DOCUMENT RESUME

ED 389 756 TM 024 389

AUTHOR van der Linden, Wim J.
TITLE A Conceptual Analysis of Standard Setting in Large-Scale Assessments. Research Report 94-3.
INSTITUTION Twente Univ., Enscnede (Netherlands). Faculty of Educational Science and Technology.
PUB DATE Nov 94
NOTE 38p.
AVAILABLE FROM Bibliotheek, Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; *Criteria; *Cutting Scores; Educational Assessment; *Evaluation Methods; Foreign Countries; *Scoring; Standards; *Test Construction
IDENTIFIERS Conceptual Analysis; Decision Theoretic Testing; *Large Scale Assessment; *Standard Setting

ABSTRACT

Elements of arbitrariness in the standard setting process are explored, and an alternative to the use of cut scores is presented. The first part of the paper analyzes the use of cut scores in large-scale assessments, discussing three different functions: (1) cut scores define the qualifications used in assessments; (2) they simplify the reporting of achievement distributions; and (3) they allow for the setting of targets for such distributions. The second part of the paper gives a decision-theoretic alternative to the use of cut scores and shows how each of the three functions identified in the first part can be approached in a way that may reduce some of the arbitrary nature of standard setting processes. The third part of the paper formulates criteria for standard setting methods that can be used to evaluate their results. (Contains six figures and eight references.) (Author/SLD)

A Conceptual Analysis of Standard Setting in Large-Scale Assessments

Research
Report
94-3

Wim J. van der Linden

faculty of
EDUCATIONAL SCIENCE
AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

A Conceptual Analysis of Standard Setting in Large-Scale Assessments

Wim J. van der Linden

To appear in Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments. Washington, DC: NAGB/NCES.

A conceptual analysis of standard setting in large-scale assessments, Wim J. van der Linden - Enschede: University of Twente, Faculty of Educational Science and Technology, November 1994. - 32 pages.

Abstract

This paper consists of three different parts. In the first part, the use of cut scores in large-scale assessments is analyzed. Three different functions of cut scores are discussed: (1) cut scores define the qualifications used in assessments; (2) they simplify the reporting of achievement distributions; and (3) they allow for the setting of targets for such distributions. The second part of the paper gives a decision-theoretic alternative to the use of cut scores, and shows how each of the three functions identified in the first part can be approached in a way which may reduce some of the feelings of arbitrariness often accompanying standard setting processes. The third part of the paper formulates criteria for standard setting methods which can be used to evaluate their results.

A Conceptual Analysis of Standard Setting in Large-Scale Assessments

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

Three Different Functions of Standards

Standard setting processes in large-scale assessments typically aim at the selection of one or more cut scores on an achievement variable. For simplicity, the case of a single cut score is analyzed. Let θ be a (content-referenced) achievement variable on which a cut score has to be selected. Figure 1 contains a graph of the distribution of the examinees in the population on this variable

Figure 1 about here

along with a possible cut score, θ_C, used to distinguish between the two intervals with the qualifications Satisfactory and Unsatisfactory. The selection of this cut score serves the following three functions:

1. Definition of the qualifications. If the test score variable is content referenced in the sense that for each level of the variable the domain of achievements mastered by the examinees is specified, cut scores define the qualifications in terms of the behavior of examinees. An excellent way to map achievements on a variable is using response functions from item response theory (IRT). For the domain of achievements represented by the test, these functions specify the types of problems that can be solved with a given probability of success for each possible level of the variable. An early application of response functions to map achievements on a score scale is the *KeyMath* test for the diagnosis of arithmetic skills (Connolly, Nachtman & Pritchett, 1971), but the origin of the idea to use the content of test items to define a variable underlying a test goes back to Thurstone (1925, Fig. 6).

Contrary to a popular point of view, the point taken here is thus not that the meaning of cut scores is defined by what now is generally known as performance standards. It is the achievements of examinees on the test items that provide the scale of test scores with a behavioral interpretation. The role of cut scores is only to link qualifications in assessments studies to these scales, and hence to their behavioral interpretation. Performance standards are verbal descriptions of achievements which form an important step in the process of specification leading to the domain of test items represented in the test and selecting the cut scores. However, once the domain has been realized and the cut scores selected, performance standards lose their (operational) meaning. From that point on, reversely, it is the domain of test items and the cut scores that define the (empirical) meaning of performance standards.

2. Reporting assessment results. If standards in the form of cut scores are available, they can be used to simplify the reporting of assessment results. The typical statistics used to report such results are estimates of the percentages of examinees with achievements in the intervals defining the various qualifications. In Figure 1, the shaded area represents the percentage of examinees with the qualification Satisfactory.

From a statistical point of view, the process of standard setting is only a form of *data reduction*. Each cut score dichotomizes the distribution of scores along the full scale. In so doing, information on the relative numbers of examinees on each side of the cut score is retained but all other information on the shape of the distribution is sacrificed. The fact that the concept of a distribution of scores on a continuous scale is a notion difficult to understand by the general public is one reason for this data reduction, but certainly not a sufficient one. A more important reason has to do with the last function of cut scores identified here.

3. Setting targets for achievement distributions. Finally, the presence of cut scores offers the possibility to set targets for the outcomes of educational policies in terms of the achievements of the examinees. Typically, such targets are set as upper and lower bounds to the percentages of examinees in the population meeting the various qualifications.

In public discussions of standard setting for large-scale assessments, often no clear distinction between the definitions of qualifications and the setting of targets is made. One reason for this omission might be the fact that (high) qualifications invariably are perceived as a challenge by some students and teachers, and therefore have a tendency to serve as *de facto* standards at an individual level. However, in large-scale assessments, targets are set for *distributions* of achievements, and it is onl} bounds to the percentages of students meeting the qualifications that can serve as targets. The tendency to automatically perceive qualifications as targets is stronger if the terms used to communicate the qualifications already have an everyday meaning loaded with positive or negative emotions. Looking for terms without an emotional loading is therefore an important activity in standard setting processes. In this respect, NAEP has done a respectable job selecting such neutral terms as Basic, Proficient, and Advanced. Nevertheless, it seems a permanent task of assessment specialists to remind the general public of the fact that qualifications used in assessment studies have no meaning over and above the behavior of examinees classified by cut scores on achievement scales, and that targets are always to be set for distributions of achievements.

It is important to observe that the exact location of a cut score is not always important when setting targets. For a given distribution of achievements, a lower cut score leads to a higher percentage of examinees above the cut score. As a consequence, different combinations of cut scores and percentages may be met by the same distribution, and, hence, imply the same target.

A Decision-Theoretic Approach to Standard Setting

The feelings of arbitrariness in standard setting referred to in the introductory section of the paper stem from the fact that cut scores have an "all-or-none" character but that, nonetheless, their exact location can never be defended sufficiently. Examinees with achievements just below a cutting score differ only little from those with achievements immediately above this score. In spite of this fact, the personal consequences of this small difference may be tremendous, and it should not come as a surprise that such examinees can be seen as the victims of arbitrariness in the standard setting procedure. These feelings of arbitrariness are reinforced if it is noted that the procedure contained such random events as the selection of judges, test items, or experts.

On the other hand, as explained above, the ultimate goal of large-scale assessments is to set targets for achievements. Therefore, policy makers as well as consumers of education are not served well with qualifications phrased in words that have only a vague relation to the achievement variable in question and can not be used to set clear-cut targets.

In conclusion, it seems as if we need standards that are a little "softer" than the cut scores now generally in use, but which, nevertheless, allow for unequivocal decisions about the quality of education. A decision-theoretic approach to the standard setting problem could be a first step towards this end, though, as will be elucidated below, no procedure will ever remove all subjectivity from standard setting. For each of the three functions of standards in the previous section, a decision-theoretic alternative not based on the use of cut scores on tests is given.

Defining Qualifications for Large-Scale Assessments

The first function of cut scores could be restated by observing that, at a more formal level, a cut score is nothing but an instance of *a rule to link a set of possible qualifications to an achievement variable*. Other rules are possible, and may even be better. An example of a more general rule is the one given in Figure 2, where the two possible qualifications used in the previous figure are

Figure 2 about here

graphed as a *continuous function* of θ.

Several interpretations of these functions are possible, including the following four suggestions:

1. A decision-theoretic interpretation would view the two functions as representations of the utilities involved in assigning the qualifications to the various levels of the achievement variable θ. The term utility is used here in its technical sense as a measure summarizing all the costs and benefits involved in a decision outcome. Empirical estimates of utility functions are easier to obtain for decisions with immediate personal or institutional consequences, such as selection decisions or mastery decisions for certification purposes, than for large-scale assessments where the consequences of the decision which qualification to assign to the achievement distribution of a population of students are indirect and involve only long-term costs and benefits. Empirical estimates of utility functions for selection and mastery testing problems have been studied in van der Gaag (1990) and Vrijhof, Mellenbergh and van den Brink (1983). A more general treatment of utility measurement is given in Verheij (1992). Nevertheless, some formal properties of utility functions for large-scale assessments are obvious. For instance, the utility associated with the decision to assign the qualification Satisfactory should be an increasing function of the achievement variable whereas a decreasing function is needed to model the utility associated with the decision to assign the qualification Unsatisfactory. For the conceptual

analysis in this paper, these characterizations will do. To actually apply decision theory to large-scale assessment, the shapes of these functions have to be specified further.

2. It is also possible to view large-scale assessment as an attempt to evaluate an achievement distribution on a rating scale. In this view, qualifications such as Unsatisfactory and Satisfactory correspond with the categories in the rating instrument. In IRT models for the analysis of ratings, the first step is always to scale or locate the categories on the variable measured by the instrument. The results from this stage are then used to rate objects on the same variable. This practice is perfectly consistent with the idea that in large-scale assessment the first function of standards is to define the possible qualifications in terms of the levels of an achievement variable, and that next these qualifications are used to evaluate the distribution of a population of students on this variable. The response functions in the common IRT models for rating scale analyses with two categories are probability curves with the same decreasing and increasing shapes as the curves in Figure 2. If this probability interpretation is adopted, the curves are each others's complement in the sense that for all levels of the achievement variable the sums of the two probabilities are equal to one.

3. Suppose a population of judges has used one of the item-centered judgment methods for standard setting, for example, an Angoff method. However, the task has not been to provide probabilities of success for a "minimally competent" examinee but for an examinee with "satisfactory achievements". The increasing curve in Figure 2 could then represent (a smoothed version of) the cumulative distribution function of the results

from this experiment for the population of judges, while the other curve is the decumulative distribution function for an experiment for examinees with "unsatisfactory achievements".

4. The last interpretation suggested here is the one of an experiment with an examinee-centered judgment method in which one group of examinees is deemed to represent a satisfactory achievement level, and the other an unsatisfactory level. The two curves in Figure 2 could then represent (smoothed versions) of the cumulative and decumulative distribution functions of the achievements of the two groups, respectively.

Each of these interpretations seems equally possible. However, to remain consistent with the decision-theoretic perspective taken in this paper, the first interpretation is followed.

If more than two qualifications are needed, additional utility curves have to be introduced some of which take a shape different from the ones in Figure 2. Suppose the qualifications Below Basic, Basic, Proficient, and Advanced have to be mapped on a NAEP achievement scale. Figure 3 offers a set of utility curves for this problem. The two curves for Basic and Proficient are bell shaped since assigning these qualifications to very low or high achievements obviously are wrong decisions, which thus represent low utilities. Since Proficient has a more favorable meaning than Basic, the curve for the former has its location to the right of the latter. The curves for the qualifications Unsatisfactory and Advanced

Figure 3 about here

decrease, respectively, increase in the values of the achievement variable. Note that the usual IRT models for rating scales with more than two categories produce families of response curves with the same shapes.

Comparison Between Cut Scores and Utility Functions

It is now possible to point at the fact that cut scores are special cases of the utility functions defined above. Figure 1 follows from Figure 2 if the curves in the latter are taken to represent a *threshold utility function*. Figure 4 illustrates this reduction for a 0-1 threshold utility function. The figure shows that the cut score is the achievement level at which the function for the qualification. Satisfactory jumps from zero to one while the one for Unsatisfactory falls from one to zero. It is this point of discontinuity, along with the assumed constancy of utility over the intervals to the left and the right of this point, which gives cut scores their all-or-nothing character.

Figure 4 about here

It should be noted that the continuous utility functions in Figure 3 have points at which their curves cross, and that these points define intervals for which one of the qualifications has largest utility. However, these point are no cut scores. Neither do the points at which the functions have maximal values have any decisive meaning. As will be clear below, when assigning qualifications to achievement distributions, it is the *full* shape of the utility functions which counts. The results are therefore remarkably robust with respect to the values of these curves at individual points along the scale.

Assigning Qualifications to Population Distributions

Let $g(\theta)$ be the density function representing the distribution of the achievements, $\boldsymbol{\Theta}$, of a population of examinees in an assessment study. The question how to report assessment results is now reformulated as: If one qualification has to be assigned to this achievement distribution, which qualification is best?

An obvious criterion for assigning an optimal qualification is the one of *maximal expected utility* common in (empirical) Bayesian decision theory. Depending on whether or not $g(\theta)$ is known, various statistical implementations of this criterion are possible. We will present the major implementations for the case of two qualifications in Figure 2, but the theory applies equally well to cases with any number of qualifications.

$g(\theta)$ known. The following notation is used for the two utility functions: $f_U(\theta)$ (Unsatisfactory) and $f_S(\theta)$ (Satisfactory). The criterion of maximum expected utility tells us that if only one qualification is to be assigned, an optimal choice is one with the largest value for the expected utility. The expected utilities of the two qualifications are calculated with respect to $g(\theta)$ as:

$$E[f_U(\Theta)] = \int f_U(\theta)g(\theta)d\theta, \tag{1}$$

and

$$E[f_S(\Theta)] = \int f_S(\theta)g(\theta)d\theta. \tag{2}$$

According to the criterion, the qualification Satisfactory is an optimal assignment if:

$$E[f_S(\Theta)] \geq E[f_U(\Theta)], \tag{3}$$

and the qualification Unsatisfactory is optimal otherwise.

In most assessment studies, interest exists in evaluation of the achievements of certain subpopulations. Examples are subpopulations defined by social-economic background, race, gender, or subpopulations existing of, say, the top 5% or bottom 20% of the total population. For each possible subpopulation, the above procedure can be repeated to determine the best qualification. Also, it is possible to extract more information from the utility distributions associated with (sub)populations, and to report, for example, which qualification is second best. These as well as other obvious refinements are not further pursued here.

$g(\theta)$ unknown. If $g(\theta)$ is unknown, then a Bayesian approach could be used to estimate it from a distribution of observed test scores. Let $g_0(\theta)$ be a prior for this unknown density and $h(x|\theta)$ the density that models the conditional distribution of the observed scores given $\Theta=\theta$. (The case where response data can not be reduced to a sufficient statistic X will not be addressed here.) From Bayes theorem, it follows that the posterior density, $k(\theta|x)$, is given by:

$$k(\theta|x) = \frac{h(x|\theta)g_0(\theta)}{\int h(x|\theta)g_0(\theta)d\theta}. \tag{4}$$

A large range of options is available to choose the prior $g_0(\theta)$ in Equation 4. For example, it is possible to choose a noninformative prior, to use an informative prior based on previous knowledge about $g(\theta)$, to estimate $g(\theta)$ along with the parameters in $f(x|\theta)$ in an "empirical" Bayes fashion, or to follow an hierarchical Bayes approach. Also, collateral empirical information could be used to improve our estinate of the posterior density. More information on each of these approaches is available in standard textbooks on Bayesian statistics.

If an estimate of the posterior density is obtained, it could be used to obtain an estimate of the marginal density $g(\theta)$ through:

$$g_1(\theta) = \int k(\theta|x)p(x)dx, \tag{5}$$

where p(x) is the density of the observed scores. If $g_1(\theta)$ is substituted in Equations 1-3, we have estimates of the expected utilities for both qualifications, and the maximum expected utility criterion is now applied to these estimates.

Formulating a Target for an Achievement Distribution

The maximum expected utility criterion tells us only what the best qualification for an achievement distribution is but not whether the targets are met. However, if the only decision problem is which qualification to assign, setting a target for the outcome is as simple as choosing the qualification that should be best for the population being assessed. If, in addition, the achievements of subpopulations are assessed, targets have to be set for their distributions too. For some subpopulations the targets will be the same but for others it makes sense to set different targets. Some fictitious examples of targets with the format suggested here are:

1. The qualification for the achievements in English of the national populations of students should be Proficient, with Advanced rather than Basic as second-best qualification;
2. The top 20% of the national population should have achievements in mathematic which qualify as Advanced;

3. The bottom 10% of the subpopulation of children of first-generation immigrants should have achievements in English with the qualification Basic; and

4. For all possible subpopulations and subjects the qualifications should not show any differences with respect to gender.

Again, obvious refinements of these suggestions will not be pursued here; the main purpose of this paper is only to outline an alternative to the current practice of large-scale assessment.

Standards for Standard Setting

In the final part of this paper, criteria are formulated which can be used to discriminate between better and worse standards. "Standard" is used here as a generic term; there is no need to choose between experiments to establish (the parameters of) utility functions and the experiments currently used to select cut scores. Some of the criteria are derived from the statistical literature whereas others address practical issues or are more empirically oriented. This presentation is only a first attempt. With the growing international interest in the potentials of large-scale assessment for enhancing the quality of education, it would be worthwhile to take a coordinated action to get to a more elaborated list.

1. Explicitness. The criterion of explicitness is not new, and in fact applies to any research activity or scientific procedure. The criterion stipulates that all steps in a standard setting experiment be based on explicit definitions and procedures. First of all, the motivation of this criterion is communicative. If this criterion is not met, a standard setter can never communicate his or her results in

a meaningful way. But there are technical reasons for this criterion too. Without an explicit definition of the steps in the standard setting procedure, it would never be possible to apply any of the criteria below. For example, without this information it would be impossible to determine if the procedure has been subject to possible inconsistencies, or to replicate the procedure to estimate its statistical properties.

2. Efficiency. This statistical criterion is defined with respect to the variability of the results from a standard setting procedure across replications. The lower the variability, the more efficient the procedure.

An important step in designing an experiment to estimate the efficiency of a standard setting procedure is the determination which aspects of the procedure are allowed to vary across replications and which not. Generally, those aspects that are allowed to vary are irrelevant and should not introduce any variability in the resulting standards. On the other hand, the basic aspects of the method should be kept fixed. If they nevertheless do vary, then variability in the outcomes is to be expected and not necessarily a bad thing.

Examples of irrelevant aspects of standard setting procedures are situational factors such as occasion and location. It would be difficult to accept the possible impact of these factors on the results of the procedure. In other words, the procedure should be efficient with respect to variation on these factors.

However, if a method is based on judgment of the contents of test items, then the question whether items are allowed to vary across replications cannot be answered without considering the scale of the achievement variable. If the achievements are scored on the number-right scale, the properties of test items are not irrelevant, and variability of results due to item sampling is to be expected. Items do differ in their psychometric properties and a standard-setting procedure that does not reflect the impact of such properties on the achievements of the examinees, would even be a bad method. On the other hand, if sampling is from a pool of items calibrated on the same achievement scale using an IRT model which allows for all of the differences between the items, and the standard is set on this scale, then the procedure should have high efficiency with respect to item sampling.

It is a common observation that results from standard setting experiments show variation across methods. This variation is to be expected. Each method instructs its subjects to a different task. At a more general level, as emphasized by the S-O-R paradigm in psychology, for any type of stimuli it holds that responses from subjects depend on the properties of the stimuli as much as they do on the properties of the subjects. For the standard setting process, the dependency is as depicted in Figure 5.

Figure 5 about here

The same relation is well known to test theorist who have struggled for decades to find a way to separate the properties of test items from the abilities of the examinees. They now use IRT models to calibrate the properties of the items before they are used as a measurement instrument, and then use these calibrations to equate test scores from different instruments. A similar development should take place in standard setting. The right approach is not to view variation between methods as error or random noise, which is only there to be averaged out, but to calibrate these methods and use the calibrations to equate results from one method to those from another.

By symmetry, the argument above also applies to variability in standard setting results across judges. Such variability is to be expected because judges have their own views about what to expect from education. If judges have worked carefully and meet all of the other criteria in this list, then a standard setting method should allow for differences in views between judges instead of suppressing them as random noise. Obviously, to get practical results, standards set by different judges always have to be combined into a single standard. The point made here, however, is that the question what is a good standard setting method should be separated from the question how to optimally combine different standards into a single standard. Decision-theoretic approaches to the latter question also exist. The question is then approached as an instance of the problem how to best represent a distribution of numbers by a single number, or how to combine collective choices into a single order of preferences.

3. Unbiasedness. The criterion of unbiasedness is another statistical criterion in this list. Generally, a method is unbiased if it produces estimates which on the average are equal to the true parameter value. The criterion is discussed here because some policy makers or educators seem to believe that true standards do exist independently of methods and judges, and that the only thing a standard setting method should do is to mirror these standards as accurately as possible. This view is not correct.

First of all, standards do not exist without achievement variables, and those variables are no natural quantities but human inventions used to score examinees in tests. Further, as already pointed out, judges have different views about what to expect from education, and different views entail different standards. Also, as argued in the previous section, even for a single judge, standards do not exist independently of the method used to set them. The correct view is to see standard setting methods as methods to *set* true standards--not to reflect them.

The belief reminds us of another classical struggle in test theory, namely the one with the concept of true score. During the first half of this century, many theorists behaved as if for each examinee there existed a numerical score which exclusively represented his or her true ability on the variable measured by the test. This view had the implication that some tests were less biased than others but did not imply any suggestions as to how to estimate bias in tests. The view has completely been left, and currently no test theorist believes in this so-called concept of a *Platonic* true score (Lord & Novick, sect. 2.9). True scores are now generally defined as expected or average test score observed across replicated administrations of the same test with the same examinee. In this approach, test scores are unbiased by definition. Standard setting theorists should learn from this experience, and follow the same direction.

4. Consistency. Though it is customary to speak about data in standard setting experiments, it is important not to forget that these data are judgmental. Examples of judgments in standard setting experiments are: estimates of subjective probabilities in an Angoff experiment and ratings of expertise in an experiment with an extreme-group method. However, judgments can be inconsistent in the sense that two or more judgments contradict each other or that their combination contradicts reality. The following three examples illustrate possible inconsistencies:

1. From paired-comparisons experiments in scaling, it is known that judges may display intransitivities. Intransitivities would occur in a standard setting experiment if a judge rated Subject A as more proficient than B,

B as more proficient than C, but A as less proficient than C.

2. In an Angoff experiment, it is possible that judges specify a high probability of success for a difficult item and a low probability of success for an easy item for the same "borderline examinee". Both probabilities imply standards that can never exist at the same time. Such inconsistencies were frequently observed in an analysis reported in van der Linden (1982).

3. The same type of inconsistency may occur at test level if two tests, A and B, have high correlation between the scores X_A and X_B because they measure the same variable. If both tests are used to select cut scores in a standard setting experiment, and some of the cut scores for test B can not be predicted from those on test A by the regression function of X_B on X_A, then these cut scores are inconsistent.

Note that the last criterion need not hold for test scores measuring *different* achievement variables. The fact that different achievements variables correlate highly for a given curriculum does not imply, for example, that if a judge sets a high standard on one variable, he or she should also set a high standard on the other.

A more general interpretation of the notion of consistency for use in evaluation research along the lines of the definition above is given in van der Linden and Zwarts (in press).

5. Feasibility. This criterion is of a more practical nature and deals with the aspect of standards earlier identified as target setting for achievement distributions. Only a rather loose description will be given. Nevertheless, the criterion seems to focus on an issue that is on the mind of many critics of standard setting results. Two definitions of this criterion are possible, one for

standards in the form of cut scores and another for standards defined as utility functions:

1. In a cut-score based approach to large-scale assessment, a target is feasible if the resources are available to meet it;
2. In a decision-theoretic approach to large-scale assessment, a target is feasible if it is based on utility functions that incorporate a realistic estimate of the costs needed to realize the levels of the achievement variable.

It is here that the practical meaning of a decision-theoretic approach becomes fully clear. When following a cut-score based approach, it is easy to forget the idea of feasibility, and to ignore the pains and costs it may take to realize a target. However, an attempt to establish utility functions would directly address such issues as costs of resources and their impacts. This fact explains why estimating utility functions is generally more difficult than selecting cut scores.

The notion of feasibility is particularly important if standards have to be set with respect to more than one achievement variable. If the resources are fixed, trade-offs between achievement distributions on different variables is typical. For example, it would not be too difficult for schools to produce high achievements in geography if all other subjects could be dropped. Neglecting such trade-offs may be the explanation why standard setters, no matter their expertise in the pertinent domain of content, often show a tendency to set standards unrealistically high when confronted with the task to address a single achievement variable.

6. Robustness. Some standard setting experiments provide their subjects with information on properties of the items or on the ability distributions of

reference groups. This information is mostly in the form of statistical estimates. In addition, estimates of item properties may be needed to calculate standards from the judgments by the subjects in the experiment. The criterion of robustness is suggested to deal with the possibilities of errors in these estimates.

Generally, a standard is robust if minor changes in the data used in the experiment do not lead to changes in it. Robustness is a welcome property because it tells us that uncertainty about certain relevant aspects of reality is not critical to the results of the experiment. Robustness of standards can be assessed through a series of analyses in which changes are made in the values of the estimates, and their effects on the behavior of the standards are ascertained.

It is important to note the similarities and dissimilarities between the criteria of efficiency and robustness. Both criteria are based on the idea of replication. To estimate the efficiency of a standard setting experiment, apart from possible irrelevant aspects, its procedure is left intact. In a robustness study, the procedure is also replicated exactly but changes are made in the empirical data presented to the subjects. If these data play a role only in the calculation of the standard from the judgments in the experiment, robustness analysis can take the form of computer simulation in which standards are recalculated from data with a simulated error term.

An example of robustness analysis in an evaluation project based on the outcomes of large-scale assessment is given in van der Linden and Zwarts (1994). In the project, the definition of the standards supposed the presence of an intact item pool which was, however, reduced slightly due to pretesting of the pool. A simulation of several item analyses procedures showed that the effects of item removal on the standards seemed to be negligible.

Evaluation of Standards

Each of the criteria in the previous list should constrain the choice of standards. A possible sequence of constraints is depicted in Figure 6. The Venn diagram shows the set of all possible outcomes of a standard setting experiment, not all of which necessarily meet each of the above criteria. An empirical check is neede̞d to demonstrate that the proposed standards belong to the subsets of consistent, efficient, and robust outcomes. In addition, the body of knowledge and insights produced by educational research should indicate that the proposed standards are in the subset of feasible outcomes. An outcome meeting the whole list of criteria is in the shaded area in the diagram.

Figure 5 about here

Discussion

It is surprising to note that in some discussions on standard setting not much attention is paid to lessons learned in the history of test theory. Psychometrically, a standard setting method is nothing but an instrument to elicit responses from subjects from which an estimate of a quantity is inferred. The same formal description holds for an achievement test, an attitude instrument, or a rating scale. It was suggested earlier that the relation between standard setting and rating is particularly close since both activities have an aspect of evaluation. The idea that standard setting methods should reflect true or Platonic standards, existing independently of standard setting methods and subjects, seems to underlie some of the arguments in standard setting discussions but has not proven to be fruitful in the history of test theory. This idea leads to the belief that properties of standard setting methods only add random error to the true standard, and that the best method to get an "unbiased estimate" of this standard is to average out differences between methods. The same idea supports the practice of averaging standards over judges to get rid of their "idiosyncrasies". The question how to separate the impact of methods from the evaluation of judges is the same as the one how to separate the properties of test items from the abilities of examinees. At a more abstract level, the problem is known as the problem of parameter separability or the problem of nuisance parameters in statistics. Standard setters should look deeper into the analogies between problems in standard setting and test theory or scaling, and profit from the solutions reached there.

This paper began by observing that in the setting of standards for large-scale assessment interaction of empirical information and subjective choice is typical. The diagram in Figure 5 shows how these inputs interact in an ideal standard setting process. Each of the subsets in the diagram is defined by *empirical information* from research for which rigorous methodology exists. However, it is left to the *subjective evaluation* of the judges to prefer one possible outcome in the intersection of the subsets over the others.

Finally, it is claimed that the feelings of arbitrariness typical of standard setting processes can be reduced further if the definition of standards should include the class of continuous utility functions. As demonstrated in Figure 4, this step would only mean that the rigid form of the threshold utility function is relaxed to allow utility functions with such continuous shapes as in Figure 2-3.

The impression exists that threshold functions as in Figure 4 are no obvious candidates for use in assessment studies, and that if one is met the precise location of the jump seldom can be defended convincingly. Since assessment results are extremely sensitive to the location of this jump, a cut-score based approach is bound to be dogged by criticism and feelings of arbitrariness. As pointed out earlier, a decision-theoretic approach with continuous utility functions has no individual points on the achievement scale with a dramatic impact the results of the assessment. On the other hand, establishing the realistic shapes of such utility functions is not a sinecure. But the criterion of feasibility reminds us that we should try establishing them.

References

Connolly, A.J., Nachtman, W., & Pritchett, E.M. (1971). KeyMath diagnostic arithmetic test. Circle Pines, MN: American Guidance Service.

Lord, F.M., & Novick, M.R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Thurstone, L.L. (1925) A method of scaling psychological and educational tests. Journal of Educational Psychology, 16, 433-451.

van der Gaag, N.L. (1990). Empirische utiliteiten voor psychometrische beslissingen [Empirical utilities for psychometric decisions]. Unpublished doctoral dissertation, University of Amsterdam, Department of Psychology, The Netherlands.

van der Linden, W.J. (1982). A latent trait method for determining intrajudge inconsistencies in the Angoff and Nedelsky techniques of standard setting. Journal of Educational Measurement, 19, 295-308.

van der Linden, W.J. & Zwarts, M.A. (in press). Robustness of judgments in evaluation research. Tijdschrift voor Onderwijsresearch, 20.

Verheij, H. (1992). Measuring utility: A psychometric approach. Unpublished doctoral dissertation, University of Amsterdam, Department of Psychology, The Netherlands.

Vrijhof, B.J., Mellenbergh, G.J., & van den Brink, W.P. (1983) Assessing and studying utility functions in psychometric theory. Applied Psychological Measurement, 7, 341-357.

Figure Captions

Figure 1. Distribution of student on an achievement variable with cut score, θ_C defining two possible qualifications.

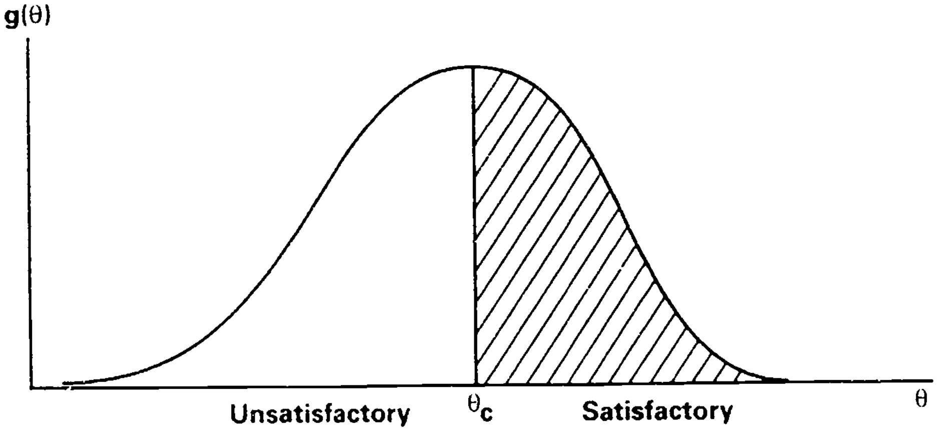
Figure 2. Continuous utility functions used to link the exemplary qualifications to an achievement variable.

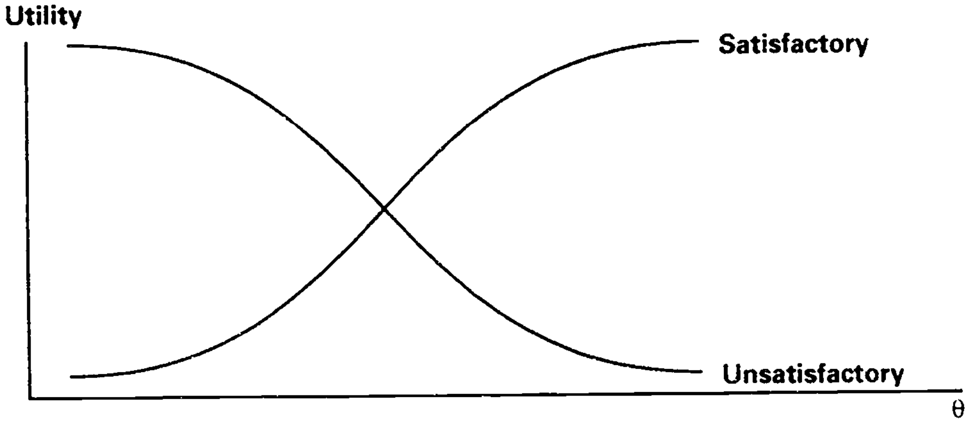
Figure 3. Example of utility functions for the four qualifications used in NAEP.

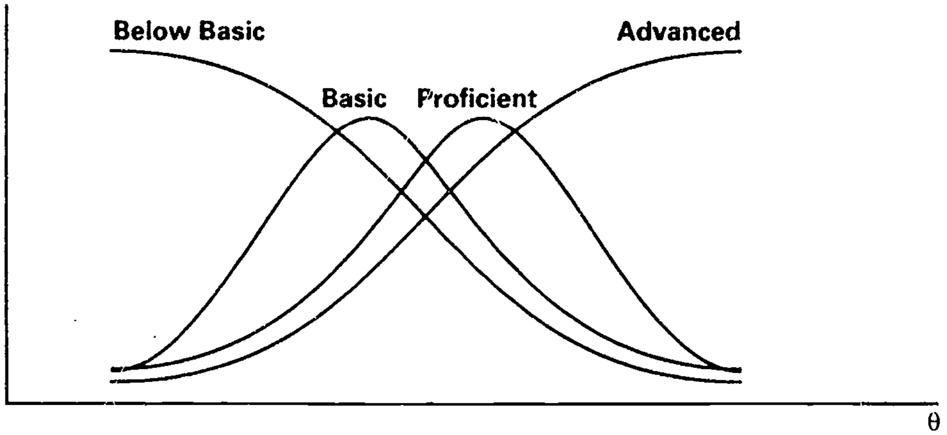
Figure 4. A cut score, θ_C, represented as a threshold utility function.

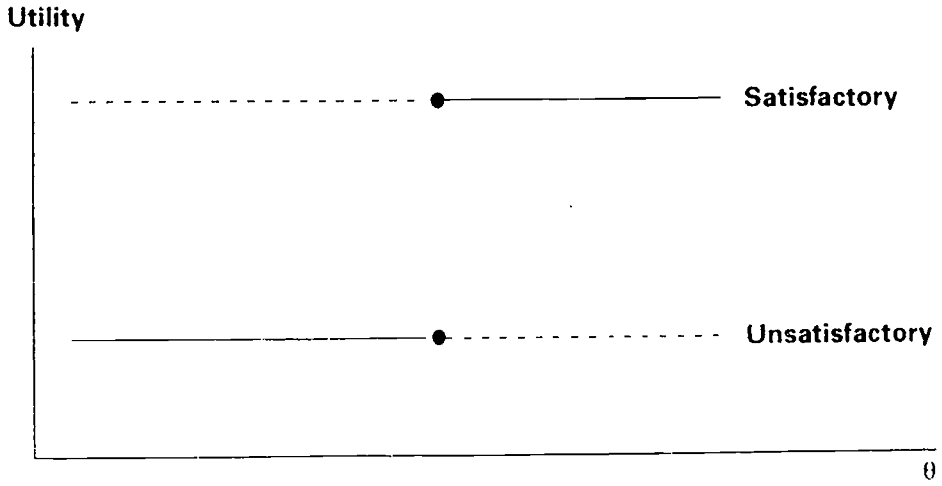
Figure 5. Interaction between judge, task, and standard.

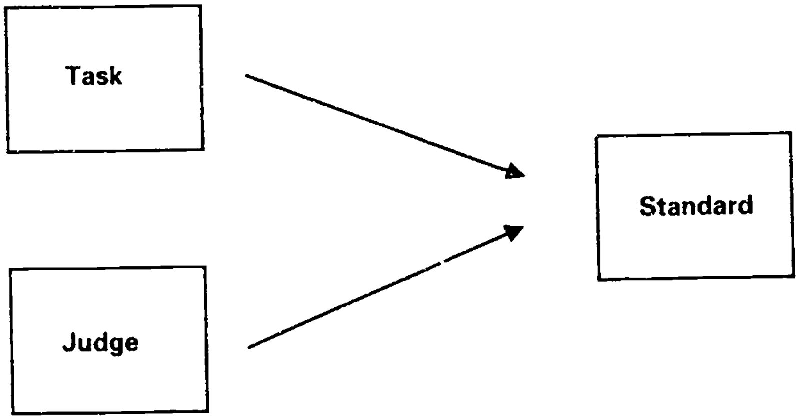
Figure 6. Venn diagram with possible outcomes of a standard setting experiment.

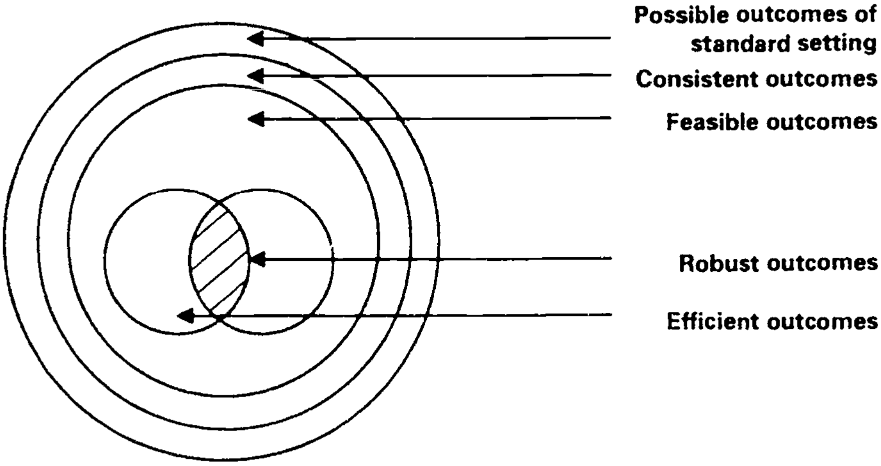

g(θ)
Unsatisfactory
θ_c
Satisfactory
θ


Utility
Satisfactory
Unsatisfactory
θ


Below Basic
Basic
Proficient
Advanced
θ


Utility
Satisfactory
Unsatisfactory
θ


Task
Judge
Standard


Possible outcomes of standard setting
Consistent outcomes
Feasible outcomes
Robust outcomes
Efficient outcomes

Titles of recent Research Reports from the Department of
Educational Measurement and Data Analysis.
University of Twente, Enschede,
The Netherlands.

RR-94-3 W.J. van der Linden, *A conceptual analysis of standard setting in large-scale assessments*

RR-94-2 W.J. van der Linden & H.J. Vos, *A compensatory approach to optimal selection with mastery scores*

RR-94-1 R.R. Meijer, *The influence of the presence of deviant item score patterns on the power of a person-fit statistic*

RR-93-1 P. Westers & H. Kelderman, *Generalization of the Solution-Error Response-Error Model*

RR-91-1 H. Kelderman, *Computing Maximum Likelihood Estimates of Loglinear Models from Marginal Sums with Special Attention to Loglinear Item Response Theory*

RR-90-8 M.P.F. Berger & D.L. Knol, *On the Assessment of Dimensionality in Multidimensional Item Response Theory Models*

RR-90-7 E. Boekkooi-Timminga, *A Method for Designing IRT-based Item Banks*

RR-90-6 J.J. Adema, *The Construction of Weakly Parallel Tests by Mathematical Programming*

RR-90-5 J.J. Adema, *A Revised Simplex Method for Test Construction Problems*

RR-90-4 J.J. Adema, *Methods and Models for the Construction of Weakly Parallel Tests*

RR-90-2 H. Tobi, *Item Response Theory at subject- and group-level*

RR-90-1 P. Westers & H. Kelderman, *Differential item functioning in multiple choice items*

Research Reports can be obtained at costs from Bibliotheek, Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.

Faculty of
EDUCATIONAL SCIENCE
AND TECHNOLOGY

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands