items of which 12 are from the routing tests. It means therefore that these candidates will spend approximately one third of their time on the routing tests. The developing team at Vox thoroughly discussed whether to include the performance on the routing tests or not, when making a decision on the candidate's performance. If one chooses to use the routing tests only as a classification instrument, more items would have to be added to the main tests. This would result in longer tests which would take more time. Vox wanted as many

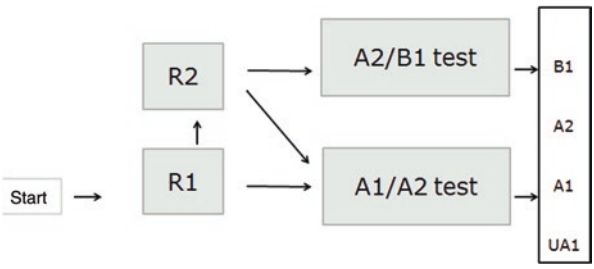


Fig. 11.1 Vox' multistage listening and reading tests

Table 11.3 Test construct and test format – listening and reading tests

Test construct: What the items measure	Item formats
Listening test	*Listening*
Understanding words and some details (A1)	Click on an item in a picture
Understanding details (A2–B1)	Click picture (match a prompt with one of four pictures
Understanding main points (all levels)	Multiple choice items
Inferencing (all levels)	Move key words (click and drag)
Text coherence (B1)	
Reading test	*Reading*
Understanding words and some details (A1)	Click on an item in a picture
Understanding details (A2 – B1)	Click picture (match a prompt with one of four pictures)
Finding information (all levels)	Click text (match a prompt with one of four texts)
Understanding main points (all levels)	Multiple choice items
Inferencing (all levels)	Click on a word/name
Understanding words in a context (A2–B1)	Move paragraph (click and drag)
Text coherence (B1)	

candidates as possible to have a real opportunity to take the tests without being stressed or over-challenged. In the end, the concern for the low-level candidates made Vox decide to include the candidate's 7performance on the routing tests in the final result. The test items measure different aspects of listening and reading through different item formats. Table 11.3 gives an overview of what the listening and reading items measure and of the different item formats.

11.6 Taking Test Takers with Little School Background into Account

Since approximately 20% of the students in schools have little or no educational background, it was necessary to have a particular focus on this group when developing *Norskprøven*. When the work started, the test developers were in contact with schools and teachers who taught Norwegian to the stream 1 group. In 2012 and 2013, teachers participated in workshops where they discussed how to tailor the A1/A2 tests to these students. In the autumn of 2012, a small study which focused on digital listening items and stream 1 and 2 students was set up. The aim was to ensure that these students were able to show their understanding of Norwegian when responding to the listening items (Moe 2013). The result of this preparatory work was that the A1/A2 reading and listening tests have a reduced number of items and a restricted number of item formats.

In addition, three item formats use pictures, and the students give their answers by clicking on pictures. Students do not have to write when answering the listening and reading tests, they click to give their answers (or click and drag at the A2/B1 test). Sample tests are freely available online throughout the year in order to make sure students and teachers know the formats and how to answer the items (Vox 2016c).

Particular attention was given to the A1/A2 listening test. Teachers stated that they assessed the listening competence of some of their stream 1 students to be A2 and for some students even higher. However, the results of former *Norskprøve 2* and *3* did not show this; probably, because students had to read a lot (questions and alternative answers) and, for some of the answers, they had to write. In the new listening tests no writing is involved. The students hear the questions, before listening prompts are given, and the questions also appear on the screen. In addition, a decision was made that all alternative answers on traditional multiple choice questions were to be in numbers (for instance, kilos, centimetres, months, and weekdays) to reduce reading difficulties as much as possible.

11.7 Piloting and Test Construction

11.7.1 The Design of the Pilot Data

The first piloting took place in spring 2013. Eighteen pilot booklets were constructed for listening according to an incomplete design: 9 A1/A2 booklets and 9 A2/B1 booklets. The same system was used for piloting reading items. Each item appeared in three booklets. The (same) A2 items were piloted both in the A1/A2 booklets and in the A2/B1 booklets. The A1/A2 booklets contained 26 or 27 items; the A2/B1 booklets contained 37 items. The teachers chose the level of the listening and reading booklets for their students: A1/A2 or A2/B1. While 3645 students took a listening pilot test, 3781 took a reading pilot test. For both skills, reading and listening, 192 items were piloted.

It is useful to draw the attention here to an important distinction: targeted testing and multistage testing. For the data collection in the pilot, targeted testing is used, in the sense that external information on the proficiency level of the candidates was used to assign them an easy (A1/A2) or a hard (A2/B1) test booklet. However, this information - the professional judgment of the teacher - is not part of the test, and is not used in the parameter estimation of the IRT-model (see below) nor in any decision about the 'certified' level of candidates. The main purpose of the analysis is to construct a series of tests (routing tests and main tests) which will be used in the field to test and make decisions on future candidates as to their level in the CEFR-system. This will be done in the future by multistage testing: candidates will start with one or two routing tests to determine their approximate proficiency level, and depending on the performance on these routing tests, they will take a main test of appropriate difficulty. In order to be able to assign them to a level of the CEFR, all their answers – routing test (s) and main test – will have to be taken into account in the final decision.

11.7.2 The Analysis of the Pilot Data

Classical Test Theory (CTT) is not very useful for the analysis of the pilot data because of the incompleteness of the design. In this context, it is very convenient to use the Item Response Theory (IRT) method. Here, we give an overview of the essential features of the theory and the analysis procedure, and we will point to the results of the analysis which play a role in the standard setting. At the heart of the theory is the conception that the trait one wants to measure (an ability, an aptitude, an attitude) can be represented as an unbounded continuous variable[2], but that this variable cannot be observed directly, and therefore it is called a latent variable or latent trait. In this chapter, we will refer to this latent variable as 'proficiency'. A person's proficiency is assumed to have some value for this variable, and to measure means to find this value (exactly or approximately). Another, but equivalent way, of this basic conception is to say that the latent variable corresponds to a line, and individual persons can be represented as points on this line or continuum. Measuring then, is finding as accurately as possible the position of a person on the line.

To know something about the position of a person on the latent continuum, one can use items and the answers to them. In an IRT model, the relation between the latent variable and the answers to the items is described in a formalised way, mainly with the aid of a mathematical formula. This is explained using Fig. 11.2. The horizontal axis represents the latent variable. For each item there is a so-called item response function that expresses the probability of a correct response as a function of the latent variable. The two curves in the figure are the graphs of two such functions. There are a number of important features of these curves which are briefly commented[3]:

1. The curves are always increasing, also for values of the latent variable not shown in the figure. This implies that for no value of the latent variable the probability of a correct answer – a number in the interval [0, 1] – will be exactly zero or exactly one.
2. Although the two curves do cross, we notice that the *dashed curve* is more to the left than the *solid* one. In psychometrics, one concentrates on a specific value of the probability: 0.50, the value that says that a correct and an incorrect response have the same probability. By following the dashed lines in the figure, one sees that less proficiency is required (−0.75) for a fifty-fifty chance on a correct response in case of the dashed curve than for the same chance with the solid curve, which requires a value of 0.85. But this is the same as saying that the dashed curve represents an easier item than the solid curve. The amount of proficiency needed to have a probability of a correct response of exactly 0.50 is called *the difficulty value of the item*.

[2] This means that the variable can take any value from -∞ to +∞.

[3] The three-parameter logistic model is left out of consideration, because it has never been used in any analysis of the data for the Norwegian tests.

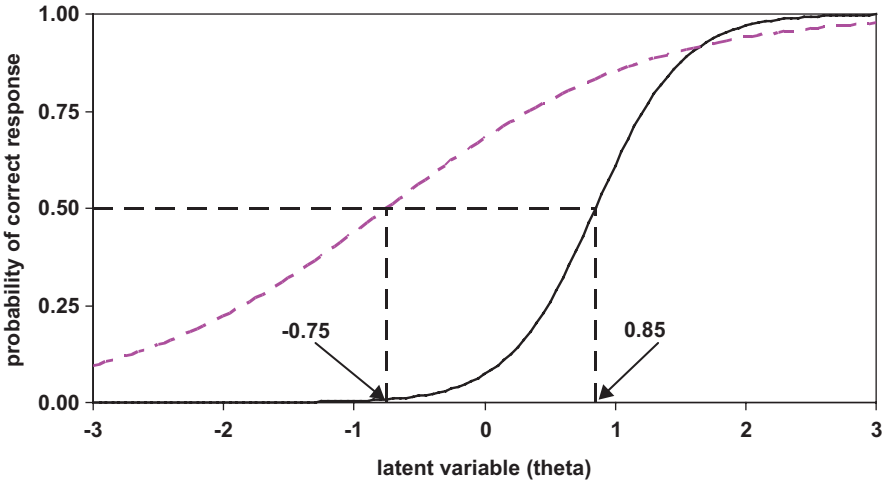


Fig. 11.2 Two item response functions

3. The dashed curve is 'flatter' than the solid curve,[4] meaning that (in the neighbourhood of the difficulty value) the item with the flatter curve discriminates less well than the item with the steeper curve, because in this neighbourhood the flatter curve changes values at a lower rate than the steeper one. The following is an example: consider two values of the latent variable, −0.80 and −0.70, both close to the difficulty value of the item represented by the dashed curve. For these two values the probability of a correct response is 0.4875 and 0.5125 and their difference is 0.5125–0.4875 = 0.025, i.e., 2.5% on the probability scale. If we do the same for the other item, choosing the values at the same distance from the difficulty value of 0.85, i.e., 0.80 and 0.90, the probabilities of a correct response are 0.4625 and 0.5375, respectively, giving a difference of 0.075 or 7.5% on the probability scale. The discrimination or discriminatory power of an item is expressed by a positive number.

In IRT models, one tries to describe all these features with a single formula. Such a formula, which has been used successfully in many applications is this one:

$$f_i(\theta) = P(X_i = 1|\theta) = \frac{\exp\left[a_i(\theta - \beta_i)\right]}{1 + \exp\left[a_i(\theta - \beta_i)\right]} \tag{11.1}$$

Here are some explanations to understand the formula:

1. For each item there is a rule that says how the performance on the item is dependent on the latent variable, which is symbolised by the Greek letter theta (θ). The symbol for this function is f_i where the subscript i refers to item i.

[4] Technically, 'flatter' means that at the point where the curve corresponds to a probability of 0.50, the first derivative of the function is smaller.

2. The meaning of this function is expressed by the middle term of (1): it is the probability that the score on item *i* equals 1, i.e., that item *i* is correctly responded to, conditional on the value of the variable θ.
3. The right side of Eq. (11.1) explains what this function looks like. The expression exp.[] in the fraction denotes the exponential function, and exp.(x) means the same thing as e^x, where e stands for the mathematical constant 2.71828…. In the fraction of (1) there are two symbols that carry the subscript i: a_i and β_i. They are place holders for a number, which is now left unspecified. These place holders are called *parameters* and they are to be read as follows: each item *i* in the test has some difficulty value which we represented as β_i. Likewise each item *i* in the test has some discrimination value (a positive number), which we write here symbolically as a_i. In Fig. 11.2, we have chosen two examples: for the dashed curve the difficulty parameter has the value −0.75, and the discrimination parameter has the value 1; for the item with the solid line, the difficulty parameter is 0.85 and the discrimination is 3.

In real life applications such as the pilot data for *Norskprøven*, the psychometric analysis has to accomplish two essential tasks: 1) from the data that were collected, the value of the a- and β-parameters have to be estimated and 2) a well-funded judgement has to be made in order to answer the question as to whether the relation between latent variable and response probability can be described reasonably accurately by a function that takes the form of the right side of Eq. (11.1). To answer the first question – parameter estimation – highly technical considerations come into play, which cannot be discussed in this article. Let it suffice to say that there exists good and efficient software to estimate the parameters. To fulfill the second task, one needs a trustworthy tool that in most cases consists of a statistical test. In the present context it suffices to say that on statistical grounds, nine listening items were eliminated from the original set of 192 items and none of the reading items.

11.7.3 The Construction of the Tests

From the set of items which remained after the estimation procedure, six tests were constructed: an easy and a hard routing test (R1 and R2), two tests at level A1/A2 (referred to henceforth as easy main tests) and two at level A2/B1 (hard tests). Full advantage was taken from the fact that the items had been constructed by an experienced team who were very familiar with the CEFR. When the team developed the tests, the selection of the items was guided by following principles:

1. For the tests R1 and the two easy main tests, most items had to be constructed for the A1 and A2 level; for R2 and the two hard tests most items had to be at the A2 and B1 level.
2. The main tests have to contain items which cover all aspects of the testing construct, and which were represented quite well by using different item formats. See Table 11.3.

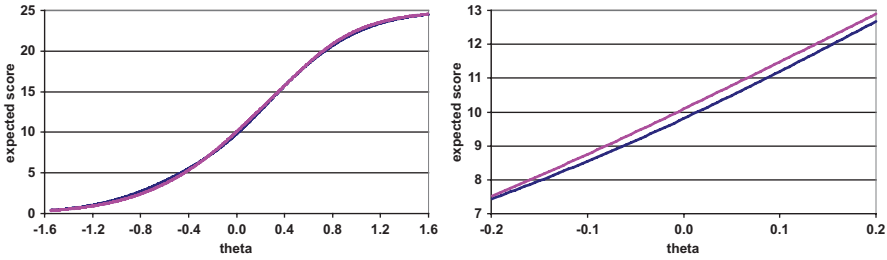


Fig. 11.3 Item response functions for the two A2/B1 reading tests

3. The two main tests at the same level should be as parallel as possible. To reach this aim, each time an item was allocated to a test, a similar item (same format, approximately same discrimination and same difficulty) was added to the parallel test. In a few cases items were allocated to both tests.
4. Items with high discrimination are preferred to items with low discrimination, as high discrimination contributes more to the accuracy of the measurement.

An important tool in the construction of the tests, especially in selecting items for parallel tests is the test response function**.** This function expresses the relation between the value of the latent variable and the expected score on the test.[5] In Fig. 11.3, the two test response functions for the two hard tests of reading are displayed as an example. The right-hand panel is a detail of the left-hand panel for θ-values in the range $[-0.2, 0.2]$. The maximum difference (in vertical direction) between the two curves is about 0.3 score points. Given the restrictions on the test construction and given the moderate number of items to choose from, this is as close to perfect parallelism we could come. Test response functions play an important role in the method of standard setting that was used, as will be discussed in the next section.

11.8 Method

The method used for standard setting is a generalisation of item-mapping strategies (Cizek and Bunch 2007, p. 155) which leans heavily on a graphical representation of item and task difficulty. The basic idea behind the method is two-fold, and is best understood if one compares it to the classical (modified) Angoff method. In the latter method, all panel members have to give an estimate of the probability of a correct answer from a minimally competent person (or borderline person). Such a task is complex as it implies a multifaceted judgment from the panel members: they have to have a (stable) idea of the minimally competent person, they have to have a good

[5]The test response function expresses the regression of the test score on the latent variable. Technically, it is the sum of the item response functions of the items in the tests. Because the graphs of the latter are not straight lines, their sum cannot be a straight line either.

idea of the relation between the intended competence and the specific requirements of the items, and they have to have (maybe implicitly) an idea about the difficulty of the items. And to a lesser (although not unimportant) extent they have to have an idea about the consequences of their decisions.

All these factors will have an influence on the stability (the standard error) of the standard. It may be helpful, therefore, to provide the panel members with information one already has. A good example is the difficulty of the items. If one has statistically reliable information on the difficulty of the items, it is a clear omission if one would not convey this information to the panel members. The crucial question, however, is how to do this. Another aspect of the Angoff method, which might lead to criticism, is the difficulty of the task of standard setting in combination with the number of items in the test: the panel members are asked to estimate the probability of success on each and every test item, which in a long test may become quite boring and difficult.

In the bookmark method (Mitzel et al. 2001; Cizek and Bunch 2007), only applicable if IRT estimates of the difficulty of the items are available, the difficulty estimate of the items are provided directly as numbers. In the Cito[6] variation on the bookmark method (Council of Europe 2009; Van der Schoot 2009), this information is conveyed graphically. In all these methods explicit reference to a latent competence is present. However, one can circumvent the use of the latent variable in the following way:

In every unidimensional IRT-model one can construct the response function for any subset of items which have been calibrated jointly. A special case arises if one of the subsets is the whole test, and the other one a subtest consisting of 4–6 items, as is shown graphically in Fig. 11.4. The dashed curve is the response function for the hard routing test, R2. Its expected scores have to be read from the left vertical axis. The solid line is the response function of a test consisting of R2 and one of the hard main tests, indicated here by H1. Its expected score is indicated on the right-hand vertical axis. The combination of R2 and H1 is the test we used to set the standard for A2/B1. In the figure it is indicated that an expected score of 3 on R2 corresponds to an expected score on the whole test of 13.71, because both are the expected scores for the same θ-value (0.066).

One can do the same thing for any score on the subtest, and for any subtest of the whole test. The correspondences were used to construct the response form for the panel members, as is shown in Fig. 11.5. The horizontal axis represents the score on the whole test, and each one of the five horizontal lines represents a subtest; the five subtests jointly represent the whole test. The placement of the numbers, the scores on the subtest, represents the exact relation between expected score on the test and on the subtest. The first (topmost) subtest is R2; the bullet with the '3' above it is right above the value 13.71 on the horizontal axis, showing the same correspondence as in Fig. 11.4.

[6] CITO is the name of the Dutch Institute for Educational Measurement. Originally, CITO was an acronym for 'Central Institute for Test Development' ('Development' in Dutch is 'Ontwikkeling').

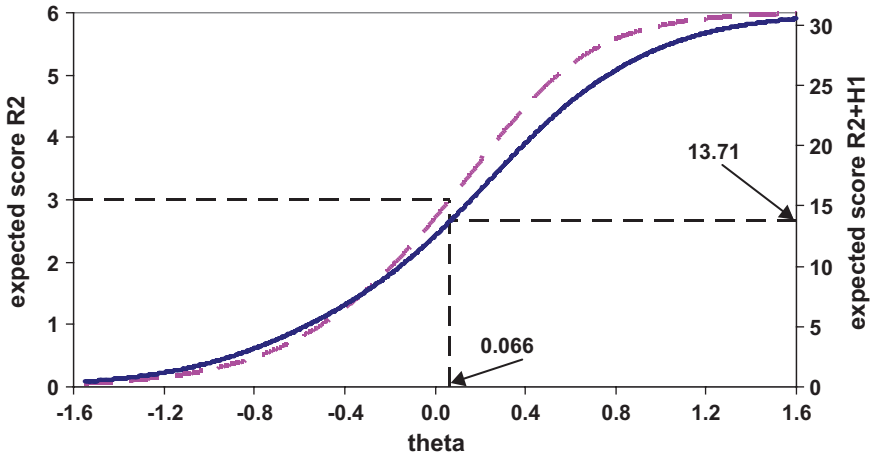


Fig. 11.4 Correspondence between expected scores on the test and on a subtest

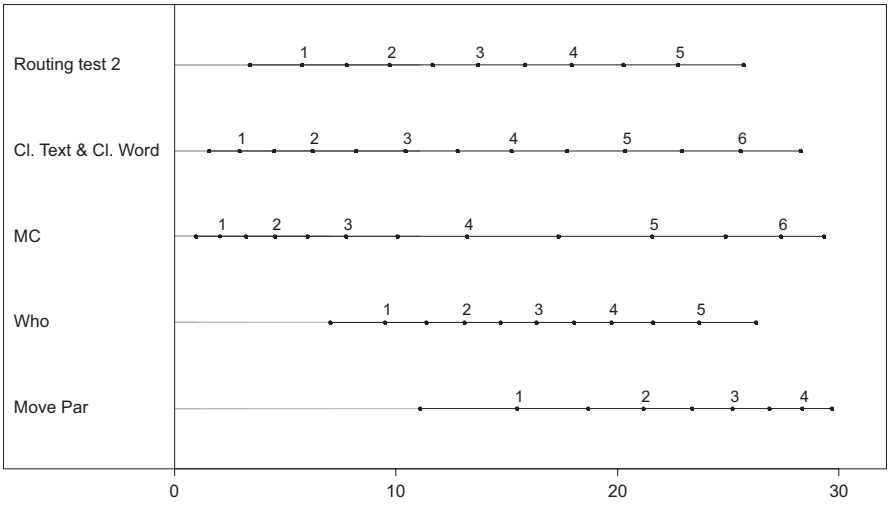


Fig. 11.5 The answer sheet for the panel members

The question asked of the panel members is to indicate on the sheet the expected score of the minimally competent person for each of the five subtests; scores are given up to ¼ of a point. The sum of the indicated expected values across subtests is the individual standard set by a panel member. The definition of the subtests in this method is in principle free. We used the routing test as one subtest; the other subtests consisted of items with the same format, see Table 11.3.

An important advantage of using a well-funded IRT model for the pilot data is that an arbitrary subset of the items can be used to do the standard setting and this standard can be transferred to any other subset that was calibrated to the same scale.

Figure 11.4 explains this principle of transfer: taking a standard defined on one subset, we find via the test response **X** function of this subset the associated latent value (θ), and using the response function of another subset, we find the expected score corresponding to this value of θ. This expected score is the standard for the new subset.

One might wonder whether the standards set using pilot data are valid in the context of a high stakes test administration. There are two aspects associated with this problem, a theoretical one and a practical one. Both aspects are discussed briefly. Theoretically one can say that the results of the analysis on the pilot can be used in a high stakes administration if the measurement model that was valid in the pilot is still valid in the high stakes circumstances. This means roughly that (1) the difference between the difficulty parameters of any two items remains the same and (2) that the ratio between the discrimination parameters of any two items remains the same. Finding out if this is the case, is an important aspect in the development of a well-founded testing situation. It is part of the self-monitoring system of the agency that develops and administers the tests. The possibility that the average performance in the pilot is different (mostly lower) than in the high stakes situation is irrelevant, as the distribution of the latent variable is not a part of the measurement model.

The practical aspect is this: after the pilot, standards have to be set, as taking a decision on the level is part of the high stakes test administration. So, one has to assume - be it provisionally - that the conditions described in the previous paragraph are fulfilled. There is certainly a risk involved in doing this, but the important feature is that the standard setting results can be revised and adapted if necessary.

11.9 Standard Setting

11.9.1 Selecting the Judges

Nineteen judges were selected to take part in the standard setting process. All judges were well acquainted with the CEFR. Thirteen of them had a background in language teaching, even though not all of them were practicing teachers at the time of standard setting. The group of judges included language teachers, language test developers, curriculum experts and/or researchers. Some details are given in Table 11.4.

Even though all the judges were well acquainted with the CEFR, they had to complete a few tasks prior to the actual standard setting event in order to be prepared for the job. The purpose of the tasks was to make them focus specifically on what characterises the descriptions of the levels A1, A2 and B1, and the difference between the three levels. Two weeks before the event, they received a list of CEFR descriptors covering the levels A1 to B1, 15 listening descriptors and 18 reading descriptors. Table 11.5 gives examples of some of the listening and reading descriptors that were sent to the judges.

No level is indicated for the different descriptors, and the sequence is random. The task of the judges was to assign each descriptor to one of the CEFR levels A1, A2 or B1. One week before the standard setting session, they got feedback on their

Table 11.4 Judges' current position

Current occupation	Judges																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Language teacher													1	1	1	1	1	1	1
Researcher	1			1										1					1
Curriculum expert									1										
Language test dev.	1	1	1	1	1	1	1	1		1	1	1	1	1			1	1	1

Table 11.5 Example of pre-standard setting task

No	**Listening descriptors**	**CEFR level**
1	Can catch the main point in simple messages, provided speech is slowly and clearly articulated.	
2	Can generally identify the topic of discussion around him/her, when it is conducted slowly and clearly.	
3	Can understand main points of radio and TV programmes on relevant issues or topics of personal or academic interest.	
No	**Reading descriptors**	
1	Can understand important key points in short public letters and documents.	
2	Can infer the meaning of unfamiliar words from the context in short texts on everyday topics, based on an understanding of the overall content.	
3	Can get an idea of the content of simpler informational materials and short descriptions, especially if there is visual support.	

first descriptor assignments. At the same time, they were asked to assign another list of listening (15) and reading descriptors (18) to CEFR levels. In both rounds one third of the descriptors were original CEFR descriptors (Council of Europe 2001), one third were DIALANG[7] descriptors (in Appendix C in Council of Europe 2001) and one third was from *The National curriculum in Norwegian and Social Studies for Adult Immigrants* (Vox 2012).

During the week before the standard setting event, all judges completed routing tests 1 and 2 as well as the A1/A2-tests and the A2/B1-tests of listening and reading. Two days before the judges received feedback on their second descriptor assignments including information about the standard setting event with special focus on the borderline student The two-day standard setting event started with an introduction of approximately two hours which aimed to 1) make clear the link between the tests and the CEFR, and 2) introduce the judges to the standard setting task and procedures.

[7] DIALANG is a language diagnosis system developed by many European higher education institutions. It reports your level of skill against the Common European Framework (CEFR) for language learning.

11.9.2 Setting the Standards

To construct the answer sheets, see Fig. 11.5, we had to define which tests we would use, as at the moment of the standard setting we only had pilot data collected in an incomplete design with 192 items, both for reading and listening. The tests that were constructed were meant to be applied in a real testing period which would begin after the standard setting event, as the results of the standard setting were needed as the concluding part of the test administration.

The tests that were defined were:

- The routing test R1 and the first A1/A2 main test for the standard setting at the A1/A2 level.[8]
- The routing test R2 and the first A2/B1 main test for the standard setting at the A2/B1 level
- For the cross validation round, two (partially) new tests were defined: the routing test R1 together with the other A1/A2 main test and the routing test R2 together with the other A2/B1 test.

This means that four different tests for reading, and four different tests for listening, were used, and that a total of eight standard settings had to be done in two days. The two main tests at each level and for both skills were pairwise parallel: they had the same distribution of item formats and their test response functions (based on the pilot data) were barely distinguishable. For the standard setting two rounds were planned.[9] During each round the panel members had a booklet with the items, arranged per subtest as indicated on their answer sheet. It was stressed throughout that confidentiality was safeguarded, and that all panel members had to give their judgment independently of each other.

To grant confidentiality, the following procedure was maintained throughout: the panel members had to write their ID (a number given to them before the session started) and after they finished filling out the answer sheet (circling a bullet on the line representing each subtest, or putting a cross between bullets to indicate an expected score ending with .25 or .75; see Fig. 11.5), they handed in their answer sheet. Their chosen scores were copied onto an Excel spreadsheet and automatically added to give their individual standard. This was also written on their answer sheet and the sheet was returned. On the Excel sheet a graph showing the results for the whole panel was automatically updated and shown after all panel members had

[8] To avoid misunderstandings, the intended standards are indicated with a pair of levels and the standards as set with a pair of numbers. If the intended standard is A1/A2 and the standard is set as 16/17, this means that 17 is the lowest score to be considered as at level A2 and 16 is the highest score indicating that the level A2 has not been reached.

[9] The sessions started with reading, and an extra round, called round zero, was organised, just to check whether the procedure had been understood.

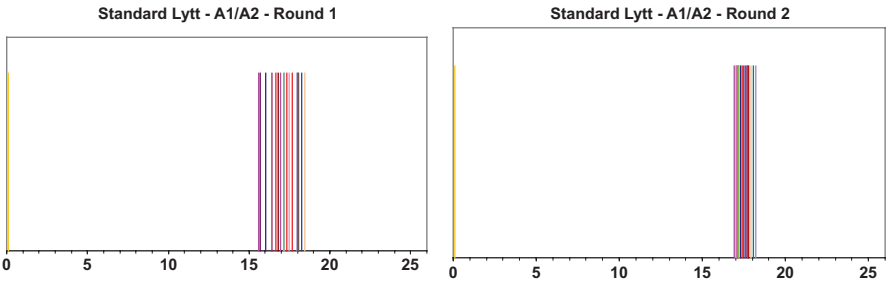


Fig. 11.6 Graphical feedback given to the panel

finished their work.[10] Two examples are shown in Fig. 11.6 of the listening (*lytt* in Norwegian) test for the rounds 1 (left) and 2 (right). Together with this graphical display the average, standard deviation and standard error of the mean were given. As the panel members had their own answer sheet with their individual standard, they could easily locate themselves in the distribution of individual standards without revealing their identity to the others.

After the first round, small groups were formed to discuss the results. The intention of this discussion was to eliminate possible misunderstandings about the CEFR or about the meaning of the lines on the answer sheet. It was discovered by one of the panel members that in order to be fully consistent with the empirical difficulty, all chosen expected scores for the subtests had to be located on a single vertical line on the answer sheet. This principle was discussed and explained if it was not understood, but it was not imposed as a requirement. Panel members were observed during their work, and nobody took the reverse way, by defining a vertical line and then placing their marks near to that line.

After the second round, some impact information was given, although with clearly expressed reservation. From the pilot data the distribution of the competences in the three streams had been estimated, assuming the candidates in the pilot were representative for the population. This is highly questionable because (1) the population taking these test is volatile (think of the various fluctuations in the number of refugees from different parts of the world), but (2) also because possible candidates not involved in formal teaching of Norwegian were not represented in the pilot sample. Using these estimates one can easily compute the proportion of candidates in each of the three streams belonging to the levels A1, A2 or B1 in a two-step procedure, which was illustrated graphically after round two. In step one the test response curve is used to find the competence of the minimally competent

[10] To avoid that two or more lines would hide each other, the individual standards were slightly adjusted so as to make all lines visible. The colours (online version) or different grey saturations (paper version) of the lines have no meaning: they simply follow the colour assignment of EXCEL.

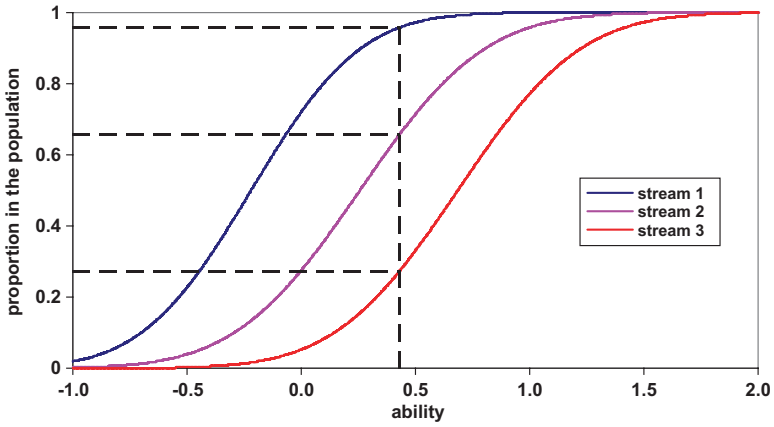


Fig. 11.7 Cumulative distribution of reading proficiency in stream 1 (left curve), 2 and 3 (right curve)

person, and in step two, the cumulative distribution functions (in each of the three streams) were shown so that the proportions at each level could readily be read from the vertical axis of the graph. An example[11] is given in Fig. 11.7.

The three curves represent the (estimated) cumulative distribution of the proficiency in reading Norwegian for the streams 1, 2 and 3 respectively The proficiency corresponding to the A2/B1 standard is 0.43 (the place where the vertical dashed line touches the horizontal axis). From this place we see, following the dashed lines, that in stream two over 60% (the exact percentage[12] is 65.7%) of the stream-2 population does not reach the A2/B1 standard. The panel were then asked if, given these results, they felt the need to reconsider the standards they had set. A third round was not deemed necessary in any of the four cases (two levels, two skills). It was agreed therefore that the standard proposed by the panel would be the average judgment of round two.

After the second round and a plenary discussion, the panel members were given the test booklet for the cross validation and the accompanying answer sheet. They were told that the first test (the routing test) was the same as before, but that the other subtests were different from the ones they had used during the standard setting. They were required to set their individual standard as before, having no discussion with the other panel members and working strictly alone. After finishing this procedure, the results were presented as in the first two rounds, and an overview of

[11] The way to find the distribution of the latent variable is quite complicated and not discussed in detail in this chapter. Technical details can be found in Verhelst and Verstralen (2005).

[12] The exact percentages were displayed jointly with the graphical display. For the streams 1, 2 and 3, they were 95.9%, 65.7% and 27.3%, respectively.

the results was presented. After the last standard setting on the second day, a short questionnaire was administered with questions about the clarity of the explanations and the method, and about agreement with the results.

11.10 Main Results

In Table 11.6, the main results for listening are summarised. The results for reading are displayed in Table 11.7. In general, the average standard is a fractional number, while the operational standard is an integer valued score. The pairs of numbers in the last rows of Tables 11.6 and 11.7 appear as rounding down and rounding up of the average standard, and they are remarkably stable across all rounds, including the cross validations. The differences for the A1/A2 standards in reading are caused by the fact that the average in the first and second round are very close to the unit. For the A2/B1 standard in reading there is also a (genuine) difference of one score point in the operational standards.

Table 11.6 Main results of the standard setting: listening

	A1/A2			A2/B1		
	rnd 1	rnd 2	CV[a]	rnd 1	rnd 2	CV
#items[b]	26	26	26	29	29	29
n[c]	17	17	17	19	19	19
Average[d]	17.18	17.49	17.51	19.00	19.38	19.83
SD	0.86	0.37	0.87	1.34	0.64	0.68
SE[e]	0.21	0.09	0.21	0.31	0.15	0.16
Standard	**17/18**	**17/18**	**17/18**	**19/20**	**19/20**	**19/20**

[a]CV = cross validation
[b]#items: number of items in the test
[c]n = number of panel members
[d]average of individual standards
[e]SE = standard error (= SD/$\sqrt{n}$)

Table 11.7 Main results of the standard setting: reading

	A1/A2			A2/B1		
	rnd 1	rnd 2	CV	rnd 1	rnd 2	CV
#items	26	26	26	31	31	31
n	17	17	17	19	19	19
average	17.09	16.97	16.59	20.50	20.80	21.13
SD	0.66	0.64	1.44	1.40	0.64	0.94
SE	0.16	0.16	0.35	0.32	0.15	0.22
standard	**17/18**	**16/17**	**16/17**	**20/21**	**20/21**	**21/22**

11.11 Standard Setting and Equating

A standard setting is an expensive event as it requires – besides training and preparation – two full days of work from a panel of almost 20 people plus the staff. Later on, two extra days were needed for setting the standards at the level B1/B2. As the test for Norwegian is administered twice a year, it is impossible to organise two, or even one, event like this per year. So the solution is to use the strength of the IRT-measurement model to transfer the standards set to new tests while maintaining the minimal competence at the same value. A minimal requirement to do this is that the items used to set a standard and new items are calibrated jointly. The standard is then transferred by IRT-equating using the following steps:

- Using the established standard (in the score domain) of the old test and the test response function of the old test we find the minimal competence.
- Using this minimal competence and the test response function of the new test, we immediately find the (equated) standard of the new test.

This is all very simple and straightforward if linear tests are used, but the real tests used are multistage tests, and the question is whether the conditional test response curve should be used, given the design as applied, or the usual 'unconditional' one, ignoring the multistage character of the test. Here is an example to show the difference between these two test response functions. Suppose a candidate has a score of less than four points on the first routing test. In that case, the test administration system will administer him/her an easy test at the A1/A2 level, comprising 20 items. So this candidate has answered 26 items and the unconditional test response function (the usual one) will express the expected score on this test as a function of the underlying latent variable (θ). The range of this function[13] is (0, 26). But if we take the dynamic design into account, we know that a candidate having taken routing test 1 and then an easy test of 20 items, has made at least three errors in the routing test, so that his total score cannot exceed 23. This means that the range of the conditional test response function[14] is (0, 23). This situation is illustrated in Fig. 11.8. Suppose the minimal competence is known to be 0.07, then using either the unconditional or the conditional response function will lead to different operational standards; see the two numbers along the vertical axis.

Until now we have not found a convincing rational argument to choose one over the other. There were two arguments, however, which made us change from the conditional function (that we chose initially as the 'necessary natural choice') to the unconditional one. The first argument is an empirical one: using the conditional response function to set the standard leads to a very high number of candidates classified as B1, even among those who did not take a A2/B1 main test, which made us doubt our initial 'natural choice'. The second argument is an argument *ex absurdo*.

[13] The range of a function is the set of values which can appear as a function value.

[14] The computation of the conditional test response function is more complicated than for the unconditional one. Technical details how to do this are not discussed in this chapter.

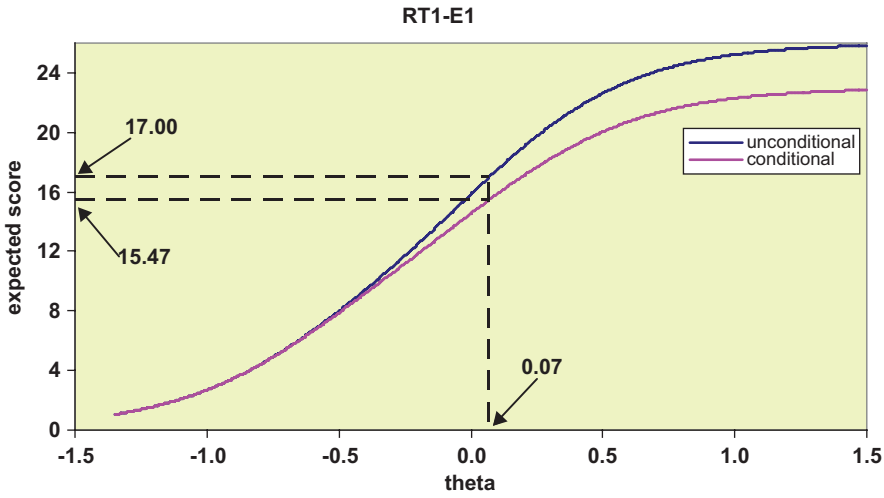


Fig. 11.8 Conditional (lower curve) and unconditional (upper curve) test response functions

Suppose there are two candidates, A and B. Candidate A takes a linear test consisting of the 26 items from the first routing test and an easy A1/A2 test. Candidate B is subject to the dynamical design as described above and happens to answer exactly the same items as candidate *A*. *Now* suppose that both candidates obtain a score of 16, then - see Fig. 11.5 – we should conclude that candidate A did not reach the requirement for the A2 level, while candidate B is clearly at the A2 level, (when the conditional curve is used).

This may appear as unfair (and maybe it is), as in the eyes of the candidates they just took the same linear sequence of items (and therefore we should treat them in the same way). But fairness – as an ethical category – does not necessarily lead to the best decision (where 'best' is specified in some way, but it certainly belongs in a rational category) as is clearly shown in the case where (empirical) Bayesian estimators are used to estimate one's latent ability. If there are two distinct populations (in terms of competence distribution), then a Bayesian estimator given a response pattern $\boldsymbol{x}$ will arrive at a higher estimate for someone of the 'higher' population than for a candidate of the 'lower' population.[15] So a really convincing rational argument in favour of one of the two response functions is still lacking. Our hope to provoke an answer from the audience at the Standard Setting conference in Oslo was in vain. Meanwhile, we will have to find comfort in empirical findings.

[15] And the posterior standard deviation will be smaller when the difference between the two populations is taken into account in the prior than when it is ignored.

11.12 Concluding Remarks

This study set out to link the listening and reading tests of *Norskprøven* to the *Common European Framework of Reference* by setting cut scores for the levels A1/A2 and A2/B1. The main aim of this chapter has been to document the quality of the procedures and to discuss how to take the multistage test design into account when deciding on cut scores.

When selecting judges for the standard setting event we were careful to choose judges who were familiar with the CEFR levels, since this was what the tests were going to be measured against. To prepare the judges and to help them focus their attention on the descriptions of the levels, they had two tasks in advance whereby they assigned a total of 66 listening and reading descriptors to CEFR levels. In addition, the standard setting seminar started with an introduction to (1) what characterised items mirroring A1, A2 and B1 competence as well as to (2) the task they were about to do. The relatively low standard deviation and standard error indices show that the judges were quite in line when agreeing on the cut scores, and also that the cut scores set by the group of judges are reliable. We have also discussed whether and how to take the multistage design of the tests into account when determining cut scores.

We did not use theoretical or rational arguments to base our choice of a conditional or an unconditional test response on, we chose to use the unconditional test response due to empirical and ethical considerations. If we had chosen the conditional test response curve, a rather large fraction of the students taking the A1/A2 test would have test results showing they had a B1 competence.

Test results and cut scores have to be monitored carefully during the next test administrations to see whether they function as intended. One reason for this is our decision to base the cut scores on the unconditional response curve. Another reason is that the standard setting was based on piloting data. Often items 'become easier' in real tests than they were in piloting. The reason for this is, of course, that real test results are more important to test takers than piloting results. To ensure the quality of the CEFR standards decided on for *Norskprøven*, we will recommend a new round of standard setting in the not-too-distant future.

References

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting*. Thousand Oaks: Sage.

Council of Europe. (2009). *Relating language examinations to the common Europeanframework of reference for Languages: learning, teaching, assessment (CEFR): A manual*. Strasbourg: Council of Europe.

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

Hasselgreen, A. (2003). *Bergen 'Can Do' project*. European Centre for Modern Languages. Strasbourg: Council of Europe Publishing.

Hasselgreen, A., Kaledaité, V., Maledonado-Maratin, N., & Pizorn, K. (2011). *Assessment of young learner literacy linked to the Common European Framework of Reference for Languages*. Strasbourg: European Centre for Modern Languages.

Lancaster University. (2015). *Information about DIALANG*. http://www.lancaster.ac.uk/research-enterprise/dialang/about. Retrieved June 2016.

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: Lawrence Erlbaum.

Moe, E. (2013, May). *Listening test items – For N2 students with no, little or some formal education*. Presentation held at the 10th annual EALTA conference, Istanbul, Turkey. http://www.ealta.eu.org/conference/2013/programme.html. Retrieved June 2016.

Norwegian Ministry of Labour and Social Inclusion. (2009). *Høringsnotat om endringer introduksjonsloven og statsborgerloven med forskriften og utlendigsforskriften*. http://www.regjeringen.no/no/dokumenter/horing—endringer-i-introduksjonsloven /id568147//. Retrieved April 2016.

Statistics Norway. (2014). *Norskopplæring for voksne innvandrere, 2013*. http://www.ssb.no/utdanning/statistikker/nopplinnv/aar/2014-09-24. Retrieved March 2016.

Statististics Norway. (2016). *Norskopplæring for voksne innvandrere, 2014*. http://www.ssb.no/statistikker/nopplinnv/aar/2014-09-24. Retrieved March 2016.

Van der Schoot, F. (2009). Cito variation of the bookmark method. In: Council of Europe, *Reference supplement to the manual for relating language examinations to the common European framework of reference for languages: Learning, teaching, assessment. (Section I)*. Strasbourg: Council of Europe.

Verhelst, N. D., & Verstralen, H. H. F. M. (2005). *Structural analysis of a univariate latent variable (SAUL): Theory and a computer program*. Arnhem: Cito.

Vox. (2012). *Læreplan i norsk og samfunnskunnskap for voksne innvandrere.* (The National curriculum in Norwegian and social studies for adult immigrants). http://www.vox.no/contentassets/.../laereplan_norsk_samfunnskunnskap_bm_web.pdf. Retrieved March 2016.

Vox. (2016a). *About Vox*. http://www.vox.no/English/About-vox. Retrieved March 2016.

Vox. (2016b). *Statistikkbanken*. http://status.vox.no/webview/index.jsp?catalog=http://status.dmz-vox.local:80/obj/fCatalog/Ctalog25&submode=catalog&language=no&mode=documentation&top=yes. Retrieved March 2016.

Vox. (2016c). *Eksempeloppgaver norsk*. http://enovate.no/voxdemo/norsk. Retrieved March 2016.

Chapter 12
Standard Setting in a Formative Assessment of Digital Responsibility Among Norwegian Eighth Graders

Ove Edvard Hatlevik and Ingrid Radtke

Abstract Students' digital responsibility is an important topic in a digital society. Since 2016, a learning supportive 50-item test in digital responsibility is available for Norwegian eighth-graders. Rasch model was used to select tasks to the test. Our study addresses experiences from using two standard setting methods, *Angoff* and *bookmark*, to set the cut-off scores and to define the proficiency levels of digital responsibility. In this process, as this is a learning supportive assessment, the decision makers had to take into consideration both the implied expectations from teachers who would use the assessment in their classroom practices and the results of psychometric analyses. A sample test with 1026 students from 26 schools was used to define three proficiency levels. The use of two standard setting methods, *Angoff* and *bookmark*, gave different results, and this difference addresses uncertainty about where to set the cut-off score. What are the validity arguments for considering different expectations when setting cut scores?

Keywords Digital responsibility • Digital competence • Standard setting • Validity • Learning supportive assessment • Formative assessment • Angoff • Bookmark • Cut off scores • Student-centered • Test-centered

O.E. Hatlevik (✉)
Department of Teacher Education and School Research, The University of Oslo, Oslo, Norway
e-mail: oveedv@icloud.com

I. Radtke
Skills Norway, Oslo, Norway
e-mail: ingrid.radtke@kompetansenorge.no

S. Blömeke, J.-E. Gustafsson (eds.), *Standard Setting in Education*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-50856-6_12

12.1 Introduction

After the so-called "Pisa shock" in 2001, Norway has adopted an Assessment for learning strategy (Black and William 1998) where formative assessment has become a priority on the national agenda. (Tveit 2014) While different assessment practices and forms have emerged, the teacher judgement in the classroom still plays a key role. At the same time, as a comparative OECD (2011) study also points out, a challenge remains in regard to the competence goals in the national curriculum, which are not specific enough to guide teaching and assessment performance.

During the last decade, besides national tests and mapping tests, *learning supportive tests* have been introduced in the Norwegian school system. A learning supportive assessment is a special type of formative test. The Norwegian Directorate for teaching and learning has introduced learning supportive assessments to help teachers assess basic skills in different subjects. The aim of these assessments is to inform schools and teachers about the mastery level the students have achieved, and which areas need to be improved. The results should be used to give feedback to the students and to customize training in the classroom.

In 2014, a process was initiated to develop learning supportive tests in ICT literacy for eighth graders. Since 2006, ICT literacy has been a basic skill in the Norwegian educational system. Digital responsibility can be identified as part of ICT literacy, together with search and process information, produce and communicate. Digital responsibility can be understood as being able to protecting your own and someone else's information in addition to "use digital tools, media and resources in a responsible manner, and being aware of rules for protecting privacy and ethical use of the Internet" (Norwegian Directorate for Teaching and Education 2012). The national curriculum contains explicit competence aims dealing with digital responsibility.

The main purpose of this test is to provide the teachers with valid information about the students' proficiencies in digital responsibility, and what can be seen as a side effect, to specify the competence goals. When trying to use the test results to make decisions about students or groups of students, a standard setting is required.

Setting performance standards and cutting scores is about making decisions and, equally important, it functions as a means to give an orientation to students, teachers and policy-makers about the standards that should be reached through teaching and learning. This applies, especially, to learning supportive assessment in digital competence, because the aim of a learning supportive assessment is mainly to improve classroom practices and performances.

Standard setting can be defined as "a translation of policy decisions […] through a process informed by expert judgment, stakeholder interests, and technical expertise" (Tiffin-Richards and Pant 2013). Mehrens and Cizek (2012) point out that setting performance standards has mostly to do with the need to make decisions. How are the decisions being made and what kinds of arguments are most trustworthy? Standard setting methods are grouped into *student-centered* and *test-centered* methods (Jaeger 1989), where a group of experts, after a rather long process,

determines the final cut score. Though little work has been done yet to monitor the content of the group discussions. (Deunk et al. 2014) By using two different methods for standard setting the uncertainty about how to reach a final decision about the cut scores increases when the two methods give different results.

Experts, stakeholders, teachers and psychometric analysts have different interests and perspectives. These differences can be of importance as to how they are analyzing und making their judgment about when students have reached a certain performance level. As different systems with different rationales are taking part in setting cut scores, arbitrariness is becoming more visible.

A possible context for explaining these differences comes from a theory about the differentiation of function in society. According to Luhmann (Afzar 2006), modern societies are differentiated into sub-systems that work autopoietically by producing connecting events. Following this we would expect that stakeholders, school leaders, teachers and experts bring different perspectives and arguments when participating in a standard setting process. We would also expect that transferring the results of a standard setting process back to the instructional part of the educational sector could be a challenge in terms of acceptance and usefulness.

Beside the procedural aspect of setting standards and concluding on cut scores, the aspect of how the results of the standard setting are perceived by stake-holders, especially teachers, in the case of learning supportive assessment, has received little attention (Pant et al. 2009).

The objective of this chapter is to describe how to deal with uncertainty when two standard setting methods were applied to a learning supportive assessment in digital literacy, and to discuss validity arguments for considering different expectations when setting cut scores.

12.2 Background

The background section contains different information introducing the concept of ICT literacy and standard setting used in this chapter. First, the concept of ICT literacy and, as a sub-category, digital responsibility is defined and targeted at the educational system. Second, standard setting is introduced presenting the judgmental process, the methods used and how to understand validity when scrutinizing standard setting. Finally, the study and the research questions are presented.

12.2.1 ICT Literacy

Since 2006 Norwegian students' ability to use ICT has been considered a fundamental literacy in the Norwegian curriculum, together with reading, writing, numeracy and oral skills (Norwegian Directorate for Education and Training 2012). There are both descriptions available of how to understand ICT literacy in each of the

subjects in the national curriculum, and in the competence aims at 2nd, 4th, 7th and 10th grade in primary and lower secondary school. One challenge with these competence aims is that they are rather vague and open to the teachers' interpretation (OECD 2011).

Recently, the European Commission has initiated a project to examine the concepts and frameworks used to describe students' use of digital technology in a school context (Ala-Mukta 2011; Ferrari 2013). There are both international studies (Binkley et al. 2012; Educational Testing Service 2001; Fraillon et al. 2014) and studies from countries like Australia (Ainley et al. 2007), Chile (Claro et al. 2012), Italy (Calvani et al. 2012), Korea (Kim et al. 2014) and Norway (Hatlevik et al. 2015), examining how capable students are of using ICT.

A literature review shows that several different concepts are used in order to describe students' ICT capabilities, for example ICT literacy (Fraillon et al. 2014), digital competence (Calvani et al. 2012; Krumsvik 2011), digital skills (Zhong 2011; Matzat and Sadowski 2012), and Internet skills (Kuhlmeier and Hemker 2007). There are some similarities among the definitions of ICT literacy, digital competence and Internet skills (Ferrari 2012). First, there is a description of the context where technology is used (i.e., digital, ICT, internet), and, second, there is a description of learning domains (i.e., skills, competence, literacy). When it comes to learning domains, it seems that literacy and competence are broader concepts compared with skills (Ferrari 2012). According to Søby (2013) skills are dealing with the more technical aspects; but literacy includes skills, knowledge, and attitudes (Ferrari 2013; Krumsvik 2011).

12.2.2 Digital Responsibility

As mentioned in the introduction, digital responsibility can be understood as part of the national curriculum. Digital responsibility means that students are capable of making judgements and being responsible about how they use digital technology (i.e., tools and resources) in their school activities. Digital responsibility, is a subcategory of ICT literacy, that overlaps with being able to interpret digital information critically and knowing about the copyright rules when it comes to published material (i.e., music, pictures, films, etc.).

Recent research shows that different concepts are used to describe students' capability of critical awareness and making responsible judgements online, for example, *safety* (Ferrari 2013), *personal and social responsibility (*Binkley et al.*)*, *using information safely and securely* (Fraillon et al. 2014), and *understanding human, cultural and societal issues* (International Society for Technology in Education 2007). The content of digital responsibility seems to be partly covered in some of the international frameworks of ICT literacy and digital competence. According to Binkley et al., the need to develop *personal and social responsibility* is among the 10 skills that are important for the twenty-first century. In their framework of two strands and seven aspects, Fraillon et al. (2014) emphasize *using*

information safely and securely as one of the aspects. Ferrari (2013) defines safety as one of five areas of digital competence. In her opinion safety can be understood as protecting devices, personal data, health and the environment.

12.3 Standard Setting as a Judgmental Process and Its Validity

Standard setting is about the procedures and judgments leading up to making decisions about how to use results of a given test. The decisions made through the standard setting process "is the conceptual version of the desired level of competence" (Kane 1994). According to Deunk et al. (2014), we assume that "the cut-off score corresponds to the performance standard […, and] that the performance standard is appropriate given its intended use" (p. 79). Overall, the standard setting process is about linking the content standards to the performance standards and the cut-off scores. Some authors have characterized the overall nature of standard setting as a procedure and a judgmental process. Hambleton (2001) points out that it is "mainly a judgmental process" and that for that reason the procedural validity evidence receives much focus. In this judgmental process, the judges, often - also named *experts* or *expert group* - play a vital role. Their role is to decide what level of competence that students show qualifies them what level of students' competence qualifies for a certain level of performance.

In order to aid the experts in the decision-making process, several methods were developed. Classifications of the methods exist, and differentiate between *test-centered* and *student-centered* methods (Jaeger 1989). The test-based methods focus on the content of the items and the linking to the overall content framework. It is therefore also a *criterion-referenced* method, since it relates to the standards that are set. Among these methods, we find, for example, the Angoff and bookmark method. The student-centered methods, conversely, are based on the actual data from the population that was tested, and refer to a norm that was set in relative terms (*norm-referenced measurement*). These types of standard setting method include the contrasting groups approaches and the borderline groups survey. What all methods have in common is that they try to combine both the results of psychometric analyses and human judging. Lately, human judging has been questioned, due to its arbitrariness because of human reasoning which is vulnerable to all kinds of aspect concerning both the procedure of standard setting and the psychology of group discussions (Pant et al. 2009).

Much emphasis has therefore been put on the validity of the whole process and on documenting the steps and decisions made in the process.

In order to use the test and trust the decisions about the cut-off scores, the process of standard setting has to be considered valid by experts, stakeholders, teachers and parents. The elements used in the standard setting evaluation derive from different areas (Deunk et al. 2014; Pant et al. 2009). There are procedural, internal and

external elements of standard setting (Deunk et al. 2014), in addition to consequential validity evidence (Pant et al. 2009). Most emphasis has been put on the internal and the procedural validity evidence. Here, one has to assure that the process of standard setting including selection and training is taken care of properly and that the process shows consistency in the judgement by the expert group.

It seems that there has been less focus on external validity and consequential validity. External validity builds on other sources of evidence, such as other standard setting methods or other external sources of information, such as different standard setting methods and similar tests and studies. As Deunk et al. (2014) also mentions, "external evidence is generally mainly used to check the validity of the performance standard: whether the standard is appropriate given its intended use." Consequential validity takes into consideration whether recommended cut scores are feasible and are in alignment with the performance standards (Pant et al. 2009). Therefore, using two standard setting methods is a means to enhance sources of external validity. In the literature, in studies where the two different methods are used, the aim is mainly to see whether different methods produce similar cut scores (Hsieh 2013). Other sources on external validity for standard setting for an assessment in digital responsibility would be other studies on digital competence and documented experience with digital competence in Norway. For consequential validity in a learning supportive assessment, the teachers' perception of the cut scores and consequences for the teaching process in the classroom are important sources.

There are several methods for setting cut-off scores in order to define proficiency levels. In this chapter, we are presenting the two test-centered methods, which we used, the Angoff and the bookmark method.

12.4 Application of the Angoff Method

There exist many variations of the Angoff method, tracing back to William Angoff (1971). We are starting with what is described as "the traditional Angoff standard setting procedure" (Plake and Cizek (2012). A group of experts is chosen to be members of an expert group, and this group is going through the test in order to judge the items. An important issue is therefore to decide what kind of qualifications to set when selecting people for the expert group. The judges have to estimate the difficulty of the test items, taking in mind 100 students on the borderline between two proficiency levels. For each item the judges have to estimate the probability in percentage whether the 100 borderline students can answer the item correctly. The cut-off points of the three levels are created by summing estimates over items and averaging them over judges.

This traditional Angoff method variation has often been replaced with the Yes/No Angoff method which expert groups often find less challenging; though a

disadvantage of this method is, as Cizek and Bunch (2007) point out, that it produces a bias due to the fact that the probability of correct response at the cut score is greater than 0.5.

The overall procedure with this method is that the expert group receives a booklet with all the items in the test, which are presented in order of appearance during the test administration. They have to go through all the items in detail and evaluate the difficulty of each item. If the members of the expert group are not fully familiar with the content of the items beforehand, this method enables them to read each item before they decide on the difficulty level.

However, the Angoff method has been criticized because it is time consuming for the expert group to go through all the items, and it can be a cognitive challenge for the members to judge the performance of 100 borderline students on each cut-off score in all items. The bookmark method was therefore developed on the background and the criticism on the Angoff method.

12.5 Application of the Bookmark Method

The bookmark method has several advantages. First, it seems to reduce the challenging task for the expert group since the Angoff method requires a judgement of the p value of each item. This is not any longer required and it also has practical advantages as it is easy to implement (Mitzel et al. 2001; Cizek and Bunch 2007; Cizek 2012a). Instead, since usually items have gone through pilots before the standard setting takes place, an overview of the difficulty of the items is already available, and if an IRT analysis has been conducted, one makes use of the same source of information for two purposes. The use of IRT analyses for the standard setting is the natural extension of the advanced data analyses of the pilots. Second, with the bookmark method it is possible to make use of this information and to hand out the so-called ordered item booklet to the expert group.

The bookmark method entails the preparation of ordered item booklets. Before the employment of the bookmark method, item response theory was applied to estimate the difficulty of the fifty items in the test of Digital Responsibility. The ordered item booklet includes the set of test items, which are ordered according increasing difficulty levels.

This characteristic of this method makes the task somewhat easier for expert groups as they can see how the items worked out during the piloting study.

In order to conduct the method, the judges receive an ordered booklet with items and have to decide about the cut-off scores. In detail, the task is to search and select the first item (the bookmark) that is on the next proficiency level.

12.6 Study Design and Research Questions

In 2013 The Norwegian Directorate for teaching and learning contracted the Norwegian Centre for ICT in education to develop a test in digital responsibility. The competence aims in the subject curriculum after 7th grade, and the more general descriptions of ICT literacy in the framework of digital skills were used to define the construct. The construct was operationalized into six themes: safety, protection, digital bullying, green data, copyright and the use of digital sources. Items were formulated on the basis of these six themes.

The items were piloted in three steps. In the first pilot, we started with approximately 70 items placed in two booklets. The data were analyzed and, based on the outcome, items were kept or removed. In addition, new items were developed. In the second pilot we tested approximately 120 items plus 20 anchoring items placed in five booklets. Based on analyses we selected 60 items to be part of the final sample test with the purpose of standard setting of mean and cut-off scores. The final sample test was carried out with 1026 students from 26 schools. The schools were randomly selected among all schools in Norway, but the selected schools could choose to participate. This could lead to some systematic bias, (i.e., due to the interest of the schools). After the final sample test, on basis of the items analyses, the number of items was reduced to 50 items. It was decided to distinguish among three different proficiency levels in the final test.

Rasch-model (meaning a "1PL model") was used to analyze items from all stages of the study. Rasch-model is the ground model of item-response-theory. It goes back to the Danish scientist Georg Rasch's work. The Rasch-model describes the probability that a person answers a specific item correctly is depending on the person's ability (Crocker and Algina 2008; Embretson and Reise 2009). In 2015, the Norwegian Centre for ICT in education invited ten experts to participate in a standard setting workshop. The workshop was arranged as a one-day event including the use of two standard setting methods. Given that the assessment is a formative low-stake assessment and the amount of resources available, a small group of experts was invited to judge each method in one round and to reflect on the process afterwards. As the Centre had little experience with the use of different standard setting methods, two representatives from Cito in The Netherlands were invited to help facilitating the workshop.

This chapter addresses experiences from using two standard setting methods (Angoff and bookmark) to set the final cut scores and to evaluate the proficiency levels. This chapter addresses the following two research questions:

1. How to deal with uncertainty about cut score when the standard setting methods (Angoff and bookmark) give different results?
2. What are the validity arguments for taking different expectations into consideration when setting cut scores in a learning supportive assessment?

12.7 Method

In this paragraph, the work with the standard setting and decision-making process is documented. The paragraph starts with the application of the two methods, and then describes the outcome of the methods and how the assessment group at the Centre for ICT in education dealt with making the final decision.

12.7.1 Participants and Procedures of Standard Setting

When we chose the judges for the workshop, we thought it is important that their background and qualifications should meet the requirements. These experts had either experience with teaching digital competence, or research in the field of digital competence or experience with the developing and piloting of items. Stakeholders as representatives of the Directorate of Education or the Department of Education were not included at this stage, albeit it was discussed.

The workshop started with a short presentation introducing the concept of standard setting. Then the three competence levels in the assessment were introduced. They defined the knowledge and skills that students must demonstrate in order to meet a certain level of competence. It was emphasized that the competence levels describe which competence elements the test would assess on each of the three levels and therefore the description was very close to the content of the test.

After that, the expert group members were introduced to the two methods that should be used during the workshop, the Angoff method and the bookmark method. Both methods include the concept of a borderline student. The borderline student is a student (or group of students) that has just enough digital competence to meet the defined performance standard. The expert has to take the borderline student in mind, when making a judgment about each item or the cut score between two performance standards in the assessment. A standard setting exercise with 10 items was conducted using the Angoff method before starting on the 50-item test in digital responsibility.

It was decided to work with standard-referenced methods and to use two test-centered procedures on the 50-item test in digital responsibility. These choices made most sense for a learning supportive assessment because a judgment on the basis of the content of each item was the most important aspect. For the procedure, the experts had to go thoroughly through each item, consider the difficulty of the content and format of items and decide on the basis of the performance levels whether a borderline student would be qualifying for a certain performance level. It was decided to use both the Angoff method and the bookmark method and in the first case, the expert group had to apply the Angoff method followed by the bookmark method.

Table 12.1 Distribution of score sums.

Distribution of score sums[a]	
Mean	25/50
Min.	0/50
Max.	44/50
p5	17/50
p25	20/50
p50	24/50
p75	30/50
p95	37/50

[a]p = per cent ([comment 20] e.g. P5=17/50 means that 5 per cent of students did get a score equal or below 17 scores out of 50 scores)

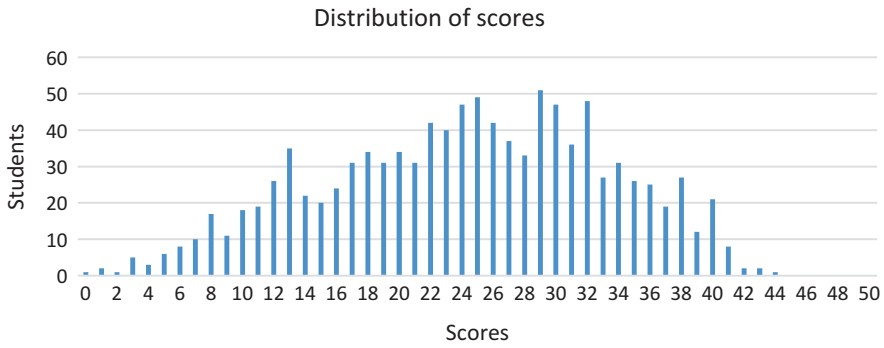


Fig. 12.1 Distribution of scores in relation to students.

12.8 Results

In this section the results of the analyses of the data concerning the distribution of scores based on the Rasch-model is presented together with results of the workshop and the decisions about setting cut-off scores.

12.8.1 Distribution of Scores

Table 12.1 shows that 25 scores are the mean and that 50 per cent of students managed between 20 and 30 points. The maximum score that the students reached was 44 scores out of 50 scores and there are only 5 per cent of the students who reached more than 37 scores.

Figure 12.1 shows the distribution of scores in relation to the students. The max score of the test was 50 points. In addition, we identified a rather strong group of

Table 12.2 Results of two standard setting methods.

	First item at level 2	First item at level 3
Angoff method	22	37
Bookmark method	17	39

students managing between 22 and 32 scores. Figure 12.1 shows that 44 per cent of the students are in this range (from 22 to 32 scores). Our analysis is based on the data from a pilot study. One could assume that the overall performance would increase from the pilot study to the final test.

12.8.2 Results of the Standard Setting

The expert group used both the Angoff method and the bookmark method. When evaluating the cut-off scores between level 1 and level 2 from the two methods, the expert group suggested a cut-off score after 16 right answers, using the bookmark method, and after 22 right answers, using the Angoff method. If one would go for the lower cutting score from the bookmark method, about 21.5 per cent of students would belong to the level 1 versus 36.6 per cent after the Angoff method.

The cut-off scores between level 2 and level 3 for the Angoff method and the bookmark method were more similar. After the Angoff method, level 3 starts on 37 points whereas with the bookmark method level 3 starts at 39 points (Table 12.2).

The results show that the two methods did **not** give exactly the same results. How to make decisions about cut-off scores when the methods provide different results? This is about the internal validity evidence because one could assume consistency between two methods of standard setting (Pant et al. 2009).

12.8.3 Making Decisions About Cut-Off Scores

The two methods gave different results. A few days after the workshop members of the assessment group at the Centre for ICT in education sat down in order to discuss and define the cut-off scores for the test on the basis of the results of the workshop and the overall test results.

The reflection on the various arguments for the cutting scores resulted in one cutting score defining the level 2 from 19 correct answers and level 3 from 40 correct answers. This result may be somewhat surprising because for the level 3 the final cut score is higher than what the expert group recommended. A crucial factor for this has been a careful examination of the content of the assessment and an assumption about the perception of the teachers in the classroom. Taking these

considerations into account, a strong test-based argument that was absolute in terms of the content of the assessment was applied.

Additionally, a level 0 was also recommended. That would apply when students managed less than ten items in the assessment. In that case, the amount of information the teachers would get is not sufficient to make any judgments. Within that range, one would assume that different causes might be responsible. It could be technical failure of the assessment, motivation or time problems when taking the assessment. Here rather a student-centered approach was applied, taking into consideration the problem with extreme low scores.

12.9 Discussion

This paragraph contains the discussion of two research questions. The first one deals with uncertainty when using standard setting methods. The second one discusses different validity arguments, when trying to reach a consensus for the final decision making about cut-scores.

12.9.1 Combining Test-Centered and Students-Centered Methods

The first research question addresses how to deal with uncertainty of the cut score when the standard setting methods (i.e., Angoff and bookmark) give different results.

The expert group reflected about the experience using the two methods. Not surprisingly, the exercise with the Angoff method proved to be more difficult as the expert group had to estimate the difficulty for each item. However, they reported that the instruction was perceived to be useful, as the expert group had been informed to use their initial estimate of the p-value for each item. The workshop ended without concluding which of the suggested results of the two standard setting methods should be applied. This decision was referred to the Centre for ICT in education. The assessment team at the Centre had to take in mind that this is a learning supportive assessment whose use is primarily to assist teachers to provide more accurate and customized teaching. In addition, it gives teachers an overview of the state of the art in digital competence and of the field and level of competence that is expected from students.

Student-centered methods came therefore explicitly to the fore, which is something that Cizek (2012a, b) regards as a natural part of any standard setting methods. Following his thought, any standard setting procedure requires to take into consideration information about both the test content and the test candidates. While during the workshop expert group was asked to concentrate on the content, for further

decision-making, relevant test-data and student-centered information were used to give further information for the decision-making process.

In addition, experiences from other surveys and research on digital competence among students in Norway and policy developments concerning digital competence in schools were taken into consideration. Finally, as this is a learning supportive assessment it is important that the teachers can use the results of the assessment in their classroom practice. It was therefore considered how the standards and the descriptions of the proficiency levels could facilitate the teachers' practice.

This led to four main strands that formed the decision-making:

1. The number of students that would be located at each of the levels.
2. The number of items qualifying for a certain level.
3. The concern was whether items located around the cut score would qualify for the lower or higher performance level (i.e., bookmark method).
4. Taking into consideration the alignment of different systems, that look at the teachers' role and the classroom practices versus test development practices.

12.9.2 Validation Arguments that Support Decision Making

The second research question addresses the validity arguments for taking different expectations into consideration when setting cut scores. This could be one way to deal with uncertainty about different cut score.

The argumentation for the final decision on the cutting points was mainly based on external and consequential validation (Pant et al. 2009). The aim was to integrate the different results of two standard setting methods into one judgment and final decision. In order to come up with a valid decision in this judgment, further evidence was mobilized and arguments for the four strands were discussed.

12.9.3 Argument 1: Distribution of the Students by Three Competence Levels

This argument starts by taking into consideration the number of students on the three competence levels. When it comes to the number of students located at level 1, it should not be too many, because then the assessment could be viewed as less valid in relation to the students' competence.

Therefore, argument 1 deals with the external validity evidence (Pant et al. 2009), because the teachers can compare the consequences of the cut-off scores with the information they have about their students (i.e., grades and teacher-made assignments or assessments). For example, research indicates a high correlation between grades in subjects and performance on ICT literacy assessments (Hatlevik et al. 2013).

When it comes to the distribution of students by the three competence levels, there is data available from a large-scale survey about ICT literacy among students in Norway. The ICILS 2013 study showed that the Norwegian students are performing as good as students from other countries (Fraillon et al. 2014). The students are also experienced ICT users and schools and homes in Norway are well equipped with digital devices (Mediatilsynet 2014).

The cut-off scores between level 1 and level 2 were different for the Angoff method and the bookmark method. As mentioned above, following the Angoff method would result in more students on level 1, compared with the bookmark method: 36.6 per cent versus 21.5 per cent, respectively. One argument against having almost 37 per cent on level 1, was that this could differ from what the teachers expected regarding students' ICT literacy, and what had been found in the ICILS 2013 study (Fraillon et al. 2014). One solution could be to have the cut-off score between the two suggested cut-off scores.

The cut-off scores between level 2 and level 3 for the Angoff method and the bookmark method were quite similar. The argument in the discussion about level 2 versus level 3 was that level 3 should be exclusive and identify students who have a high level of competence. As we have seen from the analyses of the data, only a few students are highly knowledgeable when it comes to digital security and ownership of digital material. At the same time, these content areas play a vital role as part of digital responsibility.

12.9.4 Argument 2: Number of Items that Qualify for a Certain Level

Argument 2 deals with the consequential validity evidence, for it is about to which extent the "cut-score recommendations are feasible or realistic" (Pant et al. 2009).

The cut-off scores define how many items a student has to get correct in order to qualify for a given proficiency level. On the one hand a proficiency level can be understood as an objective indicator of students' achievements, but on the other hand it can be described as a symbolic representation of achievements. According to Sadler (2009), cut-off scores and proficiency levels do not exist independently, but the cut-off scores are subjectively decided on what is considered as relevant factors. One important discussion is therefore how to get the cut-off scores in alignment with what the teachers would perceive as a realistic number of items qualifying for a certain proficiency level. Following arguments from Sadler (2009), the integrity of a grade or a proficiency level is about their authenticity.

The decisions about where to set the cut-off scores, and define the scope of a proficiency level, are about what criterions should be used. Proficiency levels can be defined by assigning each proficiency level to all aggregated scores that fall within a fixed range (Sadler 2009). This could mean that when a student has managed 5 or 10 points out of 50, he or she could be located at level 1. This could also mean that

when a student has answered about half of the items correctly, he or she could be located at level 2.

However, it is more difficult to make the decision about the exact cut-off score between level 1 and level 2 or between level 2 and level 3. One challenge is that our empirical data are based on a pilot study. From observations during pilot studies and other experience, we know that there are motivational aspects involved in pilot studies preventing the best performance of students. In a non-pilot study, we could therefore expect a better performance.

Taking the results of the data-analyses about the distribution of scores the discussion about realistic cut-off scores led to the conclusion that level 2 should start at a level less than what fifty per cent of students would be able to manage. It was decided that around 20 correct answers would qualify for level 2 while level 3 should start somewhere from 37 correct answers in order to maintain level 3 as an exclusive level. As shown in Table 12.1, at least 25 per cent of the students would qualify for level 1, while only a small proportion, 5 per cent, would qualify for level 3. However, we expect performance to increase from the pilot study to the stage where the test is in real use. We therefore assume that more than 5% of the students are able to qualify for level 3.

12.9.5 Argument 3: Difficulty of Items around the Cut Score

Argument 3, difficulty of items around the cut-off score, also deals with the consequential validity evidence (Pant et al. 2009). On the one hand, it is difficult to argue that a so-called *easy item* should be at level 2, but, on the other hand, it is hard to argue that a so-called *difficult item* should be at level 2. It is therefore necessary to examine the difficulty of items around the suggested cut-off score in order to ensure a content-based evaluation of the cut-off scores.

This argument deals with the content and difficulty of items. Concerning the content, we know from other sources that certain themes within the assessment such as ownership of digital material and safety issues are generally more difficult for students (Hatlevik et al. 2013). In addition to the content areas, the items were also tagged after Blooms taxonomy resulting in a four-folded matrix containing render, use, consider and apply in order to evaluate the difficulty of the assessment task (i.e., test item).

Based on the ordered booklet from the bookmark method and the result of the Angoff method, the suggested cutting scores from the two standard setting methods were evaluated.

The discussion about the cut-off point between the first and the second level resulted in the decision that item number 18 would still qualify for a level 1-item, whereas item number 19 would clearly qualify as an item on proficiency level 2. Previous experience shows that students have special difficulties with the area of safety issues and ownership of digital material. Also taking in mind Bloom's Taxonomy, a change in the performance level is visible between the items 17 and 18. Item 17 is a drag-and-drop item which asks students to state how they think

phishing can be avoided by completing sentences with the right words. In contrast, item 18 - an item from the area of personal security - asks students to match terminologies from that content area. This requires more advanced knowledge. And the following item, number 19, investigates whether students know which statements dealing with digital tracks are right or wrong. This is an item that requires students to take in information. Item 20 then asks students what is meant by "critical use of digital sources" by means of a multiple-choice item.

For the second cut score the expert group had suggested that level 3 should either start with 37 or 39 scores. Item 37 tests the knowledge about ownership of digital material by asking the student to identify the Creative Commons symbol as the right symbol for pictures that can be used freely. This requires knowledge about both the symbol and the concept of Creative Commons. Item 38 and 39 test the knowledge about personal security. Whereas item 38 assesses the application of knowledge about the registering of personal data on social networking sites, item 39 assesses the knowledge about consequences of registration of personal email on a website. Item 40 then assesses whether the students are able to apply knowledge about the critical use of digital resources, and which aspects are important to consider a website as credible. The item format is again a true/false question. Related to the content, item 40 marks a shift towards the testing of more advanced knowledge.

12.9.6 Argument 4: Alignment of Different Systems

This argument deals with both the procedural and the consequential validity evidence (Pant et al. 2009). As mentioned earlier, the standard setting process is about linking the content standards to the performance standards and the cut-off scores. During the process, different communicative systems meet to make judgmental decisions. These consist of the expert-group, the psychometric analysts and researchers and the local schoolteachers and educational staff.

Unfortunately, even though the procedural aspect of standard setting receives much attention, little effort has been made to look behind the scenes and monitor the content of the discussion in the expert group. Interestingly, as the study from Deunk et al. (2014) shows, student-centered arguments play a vital part in group discussions. Arguments that are put forward by the expert group in standard setting workshops, can be grouped around three content topics: the first one deals with the content of the items and skills needed, the second one with individual experiences of educational practice and the third one with the consequences for the scale when placing the cut-off scores. Unfortunately, these group discussions sometimes also lack substantial content discussions, and the employment of expertise from other external sources or surveys is missing. Moreover, as Deunk et al. (2014) points out after having observed a small group discussion on cut-off scores during standard setting, sometimes content discussions do not take place, and the expert groups tend to adjust their values to those proposed by others. They conclude that the training of expert groups is of great importance to ensure valid judgements.

The cut scores of a learning supportive assessment have to be trustworthy for the teachers based on their experience and ICT literacy. As Norway has a long standing tradition for teacher-based judgments which has been supplied with all kind of assessments being built into the educational system (OECD 2011), one challenge of a learning supportive assessment is that it is up to the teachers to use the results of the test. The OECD (2011) advises that the teachers become more transparent; for example, by documenting how they are making judgments in their classrooms.

The core of a standard setting procedure relies on a combination of psychometric analyses and reasoning by expert groups. The validity of these discussions in expert groups can be questioned since they follow their own communicative roles and agreements. In addition, we see that the results of a standard setting procedure will be transferred to another subsystem, from an assessment *development-based* and *research-based* subsystem to an *educative* subsystem which is characterized by pedagogic needs and classroom practices.

12.10 Final Conclusions: Limitations and Further Research

Standard setting for a formative assessment follows the same procedural instructions, as one would expect for any high-stakes assessment. However, when it comes to validity arguments different perspectives have to be taken into consideration.

This study has some limitations. First, the data are from a pilot study. Second, there was only one round with the expert group in this project, and as the nature of the assessment is formative, no experiences are yet available about how useful the assessment will be perceived by teachers who teach digital responsibility.

This article tried to deal with the uncertainly about different results of two standard setting methods by focusing on both student-centered and test-centered arguments. For this purpose, we are trying to make the student-centered arguments explicit in order to enhance the validity of the standard setting procedures.

When it comes to the validity arguments for taking different expectations into consideration when setting cut scores in a learning supportive assessment, both external and consequential validity arguments were applied in the first case. This included looking at other studies and at the results.

One challenge of combining different systems is the tension among the systems and the lack of alignment. This is something that could affect both the internal and the procedural validity. Expert-groups do not always act as one would ideally like, and the procedural validity has to do with the transfer across different communicative systems. The teachers play a more important part in a learning supportive assessment, and therefore challenges coming from a psychometric perspective versus the usefulness in classroom practices emerge.

More research is urgently needed on the combination of student-centered and test-centered approaches to standard setting. We also need more information about how the cut-off scores are perceived and how the schools and teachers use the descriptions of the proficiency levels.

References

Afzar, A. (2006). A systems theoretical critique of international comparisons. *Norsk pedagogisk tidsskrift, 27*, 253–264.

Ainley, J., Fraillon, J., Freeman, C. (2007). National assessment program: ICT literacy years 6 & 10 report 2005. Australia: MCEETYA.

Ala-Mutka, K. (2011). *Mapping digital competence: Towards a conceptual understanding*. Luxembourg: European Union.

Angoff, W. A. (1971). Scales, norms, and equivalent scores. In R. I. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council of Education.

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., & Rumble, M. (2012). Defining 21st century skills. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17–66). Dordrecht: Springer. doi:10.1007/978-94-007-2324-5.

Black, P. J., & Wiliam, D. (1998). *Inside the black box: Raising standards through classroom assessment*. London: King's College.

Calvani, A., Fini, A., Ranieri, M., & Picci, P. (2012). Are young generations in secondary school digitally competent? A study on Italian teenagers. *Computer & Education, 58*, 797–807.

Cizek, G. J. (Ed.). (2012a). *Setting performance standards: Concepts, methods and perspectives*. Mahwah: Lawrence Erlbaum.

Cizek, G. J. (2012b). An introduction to contemporary standard setting: Concepts, characteristics, and contexts. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 3–14). Mahwah: Lawrence Erlbaum.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting*. Thousand Oaks: Sage.

Claro, M., Preiss, D. D., San Martín, E., Jara, I., Hinostroza, J. E., Valenzuela, S., et al. (2012). Assessment of 21st century ICT skills in Chile: Test design and results from high school level students. *Computers & Education, 59*, 1042–1053.

Crocker, L., & Algina, J. (2008). *Introduction to classical & modern test theory*. New York: Holt, Rinehart, and Winston.

Deunk, M. I., van Kuijk, M. F., & Bosker, R. J. (2014). The effect on small group discussion on cut-off scores during standard setting. *Applied Measurement in Education, 27*, 77–97.

Educational Testing Service (ETS). (2001). Digital transformation. A framework for ICT literacy. A report of the international ICT literacy panel. www.ets.org/Media/Tests/Information_and_Communication _Technology_Literacy/ictreport.pdf. Accessed 15 April 2016.

Embretson, S. E., & Reise, S. P. (2009). *Item response theory for psychologists*. Mahwah: L. Erlbaum Associates.

Ferrari, A. (2012). *Digital competence in practice: An analysis of frameworks. JRC technical reports*. Seville: European Commission.

Ferrari, A. (2013). DIGCOMP: A framework for developing and understanding digital competence in Europe. Luxembourg. http://ipts.jrc.ec.europa.eu/publications/pub.cfm?id=6359. Accessed 15 April 2016.

Fraillon, J., et al. (2014). *Preparing for life in a digital age: The IEA International Computer and Educational Literacy Study International Report*. Switzerland: Springer.

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: Lawrence Erlbaum Associates.

Hatlevik, O. E., Egeberg, G., Guðmundsdóttir, G. B., Loftsgarden, M., & Loi, M. (2013). *Monitor skole 2013 - Om digital kompetanse og erfaringer med bruk av IKT i skolen Læring for framtiden*. Oslo: Senter for IKT i utdanningen.

Hatlevik, O. E., Ottestad, G., & Throndsen, I. (2015). Predictors of digital comp. in 7th grade: Students' motivation, family background, and culture for professional development in schools. *Journal of Computer Assisted Learning, 31*(3), 220–231.

Hsieh, M. (2013). Comparing yes/no Angoff and bookmark standard setting methods in the context of English assessment. *Language Assessment Quarterly, 10*, 331–350.

International Society for Technology in Education. (2007). ISTE standards students. https://www.iste.org/docs/pdfs/20-14_ISTE_Standards-S_PDF.pdf. Accessed 15 Apr 2016.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). Washington, DC: American Council on Education.

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

Kim, H. S., Kil, H. J., & Shin, A. (2014). An analysis of variables affecting the ICT literacy level of Korean elementary school students. *Computers & Education, 77*, 29–38.

Krumsvik, R. (2011). Digital competence in Norwegian teacher education and schools. *Högre utbilding, 1*, 39–51.

Kuhlmeier, H. A., & Hemker, B. (2007). The impact of computer use at home on students' internet skills. *Computer & Education, 49*, 460–480.

Matzat, U., & Sadowski, B. (2012). Does the "do-it-yourself approach" reduce digital inequality? Evidence of self-learning of digital skills. *The Information Society, 28*(1), 1–12.

Mediatilsynet. (2014). *Barn og media 2014. Barn og unges (9-16 år) bruk og opplevelse av media*. Fredrikstad: Mediatilsynet.

Mehrens, W. A., & Cizek, G. J. (2012). Setting standards for decision making: Classifications, consequences, and the common good. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 33–46). Mahwah: Lawrence Erlbaum.

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: Lawrence Erlbaum.

OECD. (2011). *OECD reviews of evaluation and assessment in education: Norway*. Paris: OECD Publishing.

Pant, H. A., Rupp, A. A., Tiffin-Richards, S. P., & Köller, O. (2009). Validity issues in standard setting studies. *Studies in Educational Evaluation, 35*, 95–101.

Plake, B. S., & Cizek, G. J. (2012). Variations on a theme: the modified Angoff, extended Angoff, and yes/no standard setting methods. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 181–199). Mahwah: Lawrence Erlbaum.

Sadler, D. R. (2009). Grade integrity and the representation of academic achievement. *Studies in Higher Education, 34*(7), 807–826. doi:10.1080/03075070802706553.

Søby, M. (2013). Learning to be: Developing and understanding digital competence. *Nordic Journal of Digital Literacy, 8*(03), 134–138.

The Norwegian Directorate for Education and Training. (2012). *Framework for basic skills*. Oslo: The Norwegian Directorate for Education and Training.

Tiffin-Richards, S. P., & Pant, H. A. (2013). Setting standards for English foreign language assessment: Methodology, validation, and a degree of arbitrariness. *Educational Measurement: Issues and Practice, 32*(2), 15–25.

Tveit, S. (2014). Educational assessment in Norway. *Assessment in Education: Principles, Policy & Practice, 21*(2), 221–237.

Zhong, Z. J. (2011). From access to usage: The divide of self-reported digital skills among adolescents. *Computers & Education, 56*, 736–746.

Chapter 13
Assessment for Learning and Standards: A Norwegian Strategy and Its Challenges

Gustaf B. Skar, Ragnar Thygesen, and Lars Sigfred Evensen

Abstract Assessment for learning in low-stakes contexts raises a series of problematic issues related to standards development. This chapter discusses several such issues on the basis of two interrelated data sets on writing as a key competency across the curriculum in Norway: How may standards communicate with teachers across the curriculum? How may standards fare in local learning environments over time? And most importantly: How can a shared rhetorical community among teachers develop over time and produce reliable assessment across local contexts? This chapter uses data sets that are based on a less than usual approach. In both data sets standards were developed in close collaboration with experienced primary-grade teachers, across the country. ICC analyses (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.

Keywords Assessment for learning • Standards development • Primary school writing across the curriculum • Sustainability

---

The approach taken in this chapter has been developed within a project group that also includes Kjell Lars Berge, Synnøve Matre, Hildegunn Otnes, Randi Solheim, as well as PhD candidates Sindre Dagsland, Trine Gedde-Dahl and Jannicke O. Bakke.

G.B. Skar
Norwegian Centre for Writing Education and Research, Trondheim, Norway

R. Thygesen
Agder University, Kristiansand, Norway

L.S. Evensen (✉)
NTNU, Trondheim, Norway
e-mail: lars.evensen@ntnu.no

S. Blömeke, J.-E. Gustafsson (eds.), *Standard Setting in Education*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-50856-6_13

13.1 Introduction

This chapter presents the development and local integration of new writing assessment standards, which were operationalised as assessment rubrics. The standards and rubrics were designed in collaboration between researchers and teachers to promote standards-based assessment for learning (AfL). A number of challenges arose during the developmental phase as well as during the phase where teachers from across Norway were trained to use standards and rubrics. These challenges will be at the core of this chapter, and will be illustrated by approaches and results from the Norwegian research project – 'Standards as a Tool for the Teaching and Assessment of Writing' (the NORMs project). This introduction will briefly review the concept of AfL and literature on standards-based AfL, before returning to the challenges and the research questions for this chapter.

13.2 Assessment for Learning

In recent literature on assessment two traditions are apparent. First, there has been an increase in national and international testing (cf. PISA and TIMMS), influencing the discourse of assessment, interventions and summative outcomes. Second, there has been a simultaneous development of AfL. While the first tradition focuses on learning outcomes, the second, still embryonic tradition focuses on ongoing classroom practice and improvement in both teaching and learning through formative feedback.

Assessment for learning aims at turning day-to-day assessment into teaching and learning processes that not only monitor, but enhance student learning (Black and Wiliam 1998a, b; Wiliam 2011). Applying the principles of AfL, considerable gains in student achievement can be seen, especially in struggling learners. It begins when teachers explain achievement targets to students, accompanied by exemplary student work. Frequent assessment sessions provide students with descriptive feedback in amounts they can manage without being overwhelmed. In this way, students can chart their trajectory toward the achievement targets set in dialogue with their teachers. The students' role is to use feedback from each assessment to realise where they are now in relation to where they should be, and determine how to improve next time.

In many studies during the 1990s that focused on assessment as an integral part of instruction, the term *formative assessment* was used. Black and Wiliam (1998a) defined formative assessment as follows:

> We use the general term *assessment* to refer to all those activities undertaken by teachers—and by their students in assessing themselves—that provide information to be used as feedback to modify teaching and learning activities. Such assessment becomes *formative assessment* when the evidence is actually used to adapt the teaching to meet student needs.

In the United Kingdom, however, the Assessment Reform Group argued that the term *formative assessment* was interpreted in so many different ways that it was no longer helpful. Instead, they preferred the concept *assessment for learning*, which

they defined as the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there (Broadfoot et al. 2002). In this chapter, we shall use the term *AfL*, but supplement it with the term *formative* when we are specifically addressing feedback within the AfL tradition.

Claims about the effectiveness of formative feedback were initially reported by Black and Wiliam in their 'Inside the Black Box' article (1998a). They estimated that feedback would increase students' learning within the range of effect sizes from 0.4 to 0.7. Later, Shute (2008) in a similar review, found effect sizes between 0.4 and 0.8. In their review of 74 meta-analyses of feedback, Hattie and Timperley (2007), found an average effect size of 0.95 based on an analysis of 4157 studies.

However, AfL can be difficult to implement, and insight into AfL practice is often based on small scale studies. For example, a study from Norway (Hopfenbeck et al. 2013) explored the development of implementation strategies used to enhance the programme 'Assessment for Learning – 2010-2014' in Norwegian schools. Their main finding was that the extent to which AfL practice was implemented, and how, differed across municipalities and individual respondents. This result shows that even within the same municipality the extent to which the AfL programme was implemented differs. An internationally oriented 'State of the Field Review' of the research literature on assessment and learning (Baird et al. 2014) showed that most studies on AfL are small-scale action research designs.

The researchers behind the report commented that the effects of formative assessment on learning have been over-sold by some authors, which is unfortunate because the limited empirical studies suggest a modest, but educationally significant impact on teaching and learning (page 6). Although few quantitative studies relating to AfL have been published, a large volume of small-scale studies involving interview data with a small number of teachers and students in a few schools was found. In this context, the NORMs project is a significant contribution, in terms of both design and coverage.

13.2.1 Standards-Based Assessment for Learning

There are two kinds of standards: *content standards* and *performance standards*. The former are collections of statements that describe specific learning outcomes. The latter specify what level of performance on a test is required for a test taker to be placed in a given performance category, meaning that *standard setting* refers to the process of deriving those levels (Cizek 2012). Classification of test takers into performance categories is often operationalised by *cut scores/passing scores/minimum achievement levels* to a performance on a test, dividing the test performances into two or more categories.

Within AfL research proficiency-based rubrics have been developed, i.e., rubrics which are aligned with standards-based rating scales from the start, describing progress in terms of achievement of the standard (e.g., Brookhart 2013). Such

rubrics are coordinated with definitions of various proficiency levels, standard by standard (e.g., Basic – Proficient – Advanced). They use the same scale (the same level of achievement) for every assessment; accordingly, they document student performance in terms of proficiency level on the standard, allowing students to track progress and set goals.

Many primary education teachers, however, experience integration of assessment in response to developmental and standards-based orientations as a challenge (Cuban 2009). In Canada, for instance, AfL approaches have been aligned with standards-based education (Gardner 2006). This assessment model is characterised by pre-established educational standards along with rigorous monitoring, planning, and continuous assessment. Although the model represents a mandatory standards-based orientation on primary education, it appears that teachers do not accept it at face value; instead they negotiate it in relation to their previously established developmental teaching orientations. Pyle and DeLuca (2013) similarly found that teachers maintained autonomy to negotiate assessment practices with their own pedagogical stance, although they were committed to standards-based education. The researchers explained their findings within a socio-developmental framework of learning where classroom context, social interactions, and developmental learning continuums are vital factors.

DeLuca and Hughes (2014) report similar results from a study aiming to analyse teachers' approaches to early primary assessment. Data were collected from Canadian teachers across different school contexts. The participating teachers expressed their views on assessment, and stated three diverse conceptions and purposes within early primary education: (1) assessment as a growth trajectory, (2) assessment as a normative structure, and (3) assessment of the whole child. As to the first conception assessments were used to determine student progress along growth and developmental trajectories; these trajectories were seen as either linked to curriculum standards and expectations or to student developmental goals.

It is noteworthy that many teachers felt a need to balance students' developmental trajectories with their growth toward curriculum expectations. Typically, they blended curriculum expectations with student developmental readiness. Although academic standards were viewed and prioritised differently across various contexts of education, all teachers used assessments to measure and promote learning of academic knowledge and skills. With reference to Katz (2007), DeLuca and Hughes call attention to the fact that planning learning experiences, rather than planning for academic learning, aligns the developmental, standards-based and academic frameworks of early primary education. They also accentuate that teachers in their study held a holistic view of student development and achievement – a finding that is interesting given mandates in upper years' education aiming at parsing of academic growth and learning skill development while using standards-based analytic scoring rubrics.

13.3 Rhetorical Communities of Assessment

The concept of *rhetorical communities* entered writing assessment research with the work of the IEA project of the 1980s on writing in first language (L1) (Gorman et al. 1988). In spite of a rigorous research design, it proved difficult to reach a good *tertium comparationis* for generalising results across 14 countries. It was assumed that internationally diverse norms had played an underestimated role and interfered with the design. Purves (1986) termed the social carriers of such norms rhetorical communities. Simultaneous work within other empirical fields revealed that such communities seemed to exist even on local or specialised social arenas, as evidenced in the work of Brown and Duguid (1991), or Lave and Wenger (1991) on the concept of *communities of practice*. In this chapter, tensions between local communities of assessment practice and a nation-wide rhetorical community of assessment practice will be considered as a serious challenge for AfL using a standard.

13.4 Challenges

The remainder of this chapter will concentrate on specific challenges in developing and implementing standards for standards-based AfL. One challenge is that the standards in such cases necessarily need to be formulated in a way that is close to the discourse of ordinary, but trained teachers. A second challenge is that assessment criteria as well as other assessment tools need to be specific enough to inform 'feed forward' in class (Hattie and Timperley 2007). This need implies a tension between discoursal specificity on the one hand and a complexity that must be manageable for a national rhetorical community (Purves 1986) to develop, on the other hand. Within the NORMs project a point of balance was sought in critical discussion of drafts in workshops with experienced teachers coming from schools across the country.

Both of these challenges may be viewed as at least partially common to all forms of assessment, but a third challenge is not. For AfL to succeed across local contexts, the assessment tools need to be open and flexible enough to adapt to different learning environments. Baird et al. (2014) argue that this sensibility to learning contexts requires an approach that is qualitatively different from the ones traditionally taken in well controlled studies within the testing tradition. While the testing tradition gives priority to issues of reliability, AfL implies a closer focus on validity where teacher monitoring of student progress is the true measure. More specifically we would like to argue that adapting AfL across contexts implies a sensitivity to relational complexity that accentuates what Cicourel (1996) has termed *ecological validity*, a form of validity not frequently considered in traditional approaches to assessment. In traditional approaches contextual variability is minimised in order to avoid intervening variables. As viewed within AfL, however, intervening variables are in fact a part of any local learning environment. What then needs to be established

is a perspective not on the single variables, but on relations among variables that may appear across contexts. Such relations may be specified and thus potentially bridge the traditional distinction between internal and external validity, if taken properly into account (Evensen 2013). When this condition of contextual adaptability is met – and only in this case – a standards-based form of AfL may prove sustainable, across local contexts and over time.

Granted that local learning environments may initially differ in their norms of expected performance and assessment practices, the above premises imply that long-term investments will be needed in order to build a nation-wide rhetorical community that may foster reliable results in the end (cf. Purves 1986).

The literature review and the challenges presented above seem to boil down to adaptability as the major problem. The problem implies that ecological validity is a challenge, and we may assume that a rhetorical community will develop only slowly. But such development will eventually result in increasing reliability, seen as an indirect measure of the homogeneity of the community. This line of reasoning may be formulated as a hypothesis. If the tools (see below) work, what ought to unfold is developing consistency across time and tasks as a symptom of the rhetorical community being established. We shall return to this hypothesis.

13.5 The NORMs Project

In light of the themes unfolded in the introduction, two research questions have been formulated:

1. How did the NORMs project develop standards for AfL-purposes?
2. Did the training of teachers from across Norway lead to higher consistency, and thus indications of possibilities to form a nation-wide rhetorical community?

The NORMs project aimed at developing standards (in the project referred to as 'norms') for communicative and semiotic aspects of writing in terms of criteria for goal attainment and to investigate whether the integration of norms lead to improved teacher assessment of students' texts and improved quality of the students' writing. The project consisted of two distinct phases, paralleling the aims. In the first phase, standards for writing proficiency were developed by researchers in close collaboration with teachers. In the second phase, a large scale writing intervention took place. The intervention was carried out between 2012 and 2014, and participators were 500 teachers and 3088 students, from 20 intervention schools and 5 comparison schools across Norway, representing primary school years 3–4 and lower-secondary school years 6–7. Students entering the project in grade 4 and grade 7 participated for only one school year.

Focusing on teachers' professional development and students' writing proficiency, the project offered conceptual and pedagogical tools to teachers. The former consisted of a novel conceptualisation of writing, The Wheel of Writing (Berge et al. 2016a), and the standards that are discussed in this chapter (Evensen et al. 2016). The model and the standards are described below.

13.6 Developing Standards

Answering the first research question, the initial phase of this project aimed at developing standards in collaboration with experienced teachers, meaning that teachers' expectations of their students' writing competency were explored as a 'bottom-up' approach. As detailed in Evensen et al. (2016) a think-aloud approach to everyday assessment was used to elicit and record specific quality judgments, as pairs of primary school teachers were assessing a partially common set of texts written by students toward the end of grades 4 and 7. Following these sessions audio transcripts were analysed to locate criteria that appeared across geographically distributed schools. The analysis also revealed how the criteria were formulated by the teachers involved. Using such formulations would help bridge the gap between researchers' and teachers' ways of expressing criteria.

Each criterion was then categorised as belonging to a specific assessment domain. A holistic strategy would probably not serve a formative purpose of assessment well, and it was appropriate to choose an approach that might yield more specific information. Even holistic scoring has an underlying factor structure where domains like contents, grammar and orthography play a role (cf. the Diederich 1974 tradition). A functional understanding of writing implies for instance that the writer-reader relationship should be included as a central domain (Berge, Evensen et al. 2016). Furthermore, the Norwegian curriculum emphasises the multi-modal nature of writing. Granted these perspectives, the following domains were included as foci for assessment: Communication (the writer-reader relationship), Contents, Text structuring, Language use (lexicon, syntax and style), Orthography (with morphology), Punctuation and Use of the written language (handwriting and use of multimodal resources).[1]

The first four domains were labelled 'functional domains,' because assessing them are contingent on cultural knowledge about writing proficiency. The fifth and sixth domain were labelled 'coding domains,' because rules for orthography and punctuation transcend specific writing situations.

The full set of criteria was eventually refined by the project group to form standards for expected proficiency after 4 and 7 years of schooling, i.e. as an expression of what the majority of Norwegian students at these school years are expected by teachers to achieve (Evensen et al. 2016). See appendix for an example. To link the multidimensional assessment to the Wheel of Writing construct (Berge et al. 2016a), a primary trait approach was taken, meaning that assessment was carried out with a perspectival focus on one pre-specified aspect of the writing. This approach implied that students' texts would be assessed as viewed through a functional lens of one specified act of writing, being combined with one specified purpose for writing, both to be indicated in each assignment.

The standards were operationalised on five-level assessment rubrics, with the phrasing 'as can be expected [in this domain] from most students at this grade' as

[1] The domain *Use of the written language* is not presented or analysed further in this chapter.

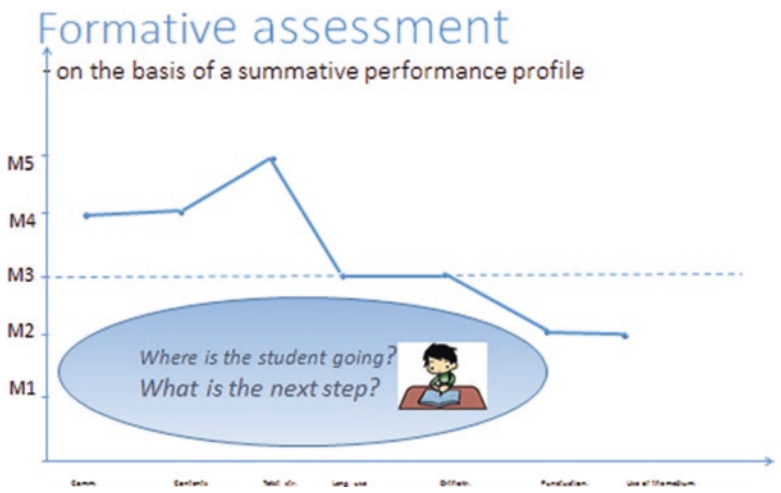


Fig. 13.1 Dummy performance profile used in project teacher training

the mid category. This mid-level equals the set standards. As an aid, annotated benchmark compositions for the class level were placed at the teacher's disposal, illustrating what kind of achievement is to be expected of most students, and what kind of achieved performance was much higher or lower than expected.

A text is assessed summatively by the teacher, focusing on where he/she is located at the moment of assessment compared to the standard. It may take the form of a performance profile that graphically represents how a student has performed across assessment domains on a given writing task. The profile should not be interpreted as a traditional mark, but contains information domain by domain as compared to the standard. The ratings are not summed into a total score, nor provided as a percentage or statistical average, but presented to the teacher as a didactic tool as a basis for assessment with a formative purpose. Each domain is related to the standard, and a selected domain may then easily be used to discuss with the student which criteria may be in focus for targeted future efforts. In this way summative assessment is translated into formative input. An example of such a profile is given in Fig. 13.1.

13.7 Training Teachers to Rate

As part of the intervention phase, the writing construct and standards were presented locally by project researchers to the staff of all participating schools, along with other assessment tools in the form of booklets, rubrics and practice sessions. In addition, selected teachers from the 20 intervention schools were sent to joint-project workshops. Two workshops were organised each year, attended by an average of 80 teachers. Due to limitations of available resources, all 500 teachers could thus not participate in each workshop. Instead, two teachers from each of the grades

3–4 and 6–7 at each school were sent upon the principal's decision. Some schools used a rotation principle from workshop to workshop, sending as many different teachers as possible, while others tended to select the same teachers every time. In this chapter, reliability data are presented from the four workshops.

The purpose of rater training is usually to minimise 'rater effects', which can be defined as 'the systematic variance in performance ratings that is associated in some way with the rater and not with the actual performance of the ratee' (Scullen et al. 2000). In the NORMs project, the goals of traditional rater training were extended. While aiming at minimising rater effects, the training also sought to maximise student learning. To this end, the training focused both on use of the assessment rubrics and on how teachers could make use of student scores in writing instruction. The training was staged in the following ways.

First, teachers were introduced to the Wheel of Writing (Berge et al. 2016a). According to this model, the writing construct is constituted by three interrelated dimensions to account for (1) writing as an act of meaning making, (2) writing as driven by a purpose and (3) writing as semiotically mediated (Mertz and Parmentier 1985) by different modalities. Teachers were introduced to this model and to ways of conceptualising their own writing instruction using the Wheel of Writing as a joint lens (Berge et al. 2016b). More importantly, the model functioned as a basis for the teachers' formulation of writing assignments to their students; which subsequently resulted in the text corpora that constitutes the basis for the analyses in the project.

Second, teachers underwent conventional rater training, c.f. the training described in Weigle (1998). The standards for the assessment domains were introduced at the workshops by the researchers who had developed them. It was demonstrated how to use the assessment tools, i.e., the assessment rubric associated with each assessment domain. A two-step rating session would follow. First, teachers used the set of criteria to assess about 10 student texts not originating from their own school. These were draft versions of student writing composed by students at participating schools. They then discussed these assessments in a rater pair with a designated rating partner. The rater pair would form a consensus decision which was reported to the researchers. After this procedure, the teacher received a new text package and eventually a new rating partner. In all, teachers would read 40 student texts which were discussed with three separate rating partners. All student texts were read by four teachers (i.e. two rater pairs), which enabled monitoring of the overall rater agreement using estimates of the intraclass correlation coefficient one-way single measure (McGraw and Wong 1996), commonly labelled ICC(1).

Third, assessment specialists from the research team gave lectures on formative assessment and how to use the Wheel of Writing to design tasks, and the assessment domain standards to conduct assessment that could inform future writing instruction. As already mentioned, teachers were also introduced to the visual aid presented in Fig. 13.1.

Fourth, teachers tested out the tools in practice. Participants devised writing tasks for use at their own school, which ask the students to write two versions: a draft version and a revised version. One draft version would be scored by the teacher for purposes of formative feedback, and a revised version would be scored and

stored in students' portfolios. Tasks and student texts, both draft and revised versions, were then sent to the research team. However, the text packages handed to teachers at the rater training only contained the draft versions.

13.8 Did the Standards Live Up to Requirements for Useful Assessment Tools?

Answering research question 2, and thus gaining insights into whether the tools provided by the NORMs project could serve as a foundation for fostering a national rhetorical community of teachers, and reliably be used for assessing a variety of task types, we conducted statistical analyses of the workshop rater data. The basic assumption was that reliability estimates would increase as teachers grew accustomed to both using and sharing the assessment tools. Such an increase would indicate progress toward the goal that participating teachers could interchangeably read a student draft and reach the same or closely similar conclusion regarding its qualities in separate domains.

The data used for the analysis were scores from *rater pairs* (c.f. above). Each student text was given two scores on the six assessment domains. However, the rater pairs were not stable entities, resulting in input data columns representing different raters. A suitable statistical technique was therefore the ICC(1) (McGraw and Wong 1996). This allows for investigation of the reliability within the group, when the rater columns do not represent stable entities. Two sets of reliability estimates were computed for each of the two subsamples of teachers, 3rd–4th grade teachers and 6th–7th grade teachers. The first estimate summed their assessment of 'functional' assessment domains, i.e., *communication, content, text structure* and *language use*. The second estimate summed 'coding' assessment domains, i.e., *spelling* and *punctuation*. The number of students that were assessed varied between the workshops. This was a result of the fact that students who entered the project in grade 4 and grade 7 only participated for one year. Table 13.1 presents the number of students that were assessed at the different workshops. The results of the ICC analysis are presented in Tables 13.2 and 13.3.

As can be seen, the number of students dropped heavily from workshop 2 to workshop 3. The number of teachers did not, but as previously stated, each participating school sent two members of staff, and not necessarily the same to each work-

Table 13.1 Data

	3th–4rd grade		6th–7th grade	
Workshop	Student texts (N)	Teachers (N)	Student texts (N)	Teachers (N)
#1, 2013, Jan	176	40	183	40
#2, 2013, May	177	40	183	40
#3, 2014, Jan	89	40	80	40
#4, 2014, May	86	40	80	40

Table 13.2 Reliability estimates for assessment of functional domains

	3th–4rd grade Teachers			6th–7th grade Teachers		
	Student texts (N)	ICC	CI 95%	Student texts (N)	ICC	CI 95%
2013, Jan	176	.526	.410, .625	183	.611	.512, .694
2013, May	177	.468	.345, .576	183	.648	.556, .725
2014, Jan	89	.525	.358, .660	80	.654	.509, .763
2014, May	86	.630	.484, .742	80	.696	.564, .794

Note: ICC = intraclass correlation coefficient one-way single measure. CI 95% = Confidence interval, 95%

Table 13.3 Reliability estimates for assessment of coding domains

	3th–4rd grade Teachers			6th–7th grade Teachers		
	Student texts (N)	ICC	CI 95%	Student texts (N)	ICC	CI 95%
2013, Jan	176	.734	.657, .795	183	.616	.518, .699
2013, May	177	.652	.558, .729	183	.513	.398, .612
2014, Jan	89	.668	.535, .769	80	.615	.458, .734
2014, May	86	.676	.543, .776	80	.551	.379, .687

Note: ICC = intraclass correlation coefficient one-way single measure. CI 95% = Confidence interval, 95%

shop. The results showed both increase and decrease in the reliability estimates over time. Thus, the analysis revealed an interesting pattern of reliability for assessment of functional assessment domains and coding assessment domains, respectively. When comparing results presented in Tables 13.2 and 13.3, it may be noted that the increase in reliability is largely associated with an increase in the consistency of assessing functional domains. For teachers in grades 3–4 the reliability increases from .526 to .630 in the project period. For teachers in grades 6–7 the increase is from .611 to .696.Turning to assessment of coding domains, the results demonstrate a somewhat surprising decrease in reliability, for both teacher samples. It may thus seem that a relatively strong rhetorical community had been present in this area at the start of the project were the reliability estimate for 3th–4rd grade teachers was .734 and for 6th–7th grade teachers was .616. At the end of the project, the levels are still acceptable, .676, for teachers in grades 3–4, but rather low, .515, for teachers in grades 6–7. We shall return to these patterns below.

13.9 Discussion

Modern validity theory stresses that validity concerns test scores and uses of test scores, not a test in itself (AERA et al. 2014; Bachman 2005; Kane 2013; Messick 1989). This relates to the practice of assigning meaning to test scores. As Kane

(2013) notes, 'test scores are of interest because they are used to support claims that go beyond (often far beyond) the observed performances.' For instance, a teacher assessing students' texts might conclude that this particular group of students needs further instruction on text structure and change instructional practice accordingly. These may or may not be more or less valid conclusions and actions, all depending on matters such as the representativeness of the text sample and the way the scoring was carried out. Validation of scores and score-based actions, then, requires empirical and analytical inquiry into the quality of the assessment process.

The test score itself is validated through analyses of scoring guidelines and consistency in using these guidelines (Bachman and Palmer 2010; Kane 2015). Reliable scores are an obvious prerequisite for any generalisation and 'extrapolation' beyond the particular moment in which the assessment was made (Kane 2015). Low reliability indicates that scores are irrelevantly influenced by, for example, rater effects. High consistency, on the other hand, is often accepted as an indication that raters share an interpretation of the assessment rubrics, and how they should be applied.

In more messy assessment situations, with new raters, a new writing construct, new assessment criteria, new rating assessment rubrics and responses to different writing prompts, i.e., the situation within the NORMs project, the reliability index might not adequately represent the status of the rater group. As was demonstrated above, on the one hand, the reliability of assessing coding domains fluctuated, but mostly remained at acceptable levels; the reliability of assessment of functional domains, i.e., the ones related to cultural knowledge about writing proficiency, on the other hand, showed increase. Untechnically speaking, the increase in reliability indicated that the tools introduced in the NORMs project could in fact strengthen a shared rhetorical community, not just in making consistent judgements, but also in basing those judgements on the same conceptualisation of writing proficiency.

The challenge of changing local communities of writing assessment practice into one nation-wide rhetorical community did indeed prove difficult. ICC(1) results demonstrate progress, but also hide remaining challenges that do not appear from the ICC results. One such challenge appeared at the school level, where local administrative problems (like reorganising the school, budget cuts with constant changes of staff, a new principal or prolonged illness with the local project coordinator) jeopardised the intervention. A second challenge appeared due to time. There was a noted pre-existence of stronger local communities of practice than expected, where their single focus tended to be on one multidimensional aspect of writing only (orthography and punctuation). In such cases adapting a wider, discourse-functional writing construct (Berge et al. 2016a) proved time consuming and seemed to somehow interfere with teacher expertise in applying their previous narrow construct. The intervention also revealed an existing community of practice at another level. As in several of the studies reviewed in this chapter, the first workshops revealed that it was difficult for many teachers to assess a student text without background knowledge about the individual writer, in terms of abilities, developmental trajectory and even home situation (one aspect of teacher ethics within adapted education).

Being forced by the sessions' design to assess only the text at hand in relation to domain-specific standards turned out to be an unfamiliar and qualitatively different

approach for these teachers. In their 2014 presentation Matre and Solheim thus observed that 'Still, in most cases a consideration of the single student's situation was emphasised more than textual qualities. Several teachers thus experienced it as problematic – and unusual – to assess the text as a text, without taking the writer into consideration.' (Matre and Solheim 2014) (our translation).

The teachers initially struggled on other accounts as well. As pointed out by Matre and Solheim (2015), even the NORMs project pilot study revealed that many Norwegian L1 teachers had a holistic approach to assessment, whereas many of their science teacher colleagues had a purely item-based analytic approach. When studying audiotaped assessment sessions within the early main stage of the project, they also found that a sub-group of project teachers 'are commenting on the linguistic surface and on local details without asking what function they have in the text' (2015). They further noticed that when discussing discourse functional domains like communication, content and text structure 'the teachers are fumbling more often' (2015).

If we assume that a move to purely text-based assessment is necessary for the assessment to have construct validity, it becomes clear that making such a move is likely to result in temporarily poorer reliability (cf. the noted ICC dip from workshop 1 to workshop 2). We thus arrive at the ironical conclusion that validity concerns may have to be given priority before reliability concerns are given equal priority, whenever a (partially) new writing construct is part of an intervention. Put differently, 'sub-surface' validity development may surface as temporarily poorer reliability, a pattern similar to a well-known one within second language acquisition where progress may be indicated by the presence of new kinds of error.

A similar challenge appears in the data when local practices of designing writing prompts proved to be vague, self-contradictory or one-sided (typically favouring narrative writing in first language contexts). Both of these challenges required considerable initial effort with the participant teachers before the intervention proper could really have much effect. A third challenge involves the intervention proper. AfL implies formulating summative insights in specified and formative feedback suggestions to individual students. Even granted the graphical tool of multidimensional student performance profiles, many intervention teachers struggled with formulating specific and evidence-based feed forward.

In sum, such learning lessons for the project group help to both locate and disentangle highly relevant contextual phenomena. Such lessons ironically underline our point that forging a nation-wide rhetorical community for assessment out of several local communities of practice does indeed require prolonged investment, in our case involving cross-curricular groups of teachers at each school. This insight raises the issue of which institution may offer such investment, as any project will be too weak to ensure long-term support. In Norway, however, such prolonged investment is institutionally provided through the new official assessment system from 2014. In this system our functional construct and standards are implemented, and writing prompts are provided from central authorities along with an empirically more refined set of assessment criteria reflected in a new and more specific rubric for each school year since 2014 (Skar et al. 2015). Even if participation in this system is a voluntary resource for primary schools, it will provide stable direction for

development, we believe, as it is the only national system currently implemented for the assessment of writing.

The learning lessons also have a methodological aspect, revealed by insights reached while the NORMs intervention was underway. As indicated by several studies reviewed in the theory section above, AfL is frequently adapted to serve multiple local constraints in the interface between curriculum and assessment. Such adapting moves were observed also within the NORMs project. It thus seems clear that AfL implies qualitative and frequently hidden elements to be operative during intervention. Identifying such elements may require questionnaires to participant teachers about their initial and post-intervention beliefs and practices. In New Zealand this approach has been one integrative methodological aspect of the AssTTle approach to writing assessment' (Parr and Brown 2015). Such elements may also require longitudinal follow-up studies to observe the sustainability of an intervention when eventually left to local communities of teachers.

So what does all of this contribute to standards development, as seen within a Nordic research context? This chapter has discussed learning lessons from a less than usual approach, where standards have been formulated in collaboration with experienced teachers. The presentation has shown that developments towards AfL have inspired looking at alternative lines of reasoning that may serve new contexts of standards-based assessment. More substantially, we hope, the discussion in this chapter has also revealed that there are a number of partially new challenges implied by AfL premises. The fundamental challenge is related to boldly facing the contextual complexity of any teaching environment where student learning is supposed to take place. This challenge seems to necessitate not only new lines of reasoning, but also a combination of quantitative and qualitative research methods that may reveal initially hidden phenomena that may affect any assessment practice, disregarding its epistemological or methodological orientation.

In the chapter we have seriously considered how standards may be adapted to teacher experience and teacher conceptions, formulating standards in close and prolonged collaboration. What is emerging from large-scale studies with co-developed standards such as the NORMs project is thus that the core issue for AfL may not be so much about *standardisation* as it is about *standards integration* in local teaching and learning environments. This issue requires approaches to ecological validity and sustainability that are not normally found within traditional approaches to standards development or standard setting.

At the same time, a final word of caution is still warranted. In high stakes contexts the approach illustrated in this chapter may probably not be advisable. In such cases the issue of reliability does indeed deserve focal attention, and allowing for local learning environments to influence the assessment would clash with equity concerns. In AfL contexts, however, this issue looks remarkably different. AfL will by necessity take place in local contexts, with the teacher expertise that is locally at hand. Within such constraints the issue is how centrally developed tools may gradually foster higher validity and reliability than what existed before. Viewed in this light, a new and multidimensional approach to standards development, like the one discussed in this chapter, may eventually decrease the role of local contextual bias and make this alternative line of reasoning a winning game rather than a losing game.

Appendix

Level	Level 1	Level 2	Level 3	Level 4	Level 5
Meaning	Very low level of mastery within the domain	Low level of mastery	**As to be expected from most students after 4 or 7 years of schooling**	High level of mastery	Very high level of mastery within the domain
DESCRIPTOR for scale used at school year 4	The writer is expected to:				
	Use some relevant principles of composition (temporal or thematic sequence, etc.)				
	Use an introduction, a main part and an ending				
	Create thematic cohesion within the various parts of the text				
	Create textual cohesion by connectors (*or*, *but*, *because,* etc.)				
DESCRIPTOR for scale used at school year7	The writer is expected to:				
	Use a variety of ways of structuring the text				
	Structure the text in a purposeful way (e.g., genre)				
	Use paragraphs as an organising principle				
	Create cohesion by a variety of connectors				

The assessment rubric for the domain 'text organisation' with descriptors for school years 4 and 7

References

AERA, APA, & NCEM. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly: An International Journal, 2*(1), 1–34. http://doi.org/10.1207/s15434311laq0201_1. Accessed 25 June 2016.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Baird, J.-A., Hopfenback, T. N., Newton, P. E., Stobart, G., & Steen-Utheim, A. T. (Eds.). (2014). *Assessment and learning: State of the field review*. Oslo: Knowledge Center for Education.

Berge, K. L., Evensen, L. S., Thygesen, R. (2016a). The wheel of writing: a model of the writing domain for the teaching and assessing of writing as a key competency. *The Curriculum Journal*, 27(2), 172–189. http://doi.org/10.1080/09585176.2015.1129980. Accessed 26 June 26 2016.

Berge, K. L., Skar, G. B., Matre, S., Solheim, R., Evensen, L. S., Ones, H., Thygesen, R. (2016b). Introducing new semiotic tools for writing instruction and writing assessment: Effects on students' writing proficiency. *Submitted*.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74. http://doi.org/10.1080/0969595980050102. Accessed 20 June 2016.

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139–148.

Broadfoot, P., Daugherty, R., Gardner, J., Harlen, W., James, M., & Stobart, G. (2002). *Assessment for learning: 10 principles*. Cambridge: University of Cambridge School of Education.

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. Alexandria: ASCD.

Brown, J. S., & Duguid, P. (1991). Organizational learning and communities-of-practice: Toward a unified view of working, learning, and innovation. *Organization Science, 2*(1), 40–57. http://doi.org/10.1287/orsc.2.1.40. Accessed 20 June 2016.

Cicourel, A. V. (1996). Ecological validity and 'white room effects': The interaction of cognitive and cultural models in the pragmatic analysis of elicited narratives from children. *Pragmatics & Cognition,* 4(2), 221–264. http://doi.org/10.1075/pc.4.2.04cic. Accessed 20 June 2016.

Cizek, G. J. (2012). *Setting performance standards: Foundations, methods, and innovations* (2nd ed.). New York: Routledge.

Cuban, L. (2009). *Hugging the middle: How teachers teach in an era of testing and accountability*. New York: Teachers College Press.

DeLuca, C., & Hughes, S. (2014). Assessment in early primary education: An empirical study of five school contexts. *Journal of Research in Childhood Education,* 28(4), 441–460. http://doi.org/10.1080/02568543.2014.944722. Accessed 25 June 2016.

Diederich, P. B. (1974). *Measuring growth in English*. Urbana: National Council of Teachers of English.

Evensen, L. S. (2013). *Applied linguistics. Towards a new integration?* London: Equinox Publishing.

Evensen, L. S., Berge, K. L., Thygesen, R., Matre, S., Solheim, R. (2016). Standards as a tool for teaching and assessing cross-curricular writing. *The Curriculum Journal*, *27*(2), 229–245. http://doi.org/10.1080/09585176.2015.1134338. Accessed 20 June 2016.

Gardner, J. (2006). Assessment for learning: A compelling conceptualization. In J. Gardner (Ed.), *Assessment and learning* (pp. 197–204). London: SAGE.

Gorman, T. P., Purves, A. C., & Degenhart, R. E. (Eds.). (1988). *The IEA study of written composition I. The international writing tasks and scorings scales*. Oxford: Pergamon Press.

Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research*,* 77(1), 81–112. http://doi.org/10.3102/003465430298487. Accessed 20 June 2016.

Hopfenbeck, T., Tolo, A., Florez, T., & El Masri, Y. (2013). *Balancing accountability and trust*. Paris: Organisation for Economic Co-operation and Development.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement,* 50(1), 1–73. http://doi.org/10.1111/jedm.12000. Accessed 26 June 2016.

Kane, M. T. (2015). Validitet [Validity]. In G. Skar & M. Tengberg (Eds.), *Bedömning i svenskämnet [Assessment in the School Subject Swedish]* (pp. 212–237). Stockholm: Natur och kultur.

Katz, L. G. (2007). Standards of experience. *Young Children,* 62(3), 94–95. http://www.jstor.org/stable/42730032. Accessed 26 June 2016.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.

Matre, S., & Solheim, R. (2014). Lærarsamtalar om elevtekstar – mot eit felles fagspråk om skriving og vurdering [Teacher talk about student texts - towards a shared meta-language on writing and assessment]. In R. Hvistendahl & A. Roe (Eds.), *Alle tiders norskdidaktiker [A mother tongue educator for all times]* (pp. 219–244). Oslo: Novus forlag.

Matre, S., & Solheim, R. (2015). Writing education and assessment in Norway: towards shared understanding, shared language and shared responsibility. *L1 Educational Studies in Language and Literature,* 15, 1–33. http://doi.org/10.17239/L1ESLL-2015.15.01.05. Accessed 25 June 2016.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46.

Mertz, E., & Parmentier, R. J. (Eds.). (1985). *Semiotic mediation: Sociocultural and psychological perspectives*. Orlando: Academic Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education.

Parr, J., & Brown, G. (2015). Learning about writing: A consideration of the recently revised baTTle: writing. *Curriculum Matters,* 11, 134–154. http://doi.org/10.18296/cm.0008. Accessed 20 June 2016.

Purves, A. C. (1986). Rhetorical communities, the international student and basic writing. *The Journal of Basic Writing, 5*(1), 38–51.

Pyle, A., & DeLuca, C. (2013). Assessment in the kindergarten-classroom: An empirical study of teachers' assessment approaches. *Early Childhood Education Journal,* 41(5), 373–380. http://doi.org/10.1007/s10643-012-0573-2. Accessed 25 June 2016.

Scullen, S. E., Mount, M. K., Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology,* 85(6), 956–970. http://doi.org/10.1037//00219010.85.6.956. Accessed 25 June 2016.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research,* 78(1), 153–189. http://doi.org/10.3102/0034654307313795. Accessed 25 June 2016.

Skar, G. B., Evensen, L. S., Iversen, J. M. (2015). *Læringsstøttende prøver i skriving 2014. Teknisk rapport* [Formative Writing Assessment Package 2014. Technical Report]. Trondheim: Nasjonalt senter for skriveopplæring og skriveforsking.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing,* 15(2), 263–287. http://doi.org/10.1177/026553229801500205. Accessed 20 June 2016.

Wiliam, D. (2011). What is assessment for learning? Studies in Educational Evaluation, 37, 3–14.

Chapter 14
How Do Finns Know? Educational Monitoring without Inspection and Standard Setting

Mari-Pauliina Vainikainen, Helena Thuneberg, Jukka Marjanen, Jarkko Hautamäki, Sirkku Kupiainen, and Risto Hotulainen

Abstract The Finnish educational system was decentralised in the 1980s and the 1990s. The school inspection system was dissolved and the municipalities as organisers of education were given responsibility for monitoring the effectiveness of education and securing that every child has equal possibilities in proceeding through the 9-year basic education consisting of primary education and lower secondary education. A national model for sample-based curricular and thematic assessments was created to ensure equity of education in different parts of the country. Unlike many other countries, Finland decided not to have a comprehensive standardised testing system, and the goals set in the national Core Curriculum were not considered as standards either. Thus, matriculation examination at the end of academic track of upper secondary education remained as the only high-stakes test, but due to extensive possibilities for subject selection and the normative approach still applied in grading the exams, it only produces a limited amount of information that can be used in monitoring the trends of pupil performance. This chapter gives an overview of educational quality monitoring during basic education in Finland, presenting first a short historical review of how the monitoring system has received its current form. Next, the national sample-based assessment system is described before introducing the local ways of monitoring the equity and functionality of basic education. These include the screening of support needs and the evaluation of the effectiveness of the provided support that has been claimed to be one of the explanations behind Finland's success in international comparisons. Finally, we will discuss whether quality monitoring without standard setting can work and what standard setting could contribute to the Finnish education system.

M.-P. Vainikainen (✉) • H. Thuneberg • J. Marjanen • J. Hautamäki
S. Kupiainen • R. Hotulainen
Centre for Educational Assessment, Faculty of Educational Sciences,
University of Helsinki, Helsinki, Finland
e-mail: mari-pauliina.vainikainen@helsinki.fi; helena.thuneberg@helsinki.fi;
jukka.marjanen@helsinki.fi; jarkko.hautamaki@helsinki.fi; sirkku.kupiainen@helsinki.fi;
risto.hotulainen@helsinki.fi

S. Blömeke, J.-E. Gustafsson (eds.), *Standard Setting in Education*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-50856-6_14

Keywords Finland • Educational Equity • Local Monitoring of Learning Outcomes • Sample-Based Assessments

14.1 Introduction

For nearly half a century, the Nordic countries have been known for their relatively similar basic education systems that are based on the idea of educational equity (Antikainen 2006; Telhaug et al. 2006). The main principle has been to provide equal possibilities for learning for every child regardless of their socio-economical background or residential area throughout the 9–10-year-long basic education. Finland still follows the model established in the Basic Education Act of 1968 that was gradually implemented in 1972–1976, having a 9-year basic education system followed by 3-year upper secondary education in either an academic or a vocational track. In 2011, Finland changed its basic education legislation to make the role of local schools even stronger than before (Thuneberg et al. 2013). In 2012, 96 percent of the comprehensive schools were run by municipalities (the Official Statistics of Finland, www.stat.fi) and followed local curricula, which were regulated by the National Core Curriculum (National Board of Education 2004a). Based on PISA 2006 data, in addition to a high average performance level, the segregation of schools was the lowest in Finland when measured both by the distribution of socioeconomic status of pupils and by their performance level in the assessment (Willms 2010). Even though the average performance level of the pupils has recently declined in both national and international assessments (Hautamäki et al. 2013), school segregation has not considerably increased (Vainikainen et al. 2016).

The equitable outcomes of the Finnish education system are, however, not a result of systematic standard setting. There are certain elements in the monitoring system that could be interpreted as standards at a national, municipal and school-level, but high-stakes population-based testing has never been a part of it at any level. Nowadays, there are national sample-based assessments and additional municipal assessments that occasionally may cover entire age cohorts, but they are also low-stakes by nature. In general, there is very little testing that is controlled by anyone else but an individual teacher and school grades given by teachers reflect almost solely their interpretation of how well the pupils have reached the goals set in the National Core Curriculum. Not even these goals can be seen as standards as there is a considerable amount of flexibility in the grading criteria related to them. Thus, no minimum criteria for acceptable performance have been defined and standard setting as it is usually defined does not exist in today's Finland.

This chapter gives an overview of the history and the current state of educational quality monitoring in Finland, focussing on the controversy of relatively high and homogeneous school-level performance despite of the clear lack of control and standard setting as it is commonly understood. First, we present a short historical overview of how the monitoring system has received its current form. Then we describe the national sample-based assessment system consisting of subject-specific,

thematic and international assessments that are typically conducted at the end of basic education. Finally, we present the local ways of monitoring and securing the equity and functionality of education that from the schools' perspective play a much larger role than the national monitoring system. This includes school- or in some cases municipal-level screening of progress and support needs and the evaluation of the provided support that has been claimed to be one of the explanations behind Finland's success in international comparisons (Sabel et al. 2011). An understanding of the school-level monitoring system is crucial for discussing whether or not a system without national standard setting can work and if more systematic standard setting could provide some benefits also to the Finnish education system.

14.2 From Centralised Control of Content to Loose Monitoring of Outcomes

During the history of Finnish basic education there have been several major changes in how the effectiveness of education has been monitored (Varjo et al. 2016). Until 1985, the system was strongly controlled by the state and all teachers were requested to participate in extensive in-service training on the obligatory contents. National and provincial school inspection was active and all textbooks were pre-examined by the National Board of Education. The very detailed curriculum was the same for all municipalities, but there were no state-level assessments of any school subjects. Thus, the input was governed centrally, except for an unsuccessful attempt to introduce a standardised testing model in major school disciplines in the 1970s, there was little control regarding the outputs (cf. OECD 2015). The national comparability of school grades given by teachers was deemed to be secured by the detailed curriculum and intensive in-service training, and the standardised test that were being developed were considered as tools for facilitating the grading. Simultaneously, the fundamental idea was to enhance the use of formative assessments in the spirit of Bloom's taxonomy and mastery learning. Grading was 'standardised' also by the use of normative scaling, where the given percentages for different grades were inspected (for example, highest grade, 10, was restricted within one school to 3–5%) as well as the means of school grades within schools. In practice, this meant that the grades usually followed normal distribution within each class. The mean level calibration was done using the standardised tests that were being developed at that time and worked with examples in different disciplines and grade-levels. These auxiliary tools were actively used for about a decade in the 1970s but they were then omitted for financial reasons. That ended the era when population-based monitoring of learning outcomes was considered as an overall goal, even though the school inspection system was still active at that time.

An important liberation from the centralised way of steering took place in 1985, when a new National Framework Curriculum was accepted. It allowed municipalities to have local applications to follow and assess educational outcomes and marks.

Only rather general recommendations on grading were given, without an explicit move away from normative to criterion-referenced assessment. No national comprehensive evaluations were executed, with the exception of participating in the early IEA-assessments. Local constraints were further liberated in the 1994 Core Curriculum that served as an obligatory core for the local curricula written by the municipalities and schools. In most European countries, school inspection is still an important instrument for educational evaluation (Gustafsson et al. 2015), but in Finland it was now completely abolished alongside the inspection of textbooks (Aho et al. 2006). Accordingly, whereas in many countries the inspection system holds schools accountable for achievements and makes these judgements about criteria and standards (Gustafsson et al. 2015), the educational legislation reform of 1998 and the Finnish model for evaluating educational outcomes that was first introduced in the mid-1990s (National Board of Education 1999) obliged the organisers of education to assess educational outcomes mainly locally. This means that the organisers of education, in most cases municipalities, were given the main responsibility for monitoring the effectiveness of education and securing that every child has equal opportunities in proceeding through basic education. Due to the minimal segregation of schools in Finland (Willms 2010) and the high education level of all teachers (Jakku-Sihvonen and Niemi 2006), no other means, such as centrally controlled distribution of teachers across different schools, have been considered necessary so far.

The 2004 Core Curriculum reintroduced stronger national criteria for performance in different school subjects to secure pupils' fair assessment across the country (National Board of Education 2004a). The subject-specific criteria complement the general guidelines covering formative assessment during the course of studies and final assessment. All assessment is to be based on diverse evidence and has also take into account pupils' work skills, even if no indication is given for the weight of these in the final grade. It is emphasised that the final assessment should relate pupils' achievement to the objectives of the basic education syllabus through the provided criteria in order to ensure the comparability of grades across schools (National Board of Education 2004a).

The criteria for performance cannot be considered strictly as learning standards, though. Rather, they serve as a tool for teachers in their grade assignment, and with this function in mind they have been further refined in the new Core Curriculum implemented in autumn 2016 (National Board of Education 2016). These criteria have been defined for each subject for critical grade levels for 'good' performance (grade 8 in the Finnish grading scale from 4 = 'fail' to 10 = 'excellent') instead of introducing descriptions for lowest acceptable performance. There are, however, (admittedly weak) standards for 'adequate' performance, the lowest acceptable grade 5, that are provided by referring to pupils who demonstrate to 'some degree' the performance level required by the criteria of good performance (National Board of Education 2004a, 246). What the notion of 'some degree' actually entails, is not specified. Additionally, the individual criteria are not binding per se, since failing to meet some of them can be compensated for by performing adequately with regard

to the other criteria. Except for a small proportion of pupils with special educational needs, everybody is to attain the level of 'adequate' performance defined in this way.

Whether or not all pupils actually reach this 'standard of adequate performance' is not supervised on a national level. Despite the changes in the Core Curriculum, the educational assessment system has not been changed considerably – although the official administration of national monitoring of learning outcomes have been reorganised several times (Varjo et al. 2016). Therefore, still in 2016, local evaluation of education is – perhaps falsely – considered sufficient in securing the achievement of the vague standards set in the form of criteria specified in the National Core Curriculum, and the national monitoring of learning outcomes consists only of sample-based national and international assessments. In other words, national information-steering in regard to learning outcomes is entirely based on low-stakes sample-based assessments that only occasionally takes place in individual schools. The occasional municipal assessments that complement the system-level monitoring may be population-based, but they are still low-stakes for both individual pupils and schools. This means that Finnish pupils and schools never face testing situations that are related to accountability.

14.3 Sample-Based Assessment System

Educational outcomes can be evaluated and monitored centrally even if education was organised according to local curricula (The Association of Educational Assessment – Europe 2012). Beside providing information about the performance level of pupils on a comparable scale, centralised assessment can be used for gaining information about equity of learning opportunities in different geographical areas or school types, and for pupils with different backgrounds. Finland makes no exception in this: beside the strong emphasis on local monitoring of learning outcomes and progress, national indicators are needed for educational policy development at a national level. However, unlike many countries that use national tests or exams for whole age cohorts to monitor the trends of performance, in Finland the national discussions and decisions are based on sample-based assessments.

The sample-based assessment system was developed gradually after the implementation of the new education system in the 1970s. After a few years' trial of implementing national standardised tests, it took almost 20 years until the Framework for Evaluating Educational Outcomes in Finland was published in 1995 and in a revised form in 1998 (National Board of Education 1999, English translation). This is despite the fact that already shortly after the implementation of the new educational system, the heterogeneity of the pupil population in comprehensive school provoked a discussion about educability (Häyrynen and Hautamäki 1977) and it was obvious that more rigorous methods were needed for measuring equity of education. The framework divided the outcomes of education into three categories: efficiency, effectiveness, and economy. Efficiency referred to the functioning of the

educational system, effectiveness in pupil-level outcomes and economy to the successful allocation of resources. The definition of the indicators of effectiveness – the most interesting part of the framework regarding standard setting – led to two kinds of practical applications: curricular assessments in key school subjects that are complemented by national thematic assessments and information obtained from international large-scale assessments.

14.4 Curricular Assessments in Key School Subjects

As the first and the most central means of the evaluation of educational effectiveness, the 1995 framework (National Board of Education 1999) introduced sample-based national assessments to the key school subjects. These assessments are not repeated each year at pre-defined grade levels, but the school subjects and the grade levels to be assessed are defined in a 4-year plan for educational assessment instead - see Ministry of Education (2012a) for the previous plan. The assessments are implemented by the government agency, nowadays the Finnish Education Evaluation Centre (www.karvi.fi/en) that was established under the Ministry of Education in 2014 to coordinate national monitoring of learning outcomes and to replace three earlier institutions. Typically, there are two to three school subjects to be assessed, and a sample of about 5000 pupils participates in each test. The sampling procedure resembles that of the international PISA-studies (OECD 2013) with within-school samples instead of full cohorts and a sufficient geographical coverage.

The assessment tasks used in the national assessments are developed by expert groups. The experts are usually teachers working part-time in the project during the item-development period. The items of the assessment tasks are developed to cover the national subject-specific goals set in the Core Curriculum (National Board of Education 2004a, 2016), which must be included in the local curricula. Larger sets of items are developed for the field trial conducted a year before the main study, and approximately a half of the pre-tested items are selected for the final test based on their statistical properties and curriculum coverage. The aim is to develop a psychometrically sound subscale for each content area defined in the Core Curriculum. In addition to the newly developed pretested items, the main study test version includes also a small number of anchor items to link the results to the earlier assessment cycles. However, the number of anchor items is usually not sufficient to monitor trends in regard to subscales.

The sample-based assessment system serves primarily the national educational policy developers. The selected approach restricts the use of the assessment results in local decision making to a minimum unless the municipality pays for extending the assessment to all its schools – something that is also possible and used by some of the large municipalities. Accordingly, evaluating individual schools or pupils against predefined standards is not possible, and the results of these assessments are mainly used in national or in some occasions regional monitoring of equity of education. Thus, just like international studies, national assessments can for instance

demonstrate gender inequalities in different regions of the country (e.g., Kupari et al. 2013) and this is exactly how the results of sample-based assessments are reported, too. In addition to the reports about national and regional results, participating schools receive their results to be utilised in their own developmental work.

Even though no clear minimum standard for acceptable performance is defined – unlike in PISA – the national sample-based assessments have included a section for checking the bi-directional distributions for test-scores and teacher-given school grades. It has been shown (Ouakrim-Soivio 2013) that there are indeed between-school fluctuations in school grades, indicating that teachers in different schools grade their pupils somewhat differently. It seems that the differences are not random but instead they indicate a negative correlation between pupils' grades and their performance in the assessments at school level. That is, pupils in high-performing schools tend to be graded more harshly and pupils in low-performing schools more leniently. Therefore, when the Core Curriculum was recently renewed (National Board of Education 2016), closer attention was paid to create more detailed prescriptions of the grades for good performance. These prescriptions for grading pupils are closest to educational standards Finland at present has.

Beside curricular assessments that have since their introduction been implemented by national agencies, the national monitoring system includes also thematic national assessments (e.g., learning to learn, ICT use) and international assessments. The reoccurring ones are usually implemented by universities or research institutes on assignment from the Ministry of Education, nowadays based on calls for tenders. In addition, the National Educational Evaluation Centre implements thematic assessments on different topics that vary from year to year. Regarding basic education, the recent themes have ranged from curriculum evaluation to immigrant children in the Finnish school system. The majority of thematic assessments implemented by the National Educational Evaluation Centre, however, focuses on pre-primary, vocational or higher education and is therefore outside the scope of this description.

The more or less regularly implemented non-curricular assessments during basic education can be divided into two categories: the national learning to learn assessments and the international assessments (PISA, TIMMS and PIRLS). Learning to learn was defined as one of the measurable indicators of the effectiveness of education in the 1995 Framework (National Board of Education 1999). It is defined as comprising general cognitive competences (thinking and reasoning skills in different contexts, problem solving, reading comprehension) that are needed in all learning and motivational beliefs which support the effective use of the competences (Hautamäki et al. 2010; Hautamäki and Kupiainen 2014). The model was created in 1996 and developed further during the following 7 years along with nationally representative large-scale assessment studies at the end of the sixth and the ninth grade and in both tracks of upper secondary education (Hautamäki et al. 2013). At the same time, the topic was also intensively discussed at European level, and learning to learn was later defined as one of the key competences for lifelong learning (Recommendation 2006/962/EC of the European Parliament and of the Council of 18 December 2006). As a result, the Finnish scales formed a substantial part of the

European learning to learn instrument piloted in eight countries in 2008 (Kupiainen et al. 2008). Nowadays, learning to learn assessments are implemented mainly on a computer-based platform. National assessments are conducted only occasionally at the end of the ninth grade (Hautamäki et al. 2013), but the method is extensively used by some of the largest municipalities in Finland, which have launched longitudinal panel studies (Vainikainen 2014) covering full age cohorts from the first grade on. These municipalities use the assessment results for providing their schools tools for identifying their strengths and weaknesses for developmental work.

The results of national learning to learn assessments, as well as the international assessments PISA and the more irregularly conducted TIMMS and PIRLS, have shown to be the most useful tools for monitoring the unfortunately declining trends of performance over the years (Hautamäki et al. 2013). The next challenge is to develop the curricular assessment system towards a direction that would enable more detailed analyses of performance trends.

14.5 Local Ways of Monitoring Equity and Functioning of Education

14.5.1 Why can We Trust the Local Authorities and Practitioners?

It might sound puzzling that between-school differences in Finland are among the smallest in the world (Hautamäki et al. 2008; OECD 2007; Willms 2010), even though there are neither any mandatory national standardised tests in basic education nor any kind of school inspectorate system[1]. Moreover, the quality of education has proved to be excellent especially in regard to the weakest learners (e.g., Kupari et al. 2012a, b) who outperform comparable groups in other countries. This is despite the fact that schools and teachers do not experience pressuring external top-down control.

In order to understand the functioning of the local ways of monitoring the equity and functioning of the education, one must realise what is behind the ability, power and possibilities of the local authorities and educational professionals. Some essential pieces of the puzzle can be clearly identified: listening to people in the field, collecting evidence for development, not for judgment, and giving free space for the teachers (Berg 1999; Berg and Wallin 1983). In the following list, these are discussed in a more detail:

[1] Unlike what has erroneously been stated about standardised test in Finland in an article of Andreasen and Hjörne (2014): 'National standardised external testing was introduced in 1998 and is monitored approximately two times during compulsory School for All pupils. Test results are being used both internally and externally, and especially their external use must be expected to have a profound impact on pedagogical practice.'

(a) *Renewing the educational norms and guidelines is a joint enterprise* – which is in contrast with the practice in many other countries (cf. Angus 2011). For example, the National Board of Education has engaged different stakeholders – teacher unions, parent and other third sector organisations, municipalities, schools and teachers – in the process of renewing the National Core Curriculum (National Board of Education 2016). The national reform of special education was organised similarly (Ministry of Education 2007; Thuneberg et al. 2014). These kinds of process reflect *a way of governance* rather than *government* (Forrest 2003). As second example, relating to the Core Curriculum of 2004 (National Board of Education 2004a), is that broad descriptions were set for the report card grade eight – the grade for "good performance". These descriptions served as a reflection surface, on which the teachers and schools could mirror their grading system. By no means were these descriptions supposed to be considered as standards, but as a way of assistance for grading in different school subjects. A third example, serving the local authorities, schools and teachers, is the "Criteria of quality" that were launched as *recommendations*, not as *norms* (Ministry of Education 2009, 2012b). These criteria first describe the quality of structures from the perspectives of leadership, personnel, evaluation and economic resources. The second part concentrates on the quality experienced by the pupil, covering different aspects from the implementation of the curriculum, to the possibilities of participation, home-school cooperation and support for well-being and learning.

(b) *Sample-based evaluations are the main instruments of knowing what is going on in the schools*. An example of this are the sample-based national assessments of ninth grade mathematics in 1998–2004 (National Board of Education 2004b). These evaluations are not conducted in order to categorise the schools into poor, better or good schools, or the pupils to less or more talented. Instead, they are elements of information steering. The aim is to provide information for the municipalities, schools and teachers on how to change and fine-tune the organisation of education and schooling and enhance more effective teaching.

(c) *Professional teachers matter*. They have earned their free space, autonomy, by showing accountability and trustworthiness, which is at the core of the research-based teacher education (Jakku-Sihvonen and Niemi 2006; Toom et al. 2010) and the master's degree as a qualification. There is evidence from various sources that a key determinant of pupil performance is the teacher quality (Gustafsson 2013), although the found effect-sizes are varying (Hattie 2009; Scheerens 2014). The quality is enhanced by teacher education, which combines research and practice and supports teachers to comprehend and develop curriculum and make evidence-informed decisions (cf. Tryggvason 2009; Tatto 2015). The Finnish teacher education takes the main components of research-based teacher education into account:1) the study program is structured on the basis of systematic analysis of education, 2) all teaching is research-based, 3) education is organised to support pedagogical problem solving by argumentation, decision-making and justification, and 4) it teaches research methods

(Toom et al. 2010). In addition, yearly in-service training is included in the collective bargaining contract, and that training can further support teachers in applying research and evidence-informed principles, methods and learning tools.

14.5.2 What can the Local Authorities and Practitioners Do to Enhance Equity and Functioning of Education?

The steering system functions by providing information, being multi-directional in nature. Local school authorities, teachers and other school personnel can engage in work and development without losing their independence and autonomy. Thus, they might more willingly integrate external goals and work also on the basis of their intrinsic motivation. However, they also need structures and scaffoldings, on which to rely. An important new structure is the three-tiered support model.

The three-tiered model was established by the reformed Basic Education Act (2010), but since the Special education strategy (Ministry of Education 2007), the municipalities were actively involved in the development of the organisation and implementation of the model and more inclusive practices in a tight cooperation with the universities and the National Board of Education. The model led to a higher systemisation of the activity by dividing the support in general, intensified and special support, of which only the last one requires an official decision. The assessment methods, support practices, monitoring of the effects, professional roles and documentation became more organised than before. The main principles of the system are that there is a strong emphasis on local schools, that the practices are flexible and that teaching is a shared rather than an individual enterprise (Thuneberg et al. 2014).

14.6 Pupil Welfare Work

Organising educational support in a way that guarantees equity in regard to how the system responds to pupils' needs would be difficult for the local authorities and teachers if they did not have appropriate structures guiding the provision of support. Local authorities are too distant from the everyday school work, the teachers too close to see the relativity of the needs and resources. That is why the responsibility lies with the multi-professional pupil welfare group working in every school (Sabel et al. 2011; Vainikainen et al. 2015). The constitution of the group varies depending on the issue to be discussed and to some extent also depending on the geographical region. Often there is the school principal, a special education teacher, a school psychologist, a social worker, a school nurse, and the teacher of the pupil or the class in question – see Vainikainen et al. (2015) for a detailed description of the work and an analysis of the regional differences. The group that typically meets bi-weekly makes hypotheses, plans and organises interventions and their follow-up,

divides roles and documents the discussion and the plans. The people involved in teaching or supporting of the pupil or class in question collaborate with the pupil and the guardians, considering support according to the three-tiered support model.

14.7 Part-Time Special Education

Part-time special education is an integral part of both pupil welfare work and teaching. It has also been an essential component of the local monitoring system since the early years of basic education. Part-time special education is provided by special education teachers who do not have a class of their own. Instead, they are increasingly working as co-teachers in addition to working with small groups of pupils in a separate classroom. Sometimes they may even teach or assess only one pupil at the time, if necessary. All pupils have access to the special education teacher's services without official decisions about support, and they systematically monitor the progress of pupils in every class to identify potential problems at an early stage. The very essential part of their role is to act as consultants for the class- and subject-teachers (Thuneberg et al. 2013.)

14.8 Screening of Support Needs

The practical tools of the three-tiered model are the *assessments* (pedagogical assessment and pedagogical evaluation) and the *plans* (learning plan and individual plan of organisation of learning). These have to be documented and monitored regularly. Before moving to the second tier intensified support, pedagogical assessment is conducted in multi-professional collaboration and a learning plan is written to document the actions and interventions to be implemented. Before an official decision about special support can be made, a more thorough pedagogical evaluation – a educational case formulation - is written by the multi-professional team, and an individual plan of organisation of learning in special support is created to serve as a basis for organising support. All the documents are prepared together with the pupil and the guardians.

Several types of screening methods are used in Finnish schools to identify support needs regarding learning or social and emotional challenges at an early stage. The first and the self-evident one is teacher observation, which is effective if done systematically by making field notes. It can be supported by co-teaching arrangements if they are organised so that another teacher can sometimes work as an observer instead of actually teaching. Discussion with the pupil and the guardians is essential. The usual summative tests that are extensively used in Finnish schools in all school subjects for all pupils offer valuable but restricted information about potential gaps in learning and progress, and they should be complemented with

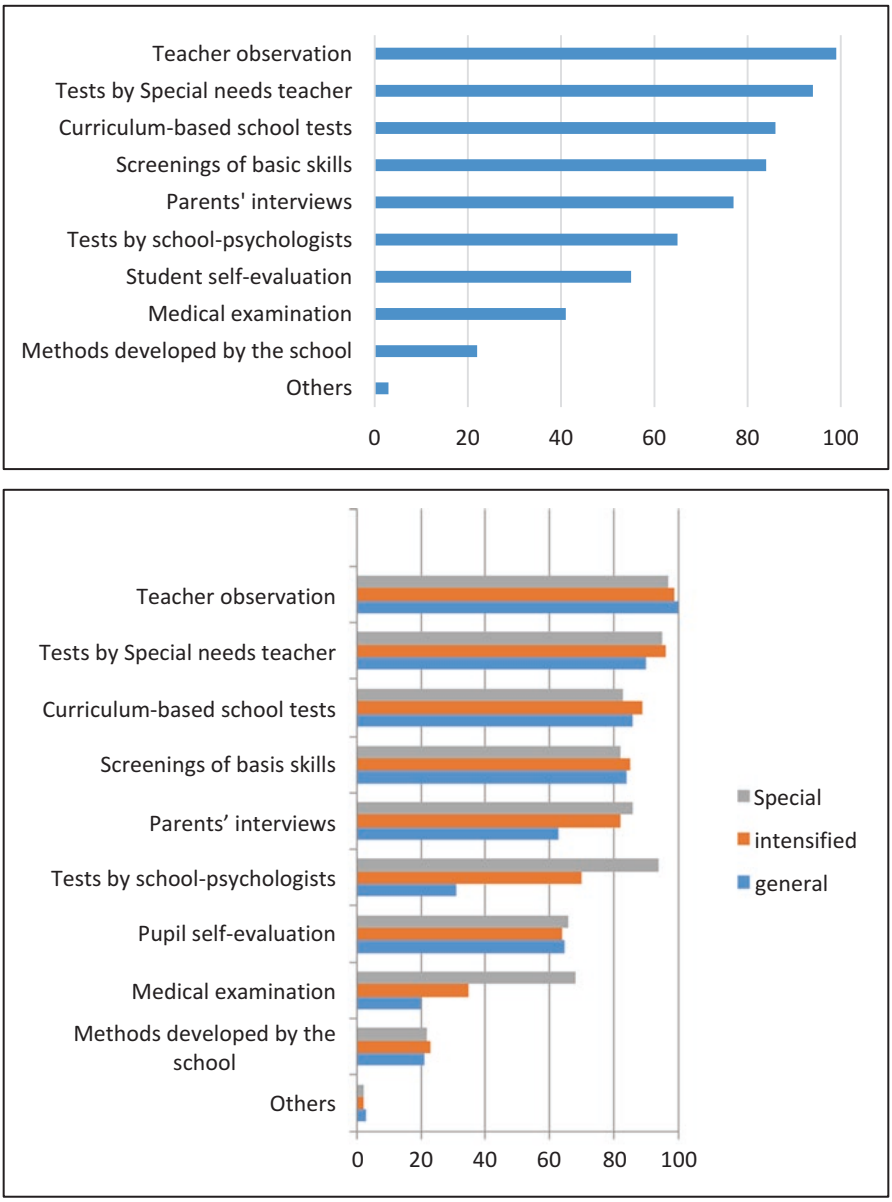


Fig. 14.1 Methods for identifying support needs in Finnish schools (upper bar: special, middle bar: intensified, lower bar: general support)

other measures. Therefore, screening tools and tests implemented by a special education teacher and sometimes also the school psychologist are needed, too. Figure 14.1 presents the main ways in which the identification methods of support

needs are used by Finnish schools according to a nationally representative principal questionnaire from 2012 (N = 1113; see Vainikainen et al. 2015 for a description of the questionnaire and the respondents). They are presented in descending order from the most used to the least used method in the stages of special, intensified and general support.

As shown in the figure, according to the principals, pupil self-evaluation was surprisingly rarely used as a method for identifying support needs. As expected, medical examination and psychological testing were mostly applied in the special support stage when the problems are more severe. Furthermore, according to the principles of the three-tiered model, expertise of teachers was more prominent than psycho-medical approaches. It has to be noted, however, that by presenting these identification methods the perspective might get skewed as the current practices aim more at applying the systems ecological theory (see Reschly et al. 2007; Thuneberg et al. 2013) rather than solely focussing on an individual child.

On the system level, the municipalities may gather data deriving from screenings conducted by special education teachers for example for monitoring how the reading or math results have developed in schools year by year. For instance, many municipalities use the reading test ALLU (Lindeman 1998) to screen second graders' technical reading and reading comprehension skills. These data are valuable for organising support for the schools, classes and individual pupils, and they can sometimes be used as means of positive discrimination by providing extra resources where needed. The municipalities also have the freedom to choose which screening methods they use systematically – or to let their special education teachers choose all their methods individually. However, it must be stressed that the results of the schools are not public, the schools cannot be identified, and they strictly only serve educational purposes.

14.9 Evaluating the Effectiveness of Local Ways in Equity and Functioning of Education

It is a justified question to ask: When the teachers are given *free space*, there are no national standardised tests in basic education, and the higher authorities provide mainly guidelines and steering by information – how can we trust that there are no schools that are hiding their poor results and drop-outs and continue their shady business in secret? We partly answered that by referring to the PISA-results: the achievement outcomes speak for themselves. However, a profound source of the secret of trustworthiness of teachers roots in the autonomy supporting atmosphere of schools, which encourages pupils and teachers to share responsibility and resiliently finish tasks and complete one's duties. Another important source is the fact that there is a long-standing cultural tradition in Finland that parents, and society as a whole, highly value education and learning.

The local ways of monitoring equity and functioning in education show a decrease in unnecessary referrals to special support - the needs are identified and met early enough to avoid that. This is apparent in the seamless continuation of the support in situations, where a pupil moves to another school, maybe in another municipality or a different part of Finland. Most importantly, they are realised in pupils' and parents' experience of participation and their general satisfaction in the provided education and support.

14.10 Summary and Conclusions

The fundamental topic of this chapter has been to describe how the Finns know what is going on in the educational system without standardised testing and national inspection. We believe that the core is in whether or not to trust school grades assigned by teachers. If comparability of grades could be assumed, there would be no need to introduce any population-based national testing or inspectorate. If there are doubts, at least a need for standard setting is established. Indeed, there have been concerns in Finland too, and some kinds of educational standard have been drawn up in the latest reform of Core Curriculum that came into effect in autumn 2016 (National Board of Education 2016). No national testing is currently planned to see how schools meet the demands of the new Core Curriculum, but it is clear that even a relatively well-functioning system would benefit from a slightly more systematic approach to standards to secure the equity of grades that pupils need to apply for upper secondary education. However, we believe that in the Finnish context, it is highly important to continue to let the teachers have control over local curriculum implementation and the selection of their teaching and assessment methods, instead of steering their work excessively through external control.

Since the 1990s, the basic national-level monitoring tools in Finland have been sample-based assessments in school disciplines, thematic evaluations, and tendered university assessments in learning-to-learn and international large-scale surveys like PISA and TIMMS. Beside other ways of reporting the results, most of these are also used to monitor the distributions of school grades in the country. Grading is a fundamental issue for schooling, not only as feedback on scholastic achievement and social adaption, but also as a topic of educational fairness, because in Finland the selection to academic and vocational schools (c. upper secondary education) takes place with school leaving credentials.

It remains to be seen how the new rules and assessment standards specified in the 2016 Core Curriculum will work. If fluctuations will continue, some other standards and new systems of assessment are bound to be introduced.

References

Aho, E., Pitkänen, K., & Sahlberg, P. (2006). *Policy development and reform principles of basic and secondary education in Finland since 1968*. Washington, DC: The World Bank.

Andreasen, K., & Hjörne, E. (2014). Assessing children in the Nordic countries: Framing, diversity and matters of inclusion and exclusion in a school for all. In U. Blossing, G. Imsen, & L. Moos (Eds.), *The Nordic education model*, *Policy implications of research in education* (Vol. 1, pp. 155–172). Dordrecht: Springer.

Angus, L. (2011). Teaching within and against the circle of privilege: Reforming teachers, reforming schools. *Journal of Educational Policy, 27*, 231–251.

Antikainen, A. (2006). In search of the Nordic model in education. *Scandinavian Journal of Educational Research, 50*(3), 229–243.

Basic Education Act. (2010). http://www.finlex.fi/en/laki/kaannokset/1998/en19980628.pdf. Retrieved June 2016.

Berg, G., & Wallin, E. (1983). Research into the school as an organization. III: Organizational development in schools or developing the school as an organization? *Scandinavian Journal of Educational Research, 27*, 35–47.

Berg, G. (1999). *Skolkultur—nyckeln till skolans utveckling. En bok för skolutvecklare om skolans-styrning*. Göteborg: Gothia.

Forrest, J. B. (2003). Networks in the policy process: An international perspective. *International Journal of Public Administration, 26*, 591–607.

Gustafsson, J.-E. (2013). Selection and evaluation of teachers in Finland and Sweden. http://slide-player.com/slide/2974824/. Retrieved June 2016.

Gustafsson, J. E., Ehren, M. C. M., Conyngham, G., McNamara, G., Altrichter, H., & O'Hara, J. (2015). From inspection to quality: Ways in which school inspection influences change in schools. *Studies in Educational Evaluation, 47*, 47–57.

Hattie, J. (2009). *Visible learning, a synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

Hautamäki, J., & Kupiainen, S. (2014). Learning to learn in Finland. Theory and policy, research and practice. In R. Deakin Crick, C. Stringher, & K. Ren (Eds.), *Learning to learn. International perspectives from theory and practice*. London/New York: Routledge.

Hautamäki, J., Harjunen, E., Hautamäki, A., Karjalainen, T., Kupiainen, S., Laaksonen, S., et al. (2008). *PISA06: Analyses, reflections, explanations*. Helsinki: Ministry of Education.

Hautamäki, A., Hautamäki, J., & Kupiainen, S. (2010). Assessment in schools – learning to learn. *International Encyclopedia of Education, 2010*(3), 268–272.

Hautamäki, J., Kupiainen, S., Marjanen, J., Vainikainen, M.-P., & Hotulainen, R. (2013). *Oppimaan oppiminen peruskoulun päättövaiheessa: Tilanne vuonna 2012 ja muutos vuodesta 2001* [Learning to learn at the end of basic education: The situation in 2012 and the change from 2001]. University of Helsinki, Department of Teacher Education, Research Reports 347. Helsinki: Unigrafia.

Häyrynen, Y.-P., & Hautamäki, J. (1977). *Människans bildbarhet och utbildningspolitiken*. Stockholm: Wahlström & Widstrand.

Jakku-Sihvonen, R., & Niemi, H. (Eds.). (2006). *Research-based teacher education in Finland: Reflections by Finnish teacher educators*, *Research in Educational Sciences* (Vol. 25). Turku: Finnish Educational Research Association.

Kupari, P., Sulkunen, S., Vettenranta, J., & Nissinen, K. (2012a). *More joy for learning. Fourth graders' reading skills and competences in mathematics and science: International PIRLS and TIMMS-assessments in Finland [In Finnish]*, *Koulutuksen tutkimuslaitos*. Jyväskylä: Jyväskylän yliopistopaino.

Kupari, P., Vettenranta, J., Nissinen, K. (2012b). *Searching for learner-centered pedagogy. Learning results of the 8th grade students in mathematics and science. International TIMMS-study in Finland [In Finnish].* Koulutuksen tutkimuslaitos. Jyväskylä: Jyväskylän yliopisto-paino. Analyses, reflections, explanations. Publication No. 44. Helsinki: Ministry of Education.

Kupari, P., Välijärvi, J., Andersson, L, Arffman, I., Nissinen, K., Puhakka, E., Vettenranta, J. (2013). *PISA 12 Ensituloksia* [PISA 2012: The first results]. Opetus- ja kulttuuriministeriön julkaisuja 2013:20.

Kupiainen, S., Hautamäki, J., Rantanen, P. (2008). *EU pre-pilot on learning to learn: report on the compiled data*. 2008-1190/001-001 TRA-TRINDC.

Lindeman, J. (1998). *Ala-asteen lukutesti ALLU*. Turku: Turun yliopisto, Oppimistutkimuksen keskus.

Ministry of Education. (2007). *Erityisopetuksen strategia* [Special education strategy]. Reports of the Ministry of Education 2007:47. Helsinki: Ministry of Education.

Ministry of Education. (2009). *Perusopetuksen laatukriteerit* [Quality criteria for basic education]. Opetusministeriön julkaisuja 2009:19.

Ministry of Education. (2012a). *Koulutuksen arviointisuunnitelma vuosille* 2012–2015 [The plan for educational assessment for years 2012 to 2015]. Opetus- ja kulttuuriministeriön julkaisuja 2012:14.

Ministry of Education. (2012b). *Perusopetuksen laatukriteerit* [Quality criteria for basic education]. Opetus- ja kulttuuriministeriön julkaisuja 2012:29.

National Board of Education. (1999). *A framework for evaluating educational outcomes in Finland*. National Board of Education, Evaluation 8/1999.

National Board of Education. (2004a). *National Core Curriculum for basic education 2004*. http://www.oph.fi/english/publications/2009/national_core_curricul_for_basic_education. Retrieved 1 June 2011.

National Board of Education. (2004b). *Summary of four national assessments of mathematics learning in the 9th grade of basic education*, 1998–2004. http://www.oph.fi/download/47697_4_matikkaa_englanniksi.pdf. Retrieved 16 Mar 2016.

National Board of Education. (2016). *Curriculum reform 2016. Renewal of the core curriculum for the pre-primary and basic education*. http://www.oph.fi/english/education_development/current_reforms/urriculum_reform_2016. Retrieved 10 Apr 2016.

OECD. (2007). *Science competencies for tomorrow's world: Results from PISA 2006*. Paris: OECD.

OECD. (2013). *PISA 2012 assessment and analytical framework*: *Mathematics, reading, science, problem solving and financial literacy*. Paris: OECD Publishing. http://dx.doi.org/10.1787/9789264190511-en. Retrieved June 2016.

OECD (2015). *Education at a glance* 2015. OECD Indicators. Paris: OECD Publishing. http://dx.doi.org/10.1787/eag-2015-en. Retrieved June 2016.

Ouakrim-Soivio, N. (2013). *Toimivatko päättöarvioinnin kriteerit? Oppilaiden saamat arvosanat ja Opetushallituksen oppimistulosten seuranta-arviointi koulujen välisten osaamiserojen mittareina.* [Do the national criteria for students' final grades work? Teachers' grading and national assessments as indicators for between-school differences]. Helsinki: Opetushallitus.

Reschly, A., Coolong-Chaffin, M., Christenson, S., Gutkin, T. (2007). Contextual influences and response to intervention: Critical issues and strategies. In S. R. Jimerson, M. K. Burns, & A. M. Van Der Heyden (Eds.), *Handbook of response to intervention. The science and practice of assessment and intervention.* New York: Springer. Recommendation 2006/962/EC of the European Parliament and of the Council of 18 December 2006 on key competences for lifelong learning. Official Journal L 394 of 30.12.2006.

Sabel, C., Saxenian, A., Miettinen, R., Kristensen, P. H., & Hautamäki, J. (2011). *Individualized service provision in the new welfare state: Lessons from special education in Finland* (p. 62). Helsinki: Sitra Studies.

Scheerens, J. (2014). Teacher training and professional development in Europe; In search of effects on educational performance. Keynote presentation at the German Rector's conference, 20–23 January, Esen. https://www.hrk-nexus.de/fileadmin/redaktion/hrk-nexus/01Logos-Hochschulen/EssenPresentation_Scheerens.ppt. Retrieved June 2016.

Tatto, M. (2015). The role of research in the policy and practice of quality teacher education: An international review. *Oxford Review of Education, 41*(2), 171–201.

Telhaug, A. O., Mediås, O. A., & Aasen, P. (2006). The Nordic model in education: Education as part of the political system in the last 50 years. *Scandinavian Journal of Educational Research, 50*(3), 245–283.

The Association of Educational Assessment – Europe. (2012). European *framework of standards for educational assessment* 1.0. http://www.aeaeurope.net/images/downloads/SW_Framework_of_uropean_Standards.pdf. Retrieved 4 Mar 2013.

Thuneberg, H., Vainikainen, M.-P., Ahtiainen, R., Lintuvuori, M., Salo, K., & Hautamäki, J. (2013). Education is special for all – The Finnish support model. *Gemeinsam leben, 21*(2), 67–78.

Thuneberg, H., Hautamäki, J., Ahtiainen, R., Lintuvuori, M., Vainikainen, M.-P., & Hilasvuori, T. (2014). Conceptual change in adopting the nationwide special education strategy in Finland. *Journal of Educational Change, 15*(1), 37–56.

Toom, A., Kynäslahti, H., Krokfors, L., Jyrhämä, R., Byman, R., Stenberg, K., & Kansanen, P. (2010). Experiences of research-based approach to teacher education: Suggestions and future policies. *European Journal of Education, 45*(2), 331–344.

Tryggvason, M.-T. (2009). Why is Finnish teacher education successful? Some goals Finnish teacher educators have for their teaching. *European Journal of Teacher Education, 32*(4), 369–382.

Vainikainen, M.-P. (2014). *Finnish primary school pupils' performance in learning to learn assessments: A longitudinal perspective on educational equity*. University of Helsinki, Department of Education Research Reports 360. Helsinki: Unigrafia.

Vainikainen, M.-P., Thuneberg, H., Greiff, S., & Hautamäki, J. (2015). Multiprofessional collaboration in Finnish schools. *International Journal of Educational Research, 72*, 137–148.

Vainikainen, M.-P., Hienonen, N., Lindfors, P., Rimpelä, A., Asikainen, M., Hotulainen, R., & Hautamäki, J. (2016). Oppimistuloksia ennustavat tekijät Helsingin metropolialueen yläkouluissa [Factors explaining the development of learning outcomes in the lower secondary schools of the Helsinki metropolitan area]? *Kasvatus, 47*(3), 214–229.

Varjo, J., Simola, H., Rinne, R. (2016). *Arvioida ja hallita. Perään katsomisesta informaatio-ohjaukseen suomalaisessa koulupolitiikassa* [To evaluate and govern: From "looking after" to steering by information" in Finnish education policy]. Publications of the Finnish Educational Research Association, 70. Jyväskylä: Jyväskylän yliopistopaino.

Willms, J. D. (2010). School composition and contextual effects on student outcomes. *Teachers College Record, 112*(4), 1007–1037.

Part III
New Methodological Approaches to Standard-Setting

Chapter 15
The Data-Driven Direct Consensus (3DC) Procedure: A New Approach to Standard Setting

Jos Keuning, J. Hendrik Straat, and Remco C.W. Feskens

Abstract Various procedures for establishing performance standards have been proposed in the literature. Among the best-known examples are the *Angoff procedure*, the *Bookmark procedure* and the *Direct Consensus procedure*. These procedures have their strengths and weaknesses. Some procedures make it possible to establish performance standards relatively efficiently and quickly, but lack empirical rigor. Other procedures do include empirical data, but are time consuming and not very intuitive. In the present study, the strengths of the aforementioned standard setting procedures were brought together in a new one: the *Data-Driven Direct Consensus (3DC) procedure*. The 3DC procedure divides the complete test into a number of clusters and uses (unlike *Direct Consensus*) empirical data and an item response model to relate the scores of the clusters to the scores of the complete test. The relationships between the clusters and the complete test are presented to the subject-area experts on a specially designed assessment form. Subject-area experts are asked to use the assessment form to indicate the score that students would be expected to achieve in each cluster if they were exactly on the borderline of proficiency. Because of the design of the assessment form, the assessment is easily allowed to be based on both content information and empirical data. This is an important difference with *Direct Consensus* as empirical information is less explicit within this procedure.

Keywords Angoff • Bookmark • Direct consensus • Empirical data • Standard setting

J. Keuning (✉) • J.H. Straat • R.C.W. Feskens
Cito, Psychometric Research Center, Arnhem, The Netherlands
e-mail: Jos.Keuning@cito.nl

S. Blömeke, J.-E. Gustafsson (eds.), *Standard Setting in Education*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-50856-6_15

15.1 Introduction

Both in psychology and education, norms are an essential prerequisite for understanding raw test scores (e.g., Linn 2000; Downing and Haladyna 2006). Traditionally, norms are derived from the distribution of raw test scores of a reference group. By means of percentiles, stanines or normal curve equivalents students are ranked with respect to how other students perform on the same test. This is called norm-referenced interpretation. Alternatively, norms may be derived from a domain of skills or subject matter to be mastered. By means of a pre-set performance standard for expected achievement it is determined how well the student performs on a test regardless of how anyone else does. This is called criterion-referenced interpretation. Over the years, the latter has become increasingly important in educational measurement (Cizek and Bunch 2007). Think for instance of the implementation of the *Common European Framework of Reference for Languages* (CEFR; Council of Europe 2001), which is now widely accepted as a standard for students' language proficiency. Performance standards are needed to decide upon the student's language level, ranging from breakthrough (A1) to mastery (C2). It is important to legitimise the performance standards that are used for separating students into performance categories. A performance standard should preferably be developed methodically, such that subject-area experts, teachers and students clearly understand the manner in which the performance standard was determined (American Educational Research Association, American Psychological Association and American Council on Measurement in Education 1999). For this purpose, numerous procedures for establishing performance standards have been developed (e.g., Cizek and Bunch 2007; Hambleton and Pitoniak 2006; Zieky et al. 2008). In the present study, we combine the strengths of existing standard setting procedures into a new methodology. The methodology is introduced and then illustrated with a practical example.

15.2 Setting Performance Standards

Subject-area experts play an important role in the development of performance standards. They give an opinion on the expected behaviour of students located exactly on the borderline of proficiency. Standard setting procedures formally guide the subject-area experts in developing new performance standards. A distinction can be made between *test-centered* and *examinee-centered* standard setting methods (Jaeger 1989; Kaftandjieva 2004; Berk 1986; Hambleton et al. 2000). In test-centered methods, subject-area experts base the performance standard on the content of the test and on the learning materials. The performance standard is independent of the testing results that students actually achieve. In examinee-centered methods, the subject-area experts base the performance standard on the work of students. The performance standard thus depends upon the achievements of the tested students. There are no clear criteria for choosing between a test-centred or an examinee-centered method

(Kane 1998). However, test-centered methods tend to be best suited for relatively straightforward (multiple-choice) assessments and examinee-centered methods for complex educational assessments involving multiple scoring rules (Kaftandjieva 2004). The method of the present study is test-centered.

Among the best-known examples of test-centered methods are the *Angoff procedure*, the *Bookmark procedure* and the *Direct Consensus procedure*. In the Angoff procedure, subject-area experts are asked to take in mind a student who is on the borderline of proficiency (Angoff 1971). They have to estimate the minimally competent student's probability of answering the test items correctly. Several modifications to the original Angoff method were proposed. Whereas the original Angoff procedure is, for instance, suited to set standards on multiple-choice assessments, the extended Angoff method was developed to set standards on constructed-response assessments containing *open-response items* (Hambleton and Plake 1995). Other suggested Angoff modifications include the use of an iterative procedure with feedback and discussion between the cycles and the adaptation of the subject-area experts' judgements into simple yes/no statements instead of probabilities of answering correctly (e.g., Busch and Jaeger 1990; Hambleton and Plake 1995; Jaeger 1978; Woehr et al. 1991).

In the Bookmark procedure, the test items are ordered from easy to difficult by using empirical information about student performance (Lewis et al. 1996, 1999). Subject-area experts are then asked to page through the ordered test items and set a 'bookmark' between the most difficult test item a minimally competent student masters and the easiest test item a minimally competent student does not master. In the Direct Consensus procedure, finally, the total test is organised into clusters based on content considerations (Sireci et al. 2000). For each cluster, subject-area experts indicate how many of this subset of test items a minimally competent student is expected to answer correctly. The passing scores of the different clusters are then summed to obtain a performance standard on the complete test. The subject-area expert report on the actual test scale (direct) and the explicit goal is to have the panel arrive at a performance standard that they can agree upon (consensus).

15.3 Methodological Issues

The different standard setting methodologies have their strengths and weaknesses. The Angoff procedure is probably the most widely used standard setting technique. The technique is relatively straightforward and generally provides acceptable results in many different situations (Berk 1986). However, critics argue that the Angoff procedure places too high cognitive demands on subject-area experts (Berk 1986; Impara and Plake 1997). The estimation of probabilities is too abstract and subject-area experts show a tendency to gradually rate the test item probabilities higher (or lower) over the course of the standard setting session. Moreover, Angoff focuses on evaluations at the level of the test item, which is particularly time-consuming. Conversely, application of the Bookmark procedure takes less time, and probably

requires less cognitive effort than Angoff. Because the Bookmark procedure uses empirical evidence to estimate item difficulty and orders the items accordingly, subject-area experts can focus on content. However, the Bookmark procedure becomes hard to apply and interpret in case of polytomous item scores (Karatonis and Sireci 2006) and sometimes subject-area experts experience difficulties in performing the task (Karatonis and Sireci 2006; Reckase 2006). The procedure does not allow the subject-area experts, for instance, to make distinctions between test items above and under the passing point while the description of the different performance levels could indeed call for such a distinction (Cizek 2001). The Direct Consensus procedure is relatively new. The procedure is designed to '… improve upon some of the perceived shortcomings of the Angoff method and to give subject-area experts more direct control in recommending where the passing score is set' (Sireci et al. 2004). Subject-area experts indeed find the Direct Consensus method more readily understandable and more time-efficient than Angoff (Pitoniak et al. 2002; Sireci et al. 2004), but empirical information is not explicitly incorporated into the process. This is a potential disadvantage of the procedure (Cizek and Bunch 2007).

15.4 The Present Study

In the present study, it was attempted to combine the strengths of existing standard setting procedures into a new methodology. The new method, which is called the *Data-Driven Direct Consensus* (3DC) procedure, aims to combine the flexibility of Angoff, the empirical rigour of Bookmark and the clarity and efficiency of Direct Consensus. The method is positioned as a variation of Direct Consensus. The method basically adds the use of empirical data to this procedure. We first introduce the newly developed technique in more detail. We describe the rationale behind the technique and show how empirical information on student performance can be presented to subject-area experts in a cohesive and easily understandable manner. We then illustrate the technique by presenting an example from a standard setting meeting that was recently conducted (Feskens et al. 2014). The one-week meeting aimed at linking the different language levels of the CEFR to the Dutch national exams. Subject-area experts set CEFR performance standards on exams and tests measuring reading and listening comprehension of English, French and German. A total of 24 performance standards were set during this meeting using the method of the present study. We end with a discussion.

15.5 Data-Driven Direct Consensus

The Data-Driven Direct Consensus (3DC) procedure assumes a test to consist of multiple items that can be divided into a number of (content-related) clusters. In reading and listening tests the clusters could, for instance, consist of items relating to

the same text or the same audio/video files. Just as in the Direct Consensus procedure, the subject-area experts are asked to indicate the score that minimally competent students would be expected to achieve in each cluster. However, the 3DC procedure adds the explicit presentation of student performance on the test items to the subject-area experts. In this way, the subject-area expert can also take into account the relative difficulty of test items in their assessments. Figure 15.1 illustratively shows an assessment form filled in by one subject-area expert. For each cluster (here, lines 1 through 6), the subject-area expert indicates the number of items a minimally competent student is expected to answer correctly. In the assessment form, empirical information is used to relate the scoring scales of the separate clusters to the scoring scale of the complete test (bottom line). If a student's raw score on the first cluster is 4, for instance, the student would be expected to achieve a raw score of 27 on the test. The relation can also be reversely interpreted. If a student's raw score on the test is 20, we would expect to find a raw score of 4 on the fifth cluster.

15.6 Prediction Model

To build the assessment form of Fig. 15.1, we need a model that relates the clusters to the complete test. An attractive option is to use a model from item response theory. Item response theory models offer excellent possibilities for predicting how students with a particular ability level will score on subsets of items from the total

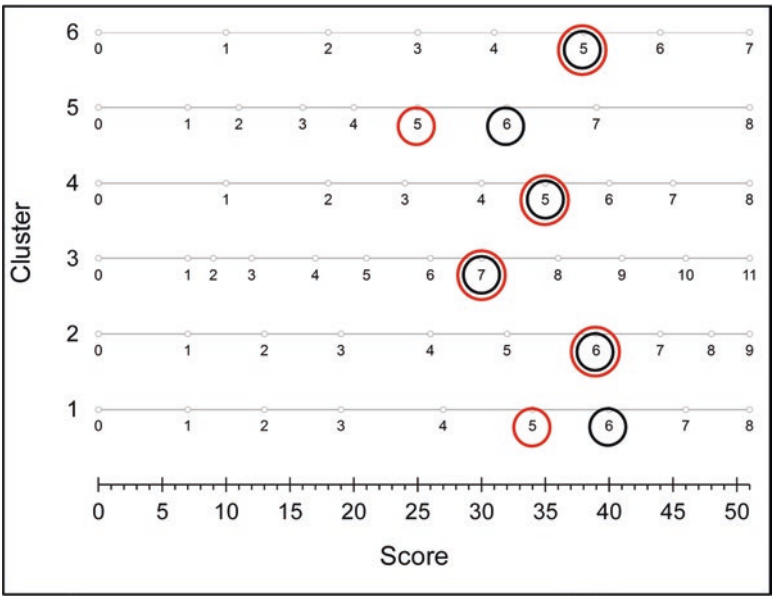


Fig. 15.1 Example of a completed assessment form

test. Different item response theory models have been proposed in the literature. In the present study, we use the One-Parameter Logistic Model (OPLM) for dichotomously-scored test items. The procedure can be extended to include polytomously scored test items, but for the purpose of a clear presentation we only discuss the dichotomous case. The 3DC procedure can also be performed with any other item response theory model or even with other types of prediction models. For a detailed description of the fundamental concepts and practical applications of the OPLM and item response theory in general, see, for instance, Hambleton et al. (1991), Van der Linden and Hambleton (1997), Embretson and Reise (2000) and Verhelst and Glas (1995). The item response function for the OPLM is

$$\mathrm{P}(x_j = 1|\theta) = \frac{\exp[a_j(\theta - \beta_j)]}{1 + \exp[a_j(\theta - \beta_j)]},$$

where θ represents the ability of a student, $a_j > 0$ is an in-advance specified discrimination index for item j, β_j represents the item difficulty, and x_j is a random variable with a value of 0 or 1. As can be seen, the model presents the probability of a correct answer ($x_j = 1$) to test item j with discrimination index α_j and difficulty parameter β_j as a function of θ. With the model it is thus possible to predict how an individual student or a group of students with a given ability level can be expected to perform on a (set of) test item(s).

Once the item response functions are estimated, simulation techniques can be used to calculate the expected cluster scores for each possible score on the complete test. In the first step, we draw N possible values for θ from a normal distribution with a mean of μ and standard deviation of σ. For N we ideally choose a large number (e.g., 100,000). The mean and standard deviation can be deduced from the sample that was used to estimate the item response functions. In the second step, item responses are generated for all θ-values. We draw a random number g from the interval [0,1] and we then evaluate the OPLM. If for item j holds that, $P(x_j = 1|\,\theta) \geq g$, the item is scored as 'correct'; if not, the item is scored as 'incorrect'. In the third step, we use the item responses to calculate an expected scoring profile for all possible scores x_+ on the complete test. The expected score for a cluster k, $k = 1,\ldots, K$, is equal to:

$$E(x_k|x_+) = \frac{1}{n_{\mathbf{X}|x_+}} \sum_{\mathbf{X}|x_+} \sum_{k_j}^{j=1} X_{jk},$$

where the summation is over all response patterns $\mathbf{x} = (x_j, \ldots, x_J)$ with $\sum_{J}^{j=i} x_j = x_+$. Finally, we built an assessment form such as in Fig. 15.1 in which the scores for the complete test are presented in relation to the scores for the different clusters. We thus do not present all results from the simulation to the subject-area experts. Only the scores that can actually be obtained in practice are included in the assessment form.

15.7 Data Collection and Analyses

After application of the 3DC standard setting procedure we have a series of passing scores for each cluster per subject-area expert that can be presented in a data matrix like Fig. 15.2. In the first column of Fig. 15.2 are the numbers (or names) of the subject-area experts participating in the standard setting procedure. These are directly followed by the passing scores for the various clusters that each of the subject-area experts recommended. Finally, the passing score for the complete test is presented. This passing score is equal to the sum of the passing scores for each cluster, $C_{\text{total}} = \sum_{K}^{k=1} C_k$, and it is not established directly by the subject-area experts. Several descriptive statistics are presented at the bottom of Fig. 15.2. First, the mode is used to indicate the passing score that is most frequently selected. Amongst other patterns, this example shows that the performance standard for the first cluster was most frequently located at score 6, and that the standard for the second cluster was

Fig. 15.2 Data matrix after the application of the 3DC standard setting procedure

Expert	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Total
1	5	8	6	9	6	34
2	5	8	6	10	7	36
3	7	9	7	11	7	41
4	7	8	6	10	7	38
5	6	5	4	7	5	27
6	6	6	6	9	6	33
7	5	8	6	9	6	34
8	6	7	7	8	7	35
9	6	8	6	9	6	35
10	6	9	6	9	7	37
11	6	9	5	10	8	38
12	7	8	7	11	7	40
13	5	8	6	9	6	34
14	7	9	6	10	7	39
15	7	9	7	10	8	41
16	6	7	6	9	7	35
17	5	8	6	9	6	34
18	8	8	6	11	5	38
19	7	8	6	10	8	39
20	5	7	5	9	7	33
Mode	6	8	6	9	7	
Frequency	7	10	13	9	9	
Passing score						36
Maximum						49

most frequently located at score 8. These scores shaded in grey reflect the frequencies associated with the cluster-specific modes. In the first cluster, score 6 was selected by 7 of the 20 subject-area experts. In the third cluster, there is more consensus among the subject-area experts: 13 of the 20 subject-area experts located the performance standard for this cluster at score 6. The results for each cluster are followed by the performance standard for the complete test, as derived from the 100 (20 × 5) individual assessments. The example presented in Fig. 15.2 reveals the performance standard for the complete test at $[\bar{C}_{\text{total}}] = 36$. After the first assessment round, the data matrix in Fig. 15.2 can be presented to the subject-area experts and used as input for discussion. The definitive data matrix then automatically follows from the second round of assessments.

The data matrix in Fig. 15.2 is not only the input for the discussion round; it also provides the foundation for the analysis of agreement between subject-area experts. An initial indication of the extent to which the subject-area experts agreed with each other is reflected in the number of cells shaded in grey. More grey indicates greater agreement between the subject-area experts. In addition, various measures can be used to analyse inter-rater agreement at multiple levels. One possibility is to compute the *Finn coefficient* for relative agreement (Finn 1970) and *Gower's similarity coefficient* for absolute agreement (Gower 1971). The Finn coefficient is particularly well suited for data with high agreement among the subject-area experts, but the coefficient cannot be determined if the assessments of all subject-area experts on all clusters are identical. Gower's similarity coefficient would be 1 in that case, but the coefficient can provide a distorted picture if not all of the assessment categories are used. It is therefore advantageous to use both measures. The consistency between the assessments of one subject-area expert with those of the other subject-area experts could be determined by means of the *ranking similarity index* (RSI). This index is calculated as the average correlation of a subject-area expert with the rest of the subject-area experts. The impact of an individual assessment on the passing score could be determined by disregarding the assessment of the subject-area expert concerned in the calculations. The result of the index shows whether the exclusion of the rating of one subject-area expert influences the performance standard based on the ratings of all experts.

15.8 Practical Considerations

Using the 3DC methodology, we first have to decide upon the number and the composition of clusters to be evaluated. It is advantageous to choose subsets of items that have a common denominator; it aids the subject-area experts in formulating a recommended passing score. Often the number and composition of the clusters is based on a test characteristic. For example, reading comprehension tests often consist of subsets of items (testlets) grouped around a common text and mathematics tests often contain items associated with different ability domains like algebra, geometry, or calculus. In that case, it is easiest to organise the items by text or ability

domain. Although a content-related clustering of test items is recommended as it may lower the cognitive burden for the subject-area experts, it is not a technical requirement for the procedure itself.

Because the group discussion and the evaluation of inter-rater agreement require a sufficient number of clusters, it is advised to use at least four and preferably five to eight clusters. Once the number of clusters is determined, the content, length and difficulty level of the different clusters must be considered. The clusters ideally vary in length and difficulty level. If the clusters would have a similar length and difficulty level, the association between the scores of the complete test and the scores of the different clusters would also always be about the same. That is, if a student with a raw score of 6 on the first cluster would be expected to have a raw score of 35 on the complete test, regardless of the specific cluster. As a consequence, subject-area experts may show a tendency to use the recommended passing score on previous clusters as a heuristic for recommending on next clusters. It is important that the assessment is based on content and the expected number correct on the complete test; it should not be possible to deduce passing scores from previous assessments. Such behaviour can be avoided by varying in difficulty level.

In addition, the length of the clusters should be chosen such that the subject-area experts have a reasonable number of scores to select from. If a cluster would comprise four score points, for instance, subject-area experts might consider only a passing score of 3 realistic as a passing score of 0 or 4 would imply a minimally competent student to have everything wrong or correct on the complete test, and a passing score of 1 or 2 could involve the expected total score to be unlikely low. Each cluster should ideally contain at least eight test items to obtain a reasonable range of number of items correct to choose from. The length of a cluster could be shortened if the items are polytomously scored.

Once the test items are divided into clusters we have to decide upon the information to present to the subject-area experts. One possibility is to use the assessment form of Fig. 15.1. The scores that students can obtain on the different clusters are then all related to a score on the complete test. This approach can readily be explained to subject-area experts. A potential disadvantage of this approach is, however, that subject-area experts are confronted with information that is rather unrealistic or self-evident. The minimum and maximum scores of the different clusters always correspond to the ends of the score scale of the complete test, for instance. Some other scores may lie below guessing level in case of multiple-choice items or may be so high that students hardly ever achieve them in practice. Another possibility is therefore to only present the scores that make sense in light of the performances of the students who take the test. In that case, the score scale of the first cluster in Fig. 15.1 might reduce, for instance, from 0–8 to 2–6. For the second cluster only scores 2 to 10 might be presented to the subject-area experts as possible performance standard. By omitting the scores that are very unlikely to be observed from the assessment form we possibly make the standard setting process easier. It usually increases inter-rater agreement. However, we must be wary of guiding the subject-area experts too much towards a particular performance standard.

15.9 Illustrative Example

The 3DC standard setting procedure was recently applied in a large-scale international standard setting meeting. The aim of the meeting was to relate the Dutch national exams to the CEFR. The CEFR is a framework of level descriptions for learning, teaching and assessing modern foreign languages. Six levels of language mastery are distinguished, ranging from breakthrough (A1) to mastery (C2). A total of thirteen exams were assessed during the meeting. In this example, we present the results for the standard setting in the final examination for German as a foreign language.

15.9.1 Materials and Procedure

The final exam for German language was compiled under the auspices of the Dutch Board of Examinations (College voor Toetsen en Examens or CvTE). The exam focused on reading comprehension. The test administration was compulsory in all Dutch secondary schools. Different test versions have been constructed for the various educational tracks in the Netherlands. This example is based on the version that was developed for pre-university education. The examination comprised 46 items, all but four of which were scored dichotomously. Students could achieve a maximum of 51 points on the examination. Prior to the meeting, the items were divided into six clusters with the following scoring scales: Cluster 1 (0–8), Cluster 2 (0–9), Cluster 3 (0–11), Cluster 4 (0–8), Cluster 5 (0–8) and Cluster 6 (0–7). Figure 15.1 shows the assessment form that was used. As can be seen, it was decided to present all scores to the subject-area experts, also the ones that were below guessing level (score ≤ 9) and the ones that were rather high in light of the mean (29.1) and standard deviation (6.6) in the population. This was done on purpose. The way test items functioned in the population of Dutch students is presented in Fig. 15.1. Of course the functioning of the test items is partially dependent on the Dutch curriculum, which might differ in the degree to which topics within reading comprehension are being taught compared to other countries. In this sense the assessment form is also a reflection of the Dutch curriculum. Given the Dutch curriculum, that is, we could expect the pre-university students to perform in the way as presented in Fig. 15.1. From the perspective of the CEFR, however, the expectations on the behaviour of students might be different. If one of the sub-skills of the CEFR would receive much attention in Dutch secondary education, for instance, students might perform – from an international point of view – unexpectedly good on that sub-skill. In that case, subject-area experts may opt for a relatively low passing score in a cluster, even if such a score would not be very realistic in light of the ability that students generally show on the complete test. The subject-area experts could therefore choose from all scores that can be achieved in practice.

A total of sixteen subject-area experts participated in the standard setting for the German language exam (Feskens et al. 2014). The panel contained a relatively high

share of subject-area experts (8) who were employed as test developers or project leaders at testing institutes or in the testing divisions of language institutes. The other panelists were employed as curriculum developer (1), research scientist (5) or teacher (2). Prior to the conference, the examination to be evaluated was sent to the subject-area experts. Each expert was asked to make a preliminary estimate of the CEFR level that was measured by the examination. Subject-area experts found the pre-university examination for German language suitable for determining whether the reading comprehension of students corresponds to the descriptors formulated for CEFR level C1. During the meeting, the passing score for C1 was established in two assessment rounds. The following question was posed to the panel: 'Which score would a student be expected to achieve on this cluster if his/her ability is exactly at the borderline of satisfactory/unsatisfactory for language level C1?' In the first round, the passing scores were marked using a black ballpoint pen. The results were then discussed. To start the discussion, the results of the first assessment round were projected, and a few subject-area experts were asked to explain their assessments of one or two clusters. In general, a question to elaborate on their first round assessment was asked to two subject-area experts located at opposite extremes of the assessment spectrum, as well as to two experts located more in the middle. In the second assessment round, the subject-area experts were once again asked to mark the passing score for each cluster on the form. In this round, the experts used a red ballpoint pen, thus clearly indicating whether they had adjusted their initial scores and, if so, where. An illustration of a completed assessment form is presented in Fig. 15.1.

15.9.2 Results

The performance standard was determined according to the 96 (16 × 6) individual assessments. Table 15.1 provides the assessment data that were collected in the two rounds of assessment. For both rounds, the recommended passing scores are first presented. Within each cluster, the recommendations that correspond to the mode are bold-faced. The passing scores for the complete test, C_{total}, are then presented. These scores are equal to the sum of the passing scores for each cluster. Finally, the behaviour of each subject-area expert is examined in relation to other subject-area experts. We both report the ranking similarity index (RSI) and the impact. The definitive performance standard for CEFR level C1 is not presented in Table 15.1 but can easily be deduced by taking the average of column C_{total}. In both rounds, the subject-area experts recommended to set the performance standard at score 28. This means that students have to earn at least 28 of the 51 points (or 54.9% correct) in order to demonstrate CEFR level C1. We can use the standard deviation of $\bar{C}_{\text{total}}$ to express the level of precision as a 90% confidence interval: $([\bar{C}_{\text{total}}] - 1.645 \times \sigma_{C_{\text{total}}}; [\bar{C}_{\text{total}}] + 1.645 \times \sigma_{C_{\text{total}}})$ = (18; 38) for the first assessment round and (20; 36) for the second assessment round. These rather wide confidence intervals suggest that the subject-area experts were relatively diverse in their estimates of the number of points a student should be expected to earn in order to

Table 15.1 Results of standard setting in the final examination for German language

	Assessment round 1									Assessment round 2								
	Clusters									Clusters								
Expert	1	2	3	4	5	6	C_{total}	RSI	Impact	1	2	3	4	5	6	C_{total}	RSI	Impact[a]
1	6	6	**7**	**5**	6	**5**	35	.40	0	5	6	7	**5**	5	**5**	33	.57	0
2	6	4	6	4	**4**	4	28	.34	0	5	4	**6**	4	**4**	4	27	.49	0
3	6	**5**	8	4	3	**5**	31	.23	0	5	**5**	8	**5**	**4**	**5**	32	.48	0
4	6	**5**	5	4	7	4	31	.10	0	**4**	**5**	**6**	4	5	3	27	.22	0
5	**4**	4	**7**	**5**	**4**	4	28	.23	0	**4**	4	7	**5**	**4**	4	28	.47	0
6	6	**5**	6	4	**4**	**5**	30	.42	0	6	**5**	**6**	4	**4**	**5**	30	.49	0
7	**4**	4	4	3	**4**	1	20	−.10	1	**4**	4	4	**5**	**4**	3	24	.47	1
8	**4**	**5**	5	4	5	4	27	.39	0	**4**	**5**	**6**	**5**	**4**	4	28	.58	0
9	**4**	4	5	4	**4**	4	25	.50	0	**4**	**5**	**6**	**5**	**4**	**5**	29	.61	0
10	5	**5**	**7**	**5**	7	**5**	34	.13	0	5	**5**	7	**5**	6	4	32	.20	0
11	2	3	3	3	**4**	4	19	.30	1	**4**	4	4	4	5	**5**	26	.44	0
12	**4**	4	6	**5**	**4**	**5**	28	.41	0	**4**	**5**	7	4	**4**	4	28	.47	0
13	5	**5**	**7**	**5**	6	**5**	33	.47	0	**4**	**5**	**6**	**5**	5	**5**	30	.56	0
14	5	**5**	**7**	6	6	**5**	34	.42	0	5	**5**	7	**5**	6	**5**	33	.47	0
15	**4**	**5**	8	**5**	6	**5**	33	−.05	0	**4**	**5**	8	**5**	6	**5**	33	.11	0
16	**4**	2	1	0	**4**	1	12	−.09	1	**4**	2	1	0	**4**	2	13	−.22	1

[a]C_{total} = passing score, *RSI = ranking similarity index, Impact = individual panellist effect*

demonstrate CEFR level C1. This can also be seen in the individual assessments. Whereas one subject-area expert argued that a student should answer 25% (13 ÷ 51) of the test items correctly in order to demonstrate CEFR level C1, another subject-area expert proposed that CEFR level C1 could not be confirmed unless a student answers 65% (33 ÷ 51) of the test items correctly.

If we consider the results for the second assessment round in detail, we see that the size of the confidence interval was largely determined by one remarkably low passing score. The suggestion of subject-area expert 16 was eleven points lower (!) than the mildest suggestion from the other 15 subject-area experts. The impact value of this assessment on the location of the definitive performance standard was +1. This means that the position of the performance standard would be 1 point higher if we eliminated subject-area expert 16 from the analysis. The same applies to subject-area expert 7. The very low RSI (−.22) confirmed subject-area expert 16 to assess differently than other subject-area experts did. A large impact and/or low RSI may give cause to eliminate the assessments of certain subject-area experts. However, restraint is advised in the elimination of assessments as 'aberrant' behaviour does not necessarily reflect unwillingness or incompetence. It might be legitimate in light of the rater's professional knowledge, function and background. In the present example, the subject-area experts were selected carefully and extensive instructions were given during the meeting (see Feskens et al. 2014). It was therefore most unlikely that 'aberrant' behaviour was the result of unwillingness or incompetence.

In that case, it could be attractive to eliminate one randomly selected maximum assessment and one randomly selected minimum assessment from the analyses, even if the statistics give no occasion to do this. This would mean that the assessments of subject-area expert 16 and subject-area expert 1, 14 or 15 would be discarded.

If we discard the assessment of two subject-area experts and basically use a trimmed mean as passing score, the definitive location of CEFR level C1 shifts from 28 to 29. If the 90% confidence interval is used to consider the uncertainty surrounding this passing score, we see that students must earn between 48% (24 ÷ 51) and 66% (34 ÷ 51) of the points in order to demonstrate CEFR level C1. This confidence interval can be regarded as quite small. It thus seemed that – on second thoughts – there was sufficient consensus within the expert panel to set the performance standard at this point. This can also be observed in the measures of inter-rater agreement: Gower's similarity coefficient for absolute agreement is .914 and the Finn coefficient for relative agreement is .913. According to the guidelines by Landis and Koch (1977), these values for inter-rater agreement can be considered good. The interim discussions had a positive impact on the agreement between the subject-area experts. In the first round of assessment, Gower's similarity coefficient (.867) and the Finn coefficient (.800) were clearly lower. Overall, it thus seemed that 3DC indeed guided the panel in arriving at a performance standard that they can largely agree upon. It is a matter of preference whether to use the mean or the trimmed mean as performance standard. During this particular standard setting meeting, out of the 24 standards that have been set, only in two cases there was a difference between the mean and the trimmed mean.

15.10 Conclusions and Discussion

In this study, we presented a methodology for standard setting based on the common concept of the performance of a hypothetical minimally competent student. The 3DC procedure has several advantages compared to other commonly used standard setting procedures. First, in separating the standard setting task into different clusters, the cognitive burden to complete the task can be reduced, especially if the clusters are relatively small. The subject-area experts are then still able to form an opinion about the minimally expected performance on the entire cluster and the focus on clusters, moreover, appears to be a cognitively easier task than the focus on single test items as is common for the Angoff and the Bookmark procedures (see also Goodwin 1999). Contrary to the Direct Consensus method, which also divides a test into clusters, the 3DC procedure presents empirical information about the relative difficulty of clusters to the subject-area experts. This can be considered a second advantage of the methodology: the empirical information can assist the subject-area experts in the evaluation of the test materials. Third, the 3DC procedure provides flexibility in the use of item types and statistical models. A variety of item types from, for example, multiple-choice to constructed-response questions – either

dichotomously or polytomously scored – can be included in the standard setting and basically any predication model can be used to construct the assessment form. Alternative procedures usually do not offer this kind of flexibility. Finally, as compared to other methods, 3DC offers many opportunities for evaluating the correspondence within and between subject-area experts. As subject-area experts are asked to set a passing score on several clusters, it is possible, for instance, to evaluate rating consistency across the clusters.

For a correct use of 3DC, the following issues must be taken into consideration. First, as in any standard setting procedure, subject-area experts need to be trained before they can start using 3DC. Although the assessment form of Fig. 15.1 is, by itself, relatively easy to understand, subject-area experts need to become acquainted with it. In the beginning, especially the less (statistically) experienced subject-area experts can find it difficult to understand the relationships between the scores on the complete test and the scores on the clusters. The understanding of the assessment form is key to a successful 3DC standard setting procedure. Second, the underlying statistical model should be valid in order to be able to meaningfully relate the cluster scores to the complete test scores (and vice versa). A thorough evaluation of model fit and assumptions is required before starting to use 3DC. Finally, application of the 3DC procedure involves making decisions on the number of clusters and the number of items within a cluster. Furthermore, it has to be decided which items to include in each cluster. It is possible to randomly allocate the items to the clusters, but a content-based allocation of items to clusters can, however, facilitate the task of the subject-area experts. Although this kind of decision may seem arbitrary at first sight, each decision can affect the standard setting outcome. It is important that the results of the decisions fit the context of the user and further research is needed to obtain a fine-grained understanding of the application of the 3DC methodology and the influence of the set-up of the assessment form.

An important development in the use of the 3DC procedure is the construction of a digital platform which can be used to conduct the standard setting. Until recently, most 3DC standard setting procedures were conducted using paper prints of the assessment form. Subject-area experts used a ballpoint pen to indicate their judgement on the assessment form and after each assessment round the ratings were collected and filled in by the moderator into an Excel summary file. These steps can now also be performed using the digital platform. In that case, each subject-area expert uses his or her own laptop and after establishing a local network with the laptop of the standard setting moderator, the assessment form appears on the laptop of each subject-area expert. The subject-area experts then indicate their judgements on the digital assessment form and by one simple action the moderator can centrally collect the judgements of all the subject-area experts. The summary file will also be filled in immediately. The digital platform has already been tested and used in several standard setting conferences. Evaluations from both the subject-area expert and the moderators were positive; the platform makes the procedure much more user-friendly. A free copy of the application can be downloaded from www.cito.com/3DC.

References

American Educational Research Association, American Psychological Association, and American Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed.), pp. 508–600. Washington, DC: American Council on Education.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion referenced tests. *Review of Educational Research, 56*, 137–172.

Busch, J. C., & Jaeger, R. M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations. *Journal of Educational Measurement, 27*, 145–163.

Cizek, G. J. (2001). 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: Lawrence Erlbaum.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks: Sage Publications Ltd.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press http://www.coe.int/T/DG4/Linguistic/Default_en.asp. Retrieved Nov 2013.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah: Erlbaum.

Downing, S. M., & Haladyna, T. M. (2006). *Handbook of test development*. Mahwah: Erlbaum.

Feskens, R., Keuning, J., Van Til, A., & Verheyen, R. (2014). *Performance standards for the CEFR in Dutch secondary education: An international standard setting study*. Arnhem: Cito.

Finn, R. H. (1970). A note on estimating the reliability of categorical data. *Educational and Psychological Measurement, 30*, 71–76.

Goodwin, L. D. (1999). Relations between observed item difficulty levels and Angoff minimum passing levels for a group of borderline candidates. *Applied Measurement in Education, 12*(1), 13–28.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics, 27*, 857–871.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8*, 41–55.

Hambleton, R. K., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport: Praeger.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park: Sage.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. N. (2000). *Handbook for setting standards on performance assessments*. Washington, DC: Council of Chief State School Officers.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*, 353–366.

Jaeger, R. M. (1978). A proposal for setting a standard on the North Carolina High School competency test. Paper presented at the 1978 spring meeting of the North Carolina Association for Research in Education, Chapel Hill.

Jaeger, R. (1989). Certification of student competence. In R. Linn (Ed.), *Educational measurement* (pp. 485–511). Washington, DC: American Council on Education.

Kaftandjieva, F. (2004). *Methods for setting cut scores in criterion-referenced achievement tests. A comparative analysis of six recent methods with an application to tests of reading in EFL.* Arnhem: Cito.

Kane, M. (1998). Choosing between examinee-centered and test-centered standard-setting methods. *Educational Assessment, 5*, 129–145.

Karatonis, A., & Sireci, S. (2006). The bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice, 25*(1), 4–12.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Lewis, D. M., Mitzel, H. C., Green, D. R. (1996). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard setting procedures utilizing behavioural anchoring*. Symposium conducted at the Council of Chief State School Officers National Conference on Large-scale Assessment, Phoenix, AZ.

Lewis, D. M., Mitzel, H. C., Green, D. R., & Patz, R. J. (1999). *The bookmark standard setting procedure*. Monterey: McGraw-Hill.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4–16.

Pitoniak, M. J., Hambleton, R. K., Sireci, S. G. (2002). *Advances in Standard Setting for Professional Licensure Examinations*. Paper was presented at the annual meeting of the American Educational Research Association, New Orleans, LA, April, 2002.

Reckase, M. D. (2006). A conceptual framework for a psychometric theory for standard setting with examples of its use for evaluating the functioning of two standard setting methods. *Educational Measurement: Issues and Practice, 25*(2), 4–18.

Sireci, S. G., Hambleton, R. K., Huff, K. L., & Jodoin, M. G. (2000). *Setting and validating standards on Microsoft certified professional examinations*, *Laboratory of Psychometric and Evaluative Research Report No. 395*. Amherst: University of Massachusetts, School of Education.

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.

Van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer.

Verhelst, N. D., & Glas, C. A. W. (1995). The generalized one parameter model: OPLM. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Their foundations, recent developments and applications* (pp. 215–238). New York: Springer.

Woehr, D. J., Arthur, W., & Fehrmann, M. L. (1991). An empirical comparison of cut-off score methods for content-related and criterion-related validity settings. *Educational and Psychological Measurement, 51*, 1029–1039.

Zieky, M. J., Perie, M., Livingston, S. (2008). *Cuts cores: A manual for setting standards of performance on educational and occupational tests*. http://www.amazon.com/Cutscores-Standards-Performance-Educational Occupational/dp/1438250304/

Chapter 16
Using Professional Judgement To Equate Exam Standards

Alastair Pollitt

Abstract The principal concern in the UK is with *maintaining* standards that already exist, rather than with setting a new standard. To ensure standards are kept 'constant' is essentially a process of comparison rather than measurement. In this chapter four examples are presented to show how Thurstone's method of comparative judgement can be used to maintain standards, especially in the more 'difficult' cases involving extended writing, performances, or other complex activities. In particular, it describes how analysis of the residuals from fitting Rasch parameters to the data can be used to monitor the quality of the equating procedure.

Keywords Comparative judgement • ACJ • Exam standards • Test equating

16.1 Introduction

In the UK, we do not set standards very often. The general presumption, amongst politicians and public alike, is that a certain grade in a certain subject represents a fairly precise level of achievement or quality of performance; the job of the examiners is to ensure that this level or quality *stays the same* from session to session. Furthermore, since there are often several agencies providing tests for the same qualification it is also their job to ensure that they all set their grade boundaries at the *same* standard. In some undefined sense there is also generally assumed to be a common standard across all the exams in all subjects, from Art to Zoology, from Mathematics to Media Studies. Even when a new qualification is introduced its standard is usually defined in terms of existing ones. In 2017, for example, a new system of grades (9–1) will replace the current system (A*-G). The Chief Regulator (Ofqual 2014a) wrote:

A. Pollitt (✉)
Cambridge Exam Research, Cambridge, UK
e-mail: alastair@camexam.co.uk

S. Blömeke, J.-E. Gustafsson (eds.), *Standard Setting in Education*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-50856-6_16

> We are being quite clear that the approach in that first year will draw heavily on statistical evidence to make sure that there are clear 'anchor points' from the old system to the new.

Here population statistics are to be used to support judgemental comparisons when there are significant changes in the system that would make simple comparisons of standard less safe. In fact, statistical predictions, based on prior test scores are regularly used to support judgement, even in more stable times.

We do not make much use of written statements as definitions of standards. Another document from the regulator reports (Ofqual 2014b):

> We have developed grade descriptors for the reformed GCSEs graded 9 to 1. ... The purpose of these grade descriptors is to give an idea of average performance at the mid-points of grades 2, 5 and 8. The descriptors are not designed to be used for awarding purposes.

In general, in UK high stakes examining 'grade descriptors' are used as *indicators* and not as *targets* or as criteria in the process of awarding grades.

16.2 Standards

We use the word *standard* in at least two quite different ways. The first is concerned with determining what kinds of things our students should be able to do at the end of any particular course of study: we call these *content standards*, and it is obvious that they must be written only after careful consideration of what's important in studying chemistry, or history, or art. Judging what is important enough to go into the content standards is not trivial and we rely on the judgement of appropriate experts for this; opinions vary around the world, of course, as to who are the 'appropriate' experts.

Performance standards, too, are essentially a matter of judgement, but here it is a judgement of how well each student has succeeded in meeting the content standards. A student's exam result may be generated in several ways, ranging from a straight judgement of whether or not they have achieved the required performance standard, to a simple count of the number of right answers they gave, or selected, to a series of questions – with many cases involving some sort of mixture of these two extremes.

There is also a third meaning, sometimes referred to as *assessment standards*, which is concerned with the level of *demands* in exams. It could be seen as a necessary complement to the performance standards, since we might expect a more demanding exam paper to result in a lower level of performance.

Before 1792, *all* educational assessment was entirely a matter of judgement. Often, a master would simply declare an apprentice 'fit to practice' (or not) after observing their work over a period of perhaps several years. In high stakes settings, such as university degree exams, a team of examiners would discuss the students and their work, and either just judge each student as 'passed' or 'failed' or agree on a complete rank order from best to worst. (Wordsworth 1877).

Then, in Cambridge University, William Farrish invented *marking*, apparently in order to prevent any one examiner from forcing his opinion on the others (Pollitt 2012a). Counting marks and adding them up replaced discussing quality and forming

a consensus, and judgement faded into the background, at least for summative assessment. Marking seemed not only more objective, but also more efficient as numbers of students in schools and universities grew rapidly through the nineteenth century.

16.3 Problems with Judgement

Today, with marking still dominant, it is *judgement* that is seen as problematic, since it is so hard to get markers to agree on exactly which number to attach to each answer they see. Four distinct kinds of problem can be identified in marking extended and complex pieces of work, four ways in which the *same* piece of work may be given different *scores* in different circumstances:

Problem	Cause	Concern
Disorder	Favouritism or prejudice	Avoiding bias
Severity	More or less generous	Reliability
Discrimination	More or less extreme	Reliability
Different order	Different conceptions of 'good'	Validity

It was Problem 1 that prompted Farrish to introduce secret marking and summation to give an '*automatic*' total score. But he still required every examiner to mark every script, so that every marker's prejudices would be diluted by the others: the result, in effect, was a total score for each student that reflected the consensus view of how well each student had met the overall performance standard. In later years, however, when the student numbers expanded too far, the load had to be shared out, and the other three problems grew ever larger. We can cope with Problem 1 through anonymising the scripts, but the other three are constant, and still test the ingenuity of assessment agencies to, and often beyond, the limit.

16.4 Thurstone's Method of Comparative Judgement (CJ)

In the 1920s Louis Thurstone developed what might have been the ideal answer to these problems with his method of comparing objects (here, scripts) in pairs and requiring an examiner (*judge*) only to say which of the two is the better (Thurstone 1927). Combining many such comparative judgements allows us to construct a scale that measures the relative quality of all of the scripts.

What then happens to the four problems?

1. Disappears because there are multiple scorers, and through analysis for bias
2. Disappears since only the relative quality is recorded
3. Disappears since only the relative quality is recorded
4. Who can be trusted to judge the *essential quality* of students' work?

The use of bias analysis to check the measurement is a powerful quality control tool, and it is an extension of it that is the core issue that will be addressed later in this chapter; for the moment the point is that CJ brings us back to the principle of consensus that underpinned high stakes assessment before Farrish. Problems 2 and 3 simply vanish when a judge is asked only to say 'A is better/poorer than B': this reminds us that the essential purpose of 'marks' for individual questions in a test was only ever an artificial aid supposed to help us decide how good each student's work was.

Only Problem 4 remains significant: the scale resulting from CJ will report the *quality* of each script – as perceived by the particular set of judges participating. It is therefore important to consider – and agree on – who should be trusted for this task. We will return to this issue later, but for the moment it's worth noting that test developers in an educational context have traditionally relied on the rank ordering of school students by their teachers as an important indicator of *concurrent validity* for their new tests. The first modern application of CJ in education showed how the comments made by judges as they judged, combined with the rank order they produced, seemed to confirm the theoretical concepts of *communicative competence* (Pollitt and Murray 1993). This can be taken as evidence that the CJ method can deliver good *construct validity* as well.

This chapter is directed at the issue of setting standards or, as it usually applies in UK experience, *maintaining standards* that were set in some way in an earlier time or concurrently by different exam boards. Here, Problem 4 is indeed important, but so is Problem 1 in a rather disguised way. The question is whether or not judges can make dependable comparative judgements *across* tests or exams. Generally, in the UK, grade boundaries have for many decades been set by judgements of this kind, though increasingly assisted by statistical tools and models. Currently, the most senior examiners in any school certificate examination are asked to look for the 'quality of an A' in scripts around a mark that is thought to be at or near the sought-for boundary, and so to identify exactly which total score point is to count as the minimum for each grade.

Can we improve on that? In the following sections some examples will be discussed to show how Thurstone's CJ method has been applied to setting grade boundaries in a variety of subject areas and in several countries in the last few years. First, a summary of terminology.

16.5 Terminology

Since this chapter is largely concerned with assessment in the UK, and assessment using some unfamiliar methodologies, it may be worth defining for all readers how some terms will be used. *Score* will be used as a general word for the 'result' of any test or exam, and *Scoring* will similarly be used as a generic term for any procedure that results in a score. Scoring procedures may include *counting*, as in multiple choice tests; *marking*, when many items are scored using partial credit mark

schemes; *rating*, when performances are judged against scoring rubrics; *comparative judgement*, when Thurstone's method is used.

Content standards are verbal definitions of what an exam is meant to be assessing. *Performance standards* are overall measures of how well these content standards are being met. *Script* is used to mean any recorded set of evidence that can be scored to indicate a student's performance level; most often this is a written script, but it may refer to other objects or records of performances, such as portfolios, graphics, or video recordings, if appropriate for a given course. *CJ – Comparative Judgement* is any application of Thurstone's method. *ACJ – Adaptive Comparative Judgement* is an application of CJ that uses adaptivity and regular re-estimation to improve the efficiency of CJ.

16.6 The CJ Methodology for Standard Setting/Checking

Until recently there were problems in applying comparative judgement to British exams. Examiners wrote - scribbled - on the students' written scripts to help themselves decide on a mark, and to justify the mark in case of reviews or appeals; for CJ, this writing had to be 'cleaned off' to avoid interfering with the CJ judges' decision processes. Also, a script may be anything up to 30 pages long, and is typically just one of several components in a whole examination. It was therefore expensive to photocopy all of a student's work, and only a small set of scripts could be used. For speed, convenience and efficiency the team of judges had to be brought together for the duration of the study. The first example below followed this procedure.

The internet changed everything. In particular, it brought on-line marking, which is now almost universal in British exams: all scripts are routinely scanned for that, without scribbles, and the only costs left in applying CJ are the judges' time, plus design and analysis. It is now easy to run an equating study with large numbers of scripts and judges.

In a typical standard checking application today, appropriate scripts are chosen from two examination sessions, which may be this year's and last year's, or two concurrent versions from different agencies, or from syllabuses designed to address the same content in different ways - or even from different subjects in an attempt to ensure inter-subject comparability. To set equivalent 'pass' marks as few as 20 from each may be needed, but if various grade boundaries are to be equated there will usually be 50 or 100 spread across the score range in each. As the examples will show, the selection and pairing of scripts is crucial, and this will be addressed later.

CJ data are analysed using a Rasch model which can be expressed as:

$$\log \text{odds}(\text{B is better than A}) = parameter_{\text{B}} - parameter_{\text{A}}$$

In the usual Rasch way, the judge in any comparison is 'cancelled out' from the equation, since they are the same person judging A and judging B. It is as if the two students tried the same test item and only one of them got it right: in CJ the judge

'is' the common item, and since we *never* allow the judge to say they are both 'right' or both 'wrong' every judgement contributes to separating out the scripts. CJ data can be analysed using the well-known *Facets* package (e.g., Linacre 2010). The analyses reported here used programs first written in 1994 specifically for this application of Rasch modelling; later versions have been regularly validated by comparison with *Facets*.

The equation above expresses Rasch's simple linear model in a comparative form. The 'parameter' in the equation is an estimate of the quality of the script, and the aim is to see if scripts of the same quality in the two exams were awarded the same result, be it pass, or fail, or A, B, C etc. The output from the analysis is, essentially, a list of parameters with estimated standard errors. Depending on how the scripts were chosen, the further analysis may simply use the average parameters from each exam, or may involve a linear equating method to find parameter equivalents for the key scores on each.

16.7 A Note on 'Consistency'

The main focus of this chapter will be on the *relative* consistency of different kinds of judgement, as will be explained later. But it is important first to note that Rasch analyses normally report an overall measure of the internal consistency of a set of data. This is the estimated ratio of the true variance of scores to the error variance of their estimates, and is properly described as an *alpha* coefficient (Cronbach 1951). It is often misleadingly reported as *reliability*, as for example by Facets (Linacre 2010), but this ignores the possibility of further threats to score reliability external to this particular data set. In this chapter the overall coefficient of internal consistency will be called 'the alpha coefficient'.

16.8 Misfit Methodology in CJ Equating

This chapter will consider four examples of standard setting or checking. The first is a very simple example from the 1990s, that simply compared average parameter values of supposedly 'equal' scripts from two tests. But within a few years of it - thanks to the internet - we were able to collect far more data in CJ studies, and hence also able to explore the quality of the judgements. It became feasible to use the misfit information that Rasch modeling provides to check the plausibility of the results, and hence the validity of the whole exercise.

The basic method for quantifying misfit is no different in CJ from in any other Rasch analysis. In summary: In every comparative judgement, script A meets script B. The full data generate parameter values for A and B, from which we calculate 'p', the probability that A will 'beat' B. The 'score' is either 1 or 0. And there is always

a residual, (1-*p*) or (0-*p*). The residual is then standardised by dividing it by the standard deviation, sqrt(*p**(1-*p*)). And then squared. This *squared standardised residual* (SSR) is calculated for every decision made: it will be small when the 'consistent' decision is made – that is, when the script judged better overall wins – or larger when the unexpected decision is made.

In general, Rasch analysis programs use the SSRs to show the relative consistency of test items: the weighted[1] average of all the SSRs that involve one item gives a misfit statistic for it, called the *infit*, or *weighted mean square* (described in detail in Wright and Masters 1982). A similar analysis can give misfit statistics for each student, or for each marker or judge, in each case by averaging the relevant SSRs. But they also have a less well-known use in monitoring potential bias in the data, and it is this that is particularly useful in standard equating.

The procedure is analogous to the Analysis of Variance: the total sum of the SSRs is similar to the Total Sum of Squared Residuals in ANOVA, and can be partitioned in the same sort of way to show how much misfit there was in various subsets of the data. In ANOVA, the usual first step is to calculate mean squares *Between* and *Within* groups. In the equating context the most obvious hypothesis will be that judges will be more inconsistent when comparing two scripts from *different* exams (between) than they are when comparing two from the *same* one (within). Examples 2 to 4 will show how this helps us judge whether or not the equating can be trusted.

16.9 Example 1: UK Comparability Studies Using CJ 1997

Most applications of CJ educational assessment in the 1990s were comparability studies for A Level and GCSE certificate examinations, the key exams taken, respectively, at around age 18 and 16 in England, Wales, and Northern Ireland. Usually these involved all five of the boards involved, and were intended to check, retrospectively, that the same performance standard had been set by each board, but in the first trial of the CJ method a simpler, though then more contentious, comparison was made between two formats of exam within the same board – one 'linear', the traditional end of course exam, and one 'modular', in which up to six modules were accumulated over a year or more (D'Arcy 1997).

To test the equivalence of performance standards when different assessment formats are being used, 10 students following the Modular format and 6 following the Linear format were chosen – all had scored exactly the minimum total mark for Grade A, as decided by senior examiners for that particular exam format – and ten judges compared pairs consisting of one from each set, deciding 'which candidate's work was better' – 'A' or 'B'. In each case they looked at all of the student's work – two or five written papers, plus a coursework project report. Their decisions were used to estimate parameters, and standard errors, for the scripts.

[1] The weights used are the variances of each judgement, *p**(1-*p*).

Fig. 16.1 Plot of linear and modular parameters from example 1

RASCH ANALYSIS using the PAIRED COMPARISON model (Pollitt, RPC.v12)

** A Level Biology Comparability Study (1996) Grade A
Plot of Parameter Estimates

```
 | |
 | 2|    6:M
 | |    5:M   14:L
 | |
 | |
 | |   10:M
 | 1|
 | |
 | |    2:M
 | |
 | |
 | 0|    7:M   13:L
 | |    3:M   12:L
 | |
 | |   11:L
 | |    4:M
 | -1|    1:M   8:M   15:L
 | |
 | |    9:M   16:L
 | |
```

16.9.1 Main Result

Figure 16.1 is part of the analysis print-out showing the parameters estimated for each of the 16 sets of student work; the labels indicate which format each belonged to. Note that the scripts, despite scoring the same mark, range quite widely when judged only for *relative quality*: the standard deviation of the parameter estimates was, in fact, just over 1 logit. This might be taken to mean that the judges were not very accurate in their judgements or, alternatively, that the total mark is a rather poor measure of how well each candidate met the content standards.

The main interest in these studies was just in the average quality of scripts on the two grade boundaries. The means were:

$$\text{Linear: } +0.20 \quad \text{Modular: } -0.12$$

which suggests that the borderline linear work was generally seen as a little better than the borderline modular work. For several years, this approach was used extensively for checking standards across boards in the UK. Generally, the results of these comparability exercises seemed to confirm the examiners impressions and were consistent with other statistical indicators. While this seems to validate the method, two significant problems were found that limited the wider use of CJ in assessment.

First, while it was easy to determine if any differences seen between exams were statistically significant, the experimental design did not allow these statistical measures to be translated into a more familiar metric: how many marks different are two standards if they are 0.32 logits apart? Is this really a significant mismatch?

But more importantly, as far as examinations are concerned, Thurstone was way ahead of his time. It was never practicable to handle *paper* scripts efficiently enough to make CJ a feasible alternative to traditional marking for school exams, even when photocopiers were available to help: CJ really needed the internet. The next example shows an on-line judgement system collecting data for standard equating.

16.10 Example 2: A Modern Experimental CJ Design 2015

Here was the problem: an assessment agency would like to set a common set of examinations, in several subjects, but on a global scale. The format will include some extensive and complex written exercises that will need to be scored by subject experts. The biggest problem here is that the school year runs roughly August to June in the northern hemisphere but from February to December in the southern hemisphere, meaning that two exams will be needed each year. There can be no common items, for security reasons, and there may be little stability from year to year in the entry characteristics of the candidates How can they ensure that the same performance standard is set in each?

Trials were run in 2015 to test the feasibility of equating these two sets of exams, after the first had been marked in the traditional way and simultaneously with the marking of the second. The aim was to effect a 'whole-test' equate rather than concentrating on a single grade boundary, solving the first problem mentioned above very simply; and the whole judgement and data collection process was managed via a web site, completely resolving the second.

For one trial – an Accountancy exam for age 17/18 – 42 papers from the 2013 session stood in for the 'first' exam, and 80 papers from the 2014 session stood in for the 'second' one. Eighteen experienced examiners acted as judges, and made 2054 comparative judgements, following a balanced experimental design. The first question was whether or not CJ would provide a plausible equate in this context. The graph below shows a plot of the parameter values for each test against total mark, and the two linear regression lines.

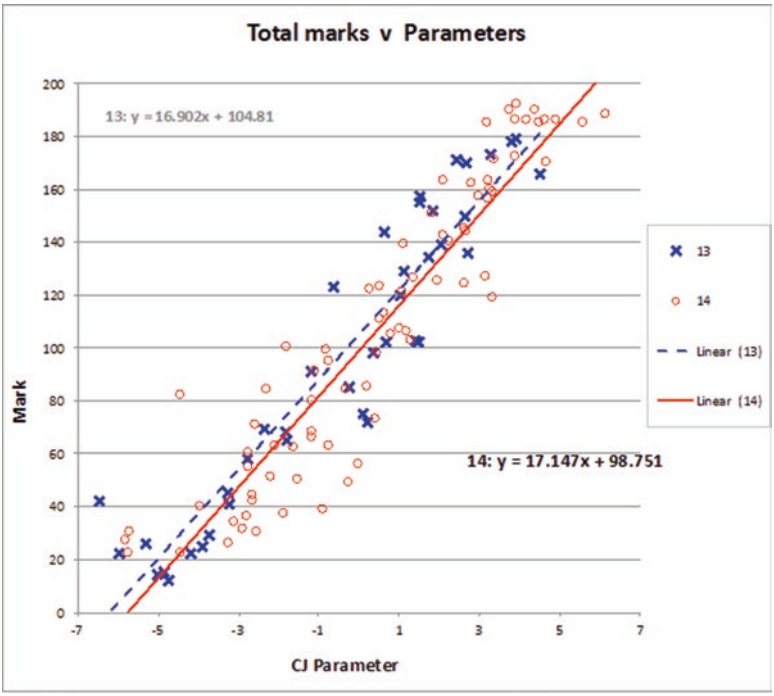


Fig. 16.2 Graphs, and regressions, of the script parameters against mark for each test

16.10.1 Main Result

The linear regression lines in Fig. 16.2 are very close to parallel, suggesting that only a simple constant would be needed to equate the two tests across the whole range, as may often be the case for examinations like this that are constructed to strict specifications. The alpha coefficient for all the judgements made was 0.95. Furthermore, the difference between the two sets of marks – approximately 7 marks – agreed with the decision that had already been made in setting the 2014 exam's grade boundaries.

The use of CJ here, however, allowed a much more detailed analysis of the quality of the judgements made by the judges. The issue was: *Were these judges able to compare scripts from these two different tests consistently enough for us to trust this simple equate*?

16.10.2 Analysing the Residuals

Table 16.1 shows the results of partitioning the residuals to show the difference between comparisons *within* one test and comparisons *between* the two tests. The columns show: the Partitions; the Number of comparisons in each; the Weighted

Table 16.1 Partitioning the total sum of squared residuals for example 2

	N	Σ w*z2	Σ w	WMS
Total	**2054**	**162.5**	**164.9**	**0.99**
Within	**1127**	**89.3**	**92.1**	**0.97**
w′in 13	243	18.1	19.0	0.95
w′in 14	884	71.2	73.1	0.97
Between	**927**	**73.2**	**72.8**	**1.01**

sum of standardised residuals; the Sum of the weights; and the Weighted Mean Square. In perfectly consistent data WMS should equal exactly 1. The overall weighted mean was 0.99. This fell to 0.97 if we look only at comparisons *within* one of the tests, and rose to 1.01 if we look only at comparisons *between* the two tests.

Thus there was a small amount of disturbance in the judgements when the examiners had to compare scripts from two different exams rather than two scripts from the same one. The increase, though, is just about the same amount as we see going from the 2013 exam (0.95) to the 2014 one (0.97), which suggests that it is not serious enough to challenge the validity of the equating procedure in this case. Comparing between two tests must always involve more uncertainty than comparing within one test, and so a small amount of increased inconsistency is inevitable. Not all studies work so well, however, as the next example will show.

16.11 Example 3: One that Did Not Work 2015

In this case, an examination in Business for age 17/18, a very different design was tested, using CJ in a way that more closely parallels what is routinely done by informal judgement in British examinations - direct 'impression' comparison of boundary scripts from two exams. The aim was to fix two grade boundaries in a single comparative exercise which would, if it worked, solve the logit/mark problem in a more efficient way than carrying out a 'whole-test' equate.

For the exercise, scripts from 2012 were combined with scripts from 2006. Ten scripts were chosen from each of the 2006 A, 2006 C, and 2012 A, 2012 C boundaries. That means the 40 scripts consisted of:

10 scoring 70, and 10 scoring 50 in 2006, and
10 scoring 72, and 10 scoring 51 in 2012.

Four judges, again experienced examiners, made 200 comparative judgements: any combination of scripts was allowed. Since all the scripts in each set of ten had the same mark score, the graph below shows their ranking within each year/exam, based on all the comparisons within and across both years and grades.

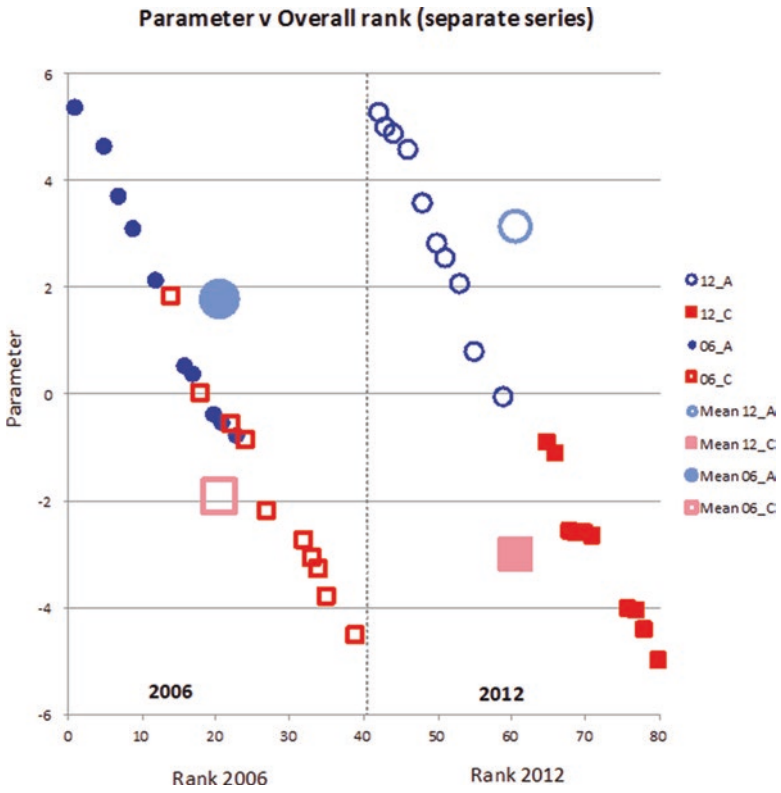


Fig. 16.3 Parameters in rank order, for each test

16.11.1 Main Result

Figure 16.3 is a double graph, showing the parameters for the 20 scripts in each exam plotted in rank order, with the means for each of Grades A and C. The graph shows, again, that all scripts with the same total mark are *not* equal in quality. The 'best' 2012_A script was rated more than 5 logits better than the 'poorest', meaning that a typical judge would be more than 99% sure to rate it as better – even though they were both marked 72. At least, though, the analysis did rate all the 2012_A scripts as better than all the 2012_C scripts that scored just 51 marks. Not so for the older 2006 exam though; there one of the C scripts (score: 50) was actually rated higher than five of the ten A scripts (score: 70). Overall, it 'seems' that the A grade boundary went up between 2006 and 2012, while the C grade boundary went down.

Table 16.2 Partitioning the total sum of squared residuals for example 3

		N	Σ w*z2	Σ w	wms
Total		200	22.80	24.30	**0.94**
Within	**Grades**	133	17.37	20.24	**0.86**
w.Y w.G		61	7.73	8.98	**0.86**
w.A		29	2.72	3.94	0.69
	w12A	16	1.57	2.34	0.67
	w06A	13	1.15	1.61	0.72
w.C		32	5.01	5.04	0.99
	w12C	17	3.01	2.75	1.09
	w06C	15	2.00	2.29	0.88
b.Y w.G		72	9.64	11.26	**0.86**
	w.A	37	5.29	6.08	0.87
	w.C	35	4.35	5.19	0.84
Between	**Grades**	67	5.43	4.06	**1.34**
b.G w.Y		28	3.29	1.63	**2.01**
	w.12	12	0.99	0.10	9.91
	w.06	16	2.31	1.54	1.50
b.G b.Y		39	2.14	2.42	**0.88**
	12A-06C	19	1.60	1.19	1.35
	06A-12C	20	0.54	1.24	0.44

16.11.2 Analysing the Residuals

Table 16.2 shows, in considerable detail, how the sum of squared standardised residuals can be partitioned to show where the judgements were less consistent, and less trustworthy, than we need. The analysis here is quite complex, but we need only to recognise that there were two main effects, in the ANOVA sense: *Year* or test with two values '2006' and '2012', and *Grade* also with two values 'A' and 'C'. The overall weighted mean square was 0.94, but there were considerable differences when the two main effects were separated. The mean square *within grades* was 0.86, while the mean square *between grades* was 1.34: this shows immediately that the judges found it much more difficult to make consistent judgements between A and C scripts than when comparing scripts given the same grade. The highest mean square of all the main sections, 2.01, came when the judges were asked to compare 'between grades, within years' – that is, when comparing an A and a C script from the same year's exam – just where we might expect them to be most sure of their decisions.

In contrast, the mean square *within grades* showed the same value (0.86) whether the comparisons were *within* or *between years*. This comparison is reassuring, since it implies that, given a more appropriate experimental design, their judgements of the same grade boundary in different years would have been highly consistent, just as in the previous example.

16.12 What Went Wrong?

The lesson of this failed study is that experimental design *does* matter when CJ is used for this sort of test equating function. It's not, however, how the scripts are sampled that matters, so much as what the judges are being asked to do – the nature of the judgement task itself.

Kahneman (2011) recounts his long-running debate with Gary Klein. Kahneman believed that intuitive decision-making was always subject to a serious danger of bias, and uncovered many examples in economics and probability contexts. Klein believed that experts could indeed come to instant, and valid, conclusions without being always able to explain or justify them (Klein 2008). A joint paper eventually described their compromise (Kahneman and Klein 2009). In his 2011 book (p243), Kahneman concluded:

> **If** the environment is sufficiently regular and **if** the judge has had a chance to learn its regularities, the associative machinery will recognize situations and generate quick and accurate predictions and decisions.

In the case of Example 3 it seems clear that the environment was not sufficiently regular. The judges apparently expected to be asked to compare 2006 scripts to similar 2012 ones, but in the event they were never told if a script was from Grade A or Grade C. When the two came from different exams/years this didn't seem to cause much trouble, but when the pairing were from the same exam it seems confusion set in: the second script was *either* at the same grade *or* from two whole grades higher or lower – a gap of either 0 or about 20 marks. This was not the kind of comparison they were expecting, and not the sort of judgement they would have been used to making. It seems that the psychological setting of the task was sufficiently disturbed or confusing to make unexpected decisions far more likely.

16.13 Efficiency and Adaptivity

The use of the internet to manage the administration of a Comparative Judgement exercise, while it overcomes the two problems described above can, however, run into one other serious difficulty – time, or cost, or inefficiency. In testing, efficiency can be measured by the amount of *Information* contributed by each 'item' of the assessment, given by

$$I = p * (1 - p),$$

where p is the estimated probability of any given outcome, such as 'A' beating 'B' in a comparative judgement. Since the maximum value of I occurs when the two scripts are exactly equal in quality – that is, 0.5*0.5 – it is convenient to define *efficiency* so that it shows how much information any comparison gives as a percentage of the maximum possible:

$$\text{efficiency} = 400 * \text{I}.$$

In the case of Example 2 above, the most successful of the ones described, 2054 judgements were made, with an average (median) time of just over 2 min – and a total time of about 70 h of judging. Yet the median efficiency of the judgements was only 17%, because a very large number of the comparisons involved scripts very far apart in estimated value.

The final example shows the use of an adaptive form of CJ, referred to as ACJ, which aims to increase the efficiency, and reduce the cost, of using comparative judgement. In addition, it looked forward to two additional possible future applications. Thurstone developed CJ as a method for measuring the 'quality' of objects, such as exam scripts, which can only be evaluated by human judgement of one kind or another. According to Laming (2004), 'There is no absolute judgment. All judgments are comparisons of one thing with another', which implies that Thurstone's comparative judgements should be more natural, and probably more accurate, than the kinds of indirect comparisons via marking rubrics that are currently used in most such exams.

It therefore makes sense to ask if CJ can be used as a scoring system instead of marking (Pollitt 2004). Further, if this is feasible, would it be possible to incorporate an 'automatic' test equate into the process, by including some scripts from an earlier exam into the regular ACJ scoring process, and using their re-estimated parameters to equate the new test to the same standard?

16.14 Example 4: A Pilot of Automatic Test Equating 2009

The context for this exercise was first language writing by pupils aged around 9–12 years in England. At the time, there were plans to implement 'single-level tests' that would simply decide whether or not each pupil had reached the next 'level' defined by the content standards, and this exercise used the pilot Level 4 Writing Tests. The study is reported more fully in Pollitt (2012b).

1000 scripts were selected for us by the government agency, each with two writing tasks. Of these, 980 were from a 'new' test, and 20 from an 'old' one that had already been used and marked in the traditional way. The two tasks in each test were quite different – one factual and the other narrative – which we thought might cause problems for judges in addition to the regular complexity of making some comparisons across different tests. Further, the nature of the factual task was very different in the two tests – persuasive in one and imaginative-descriptive in the other.

We were given 52 judges, with a variety of backgrounds: some were secondary school teachers, some from primary schools; some had experience of marking these tests and others did not. This allowed us to provide evidence concerning some ways to extend the range of eligible examiners, a serious government concern at the time.

A total of 8161 comparisons were made; after every 500 (called *a round* since each script would be judged once more, on average) the data were analysed to give updated estimates for each script's quality, and these were used in the next to avoid inefficient pairings, in a way similar to that used in computer-adaptive testing.

Note, however, that there is one significant and essential difference between adaptivity in CAT and ACJ. In both, we do not want to present a judge with two scripts that are too far apart in quality, since this would result in inefficient and expensive assessment. But in ACJ we equally do not want to present them with a pair that are so similar in quality that judging which is the better is next to impossible; judges – quite reasonably – complain if they think they are being given 50:50 choices to make, and we take pains to explain the adaptive system to them. Difficult decisions may seem to be very informative, as their efficiency may be close to 100%, but if they are effectively the result of tossing a coin, it would not be wise to treat the result as meaningful.

The adaptive algorithm needs to be quite clever to optimise the quality, rather than just to maximise the quantity, of information that goes into the analysis system.

16.14.1 Main Result

Figure 16.4 shows the 1000 parameters (hence the name 'Para-graph') and associated standard errors from the analysis, sorted into increasing order, or decreasing ranking, from about −10 to +10 logits. The scripts from the 'new' test are represented by the dark spots forming an almost continuous line through the centre of the graph, and their standard errors by thin vertical lines above and below the spots. The thickness or thinness of this plot gives a visual indication of the internal consistency of the data; in this case the alpha coefficient was 0.96. The twenty 'old' test scripts are highlighted by crosses. We found that the script selection, or perhaps the Government's policy, was rather peculiar since almost 90% of the students in the group were deemed to 'pass', as were 19 of the 20 'old' scripts, but the result did show that the method was feasible at least. For an operational testing system, transferring the 'old' standard to the 'new' test would involve, in effect, a simple 'item banking' process; the 'old' scripts would be used as 'anchor items' to calculate a shift constant to bring the 'new' scale into its correct position relative to the 'old' one.

16.14.2 Analysing the Residuals

Of course, we would always carry out the bias check to ensure that the judges were in fact unbiased in making their comparisons across the two tests. Table 16.3 shows the key part of the analysis. Because 980 of the scripts were 'New', and only 20 'Old', the story is dominated by the 'New' v 'New' comparisons; only four of the

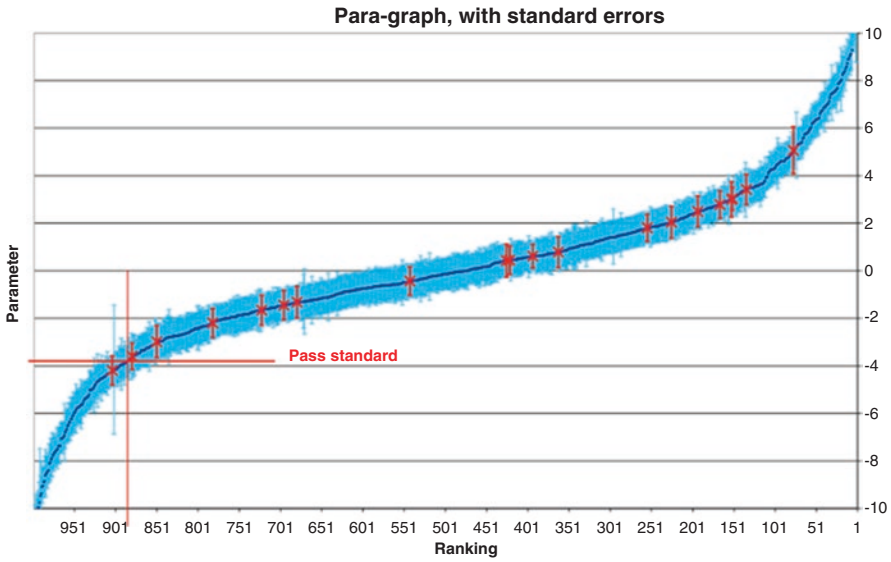


Fig. 16.4 Graph of the parameter values, with standard errors, and the standard

Table 16.3 Partitioning the total sum of squared residuals for example 4

	N	Σ w*z2	Σ w	WMS
Total	8161	1074.9	1176.9	**0.91**
Within	7828	1022.9	1122.4	**0.91**
Between	333	52.0	54.5	**0.95**

7828 'within' comparisons involved two 'Old' scripts. Once again, the weighted mean square for comparisons between tests was larger, at 0.95, than the mean within tests, at 0.91. But despite the considerable differences between the tasks in the two tests, the difference is small. More detailed analysis also showed that it made little difference what kind of history each judge had – so long as they were experienced teachers of children around ages 9–12.

And what of efficiency? The effect of using an adaptive algorithm was to raise the median efficiency of judgements to 68%, four times as high as in Example 2. This means that the same level of precision in the parameter estimates was reached with little more than one quarter as many judgements. While the principle is clear – adaptivity increases efficiency – more work is needed to determine the best kind of adaptivity to apply in different contexts.

One further analysis was interesting with regard to the interest of Kahneman and Klein in '*quick and accurate … decisions*': despite considerable variation in the speed with which judges made their decisions, we found no significant correlation between judges' speed and the misfit measures of how consistent their decisions were with the consensus of the others. Again, more study of the judgement process

in CJ would help us understand why the speed of decision-making does not correlate with quality.

In this English writing context, then, we concluded that the 'Old' standard could indeed be transferred to a 'New' test, and that judges who meet Kahneman's criterion of adequate experience can be trusted to make judgements across tasks. Of course, the same may not apply in every other case.

16.15 Discussion: Design Considerations in CJ Equating

The key principle for good CJ seems to lie in Kahneman's notion of 'regular environment'. We must design the exercise so that the judges are working within their experience as much as possible, and doing the kind of thing they are expert in - comparing pairs of scripts that are *quite* similar in quality. And they should understand in general how the script pairs are chosen for comparison, so that they know they are - usually - being given a reasonable, professional, task to perform.

The second principle concerns information: choose scripts for the study that are capable of answering the important questions efficiently. This means choosing most scripts from the parts of the ability range that are most critical, while maintaining a sufficient range of quality to allow reasonable comparison tasks and to allow any differences to be converted from logits to the familiar scoring scale.

A third principle concerns the quality of the equate. Any possible sources of bias, or distortion, should be formalised into hypotheses that can then be explored by suitable partitioning of the weighted mean squared residuals. Any bias analysis must be planned in advance.

And finally, everything depends on the consistency of the judges. Care is needed to choose an appropriate set of judges for each exercise. The final example showed that it is possible to choose judges in such a way that hypotheses can be tested about the existence of consensus.

16.16 Conclusions

1. It is often possible for test standards to be equated using CJ. This has been shown to be true for a wide range of school subjects in the UK, including sciences and mathematics from at least age 16.
2. ACJ will be more efficient than pre-designed CJ whenever equating uses scripts with a reasonably wide range of quality.
3. In general, where test equating is needed but the scoring will use traditional marking, a combination of Examples 2 and 4 is recommended: a suitable sample of scripts from both tests should cover the most important range of scores with adaptivity maximising the efficiency of the process.

4. It is sometimes possible for equating to be integrated into the scoring procedure, by 'seeding' the mix with selected scripts from earlier test(s). This is especially likely in cases involving complexity and creativity, such as design, art, or the assessment of long essays, explorations or projects.
5. One significant advantage of CJ over other empirical equating procedures is that it makes no extra demands on students: all of the equating activity is carried out by examiners after the test session. Also, if more precision is needed, this simply means asking the examiners to make some more judgements.
6. It is always wise to monitor for any systematic deviation from consistent judgement that may occur. The kind of analysis described here is very general: it can seek out evidence for any imagined source of bias in the data. At present, we are not clear how to set limits for how much discrepancy in a sub-set of the data can be tolerated.
7. It is important to understand the conditions in which the judges will be 'comfortable' in making decisions, and to ensure that no intentional or unintentional sort of deception is involved. It seems that judges may be 'led to believe' that they are looking for differences that really exist in the quality of the scripts, when there is in fact very little true variance. This is unwise; in fact, it is *never* wise to require judges to separate scripts which you believe are really of very similar quality.
8. The procedure is likely to be highly valid – if and only if there is agreement about who should constitute the panel of judges. Who are the experts in judging the quality of students' work? A long tradition of using teachers' rank orders to validate a testing system suggests that teachers similar to those who teach the students may, in general, be the most appropriate judges, but this needs to be established in fact for each individual judge by analysing their judgement record for misfit.
9. Since, in general, rank orders from scoring with CJ and with marking agree only moderately well, the question should always be asked whether CJ is more or less valid than marking as a method of scoring students' work.

References

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 15*(3), 297–334.

D'Arcy, J. (Ed.). (1997). *Comparability studies between modular and non-modular syllabuses in GCE advanced level biology, English literature and mathematics in the 1996 summer examinations*. Standing Committee on Research on behalf of the Joint Forum for the GCSE and GCE.

Kahneman, D. (2011). *Thinking, fast and slow*. New York/London: Allen Lane.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64*, 515–526.

Klein, G. (2008). Naturalistic decision making. *Human Factors, 50*(3), 456–460.

Laming, D. (2004). *Human judgment: The eye of the beholder*. London: Thomson.

Linacre, M. J. (2010). *A user's guide to Facets. 3.67.1*. Chicago: MESA Press.

Ofqual. (2014a). Setting standards for new GCSEs in 2017: Press release. https://www.gov.uk/government/news/setting-standards-for-new-gcses-in-2017. Accessed 10 Oct 2016.

Ofqual. (2014b) Guidance: Grade descriptors for GCSEs graded 9 to 1. thttps://www.gov.uk/government/publications/grade descriptors-for-gcses-graded-9-to-1. Accessed 10 Oct 2016.
Pollitt, A. (2004) *Let's stop marking exams*. Paper presented at the annual conference of the International Association for Educational Assessment, Philadelphia, June 2004.
Pollitt, A. (2012a). Comparative Judgement for assessment. *International Journal of Technology and Design Education, 22*(2), 157–170. doi:10.1007/s10798-011-9189-x.
Pollitt, A. (2012b). The method of adaptive comparative judgment. *Assessment in Education: Principles, Policy and Practice*. doi:10.1080/0969594X.2012.665354.
Pollitt, A., & Murray, N.L. (1993). *What raters really pay attention to*. Language Testing Research Colloquium, Cambridge. Reprinted in M. Milanovic & N. Saville (Eds.), (1996), *Studies in language testing 3: Performance testing, cognition and assessment.* Cambridge: Cambridge University Press.
Thurstone, L.L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273–286. Chapter 3 in L.L. Thurstone (1959), *The measurement of values*. Chicago: University of Chicago Press.
Wordsworth, C. (1877). *Scholae academicae*. London: Frank Cass.
Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Chapter 17
Closing the Loop: Providing Test Developers with Performance Level Descriptors So Standard Setters Can Do Their Job

Amanda A. Wolkowitz, James C. Impara, and Chad W. Buckendahl

Abstract Standard setting panels are tasked with recommending one or more performance level standards for assessments that are used to classify students into ability categories. These assessments are sometimes developed with the performance level descriptors known and other times without these descriptors. Based on an analysis of 11 state, educational, alternative assessments, this chapter investigates the effects on the standard setting process of developing a test both with and without these descriptors. The results suggest that the standard setting panelists are more consistent with one another and more aligned with empirical data when the items were developed with the descriptors in mind.

Keywords Standard setting • Performance level descriptors • Test development • Item development

17.1 Introduction

Standard setting panels are often in a quandary when recommending one performance standard for an assessment, and even more challenged when classifying examinees into more than two performance levels. The cognitive task of applying a content-based policy statement to an assessment score scale involves the interaction of multiple factors. When using test-centered methods such as Modified Angoff,

A.A. Wolkowitz (✉)
Alpine Testing Solutions, Orem, UT, USA
e-mail: Amanda.Wolkowitz@alpinetesting.com

J.C. Impara
Psychometric Inquiries, Buena Vista, CO, USA
e-mail: twopi@hughes.net

C.W. Buckendahl
ACS Ventures LLC, Las Vegas, NV, USA
e-mail: cbuckendahl@acsventures.com

S. Blömeke, J.-E. Gustafsson (eds.), *Standard Setting in Education*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-50856-6_17

e.g., the Yes/No modification (Impara and Plake 1997), or Bookmark (Mitzel et al. 2001) panelists must have sufficient numbers of items or measurement opportunities that can be answered correctly at each performance level to support interpretation of the classification. If there are no, or too few, items that can be answered correctly by the target examinee, the test score may not accurately differentiate between examinees at the different levels. A limited number of items at these levels will also reduce classification consistency evidence. It is, therefore, necessary to provide test developers with the performance level descriptors early in the development and validation process and direct them to ensure that they attempt to write test items that can be answered by examinees at each of the performance levels. This chapter discusses the meaning of performance level descriptors, when they should be developed, and presents a case study illustrating the importance of using performance level descriptors during the item writing stage of test development.

17.2 What Are Performance Level Descriptors?

Performance Level Descriptors (PLDs) are the descriptions used to define the categories into which examinees are classified based on their performance on an assessment (Egan et al. 2012). There are different types of PLDs that could be developed and used. *Range* PLDs are intended to characterize the knowledge, skills, and abilities of examinees within the full range of the respective category. As a subset, *threshold* PLDs are intended to define the entry point into a given category and are often the reference point for standard setting panels. *Reporting* PLDs may only include the higher-level policy definition and not be inclusive of the richness needed by test developers to support standard setting activities. For purposes of reference, our discussion of PLDs suggests that *range* PLDs would be recommended for test development with *range* or *threshold* PLDs recommended for use during standard setting studies, depending on the methodology (Egan et al. 2012).

PLDs in primary and secondary educational assessments are akin to the minimally qualified candidate description in professional credentialing. However, these assessments are often built with multiple performance level standards used to classify students into three or more categories, such as *Below Basic, Basic*, *Proficient,* and *Advanced*. In addition, these PLDs tend to focus on students who are clearly within the given category (i.e., *range*) as opposed to the student with minimum knowledge and skills required to achieve entry into each performance levels (i.e., *threshold*; Cizek and Earnest 2016).

To illustrate how *reporting* PLDs may be developed for an educational assessment for which there are multiple performance levels, consider the following for a United States, end of grade level, third grade mathematics assessment:

Level 0: Examinees achieving a score at this level have *less than a limited command* of the knowledge, skills, and abilities described in the state mathematics content standards for students in Grade 3. Students at this level *will require a substantial amount of additional academic support*.

Level 1: Examinees achieving a score at this level demonstrate a *limited command* of the knowledge, skills, and abilities described in the state mathematics content standards for students in Grade 3. Students at this level *will require some additional academic support*. Students at this level will consistently demonstrate a command of the following standards: (a list of the state mathematics content standards that a Level 1 examinee has mastered would be included here).

Level 2: Examinees achieving a score at this level demonstrate a *command* of the knowledge, skills, and abilities described in the state mathematics content standards for students in Grade 3. Students at this level *are academically prepared to continue further studies in mathematics*. Students at this level will consistently demonstrate a command of the following standards: (a list of the state mathematics content standards that a Level 2 examinee has mastered would be included here).

Level 3: Examinees achieving a score at this level demonstrate a *high level of command* of the knowledge, skills, and abilities described in the state mathematics content standards for students in Grade 3. Students at this level *are academically well-prepared to continue further studies in mathematics*. Students at this level will consistently demonstrate a command of the following standards: (a list of the state mathematics content standards that a Level 3 examinee has mastered would be included here).

These definitions not only describe the general qualifications for a particular performance level, but further specify the knowledge, skills, and competencies required for each level. Regardless of whether there is one or multiple performance standards, it is important that each PLD explains or lists the qualifications necessary to be classified at each level. These definitions support the validity of the interpretation of the test scores.

17.3 When Should Performance Level Descriptors Be Developed?

There is little, if any, debate that PLDs should be developed at or prior to a standard setting workshop. The question that remains is: When is the best time to develop these descriptors?

The PLDs provide the framework upon which standard setting panelists make judgments that differentiate the scores needed to achieve specific performance levels; see Giraud et al. (2005); Skorupski and Hambleton (2005) for additional information about how standard setting panelists use PLDs. When PLDs are developed prior to a standard setting and used during item development, the PLDs become part of the test specifications in that they indicate the level to which items should be written. As such, the resultant items conform to the test specifications that include the PLDs (Millman and Greene 1989). When the PLDs are developed early on in the

test development activities, their definitions and uses are directly linked to the test design and ultimate score interpretations. According to Plake et al. (2016), "By building the intended score interpretations into the test design, a foundation to support score interpretations is created. With the PLDs in place, the table of specifications can reflect the intended score interpretations, and item writers will have guidance on how to develop items with those interpretations in mind." Thus, developing the PLDs early on in the test development cycle supports the validity of the intended interpretation of the test scores.

If the PLDs are constructed during a standard setting workshop, then they should be constructed early on in the workshop or developed during a separate meeting just prior to the standard setting workshop so that they may be used during the judgmental process (Egan et al. 2012). However, by waiting to develop the PLDs until the standard setting workshop – after item development – there is a risk that there will be a limited number of items developed for each of the desired performance levels. When there are insufficient measurement opportunities available for panelists to consider, it adds challenge to establishing appropriate standards or cut scores for each performance level (Foley 2016; Wyse 2015). Further, the validity of the interpretations of the resulting classifications is then questionable.

In the worst case scenario, consider an assessment that has four performance levels, *Levels 1–4*, that were defined just prior to a standard setting workshop and not part of the item development process. After subject matter experts rate the items to establish the standard for each PLD, the results indicate the exact same cut score for entry into performance Level 2 as performance Level 3. These results suggest a lack of items targeting performance Level 3 and, therefore, a lack of items that can be used by the standard setting panelists to differentiate between students with the qualifications required to perform at a Level 2 versus Level 3. Thus, the assessment cannot distinguish between students in these two described performance levels.

In this hypothetical example, items were written for an assessment without targeting a specific performance level standard or without a strong consideration of the difficulty of the item for the target population. As a consequence, standard setting panelists faced the challenge of recommending achievement standards for multiple performance levels when there were a limited number of items available for certain performance levels. When this situation occurs, panelists may inadvertently introduce a classification error into their ratings by changing one or more of their item level ratings so as to force items into a level. This would ultimately lead to students being classified into a performance level to which they truly do not belong.

To emphasize the negative impact of this occurrence, consider an assessment with three performance level standards: *Beginner*, *Intermediate*, and *Advanced*. If the assessments were developed with items targeted at the three performance level standards, then students who were just able to meet the standard for entry into one of these levels would perform differently compared to those examinees who were more aligned with PLDs from the other levels. Thus, a standard setting panel would be able to identify items that naturally separated the *Beginner* from the *Intermediate* and the *Intermediate* from the *Advanced*. If, however, the items had not been developed to target these three levels, then it is conceivable that, for example, all of the

items may separate the *Beginner* from the *Intermediate* student and no items may separate the *Intermediate* from the *Advanced* student. Thus, standard setting panelists may be forced to rate one or more items as separating the *Intermediate* from the *Advanced* student, when they truly do not believe that any items accomplish this task. In this example, such a classification error in the item rating leads to a misclassification of students (i.e., classifying a student as *Advanced* when the student is truly *Intermediate*).

Classification error may also occur when setting multiple standards for exams with an insufficient number of items. For educational assessments, such errors can potentially affect the public perception and funding of schools that are held accountable for how their students perform on state assessments (Norman and Buckendahl 2008). If the standards set to interpret the assessment data are error ridden, then the degree of validity of the assessment results decreases as the interpretation of the tests scores becomes a less reliable measure of the intended purpose of the assessment.

For these reasons, it is prudent to consider the influence of PLDs at earlier stages in the test development and validation process, that is, prior the standard setting study and before or during item development. Defining the PLDs prior to item development, such as during the test design phases or early phases of the job analysis or blueprint development stage, allows items to target the different PLDs; thus, leading the way for standard setting workshops to have sufficient measurement opportunities at each level. Although the intended difficulty and actual, empirical difficulty level of the test items may not correlate perfectly for items written to target a specific PLD, developing items that are intended to focus on these target PLDs likely helps standard setting panelists be more consistent in their ratings as well as rate items to the appropriate level.

17.4 Introduction to the Case Study

17.4.1 Assessment

The assessments used in this case study were developed for use in a Southeastern U.S. state's alternate education assessments. These assessments are intended for students with the most severe cognitive disabilities. The grade level curriculum and assessment content are designed to represent the progression and continual development of knowledge and skills across the successive grade levels. Each assessment in English Language Arts (ELA) and Mathematics is aligned with the state's Extended Content Standards based on the *Common Core State Standards*[1]. The results of the alternate assessments are used to evaluate students' abilities and classify them into

[1] The Common Core State Standards are a series of academic content standards developed in the United States that have been defined for English Language Arts and Mathematics across primary, middle, and secondary grade levels.

one of four achievement levels (i.e., Performance Levels 1, 2, 3, and 4) with Level 3 designated as a goal for students having "met" the expectations of the academic content standards.

The 2009 and 2013 assessments included in this study represented ELA and Mathematics. The ELA assessments were administered to students in grades 4–8 and were each 15 items in length. The mathematics assessments were administered to students in grades 3–8 and each was also 15 items in length. The student performance data used for each content area and grade level for the 2009 and 2013 standard settings were based on between 900 and 1300 students. Each item on the 2009 and 2013 assessments was worth 2 points. However, the students could only earn 0 or 2 points on the 2009 assessments, whereas a student could have earned 0, 1, or 2 points on the 2013 assessments.

17.4.2 2009 Standard Setting Workshop and the Development of the PLDs

A standard setting workshop for the alternate assessments was conducted in 2009. For each grade span (e.g., 3–4, 5–6, 7–8), there was a panel consisting of 17–20 subject matter experts. Each panel included individuals with a variety of teaching backgrounds and included teachers who had experience with the state's Extended Content Standards, teachers who had experience working with students with disabilities, and general education teachers.

There were two goals of the workshop. The first goal was to produce a set of recommended *range* PLDs that summarized the expected knowledge, skills, and abilities of students at each level. The second goal was to elicit recommended cut scores that defined the expected performance for students within each performance level consistent with the PLDs.

For the 2009 studies, the PLDs were developed during this standard setting workshop. The PLDs corresponded to four levels of a student's command of the knowledge and skills contained in the Extended Content Standards for that particular area: Level 1 (limited command), Level 2 (partial command), Level 3 (solid command), and Level 4 (superior command). The PLDs further detailed each level by describing that Level 1 students would need academic support, Level 2 students would likely need academic support, Level 3 students would be prepared, and Level 4 students would be well prepared to be successful in further studies in this content area. The PLDs also contained specific abilities that students at the given level could demonstrate.

Because the PLDs for the assessment were developed during the 2009 standard setting workshop, the items that appeared on these assessments were not developed with knowledge of the PLDs. However, the PLDs that were established during the 2009 standard setting were available to inform future item development.

17.4.3 2013 Standard Setting Workshop and the Role of the Existing PLDs

Due to revisions to the assessments, another standard setting workshop for the alternate assessments was conducted in 2013. For each grade span, there was a panel consisting of 14–15 subject matter experts. Similar to the 2009 panels, each panel included teachers who had experience with the state's Extended Content Standards, who had experience working with students with disabilities, and who had general education experience.

Because the PLDs for these assessments had been developed in 2009, the items on the assessments that were used during the 2013 standard setting workshops were developed with the PLDs known to the test development team. Similarly, there was no need to redevelop the PLDs during the 2013 standard setting workshop. Instead, the panelists at this workshop divided into groups to review and enhance the existing PLDs. Each group was assigned one or two sets of PLDs to refine. This refinement process maintained the integrity of the existing PLDs, while also helping the standard setting panelists internalize the meaning and purpose of the PLDs. This process also helped the panelists gain a deeper understanding of the type of students included in each performance level and also helped them gain a better understanding of the differences between two adjacent performance levels.

17.4.4 2009 and 2013 Standard Setting Method

The recommended range of cut scores for both the 2009 and 2013 standard setting workshops was based on modifications of the Angoff (1971) standard setting method. In this process, panelists were presented with the assessment just as students would see it and were asked to make item-level judgments. For each item, they were asked to imagine the "target student" and make their best judgment as to the score the student would likely achieve on each item (i.e., 0 points or 2 points for 2009 assessment; 0 points, 1 point, or 2 points for the 2013 assessment).

For the 2009 standard setting workshop, the panelists followed the Yes/No method (Impara and Plake 1997) and rated each item based on whether they believed a borderline student would answer the item correctly (Yes; 2 points) or incorrectly (No; 0 points). This method was used because students on the 2009 assessment could only earn 0 or 2 points on an item. In contrast, students taking the 2013 assessments could earn 0, 1, or 2 points on an item. Therefore, the 2013 standard setting panels followed a method more closely aligned to the extended Angoff method (Hambleton and Plake 1995; Plake and Hambleton 2001) in which each panelist rated the items on a 2-point scale (i.e., panelists decided if a borderline student would earn 0, 1, or 2 points for each item).

For both the 2009 and 2013 panels, there were three groups of target students to consider: the students who were just achieving Performance Level 2 (separating Level 1 from Level 2), just achieving Performance Level 3 (separating Level 2 from Level 3) and just achieving Performance Level 4 (separating Level 3 from Level 4). By focusing on the threshold, or transition, points between the performance levels, panelists demonstrated their expectations for students with the minimum level of knowledge and skills at Levels 2, 3, and 4. These expectations were then used to represent the minimum score required for each of these levels (i.e., the cut scores).

Panelists recorded their judgments on specially designed rating forms that the facilitators of the standard settings collected and used to compute the panel-level statistics. Rating forms that included individual recommended cut scores were returned to panelists. The facilitators also shared with the panelists the group median cut scores, the range of cut scores across the panel, graphical representations of the distribution, the estimated impact if the median cut scores were used (i.e., what percentage of students would be classified at or above each achievement level), and the average item score from the 2009 and 2013 administration years for the respective panels.

In addition, each panel discussed two items for each assessment – one that was generally easier for students and one that was more difficult – to help with understanding how to apply the PLDs to the rating task. After explaining this feedback, the facilitators instructed the panelists to review their first round of ratings and, after receiving feedback, make any modifications they felt necessary in their second round of ratings. The second ratings were then used to compute the final recommended cut scores.

17.4.5 Evaluating PLDs as Related to These Standard Setting Workshops

As previously noted, the items appearing on the 2009 assessments were written without the guidance of PLDs, because the 2009 panelists developed the PLDs during the standard setting workshop. As a result, the 2009 panelists rated items during the standard setting for performance levels that had not been previously defined. In contrast, the items appearing on the 2013 assessments were written with the 2009 PLDs available to inform item development and form construction. To determine whether or not knowing the PLDs during the item development stage was advantageous to the standard setting process, the consistency of the Round 1 Angoff ratings (i.e., prior to the standard setting panelists being provided with impact data) was compared between the 2009 and 2013 panels.

The final recommended cut scores were not the focal point for comparison purposes because these decisions are influenced by multiple forms of feedback, such as item p-values, individual Round 1 cut score ratings, group Round 1 cut score ratings, and any type of group discussion about individual items prior to the Round 2 ratings.

Thus, if developing items with known PLDs benefits panelists by leading to more consistency between panelists with their *initial* ratings and more congruency with the item p-values prior to any feedback, then it can be surmised that developing PLDs at an early phase in the assessment design, prior to item development, is both an important and necessary step when developing assessments.

17.5 Effects of Using PLDs to Develop Items

17.5.1 Developing Items Using PLDs Positively Affects the Distribution of Assessment Scores

On a credentialing assessment, there is typically just one cut score. During a standard setting workshop, the purpose is to recommend a cut score that can differentiate between candidates who are and are not minimally qualified to hold that credential. If there are items that do not serve that purpose (e.g., items that are too easy or too difficult for the entire target population) then the item wastes precious real estate on the assessment. On an assessment with multiple cut scores that may be more commonly observed in educational settings, there is a parallel danger of writing items not targeted to specific PLDs. In these types of assessment, the ultimate distribution of items may be far off from the need (i.e., not maximizing items needed to differentiate among different levels of performance).

Figures 17.1 and 17.2 show the results from the case study described above. Due to the nature of these assessments, it was expected that more students would fall into Level 2 and Level 3 than into the extremes of Level 1 and Level 4. For the ELA students in 2009 (see panel A of Fig. 17.1), the percentage of students at each performance level generally increased as the performance level increased. For ELA grade 7, in particular, there was a large increase in the percentage of students achieving Level 4 compared to the other levels. The observed distribution for 2013 (see panel B of Fig. 17.1) is more closely aligned with the expected distribution of scores in that the greatest number of students tended to fall into either Level 2 or Level 3.

A similar change from 2009 to 2013 was also observed for the Mathematics assessments. Shown in Fig. 17.2, in 2009, there was not a clear trend across all grade levels as to which performance level the greatest percentage of students were classified. For example, the greatest percentage of students achieved Level 3 in grade 3, whereas the greatest percentage of students achieved Level 2 in grades 4, 5, and 8, and Level 4 in grades 6 and 7. In 2013, the distribution was more consistent with expectations. Specifically, in all grade levels, the great number of students achieved Level 2, followed by Level 3.

These results suggest that developing items with an awareness of the PLDs helped distribute the student population into more of an expected distribution with fewer students assigned to Levels 1 and 4 and more students assigned to Levels 2 and 3.

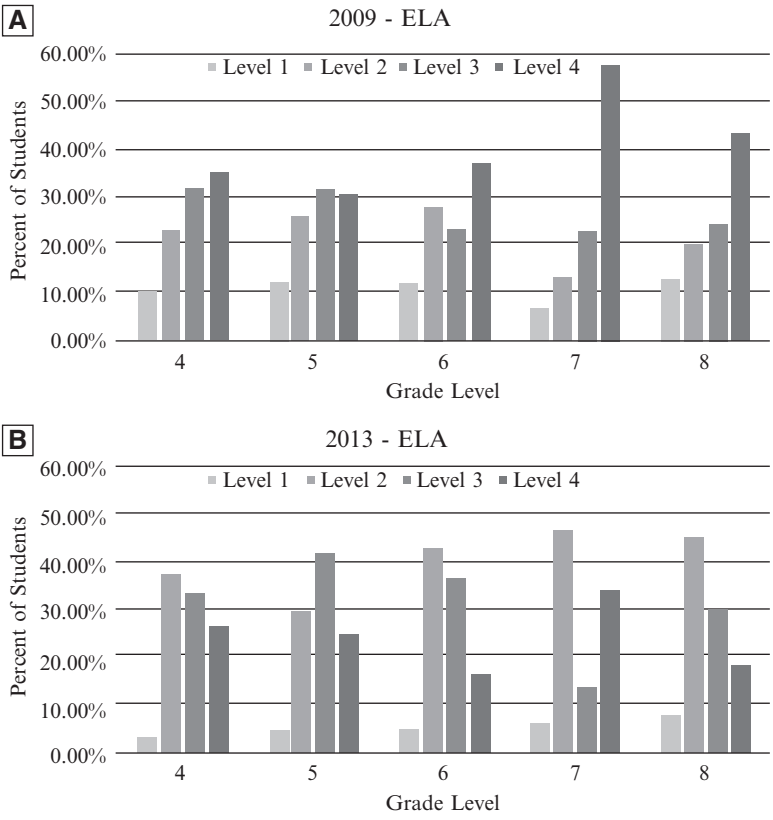


Fig. 17.1 Percentage of students at each ELA performance level by grade

17.5.2 Developing Items Using PLDs Increases the Congruence Between Item Difficulty and Standard Setting Ratings

Whether subject matter experts develop PLDs during the test design or other phases of development (e.g., job analysis, blueprint development), or at a standard setting workshop, there is an expectation that the cut score set for each performance level will increase as the difficulty of the items increase. However, to what extent is the strength of this correlation affected by whether the items were developed with the PLDs known?

Referring again to the case study, the average item rating at each performance level standard was computed across all panelists. This value was then compared to the corresponding item difficulty (p-value) for that item. The item difficulty scores were based on sample sizes between 900 and 1300 students depending on grade,

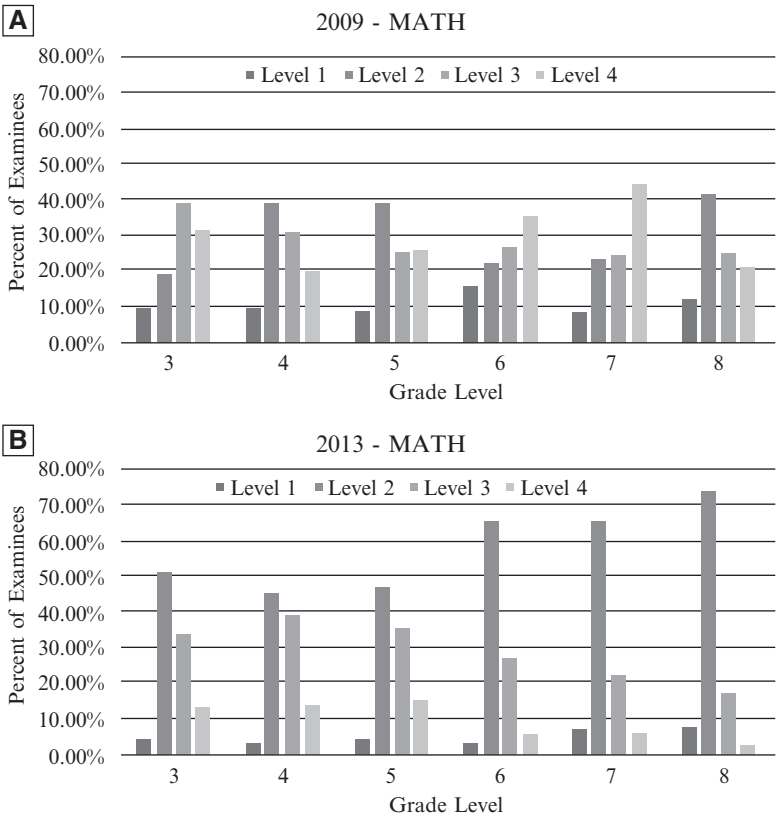


Fig. 17.2 Percentage of students at each Mathematics performance level by grade

content area, and administration year. Given that most students were expected to align with performance Levels 2 or 3, the item difficulty for the items would ideally fall within the Level 2 and Level 3 range with some items also falling into the Levels 1 and 4 ranges.

Figure 17.3 illustrates a comparison of the item difficulty of each item on the ELA grade 7 assessments to the Level 2 through Level 4 cut score ranges. The bottom of the vertical lines in Fig. 17.3 represents the (average) Level 2 cut score for each item. The top of the vertical lines represents the Level 4 cut score. The black dot is the item difficulty value. Figure 17.3 (panel A) shows the results of the 2009 version of this assessment in which PLDs were not known during item development. Seen in this figure, approximately half of the items (8 of 15) had item difficulty values that were between the Level 2 and Level 4 cut scores and the rest were either below the Level 2 cut score or above the Level 4 cut score. Specifically, the average item difficulty for one item was below Level 2, two items were between Level 2 and Level 3, six items were between Level 3 and Level 4, and five items

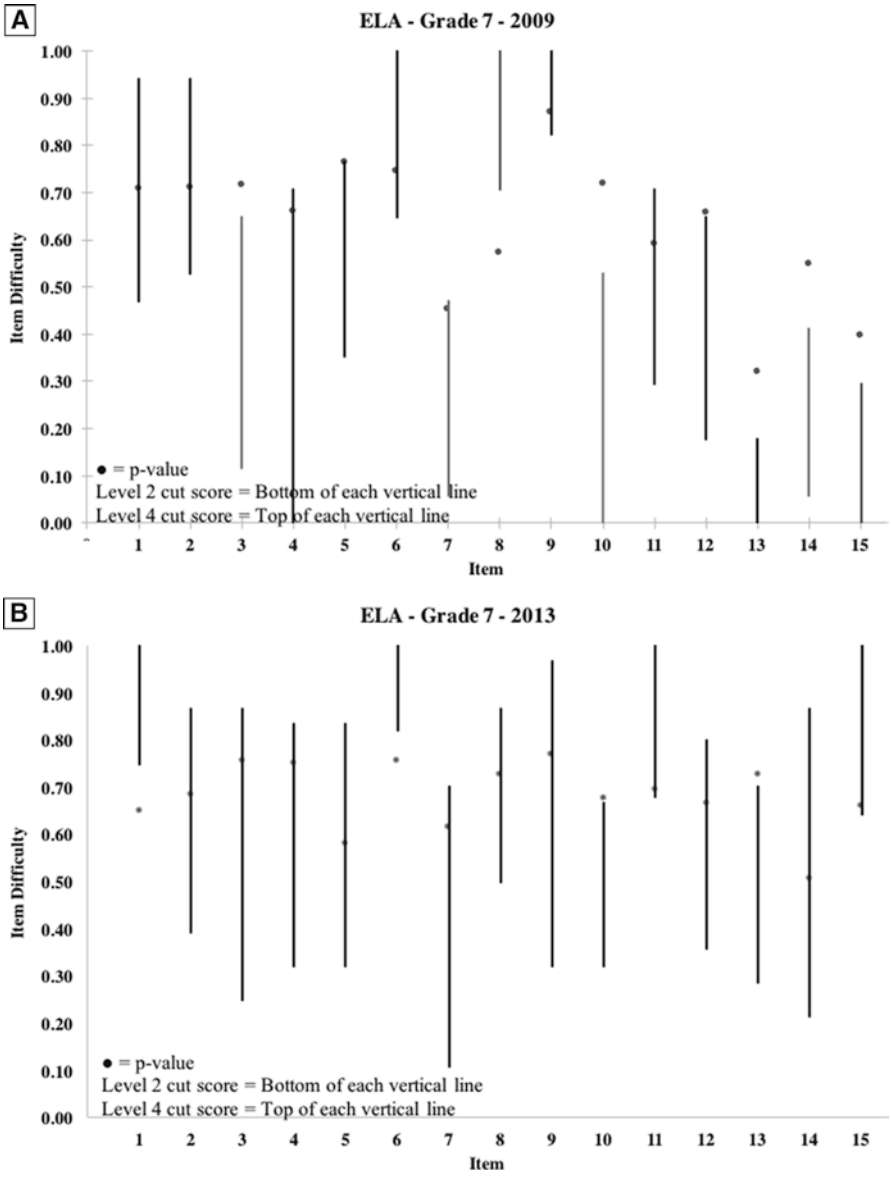


Fig. 17.3 Comparison of 2009 and 2013 item difficulty values to Level 2 and Level 4 cut scores for the ELA grade 7 assessments ((**a**) 2009 items not developed with PLDs in mind; (**b**) 2013 items developed with PLDs in mind)

were above the Level 4 cut score. This distribution of item difficulties lacks the ability to differentiate examinees who are at the different performance levels, especially between Level 1 and Level 2. Ideally, if item writers are paying attention to the PLDs, the difficulty of the items would be more strategically distributed to enable the test scores to differentiate among students across all performance levels.

Shown in Fig. 17.3 (panel B) with the 2013 assessments, when the items were developed with the known PLDs, approximately 75% of the items (11 of 15) were between the Level 2 and Level 4 cut scores and the remaining items were evenly distributed to be below the Level 2 cut score and above the Level 4 cut score. Specifically, the average item difficulty for two items was below the Level 2 cut score, six were between Level 2 and Level 3, five were between Level 3 and Level 4, and two were above Level 4. Given the expectation that a majority of students are between the Level 2 and Level 4 cut scores, this distribution of item difficulties is more aligned with the expectation and will help differentiate students according to the PLDs.

17.5.3 Developing Items Using PLDs Increases the Internal Consistency of the Standard Setting Results

When items are developed to target specific PLDs, then standard setting panelists will likely have an easier time determining the appropriate level for each item. The easier it is for the panelists to evaluate whether or not a minimally qualified candidate or a student just meeting a certain performance level will answer an item correctly, the more consistent the panelists' ratings. If the definition of the PLDs is unclear in the panelists' mind or if the PLDs did not exist prior to item development, then internal consistency of the panelists' ratings may suffer.

One way in which the internal consistency of the raters can be evaluated is by examining the range of Round 1 recommended cut scores across the panelists. More consistent raters would result in narrower ranges. Although such a comparison is sensitive to outliers within a panel, it does provide some general insight into the consistency of the ratings.

Returning to the case study, the range of recommended cut scores following Round 1 for each performance level and panel was compared. For example, Table 17.1 shows that the 2009 panels had Level 2 cut score recommendations spanning an 8-point range for Math-5 and Math-7 and a 16-point range for ELA-8. In general, the results displayed in Table 17.2 indicate that the 2013 ELA panels tended to be less consistent than the 2009 ELA panels for the Level 2 cut scores, but more consistent for the Level 3 and Level 4 cut scores. The same trend was found for the mathematics panels. Although these results are not without contradiction, the narrower range of recommended cut scores at the Level 3 and Level 4 cut scores

Table 17.1 Range of Panelists' Cut Scores after Round 1 by Assessment and Performance Level (total possible points = 30)

	Level 2		Level 3		Level 4	
Assessment	2009	2013	2009	2013	2009	2013
ELA – 4	12	21	14	18	12	*9*
ELA – 5	12	*7*	12	12	10	*7*
ELA – 6	12	14	12	*8*	12	*7*
ELA – 7	10	13	12	*9*	18	*5*
ELA – 8	16	16	18	*8*	12	*6*
Math – 3	12	*9*	20	*14*	14	*10*
Math – 4	12	*9*	14	*13*	14	*8*
Math – 5	8	12	12	15	16	*9*
Math – 6	10	19	12	15	12	*9*
Math – 7	8	16	12	*10*	12	*6*
Math – 8	10	*9*	14	*8*	8	*7*
MEDIAN (Median Abs. Deviation)	**12 (1.0)**	**13 (3.5)**	**12 (0.0)**	**12 (3.0)**	**12 (1.0)**	**7 (1.0)**

Note: *X* indicates the ranges in which the 2013 panel was narrower than the 2009 panel

Table 17.2 Percentage of Panelists Whose Individual Round 1 Cut Score was Within One Point of the Median Across all Panelist's Round 1 Cut Scores

	Level 2		Level 3		Level 4	
Assessment	2009	2013	2009	2013	2009	2013
ELA – 4	40%	33%	25%	*47%*	35%	*67%*
ELA – 5	30%	*47%*	25%	*40%*	65%	53%
ELA – 6	47%	40%	29%	27%	24%	*40%*
ELA – 7	35%	*47%*	18%	*40%*	24%	*47%*
ELA – 8	35%	33%	12%	*60%*	35%	*47%*
Math – 3	15%	*47%*	20%	*40%*	40%	33%
Math – 4	40%	*53%*	35%	*53%*	45%	*60%*
Math – 5	25%	*47%*	45%	27%	40%	33%
Math – 6	29%	*40%*	18%	*47%*	29%	*67%*
Math – 7	18%	*27%*	29%	*40%*	41%	*60%*
Math – 8	29%	20%	24%	*47%*	29%	20%
Median (Median Abs. Deviation)	30% (5.0%)	40% (7.0%)	25% (4.5%)	40% (7.0%)	35% (6.0%)	47% (13.0%)

Note: *X%* indicates that the 2013 panel was more consistent than the 2009 panel as measured by the percentage of panelists who were within one point of the median rating for the panel

support the action of developing PLDs prior to item writing activities so that they may be used to support item writing development and ease the job of the standard setting panels. It is possible that if there were more items on the assessment, then a similar trend may have been observed at the Level 2 cut score as well.

Table 17.3 Standard Deviation of Panelists' Round 1 Cut Score Recommendations

	Level 2		Level 3		Level 4	
Assessment	2009	2013	2009	2013	2009	2013
ELA – 4	2.74	5.18	3.73	4.35	2.62	*2.44*
ELA – 5	2.93	*2.14*	3.18	3.51	2.95	*2.21*
ELA – 6	2.85	3.63	3.82	*2.82*	3.28	3.72
ELA – 7	2.60	3.70	3.31	*2.66*	4.03	*1.58*
ELA – 8	3.89	4.33	4.56	*2.29*	3.10	*1.87*
Math – 3	3.43	*2.72*	4.82	*4.13*	3.34	*3.14*
Math – 4	2.91	*2.67*	3.08	3.31	3.33	*2.30*
Math – 5	2.62	3.19	3.06	3.83	4.01	*2.90*
Math – 6	2.92	4.91	2.83	4.00	2.96	*2.35*
Math – 7	2.93	4.51	3.06	*2.72*	3.09	*1.83*
Math – 8	2.54	3.02	3.94	*2.34*	2.73	*2.26*

Note: *X* indicates that the standard deviation of the 2013 panelists' recommended cut scores were smaller than that of the 2009 panelists' recommended cut scores

Another way in which the internal consistency of ratings can be evaluated is by calculating the percentage of panelists whose ratings are within one point (plus or minus) of their panel's recommended Round 1 median cut score. The higher the percentage of panelists meeting this criterion, the higher the agreement or consistency of the ratings. In addition, by using the median and the percentage of panelists who are within one point of that value, outliers will have no influence on this measure of consistency.

These calculations were performed on results from the case study to again compare the internal consistency of the 2009 standard setting panelists who did not have items targeting specific PLDs to the 2013 standard setting panelists who did. Table 17.2 displays the results. In general, the 2013 ELA and mathematics panelists had a greater percentage of ratings within one point of the median rating for the panel compared to the 2009 panelists; thus, the 2013 panels tended to be more consistent.

A third way to examine the internal consistency of the panelists' ratings is to compare the standard deviation of the ratings. The smaller the standard deviation of the ratings, the more consistent are the panelists' ratings. However, this measure is sensitive to outliers.

Table 17.3 displays the standard deviations for the case study. For the Level 3 and Level 4 cut scores, the standard deviations of the 2013 panels were smaller than that of the 2009 panels for both the ELA and mathematics panels. This trend was not observed for the Level 2 performance level cut score; however, a similar trend may have been observed had there been more measurement opportunities at this level. Overall, the results in Table 17.3 again suggest the usefulness of using PLDs during the item development stage.

17.5.4 Developing Items Using PLDs May or May Not Strengthen the Correlation Between the Item Difficulty and Standard Setting Ratings

Thus far, using PLDs prior to the item development stage has shown to help the distribution of scores on an assessment align with students' performance expectations, help standard setting panelists recommend expectations that follow empirical item performance, and help the internal consistency of the standard setting panelists' first round of ratings. A fourth way in which using PLDs during the item development stage may contribute to validity evidence of the standard setting results is found in the strength of the relationship between the item difficulty values and the average standard setting ratings at each performance level. In other words, as the p-value for an item increases, the probability that a student just meeting the standard for entry into a performance level will correctly answer the question also increases. Thus, a positive correlation might be expected between the average item p-values and the average item rating across all panelists. However, if items are developed to target the minimally qualified candidate or a student just meeting the minimum qualifications to achieve a certain performance level, then the empirical item difficulty values are likely more similar to each other and the rank order of the panelist's item ratings may be less exact than if the difficulty of the items were more spread out. Thus, items with more similar difficulty levels may result in weaker correlations and possibly negative correlations between the item difficulty value and panelist's average item cut score ratings. This does not necessarily threaten the validity of interpreting the results of a standard setting study. Instead, an appropriate investigation and explanation as to why there is a low correlation may provide additional evidence to support the interpretation of the standard setting results.

For example, in the case study, the correlation between the item difficulty and average item rating for the Level 3 cut score for Grade 6 ELA in 2009 was 0.90. In 2013, the value dropped to 0.51. The positive relationship in both years indicates that as the items became easier, the panelists believed that more students would correctly respond to the items who are just at the Level 3 entry point. Although this drop in the correlation value from 2009 to 2013 may seem undesirable, further investigation may reveal that the lower correlation value may be a result of the item difficulties in 2013 clustering around the Level 3 cut score versus being more spread out. Provided that there is a good spread of item difficulties across *all* items on the test so that there is a sufficient number of items to make cut score judgements for each performance level, the clustering of items at the different borderline performance levels is acceptable. Moreover, this clustering supports the validity of the study because there are more item-level judgements (i.e., more information) occurring at the point at which the "pass/fail" decision for the given performance level is being made.

To illustrate this point, Fig. 17.4 (panel A) shows the 2009 ELA item difficulties plotted against the average item rating for the Level 3 cut score across the panelists.

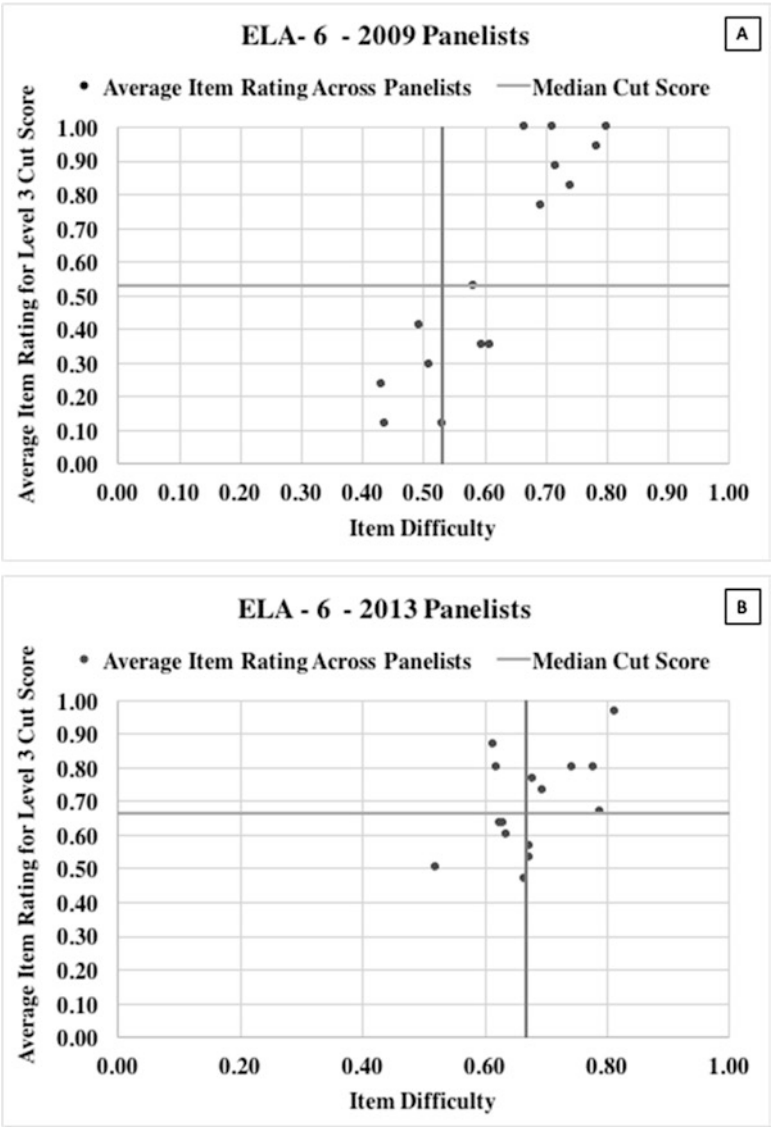


Fig. 17.4 Comparison of the item difficulty versus average Round 1 item rating

There was one item that had a difficulty rating very close to the average item rating for the Level 3 cut score. Figure 17.4 (panel B) shows a parallel graph for 2013. In 2013, there were at least four items that were very close to the average item rating for the Level 3 cut score. The clustering of items with similar difficulty likely resulted in the decreased correlation.

The wide spread of item difficulties shown in Fig. 17.4 (panel A) is needed to set cut scores for multiple PLDs. At the same time, more items that target the separation of examinees from one cut score to the next, as shown in Fig. 17.4 (panel B), is also desired. This particular case study would have benefited from a larger number of items so that both a spread of item difficulties could be achieved across the multiple performance levels while still targeting extra items at the anticipated cut score points.

17.6 Conclusion

PLDs for educational assessments or professional credentialing assessments are a necessary component of the standard setting process because they serve as the reference point for panelists' judgments. The question evaluated in this chapter is when to develop these descriptions for greatest utility. To help answer this question, a case study was presented that compared the results of developing PLDs during a standard setting to the results of developing PLDs prior to item development. The results indicated that when PLDs are known during the item development process, the recommended Round 1 cut scores from a standard setting panel lead to three advantages: (1) a more expected distribution of students at each performance level; (2) an increase in the congruence between item difficulty and item ratings, and; (3) an increase in the internal consistency of the ratings. These three findings support the validity of the interpretation of the ultimate test scores at each level.

The strength of the correlation between the item p-values and average ratings of an item across all panelists does not necessarily add to nor take away from the degree of validity of the standard setting process because the strength of the correlation is impacted by the internal consistency of the ratings and the range of the item difficulties. However, ideally, the correlation would be strong and there would be clusters of items at the targeted entry points to each performance level.

This chapter describes a case study in which the evidence provides support and recommendations for developing PLDs prior to item development and using them to inform the item development and form construction process. More generally, the findings suggest that when items are written to target one or more performance levels, standard setting panels can more readily and consistently determine which items a minimally qualified examinee will or will not correctly answer at each performance level. Given that the results presented in this chapter are from one case study, similar studies would benefit the field.

An advantage of developing the PLDs prior to item development is that the standard setting panelists may not be as frustrated in their task when there are items written at all performance levels. This reduction in frustration is accompanied with less likelihood of a classification error due to panelists trying to force items into levels into which they do not belong.

Additional advantages of developing the PLDs prior to the item development process is the development of fewer items that do not contribute to the intended classification decisions, more efficient standard setting results, and a stronger foundation upon which to build a validity argument for the intended interpretation and use of the test scores. Because item writers are targeting PLDs, they are more likely to write items that will target the different performance groups and ultimately help differentiate students on the border between two levels. In addition, the increased consistency of the first round of ratings may increase the efficiency of the second round (and possibly third round) ratings using test-centered methods like the Angoff (1971) method. Specifically, the "corrections" that are required during a second round of ratings could be substantially fewer, which would in turn have the potential of reducing the time required for the standard setting process. Finally, developing and integrating the PLDs early on in the test development cycle contributes to the validity framework of the exam because the intended interpretation of the test results are taken into consideration from the beginning of the design phase versus somewhere in the middle of the cycle. Based on these results, it is recommended that practitioners consider developing PLDs early in the test design phase to inform item development, item review, and test construction as well as standard setting.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington: American Council on Education.

Cizek, G. J., & Earnest, D. S. (2016). Setting performance standards on tests. In S. L. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 212–238). New York: Routledge.

Egan, K. L., Schneider, M. C., & Ferrara, S. (2012). Performance level descriptors: History, practice, and a proposed framework. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 79–106). New York: Routledge.

Foley, B. P. (2016). Getting lucky: How guessing threatens the validity of performance classifications. *Practical Assessment, Research & Evaluation, 21*(3). Available online: http://pareonline.net/getvn.asp?v=21&n=3

Giraud, G., Impara, J. C., & Plake, B. S. (2005). Teachers' conceptions of the target examinee in Angoff standard setting. *Applied Measurement in Education, 18*(3), 219–232.

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.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*, 353–366.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335–366). New York: Macmillan Publishing Company.

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: Lawrence Erlbaum Associates.

Norman, R. L., & Buckendahl, C. W. (2008). Determining sufficient measurement opportunities when using multiple cut scores. *Educational Measurement: Issues and Practice, 27*(1), 37–46.

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: Lawrence Erlbaum Associates.

Plake, B. S., Huff, K., Reshetar, R. R., Kaliski, P., & Chajewski, M. (2016). Validity in the making from evidence-centered design to the validation of the interpretations of test performance. In C. S. Wells & M. Faulkner-Bond (Eds.), *Educational measurement from foundations to future* (pp. 62–73). New York: The Guilford Press.

Skorupski, W. P., & Hambleton, R. K. (2005). What are panelists thinking when they participate in standard-setting studies? *Applied Measurement in Education, 18*(3), 233–256.

Wyse, A. E. (2015). The issue of range restriction in bookmark standard setting. *Educational Measurement: Issues and Practice, 34*(2), 47–54.

Chapter 18
Setting Standards to a Scientific Literacy Test for Adults Using the Item-Descriptor (ID) Matching Method

Linda I. Haschke, Nele Kampa, Inga Hahn, and Olaf Köller

Abstract Common standard setting methods such as the *Angoff* or the *Bookmark method* require panellists to imagine minimally competent persons or to estimate response probabilities, in order to define cut scores. Imagining these persons and how they would perform is criticised as cognitively demanding. These already challenging judgemental tasks become even more difficult, when experts have to deal with very heterogeneous or insufficiently studied populations, such as adults. The *Item-Descriptor (ID) Matching method* can reduce the arbitrariness of such subjective evaluations by focusing on rather objective judgements about the content of tests. At our standard setting workshop, seven experts had to match the item demands of 22 items of a scientific literacy test for adults with abilities described by performance level descriptions (PLDs) of the two proficiency levels Basic and Advanced. Since the ID Matching method has hardly been used in European standard settings, the method has not been evaluated comprehensively. In order to evaluate the appropriateness and correct interpretation of cut scores, information about the validity of standard setting methods is essential. In this chapter, we aim to provide procedural and internal evidence for the use and interpretation of the derived cut scores and PLDs using the ID Matching method. With regard to procedural validity, we report high and consensual agreement of the experts regarding explicitness, practicability, implementation, and feedback, which we assessed by detailed questionnaires. The inter-rater reliability for the panellists' classification of items was low, but increased during subsequent rounds ($\kappa = .38$ to $\kappa = .63$). The values are consistent with findings of earlier studies which support internal validity. We argue that the cut scores and PLDs derived from the application of the ID Matching method are appropriate to categorise adults as scientifically illiterate, literate, and advanced literate.

Keywords Item-Descriptor Matching method • Internal validity • External validity • Science abilities • Standard setting

L.I. Haschke (✉) • N. Kampa • I. Hahn • O. Köller
Leibniz Institute for Science and Mathematics Education (IPN), Kiel, Germany
e-mail: haschke@ipn.uni-kiel.de

S. Blömeke, J.-E. Gustafsson (eds.), *Standard Setting in Education*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-50856-6_18

18.1 Introduction

Standard setting procedures pose cognitively demanding tasks for expert panels. Cut scores defining the boundaries between proficiency levels are based on individual judgement. In widely used standard setting methods such as *Angoff* or *Bookmark* (Lewis et al. 2012; Plake and Cizek 2012), panellists have to imagine an examinee on the border between two proficiency levels. This task in itself is demanding (Ferrara et al. 2008; Shepard et al. 1993), but in the case of rather insufficiently studied populations such as adults, it becomes even more difficult. Adult surveys such as the *Programme for the International Assessment of Adult Competencies (PIAAC)* mainly focus on skills in mathematics or reading (OECD 2013). There are neither surveys testing scientific literacy in adults, nor are there any standards for the scientific literacy of adults. Hence, it seems to be extremely challenging to imagine a minimally scientifically literate person. Inaccurate assumptions might lead to imprecise cut scores.

This problem can be circumvented with alternative standard setting methods. For example, the Item-Descriptor (ID) Matching method (Ferrara and Lewis 2012; Ferrara et al. 2008) mainly focuses on the content of tests without considering the examinee. It requires panellists to analyse items with regard to abilities and skills that are needed to answer the items correctly, and then assign them to performance levels according to their similarities in cognitive demands or to existing performance level descriptors (PLDs).

The ID Matching method is rarely used in European standard setting procedures. This leads to a lack of data on the validity of using and interpreting cut scores derived from the ID Matching method (Bazinger et al. 2013; Freunberger 2013; Freunberger and Yanagida 2012). Following guidelines to validate results of standard setting methods (Kane 1994; Pant et al. 2009; Pitoniak 2003), we aim to provide evidence for procedural validity of the application of the ID Matching method as well as for the internal validity of the resulting cut scores.

In order to contrast the ID Matching method against other standard setting methods, we start with a brief overview of two widely used standard setting methods, the *Angoff* and *Bookmark methods*. We proceed with a more detailed description of the ID Matching method and then introduce the procedural, internal, and external validity aspects according to Pitoniak (2003) as well as the supplemented consequential validity reported by Pant et al. (2009). We base our results on procedural and internal validity of data from the *National Educational Panel Study* (NEPS), which is the first study to measure adults' scientific literacy in Germany (Artelt et al. 2013; Blossfeld et al. 2011). We make reference to the use and interpretation of the derived cut scores, and discuss advantages and disadvantages of the ID Matching method.

18.2 Choosing the Accurate Standard Setting Method for an Adult Scientific Literacy Test

The most widely used standard setting methods are the Angoff Method and the Bookmark Method (Cizek 2012b; Zieky 2012). In the simplest version of the Angoff Method, panellists have to decide if a minimally competent person will answer each test item correctly (Angoff 1971). The sum of items that are labelled 'yes' will make up the cut score which distinguishes between participants failing or passing the test. For instance, if the experts judge that minimally competent participants will answer at least 30 out of 100 items correctly, all participants have to solve 30 items in order to pass the test. On the one hand, the clear instructions and application makes the Angoff Method and its variations the most widely used method for standard setting procedures. On the other hand, the Angoff method is often criticised for not integrating empirical item difficulties (Buckendahl et al. 2002; Lewis et al. 2012).

The Bookmark method follows a different approach and is based on an *Ordered Item Booklet* (OIB; Lewis et al. 2012). Items are arranged by their empirical difficulty beginning with the least difficult to the most challenging item (Cizek 2012b; Karantonis and Sireci 2006; Lewis et al. 2012). In this method, the panel members imagine a minimally competent person at each proficiency level and estimate the probability of that person solving each item. They place a bookmark (i.e., a cut score) between the two items between which the response probability drops below a certain percentage, usually 67% (Mitzel et al. 2001).

The panellists' task of judging and imagining a minimally competent person and estimating probabilities is highly criticised (Shepard et al. 1993). Since there are neither scientific literacy studies for adults nor any standards for scientific literacy of adults, it is particularly challenging for experts to imagine an adult with such minimal scientific literacy. Therefore, using the two described methods for the NEPS scientific literacy test for adults seems highly inappropriate. To overcome these critical aspects, we propose the Item Descriptor (ID) Matching method for setting standards to the NEPS scientific literacy test for adults.

18.2.1 Item-Descriptor (ID) Matching Method

The ID Matching method is a test-centred standard setting method. Within the procedure panellists match the skills and abilities needed to solve an item according to performance levels (Ferrara and Lewis 2012; Ferrara et al. 2008). This main feature distinguishes the ID Matching method from the two described methods. While the Bookmark method (Lewis et al. 2012) challenges the experts to determine the probability of examinees answering an item correctly, the Angoff Method (Plake and Cizek 2012) requires the experts to estimate the ability of a hypothetical examinee whose success in mastering the items will define the basic cut score.

The ID Matching method shares most of its procedure sections with common standard setting methods such as training, practice, and iterating rounds. It can be

viewed as an answer to the critique on the methods that require panellists to estimate probabilities and imagine examinees ranging between proficiency levels. The ID Matching method aims to offer a judgemental task which is more consistent with the panellists' expertise (Ferrara and Lewis 2012). Since most of the panel members in standard setting workshops have teaching experience or are involved in educational assessments, their skills include the analysis or development of items. Therefore, it seems likely that they are well capable of matching item demands to PLDs (Ferrara et al. 2008).

The basis for the ID Matching method is an OIB in which items are ranked by their difficulty, beginning with the least difficult to the most challenging one. Within each standard setting round, the experts have to work through each item of the OIB and have to answer two essential questions:

i. *'What do students need to know and be able to do in order to respond successfully to this item?'*
ii. *'What makes this item more difficult than the ones that precede it?'*

(Ferrara et al. 2008, p. 13)

In the next step, they have to match the item demands with the expectations of the PLDs. PLDs define skills and abilities of persons specified at levels of achievement, such as *Basic*, *Proficient*, and *Advanced* (Egan et al. 2012). The underlying question at this step is 'Which PLD most closely matches the knowledge and skills required to respond successfully to this item (or score level for constructed-response items)?' (Ferrara and Lewis 2012, p. 262). The panel members document their decisions in so-called *item maps*. Item maps list the items according to their position within the OIB (see Fig. 18.1). Each row can include various additional information (e.g., item

OIB Page	Item-Descriptor Matches	
1	B	
2	B	
3	P	Threshold Region
4	B	
5	B	
6	P	
7	P	
8	P	
9	A	Threshold Region
10	P	
11	A	
12	A	
13	A	

Fig. 18.1 Item maps with hypothetical item-descriptor matches to the proficiency levels *B* Basic, *P* Proficient, *A* Advanced. *Grey* cells represent alternating matches depicting the threshold regions

format, the original location in the test or the location on the IRT scale). Ideally, the panellists match each item to one of the PLDs, whereas less difficult items are sorted to lower proficiency levels and items that are more difficult are assigned to higher proficiency levels. The cut score between two proficiency levels would then be on the first item, which is assigned to the next higher performance level.

It is also possible, that experts alternately match the items to PLDs between sequences of consensual matches (see grey cells in Fig. 18.1). These regions are called *threshold regions* and extend from:

> '(a) the first item that matches a higher performance level descriptor, just after a consistent run of matches with a lower performance level descriptor, to (b) the final item just before the first run of three matches to the next higher performance level.'
>
> (Ferrara et al. 2008, p. 10)

The cut scores between two performance levels will be located within these threshold regions. In the example in Fig.18.1, the cut scores between the performance levels Basic and Proficient would be at item number 4, while item number 8 would mark the cut score between the levels Proficient and Advanced. Ferrara and Lewis (2012) propose four options to determine a cut score. First, panellists can define an item within the threshold region as the cut score using their best judgement. Second, the cut score is set at the first item from the row of three items that are matched to the next higher proficiency level. Third, psychometricians calculate the cut score as the midpoint of the threshold region or fourth, they calculate the cut score via regression analysis.

The ID Matching method works with an iterative procedure in which the panellists match the items to the PLDs in subsequent rounds. This leads to varying cut scores during the standard setting and finally results in cut scores that can be considered as more or less arbitrary. Therefore, it is essential that the panellists receive feedback on their decisions after each round and are able to discuss this feedback with the other panel members. Moreover, the standard setting process should be evaluated in order to prove the appropriate use and interpretation of the derived cut scores and PLDs.

18.3 Validation of Standard Setting Methods

> '…'validation' is associated with a critical evaluation of the extent to which the proposed interpretations and uses are plausible and appropriate.'
>
> (Kane 2012, p. 4)

Since there is neither 'the best' standard setting method nor any 'true' cut score (Ferrara and Lewis 2012), standard setters can only provide evidence through implementation of a method and through the interpretation of the determined cut scores. Several guidelines provide suggestions on how to evaluate the appropriateness of cut scores that derive from a standard setting method (Berk 1986; Cizek 2012a; Hambleton 2001; Kane 1994, 2012; Sireci 2007).

Pitoniak (2003) integrated existing concepts of validity as well as conceptions of how to evaluate standard setting methods and provided a systematic way to gather validity information on any given standard setting workshop. Her categorical system discriminates the three evaluation elements *procedural, internal*, and *external* validity (see Table 18.1).

The *procedural validity* element aims to provide evidence regarding the experts' confidence in the results of a standard setting workshop. Aspects of this element are (a) the reasonableness of the performance standards, (b) the involvement of unbiased panellists, (c) the experts' understanding of the purpose of the performance standard, and (d) their understanding of the underlying process they are involved in (Hambleton et al. 2012). The procedural validity element is subdivided into the aspects explicitness, practicability, implementation, feedback, and documentation. The aspect e*xplicitness* covers the panellists' level of information about the purpose and procedures of the standard setting, whereas the category *practicability* takes into account whether the instructions were easily applicable for the panel members. Within the aspect *implementation* the main focus is on accuracy of the methods' application and the justification of aberrations. The category *feedback* gives information about the panellists' confidence with the procedures and with their decisions. A comprehensive *documentation* of the whole process is essential because it forms the basis for validating the interpretation and use of the derived outcomes (e.g., cut scores or PLDs).

The internal validity element concerns empirical evidence on consistency (Pant et al. 2009). In an ideal case of internal validity, a repetition of the same standard setting workshop under the same conditions would result in the same performance standards and cut scores (*consistency within method*). Since panellists differ in their level of expertise in setting performance standards or in particular standard setting methods, the variance of judgements on cut scores among panellists across rounds (*intra-panellist consistency*) as well as among panellists (*inter-panellist consistency*) needs to be evaluated. It is desirable that panellists adapt their judgements based on provided feedback and discussions with the panel and that those adaptations lead to convergence across the panel over time (Pant et al. 2009). The aspect *other measures* addresses for instance the stability of the cut scores across item types, content areas or cognitive processes.

The external validity element focuses on the comparison with other sources of evidence. Results of at least two standard setting methods are compared within the *comparison to other standard setting methods*. Such comparisons should be interpreted with caution as they often lead to differing and inconclusive results (Hambleton et al. 2012; Pant et al. 2009). In contrast, the comparison with *other sources of information* is rather informative. A cut score should be aligned to other available information about the examinees, such as passing or failing other scientific literacy tests, science grades or success in the science-related labour market.

Pitoniak (2003) primarily subordinated *reasonableness* to external validity evidence, whereas Pant et al. (2009) report it under the supplement of consequential validity evidence (see Table 18.1). Consequential validity evidence refers to the alignment of cut scores and proficiency levels with the purpose of setting those

Table 18.1 Standard setting evaluation elements – adapted and integrated from Cizek (2012a), Pant et al. (2009), and Pitoniak (2003)

Evaluation element	Description
Procedural	
Explicitness	The degree to which the standard setting purposes and processes were clearly and explicitly articulated to panellists
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 procedures were reasonable, systematically, and rigorously conducted, including the selection and training of panellists, definition of the performance standard(s), and data collection
Feedback	The extent to which panellists have confidence in the process and in the 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)
Intra-panellist consistency	The degree to which a panellist is able to provide ratings consistent with the empirical data, and the degree to which ratings change across rounds
Inter-panellist consistency	The consistency of item ratings and cut score(s) across panellists and the degree to which group ratings converge across rounds
Decision consistency	The extent to which the identified performance standard(s) can be expected to yield consistent classifications of examinees
Replicability	The extent to which the procedure can be expected to produce consistent results across samples of equally qualified panellists using the same method
Other measures	The consistency of cut scores across item types, content areas, and cognitive processes
External	
Comparison to other standard setting methods	The agreement of cut scores across replications using other standard setting methods
Comparison to other sources of information	The relationship between the decisions made using the test to other relevant criteria (e.g., grades or performance on tests measuring similar constructs, etc.)
Consequential	
Reasonableness of cut-score(s)	The extent to which cut score(s) recommendations are feasible or realistic (including pass/fail rates and differential impact on relevant subgroups)
Adequacy of reporting and reception	The extent to which resulting cut score(s)/proficiency categories are reported and interpreted in alignment with the performance standard(s)

standards. The *reasonableness* of cut scores evinces their adequacy, which means that the proportion of examinees at the different proficiency levels can be evaluated by comparing them to the distribution of examinees in other studies or concerning other competencies (Hambleton et al. 2012).

In our study we developed standards for scientific literacy of adults using the NEPS scientific literacy test for adults. Since this is the first test of its kind we could not rely on previous data or any other comparable references. In this regard, it seems neither feasible to imagine a minimally scientifically literate adult nor to estimate a response probability for such an adult. Since those tasks are required by the Angoff and Bookmark methods, we argue that they are not appropriate in this case. We chose the ID Matching method for our standard setting because it leaves out the examinees and only requires the panel to match item demands to PLDs. The derived cut scores will categorise German adults as scientifically illiterate, literate, and advanced and the PLDs will form the basis for further assessments in scientific literacy of adults. Both, cut scores and PLDs, lead to educational and political implications and therefore should be validated in order to support these proposed interpretations and uses (Kane 2012).

In our validation study we focus on the procedural and internal validity of the interpretation and use of cut scores obtained using the ID Matching method. Because of a lack of suitable external resources of information for the NEPS test, we needed to exclude external and consequential validity.

18.4 Methods

18.4.1 The NEPS Scientific Literacy Test

The German longitudinal study NEPS tracks *inter alia* the development in reading, mathematics, information and communication literacy (ICT), and scientific literacy over the lifespan from new borns to adults (Artelt et al. 2013). In 2012, a scientific literacy test for adults incorporating 20 multiple choice and 2 multiple true-false items which were administered to a German adult sample. This sample consisted of 6625 participants between the age of 27 and 69 (M = 50.12, SD = 13.92). Approximately half of the sample (51.23 %) was female. All participating adults were selected in 2009 as a representative sample by the *Institute for Applied Social Sciences* (infas). Since then, the adults have attended four waves of longitudinal assessments and have already participated in mathematics and reading tests (Haschke and Kähler in press).

The item development was based on the NEPS framework for scientific literacy that distinguishes between the two domains knowledge of science (KOS) and knowledge about science (KAS) (Hahn et al. 2013). An item example for the domain KOS can be found in Appendix A. The participants were tested at their homes via a paper-pencil-test with a maximum test length of 25 min.

In order to estimate the adults' scientific literacy, we scaled the data based on *Item-Response-Theory* (IRT; Lord 1981; Moosbrugger 2012; van der Linden and Hambleton 1997) using ConQuest (Wu et al. 2007) and technical guidelines for scaling competence tests within the NEPS project (Pohl and Carstensen 2013). We dichotomised each item and fixed the response probability at .67 as it is suggested by Ferrara et al. (2008). Missing values (e.g., not reached or omitted) were treated as non-response. We excluded adults from the scaling procedure who gave less than at least three valid responses. We scaled the data based on the Rasch-Model to arrive at *weighted likelihood estimators* (WLE) as point estimators for our persons´ ability. The resulting performance scale ranged from −5.30 to +4.26 logits ($M = -.18$, $SD = 1.03$). The test variance was 1.00 and the reliability of the test was .72 (Haschke and Kähler in press). In order to facilitate the interpretation of the scale, we transformed the WLE to a mean person parameter of 500 and standard deviation of 100. Afterwards we prepared the OIB (see Sect. 18.4.2) which was the basis for the standard setting process.

18.4.2 Setting Performance Standards with the ID Matching Method

The ID Matching method requires panellists to match item demands to PLDs. Since standards for scientific literacy for adults do not exist, we had to develop preliminary PLDs based on the NEPS scientific literacy test. Three male scientists within the field of science education with a high expertise on performance standards for science (e.g., PISA or the German educational standards) were invited to formulate these PLDs. Since the number of items in the scientific literacy test was limited, we reduced the three typically used proficiency levels Basic, Proficient, and Advanced to the two levels Basic and Advanced. The preliminary PLDs included specific abilities for the two knowledge domains KOS and KAS, respectively.

In the next step we conducted the standard setting workshop during two consecutive days in February 2015 (see Appendix B). Ferrara and Lewis (2012) advised to optimise the cognitive challenge for panellists by choosing experts according to the judgemental task. Due to the lack of experts in Germany familiar with scientific competencies or scientific standards for adults, we invited professors and researchers who either had substantial experience in the development of science tests or in the development of science standards for school students. The panel comprised seven experts, two of whom were from the respective science disciplines biology, chemistry, and physics. The seventh expert was a psychologist with a high expertise in educational assessment regarding adults. Two of the experts were female.

The standard setting workshop consisted of the essential steps proposed by Hambleton et al. (2012). In the orientation phase, we introduced the NEPS project to the panel, explained the method for assessing scientific literacy, and informed the panellists about setting performance standards in general as well as about the ID Matching method. Afterwards, the experts examined the preliminary PLDs and

practiced the ID Matching method based on item examples within a training session. The precise question was '*Which PLD most closely matches the knowledge and skills required to respond successfully to this item (or score level for constructed-response items)*?' (Ferrara and Lewis 2012, p. 262). During this phase, we invited the experts to share their thoughts with the other panel members and to discuss emerging problems and questions. The first standard setting round took place after the extensive training. The panellists worked on the OIB individually and matched each item to the PLDs via item maps (see Fig.18.1). After this first round, we gave the feedback to the experts displayed in Fig.18.2 (see results section). For each item, the experts could examine how often they allocated it to the proficiency level Basic and how often to the proficiency level Advanced. For example, during the first round, four out of seven experts (57.14 %) matched the first item to the Basic level. The remaining three experts (42.86 %) matched the same item to the next higher proficiency level Advanced. We asked the experts to discuss their decisions and to give suggestions about the precise abilities within the PLDs that are required to solve a specific item. In the second round, we repeated the procedure. We afterwards enriched the feedback and additionally presented a preliminary cut score as well as the percentages of adults on the two proficiency levels. We calculated the cut score as the mean of the threshold regions of each panellist. The experts again discussed the required abilities of items that were not yet allocated consensually. The third and last round was conducted as a panel discussion. The panellists set the final cut score based on the best judgment.

During the second day, the panellists had to discuss and summarise the abilities required by the items that were allocated to one performance level. They examined the preliminary PLDs carefully and, when necessary, extended them with regard to the allocated items. In an additional fourth round, the experts set the cut score for the Below Basic level via the Direct Consensus Method (Sireci et al. 2004). In a panel discussion, they deliberated on how many of the 22 items a minimal scientifically literate adult could master successfully. The corresponding person ability value on the performance scale served as the cut score between the performance levels Below Basic and Basic. Since a person with a performance score below that cut score did not reach the Basic level, we label this person as scientifically illiterate. In order to support the fourth round, we presented the percentages of adults on the performance levels Below Basic, Basic, and Advanced. We concluded the standard setting workshop with a final discussion about scientific standards for adults in general and summed up the standard setting workshop.

18.4.3 Examination of Validity Aspects

Validity elements according to Pitoniak (2003) and Pant et al. (2009) are procedural, internal, external, and consequential. The procedural and internal validity elements focus on the procedure of setting standards, which is the foundation for its subsequent interpretation and uses, as well as on the experts' decisions during the

process. External and consequential validity elements focus on the outcome, that is, the cut scores derived by the standard setting process and their subsequent interpretation and uses. The latter two validity aspects require external sources for comparison or other sources of information, such as correlation of the cut scores with specific outcome variables of the examinees such as school grades. Due to the lack of standards for adults in science, missing precedent standard setting procedures for adults, and access to cross sectional information only (instead of longitudinal information about the adults), we could not add substantial external or consequential criteria into our validation study. Hence, we had to restrict our validity study to procedural and internal aspects.

18.4.3.1 Questionnaire for Procedural Validity

Procedural validity consists of the five dimensions explicitness, practicability, implementation, feedback, and documentation (Pitoniak 2003). While Cizek (2012a) gives an overview of a systematic evaluation, data from standard setting procedures, some evaluation questionnaires from a pilot study of Plake and colleagues or Freunberger are available as well (Freunberger 2013; Plake et al. 2008). We translated the questionnaires into German and adapted them to the specific terms of the ID Matching method.

In preparation for the upcoming standard setting workshop, the panellists had to give information about their background. We were especially interested in their experience with setting performance standards and the ID Matching method. Within the standard setting process, the panellists had to answer questionnaires after each section (see Appendix A). In total, we administered eight questionnaires to the panel. Since only five questionnaires (introduction, training, round 1, round 2, and final) addressed the application of the ID Matching method, we only present findings from these questionnaires in this chapter.

Each questionnaire comprised 10 to 40 4-point Likert-scale items on the experts' agreement to specific statements on the validity aspects explicitness, practicability, implementation, and feedback. For instance, after the training and round one and two, a statement referring to feedback was 'I am comfortable with my ability to apply the ID Matching method'. The panellists could *strongly disagree*, *disagree*, *agree* or *strongly agree*.

First, we recoded the agreement as an equidistant scale with a range from 1 = *strongly disagree* to 4 = *strongly agree*. The centre of the scale was at 2.5. We then calculated the mean score of agreement for the four validity aspects for each panellist. Afterwards, we summarised the data in order to calculate mean scores for each element. We also determined whether the agreements were significantly below or above the centre of the scale at 2.5 via *t*-test. We interpreted a significant score above 2.5 as a clear agreement. This allowed us to investigate single problematic sections within the process. In order to examine large variances within the panel, we estimated standard deviations. As we administered some of the questions repeat-

edly, we were able to identify significant increases and decreases of agreement over time via repeated measurement ANOVAs in SPSS 19 (IBM 2010).

Due to the relatively small sample size, problems in regard to the test power may arise. When the sample size is small, effects have to be quite large in order to produce significant differences. The results are trustworthy if the ANOVA indicates significant differences. Non-significant differences might still be practically relevant. Therefore, we also report the effect size Cohens *d* in order to be able to interpret the ANOVA results. Additional open response questions such as, '*One thing that might require explanation before we move on is…*' or '*How many items out of 22 do you think you categorised confidently?*' helped us to understand differences in the panellists' agreements in a more detailed way.

18.4.3.2 Inter-rater Reliability

The internal validity described by Pitoniak (2003) comprises consistency within method, intra-panellist consistency, inter-panellist consistency, and other measures. As panellists are expected to vary within their decisions according to cut scores, the *inter-panellist consistency* can provide evidence for the consistency of agreements (Pant et al. 2009). If panellists cannot reach a consensus after several rounds, the derived cut scores should be considered as not representative (Pant et al. 2009). Therefore, we evaluated the panellists' consensual adaption of their decisions while establishing the cut score across rounds. After the first and second round we examined how often each item was assigned to which performance level by each expert and calculated the Fleiss' Kappa (κ) as the inter-rater reliability (Fleiss et al. 2003). The strength of agreement is graded as follows: poor (κ<.00), slight (.01< κ <.20), fair (.21< κ <.40), moderate (.41< κ <.60), substantial (.61< κ <.80), and almost perfect (.81< κ <1.00) (Landis and Koch 1977, p. 165).

18.5 Results

The aim of our validity study was to prove procedural and internal validity for the use and interpretation of cut scores established through the ID Matching method. In the following section we first cover the procedural aspects and proceed to internal aspects.

18.5.1 Results on Procedural Validity

First, we give an overview of the panels' characteristics in order to facilitate the later interpretation of the results. In preparation of the standard setting procedure we asked each panel member to rank their knowledge and experiences about setting performance standards and about their knowledge and experiences with scientific literacy of adults.

With regard to the topic of setting performance standards, six out of seven panellists stated that they were familiar with standard setting procedures, while only two panellists had heard of or knew the ID Matching method. Nearly half of the panel (three out of seven) had already participated in a standard setting workshop. Furthermore, all experts reported that they had experience in the development and assessment of educational assessment tests and felt competent with the application of proficiency levels.

Concerning adults as the target population, three out of seven panellists stated that they possessed extensive experience within the field of adult education, one expert reported moderate experiences, while two panel members hardly seemed to have any preliminary experiences. One expert admitted to having no previous experiences with adult education. In an open question, the panellists had the opportunity to give some examples from their experiences. We emphasise that those experiences pertain primarily to the education of teachers and students, but do not cover adults across the lifespan or from various backgrounds. One expert already developed and administered a performance test for adults. Overall, two panellists were very confident and three panel members were at least confident that they were able to appraise adults and their skills satisfactorily. The last two panellists expressed concerns that they were not able to appraise the skills sufficiently.

After each section within the standard setting process we administered the questionnaires on the procedural validity elements explicitness, practicability, implementation, and feedback. Table 18.2 depicts the calculated mean scores and standard deviations of the overall agreements to the questions on explicitness, practicability, implementation, and feedback after each step of questioning (introduction, training, round 1, round 2, and final questionnaire). With regard to explicitness, we could already observe a strong agreement after the introduction. This goes together with a relatively high standard deviation ($M = 3.40$, $SD = .91$). While the standard deviation became smaller after each round (range: $SD = .91$ to $.32$), the agreement remains high until the end of the process ($M = 3.39$, $SD = .32$). The first and the final agreement were significantly above the middle of the scale of 2.5. The repeated measurement ANOVA showed that the change of agreement from the introduction to the final questionnaire was not significant, $F(1,6) = .00$, $p = .96$ and the effect size was low ($d = .01$).

Regarding the validity aspect practicability, we did not administer the questionnaires until the end of the training. The panellists were actively involved only during

Table 18.2 Mean agreement between the panel members regarding the validity aspects explicitness, practicability, implementation, and feedback

	Introduction		Training		Round 1		Round 2		Final	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Explicitness	**3.40***	0.91	3.13	0.68	3.07	0.72	2.96	0.64	**3.39*****	0.32
Practicability	–	–	3.43	1.13	**3.40*****	0.25	**3.64*****	0.40	**3.95*****	0.13
Implementation	3.52	1.11	3.36	1.11	**3.43*****	0.32	**3.47*****	0.45	**3.94*****	0.10
Feedback	–	–	3.10	0.81	**3.10*****	0.25	**3.43****	0.66	**3.42*****	0.38

M = mean; SD = standard deviation; – = not administered. Numbers in bold indicate a significant deviation from the centre of the scale of 2.5. * = p < .05, ** = p < .01, *** = p < .001

the training and therefore able to rank the practicability of the ID Matching method. In the beginning and similar to the results regarding explicitness, the agreement was relatively high in combination with a high standard deviation ($M = 3.43$, $SD = 1.13$). The latter decreased drastically after round 1 ($SD = .25$) and was smallest after the final questionnaire ($SD = .13$). The mean agreement lay significantly above 2.5 after round 1 ($M = 3.40$), increased over time and almost reached the upper limit of 4.0 at the end ($M = 3.95$, $SD = .13$). According to the repeated measurement ANOVA, the gain of agreement from questioning after training to the final questioning was not significant, $F(1,6) = 1.41$, $p = .28$, with a medium effect size ($d = .46$) (Table 18.2).

The highest initial value was found for the validity aspect implementation ($M = 3.52$, $SD = 1.11$). Analogously to practicability, we observed an immense decrease of the standard deviation after the questioning subsequent to round 1 ($SD = .32$), whereas the agreements rose and stayed significantly above 2.5. The agreement after the final questioning was $M = 3.94$ ($SD = .10$). Regarding the initial and the final value, the gain was not significant, $F(1,6) = .93$, $p = .37$, with a medium effect size ($d = .38$).

We found the lowest initial value for the validity aspect feedback ($M = 3.10$, $SD = .81$). The relatively high standard deviation decreased over time, but stayed rather high compared to the other evaluation aspects. Simultaneously, the agreement increased during the subsequent sections to a mean of $M = 3.42$ ($SD = .38$). Again, the increase was not significant with regards to the initial and final values of agreement, $F(1,6) = .79$, $p = .41$. We observed a medium effect size ($d = .40$).

In summary, we observed high agreement at the beginning of the process according to all validity aspects, which stayed nearly constant over time and were significantly over 2.5 after the final questionnaire. The relatively high standard deviations at the beginning decreased during the process; in some cases the decrease was quite drastic.

18.5.2 Results on Internal Validity

Our second aim concerns the panellists' change in consensus throughout the procedure. Therefore, we first examined the individual item allocations to the performance levels. Figure 18.2 depicts the item allocation to the proficiency levels Basic and Advanced in percentages after the first and second round. The x-axis displays the items with increasing difficulty as they appeared in the OIB. The y-axis represents the percentages of experts who matched the respective items to the proficiency levels Basic and Advanced.

After the first round, the experts matched eight items concordantly (3 to 7, 16, 19, and 22), whereas they matched the remaining 14 out of 22 items differently. This accords with a Fleiss' Kappa of $\kappa = .38$, which implies a fair agreement. The individual cut score was calculated as the midpoint of each panellists' threshold region. The range of the individual cut scores was between item 4 and 15. The mean cut score across the panel was $M = 8.43$ ($SE = 1.71$).

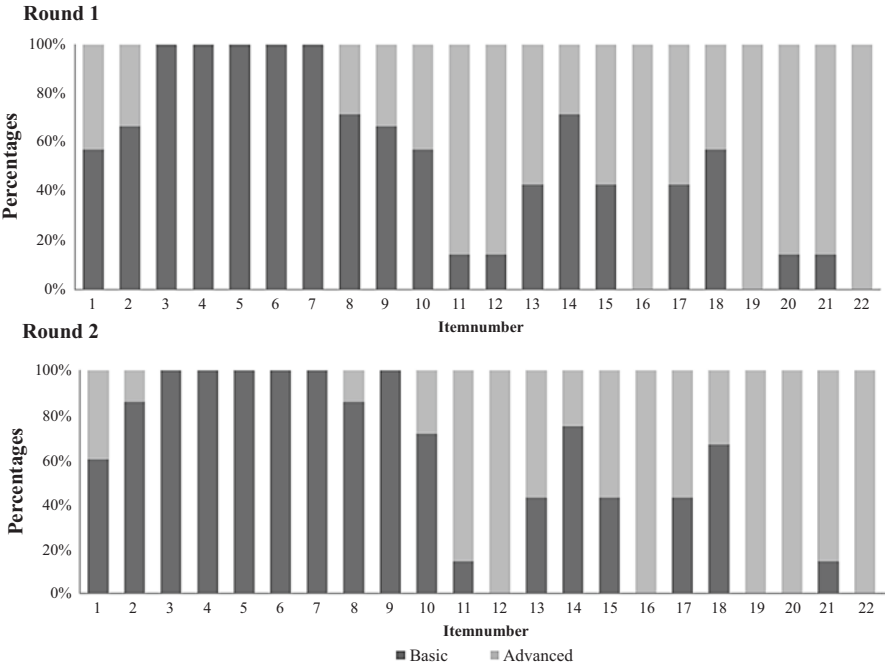


Fig. 18.2 Percentages of experts matching the items to the proficiency level Basic or Advanced after the first and second round for each item

After the second round, 11 out of 22 items were matched concordantly. The resulting Fleiss' Kappa of $\kappa = .63$ implied a substantial agreement. The individual cut scores ranged from item 5 to 16 and the mean cut score across the panel was $M = 10.5$ ($SE = 1.10$). During the third round, the panel members discussed the items that were not yet allocated consensually. They decided via best judgement that the items adjacent to item number 16 match the PLDs of the Advanced level more closely than less difficult items. Consequently, they set the cut score for the proficiency levels Basic and Advanced at item number 16.

18.6 Discussion

Our standard setting was conducted in order to establish science standards for adults and to determine cut scores to categorise adults as scientifically illiterate, literate, and advanced. With the presented validity study, we provide evidence that using and interpreting the cut scores derived from the application of the ID Matching method is procedurally and internally valid.

Since we were concerned that a panel of experts would encounter problems imagining a minimally scientifically literate adult, which is necessary for widely

used procedures such as the Angoff Method or Bookmark Method (Lewis et al. 2012; Plake and Cizek 2012), we chose the ID Matching method to set standards for a scientific literacy test for adults. The ID Matching method circumvents this potential problem by focusing on item demands. The panellists matched the requirements to solve items to the abilities described within each PLD. Although Ferrara et al. (2008) advised to choose panellists whose expertise is closely aligned with the test group, our experts mainly had experience with assessment of students or teachers, not with the assessment of adults. Therefore, the expertise of our panellists about adults was quite limited. However, they all had considerable experience with item development or standard setting. Additionally, the panel was balanced regarding the domains, consisting of two experts from each science didactics field which covers a broad expertise in science. Keeping in mind that adults' scientific literacy in Germany has never been tested before, we argue that the selected panel matches the judgemental task of the ID Matching method quite sufficiently.

The success of this vigilant recruitment of the experts is reflected in the results on the validity aspects. Regarding the evaluation element procedural validity, we found a highly satisfying agreement between the panellists according to the aspects explicitness, practicability, implementation, and feedback. We already found relatively high values at the beginning of the questioning in all four aspects, especially regarding explicitness and implementation. At the end of the procedure, the agreement was significantly above the average score of 2.5 for all four aspects. Due to high standard deviations, the initial values were not significantly above the centre of the scale. The reason for this variance could be explained from the panellists' varying experiences regarding standard settings. The three experts who already attended previous standard setting workshops might have been more familiar with certain terms and therefore might have experienced the information given during the orientation as more comprehensive than other members of the panel. This argument is applicable to the aspects practicability, implementation, and feedback as well. The results of the questioning subsequent to round 1 of the standard setting already show that even though the agreements stayed at about the same level, the standard deviations decreased noticeably. This indicated that the panellists actually reached the same level of expertise over time. Since we found high values in the beginning we did not expect significant gains over time in the repeated measurement ANOVA.

However, the middle effect sizes confirm that the differences between the agreement at the beginning and at the end of the standard setting are practically relevant. Besides, regarding practicability and implementation the agreement nearly reached perfect agreement. We argue that the application of the ID Matching method has been a success. The experts were also very confident with their cut score recommendations. Three experts stated that they strongly agree and four stated that they agree that the cut scores are reliable. In summary, we therefore argue that our evidence on procedural validity supports the interpretation and use of the cut scores in order to categorise adults according to their scientific abilities as scientifically illiterate, literate, and advanced. Even though our experts were not too familiar with adult literacy, they reached a broad consensus throughout the process.

Regarding internal validity, the results are more diverse. After the first round, the experts had already matched eight items concordantly. Those items were not discussed by the panel and consequently were not matched differently after the second round. Although four experts agreed and one expert strongly agreed that they adapted their decisions due to the influence of their colleagues within the second round, only three more items (9, 12, and 20) were matched consensually after the second round. The disagreement about the remaining eleven items continued until the last round. For instance, three panellists already and unexpectedly matched the first two items to the Advanced proficiency level. Typically, we would assume that the simplest item belongs to the lowest level and that the experts would follow this assumption. Ferrara and Lewis (2012) reported that experts sometimes find that items are misplaced within the OIB. The discussion after the first round confirmed this problem. The panellists expressed that they perceived a strong discrepancy between the empirical difficulty of items 1 and 2 and the abilities both items seem to require in order to solve them. The problem arising from this circumstance is that the threshold region for those three panellists expanded and the cut score (as the mean of the threshold region) was biased downwards. We propose two possibilities to deal with such problems. First, since there is a large imbalance of expected and actual difficulty, there seem to be general problems with the construction of those items. Therefore, one should consider deleting the items from the standard setting process. Second, one should consider redefining the threshold region or the cut score. In our study, we defined the cut score as the midpoint of the threshold region. However, it is also possible to define the cut score as the first item in a row of three matching the next higher proficiency level (Ferrara and Lewis 2012).

In our study, this would have resulted in a cut score after item 13 for both rounds, matching the final decision of panellists more closely. The inter-rater reliability points to an acceptable consensus of the panel after the second round which aligns with findings of other studies applying the ID Matching method. Freunberger (2013) report inter-rater reliabilities within a range of $\kappa = .24$ to .43 after three subsequent rounds with 23 experts and Bazinger et al. (2013) report an inter-rater reliability of $\kappa = .46$ after a third round with 14 panellists. Regarding the increase of the Fleiss' Kappa from a fair agreement to a substantial one ($\kappa = .38$ to .63) after two rounds during the standard setting, we argue that the final setting of the cut score after the third round took place in full consensus. To sum up the internal validity evidence, we see a strong support for using the cut score in order to categorise adults as scientifically illiterate, literate, and advanced literate.

18.7 Limitations and Implications

The NEPS scientific literacy test for adults is the first of its kind in Germany and internationally. Until now, there was a lack of knowledge about the scientific abilities of adults and of science standards for adults. Our standard setting workshop was

a first attempt to foster research within the field of scientific literacy and science education of adults. Nevertheless, some limitations need to be considered.

First, it is desirable to divide heterogeneous groups such as adults into multiple proficiency levels in order to recommend educational and political implications. Since we only had 22 items, we had to limit the number of proficiency levels in order to create satisfying discriminations between the proficiency levels and meaningful PLDs. Since our panellists only had to match the item demands to the two PLDs of the proficiency levels Basic and Advanced, this reduction made the task a lot easier for them. We advise to repeat the standard setting with more cut scores after the next NEPS assessments of adults´ scientific literacy in 2020, when more test items are available. Second, the panel consisted of a small number of experts. A small panel can cause high standard errors or limit potential discussions. Although we did not experience this problem within our study, we recommend using a larger panel for the next standard setting workshop. Third, we only had access to information about the procedural and internal validity evidence. In order to give comprehensive judgements about the ID Matching method and the use and interpretation of its results (i.e., cut scores and performance standards), external and consequential validity needs to be investigated as well. We want to show two possibilities to approach these two validity aspects. In a future standard setting for adults' scientific literacy, two alternative methods (e.g., Angoff method or Bookmark method) can be conducted simultaneously to the ID Matching method in order to compare results in terms of external validity. Moreover, assessing the scientific literacy of adults in 2020 will provide longitudinal information about the examinees, such as occupational success. This information might then shed more light on the consequential validity of interpretation and use of cut scores and performance standards.

Appendices

Appendix A: Item Example

Example of an item measuring adults' scientific literacy. The item was part of the pilot study and was excluded from consecutive studies due to the restricted number of test items. Translation by Ulrike Hemstock, IPN.

☐	The anti-bodies last only for a maximum of one year.
☐*	The influenza virus can change within this time frame.
☐	Vaccination techniques evolve rather quickly these days.
☐	The memorized information about the virus will be deleted by this time.

Immunisation Protection

Wintertime is influenza time. To avoid an infection and its ramifications doctors recommend getting vaccinated. During an infection the influenza viruses attack body cells. The human body reacts to the attack by building anti bodies. These mark the infected cells and destroy them. The virus' information is memorized, enabling the immune system to detect a new infection earlier and to react faster.

Doctors recommend annual vaccination against influenza.

Why should the vaccination against influenza be repeated each year?

Check the right answer! Please check one box only!

Appendix B: Elements of the Standard Setting Workshop Using the ID Matching Method

Introduction	The panel received information about the NEPS project, the method for assessing scientific literacy, setting performance standards in general as well as about the ID Matching method.	1h	
Training	The experts examined the preliminary performance level descriptors (PLDs) and practiced the ID Matching method with item examples.	1h	Questionnaire
Round 1	The panellists worked through the Ordered Item Booklet (OIB) and matched the items to the PLDs. The precise task was, *Which PLD most closely matches the knowledge and skills required to respond successfully to this item (or score level for constructed-response items)*?	45 min	Questionnaire
Feedback	The first feedback showed to which performance level (and how often) each item was assigned (see figure 2). The experts were asked to discuss their decisions.	20 min	Questionnaire
Round 2	Repetition of round 1	45 min	
Feedback	In addition to the item allocation, the current cut score was calculated and the corresponding distribution of adults on the proficiency levels were shown.	15 min	Questionnaire
Round 3	The panel set the final cut score to separate the performance levels Basic and Advanced.	2h	
Finalising PLDs	The experts finalised the PLDs according to the matched items.	2h	Questionnaire
Round 4	Via the direct consensus method, experts decided how many items an adult had to master successfully in order to set the cut score for the Below Basic level.	30 min	Questionnaire
Final discussion	The experts summed up the standard setting workshop.	30 min	Questionnaire
			Questionnaire

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores, pp. 508–600.

Artelt, C., Weinert, S., & Carstensen, C. H. (2013). Assessing competencies across the lifespan within the German national educational panel study (NEPS). *Journal for Educational Research Online, 5*(2), 5–14.

Bazinger, C., Freunberger, R., & Itzlinger-Bruneforth, U. (2013). *Standard-Setting Mathematik 4. Schulstufe: Technischer Bericht*.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56*(1), 137–172.

Blossfeld, H.-P., Rossbach, H. G., & von Maurice, J. (Eds.). (2011). *Education as a lifelong process: The German national educational panel study (NEPS) [Special issue]*, *Zeitschrift für Erziehungswissenschaft. (14)*. Wiesbaden: VS Verlag fur Sozialwissenschaften.

Buckendahl, C. W., Smith, R. W., & Impara, J. C. (2002). A comparison of Angoff and bookmark standard setting methods. *Journal of Educational Measurement, 39*(3), 253–263.

Cizek, G. J. (2012a). The forms and functions of evaluations of the standard setting process. In G. J. Cizek (Ed.), *Setting performance standards. Foundations, methods, and innovations* (2nd ed., pp. 165–178). New York: Routledge.

Cizek, G. J. (Ed.). (2012b). *Setting performance standards: Foundations, methods, and innovations* (2nd ed.). New York: Routledge.

Egan, K. L., Schneider, M. C., & Ferrara, S. (2012). Performance level descriptors: History, practice, and a proposed framework. In G. J. Cizek (Ed.), *Setting performance standards. Foundations, methods, and innovations* (2nd ed., pp. 79–106). New York: Routledge.

Ferrara, S., & Lewis, D. M. (2012). The item- descriptor (ID) matching method. In G. J. Cizek (Ed.), *Setting performance standards. Foundations, methods, and innovations* (2nd ed., pp. 255–282). New York: Routledge.

Ferrara, S., Perie, M., & Johnson, E. (2008). Matching the judgemental task with standard setting panelist expertise: The item-descriptor (ID) matching method. *Journal of Applied Testing Technology, 9*(1), 1–20.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). The measurement of interrater agreement. In J. L. Fleiss, B. Levin, & M. C. Paik (Eds.), *Wiley series in probability and statistics. Statistical methods for rates and proportions* (3rd ed.). Hoboken: J. Wiley.

Freunberger, R. (2013). *Standard-Setting Mathematik 8. Schulstufe.*

Freunberger, R., & Yanagida, T. (2012). Kompetenzdiagnostik in Österreich: Der Prozess des Standard-Settings. *Psychologie in Österreich, 5*, 396–403.

Hahn, I., Schöps, K., Rönnebeck, S., Martensen, M., Hansen, S., Saß, S., Dalehefte, I., & Prenzel, M. (2013). Assessing scientific literacy over the lifespan: A description of the NEPS science framework and the test development. *Journal of Educational Research Online, 5*(2), 110–138.

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: Lawrence Erlbaum Associates.

Hambleton, R. K., Pitoniak, M. J., & Copella, J. M. (2012). Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results. In G. J. Cizek (Ed.), *Setting performance standards. Foundations, methods, and innovations* (2nd ed., pp. 47–76). New York: Routledge.

Haschke, L. I., & Kähler, J. (in press). *NEPS technical report for science-scaling results of starting cohort 6: adults: NEPS working paper.*

IBM. (2010). *IBM SPSS Statistics for Windows, Version 19.0*. Armonk: IBM.

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

Kane, M. (2012). Validating score interpretations and uses. *Language Testing, 29*(1), 3–17.

Karantonis, A., & Sireci, S. G. (2006). The bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice, 25*(1), 4–12.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Schulz, E. M. (2012). The bookmark standard setting procedure. In G. J. Cizek (Ed.), *Setting performance standards. Foundations, methods, and innovations* (2nd ed., pp. 225–254). New York: Routledge.

Lord, F. M. (1981). Standard error of an equating by item response theory. *ETS Research Report Series, 2*, 463–471.

Mitzel, H. C., Patz, R. G., & 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: Lawrence Erlbaum Associates.

Moosbrugger, H. (2012). Item-Response-Theory (IRT). In H. Moosbrugger & A. Kelava (Eds.), *Tetstheorie und Fragebogenkonstruktion* (2nd ed.). Berlin/Heidelberg/New York: Springer.

OECD. (2013). *OECD Skills outlook 2013.* Paris: OECD Publishing.

Pant, H. A., Rupp, A. A., Tiffin-Richards, S. P., & Köller, O. (2009). Validity issues in standard-setting studies. *Studies in Educational Evaluation, 35*(2–3), 95–101.

Pitoniak, M. J. (2003). Standard setting methods for complex licensure examinations (dissertation). University of Massachusetts Amherst.

Plake, B. S., & Cizek, G. J. (2012). Variations on a theme: The modified Angoff, extended Angoff, and yes/no standard setting methods. In G. J. Cizek (Ed.), *Setting performance standards. Foundations, methods, and innovations* (2nd ed., pp. 181–200). New York: Routledge.

Plake, B. S., Impara, J. C., Cizek, G. J., & Sireci, S. G. (2008). *AP standard setting pilot studies final report*. New York, NY.

Pohl, S., & Carstensen, C. H. (2013). Scaling of competence tests in the national educational panel study – Many questions, some answers, and further challenges. *Journal of Educational Research Online, 5*(2), 189–216.

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

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher, 36*(8), 477–481.

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.

van der Linden, J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. (2007). *ACER ConQuest*. Camberwell: Australian Council for Educational Research.

Zieky, M. J. (2012). An historical overview of setting cut-scores. In G. J. Cizek (Ed.), *Setting performance standards. Foundations, methods, and innovations* (2nd ed.). New York: Routledge.