DOCUMENT RESUME

ED 310 139 TM 013 722

AUTHOR	van der Linden, Wim J.
TITLE	A Latent Trait Method for Determining Inconsistencies in the Use of the Angoff and Nedelsky Techniques of Standard Setting. Twente Educational Report Number 12.
INSTITUTION	Twente Univ., Enschede (Netherlands). Dept. of Education.
PUB DATE	May 81
NOTE	32p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Error of Measurement; Evaluators; Foreign Countries; *Latent Trait Theory; Mathematical Models; Probability; *Specifications
IDENTIFIERS	Angoff Methods; *Misspecification; Nedelsky Method; *Standard Setting

ABSTRACT

It has often been argued that all techniques of standard setting are arbitrary and likely to yield different results for different techniques or persons. This paper deals with a related but hitherto ignored aspect of standard setting, namely, the possibility that Angoff or Nedelsky judges misspecify the probabilities of the borderline student's success on the items because they do not use the psychometric properties of the items consistently. A latent trait method is proposed to estimate such misspecifications, and an index of consistency is defined that can be used for deciding whether standards are set consistently enough for use in practice. Results from an empirical study are presented to illustrate the use of the method in a typical educational situation. The results indicate that serious errors of specification can be expected and that, on the whole, these will be considerably larger for the Nedelsky than for the Angoff technique. (Four data tables are provided.) (Author)

TWENTE EDUCATIONAL REPORT NUMBER 12

A LATENT TRAIT METHOD FOR DETERMINING INCONSISTENCES IN THE USE OF THE ANGOFF AND NEDELSKY TECHNIQUES OF STANDARD SETTING

W.J. van der Linden

TWENTE UNIVERSITY OF TECHNOLOGY
DEPARTMENT OF EDUCATIONAL TECHNOLOGY
P.O. BOX 217
7500 AE ENSCHEDE THE NETHERLANDS

COLOFON :

Publisher	: Department of Educational Technology Twente University of Technology
Typist	: Paula Achterberg
Reproduction	: Central Reproduction Department of the University
Edition	: 60

A LATENT TRAIT METHOD FOR DETERMINING INCONSISTENCES IN THE USE OF THE ANGOFF AND NEDELSKY TECHNIQUES OF STANDARD SETTING

W.J. van der Linden

Twente Educational Report No. 12. Enschede, The Netherlands: Twente University of Technology, Department of Educational Technology, May 1981

Abstract

It has often been argued that all techniques of standard setting are arbitrary and likely to yield different results for different techniques or persons. This paper deals with a related but hitherto ignored aspect of standard setting, namely the possibility that Angoff or Nedelsky judges misspecify the probabilities of the borderline student's success on the items because they do not use the psychometric properties of the items consistently. A latent trait method is proposed to estimate such misspecifications and an index of consistency is defined which can be used for deciding whether standards are set consistently enough for use in practice. Also, results from an empirical study are given which illustrate the use of the method in a typical educational situation. The results indicate that serious errors of specification can be expected and that, on the whole, these will be considerably larger for the Nedelsky than for the Angoff technique.

A LATENT TRAIT METHOD FOR DETERMINING INCONSISTENCIES IN THE USE OF THE ANGOFF AND NEDELSKY TECHNIQUES OF STANDARD SETTING

An important problem in mastery testing is the setting of cutoff scores. Two separate cutoff scores must be determined—the true cutoff score and the cutoff score on the observed score variable. The problem of setting the former is usually called a standard-setting problem. Several techniques for solving this problem are available. Cutoff scores on the observed score variable can be determined best using a decision-theoretic approach.

In objectives-based instructional programs, which is the area where mastery testing typically is applied, standards are the translation of the learning objectives into cutoff scores on the true score scale. They constitute the predetermined levels that the student's true performance must exceed to be granted the mastery status and to be allowed to proceed with the next instructional unit. Or, in other words, standards are the expected test performances of a student who just meets the requirements formulated in the learning objectives. There is an extensive literature on standard-setting techniques which recently has been reviewed by several authors. Glass (1978) gives a critical review and reminds us of the fact that all standard-setting techniques ultimately rest on some (arbitrary) judgement. Other reviews that can be recommended are those by Hambleton (1980), Hambleton, Powell, and Eignor (1979), Jaeger (1979), and Shepard (1980a, 1980b). In this paper, the emphasis will be on the Angoff (1971) and Nedelsky (1954) standard-setting techniques. These techniques are

commonly classified as techniques based on judgment of test content.

Once a standard has been set, the cutoff score on the observed score variable must be determined. This can be done optimally within the framework of (Bayesian) decision theory. In such an approach, an explicit loss function weighting the consequences of the outcomes is chosen and the cutoff score on the test is selected such that the expected loss is minimal.

Occasionally, the necessity of determining a separate cutoff score on the test seems to be forgotten. This is the case, for example, in all those applications of the Nedelsky technique in which the resulting standard is used as a cutoff score against which the observed test scores can be judged directly. However, this amounts to assuming that the students' observed test scores are equal to their true scores, i.e., that the test is free of measurement error, which seems no realistic assumption. Also, from a decision-theoretic point of view it can be shown that using the same value for the true and observed cutoff scores entails the adoption of implicit loss functions. It may be wondered whether one of these loss functions would be chosen if an explicit choice had been asked for. Occasionally, decision-theoretic approaches are classified as standard-setting techniques. Such approaches are however techniques for minimizing the consequences of measurement error when the standard or true cutoff score is transformed into a cutoff score on the test. They ought to follow each time a standard-setting technique is used. For further particulars about the use of decision models in mastery testing, see van der Linden (1980a).

It has been argued that all standard setting is arbitrary (Glass, 1978; Shepard, 1979, 1980a, 1980b). This is correct since standards ought

to reflect learning objectives and these ultimately rest on value judgments and norms. Moreover, the various standard-setting techniques available differ, more or less, in the conception of mastery underlying the way standards are obtained. Thus, different results can be expected both for different techniques and for different persons using the same technique. This has been demonstrated in investigations by Andrew and Hecht (1976) and Brennan and Lockwood (1980). That all standard-setting is ultimately arbritary has seduced Glass to the pessimistic conclusion that we should abandon the use of these techniques. However, as Hambleton (1978) and Popham (1978) have put forward, arbritariness does not necessarily have a negative sense. There are many other instances in which arbitrary choices have to be made and deliberate, defensible results are obtained. What should be avoided is capricious standard setting, that is, standard setting in which the learning objectives are incorsistently translated into the true cutoff score and, in fact, erratic standards of mastery are obtained.

In a sense, the present paper addresses the second, negatively loaded definition of arbitrariness. Its concern is not with comparisons of differences in standard setting between persons or techniques. Such differences can be expected when persons bring different evaluations of learning objectives or techniques are based on different conceptions of mastery. Rather, the interest is in the occurrence of inconsistencies when a person uses the Angoff or Nedelsky technique to set a standard for a given test. Inconsistencies arise when the person's judgments are not in agreement with the properties of the test items and judgments of different items in fact imply different standards. An example is a person using the Angoff technique and assigning a low probability of success to an easy item but a

large probability to a difficult item. These two judgments are clearly inconsistent: the former implies a low standard whereas the latter indicates that a high standard should be set. Inconsistencies can also be due to an inadequate use of other item properties. Further examples will be given below.

Obviously, the less serious the inconsistencies, the better the standard. It seems reasonable to require that standard-setting procedures must satisfy a certain degree of consistency before its results can be used. So far, no attention has been paid to the occurrence of inconsistencies in using the Angoff or Nedelsky techniques, and their results are employed without checking their quality. This may be due to the fact that classical test theory does not provide satisfactory methods for analyzing such inconsistencies. It is the intent of the present paper to show how latent trait theory (Birnbaum, 1968; Lord, 1980; Wright & Stone, 1979) does provide such a method and how a simple index of consistency can be defined which can be used for deciding whether standards are consistent enough for use in practice. Before doing this, the Angoff and Nedelsky standards will be discussed following a slightly different notation so that some of their properties can be indicated and the possibility of a latent trait analysis becomes obvious. The paper also presents results from an empirical investigation which demonstrate how the method should be used and shed some light on the question how consistent the Angoff and Nedelsky techniques are when used in a typical educational situation. In the final section, some additional comments on the method and its results are made.

The Angoff and Nedelsky Techniques

Although the Angoff technique was introduced only in a short footnote (Angoff, 1971, p. 515), it has become one of the best known and widely used methods of standard setting. It is suited for dichotomously scored items and consists of the following few steps: A content specialist is asked to imagine a student just meeting the requirements as formulated in the learning objectives. This may be a hypothetical as well as a real student. Keeping this borderline student in mind, he/she is requested to inspect the test item by item and to specify for each item the probability that the student will answer it correctly. The standard is equal to the sum of the probabilities.

Let $P_i(+|\theta)$ be the probability that a student with mastery level θ answers item i correctly and let θ_c denote the mastery level of the borderline student whom the content specialist has in mind. In the Angoff method, $P_i(+|\theta_c)$ is specified for each of the n items in the test and the standard is defined by

$$(1) \qquad \sum_{i=1}^{n} P_i(+|\theta_c).$$

Note that $P_i(+|\theta)$ is not only the probability of success but also the expected item score for a student with mastery level θ, since it holds that

$$(2) \qquad E(u_i|\theta) \equiv 1.P_i(+|\theta) + 0.\left[1 - P_i(+|\theta)\right],$$

where $u_i = 0, 1$ denotes the item score. In classical test theory, the true number-right score for a fixed person, τ, is defined as the expected value of his/her observed test score, $X = \sum_{i=1}^{n} u_i$. From (1) and (2), it follows that

$$(3) \qquad \sum_{i=1}^{n} P_i(+|\theta_c) = \sum_{i=1}^{n} E(u_i|\theta_c) = E(\sum_{i=1}^{n} u_i|\theta_c) = E(X|\theta_c) \equiv \tau_c$$

Thus, the Angoff technique translates the performances of a borderline student, who just meets the learning objectives, in a true cutoff score for the given test. This cutoff score can subsequently be used to determine an optimal cutoff score on the observed score variable. For future reference, it is noted that the relation given in (3) is known as the test characteristic curve (Lord, 1980, p. 49).

The Nedelsky technique was introduced some twenty-five years before the Angoff technique became known (Nedelsky, 1954). It is also based on judgment of test content and uses the same setting of judges who, imagining a borderline student, are requested to go through the test item by item. However, it can only be used for multiple-choice items and is based on an all-or-none model with respect to the item alternatives. It assumes that a student knows which alternatives are incorrect and guesses between the remaining alternatives (if more than one are left). It is the task of the judge to indicate for each item the alternatives for which the borderline student knows that they are incorrect. The Nedelsky standard is then set equal to the sum of the reciprocals of the numbers of remaining alternatives of the items.

Let q_i denote the number of alternatives of item i and suppose that

$k_i^{(c)}$ is the number of alternatives for which the judge indicates that a student with mastery level θ_c knows that they are incorrect. According to the model underlying the technique, the probability that this student asnwers item i correctly is equal to the reciprocal of the number of remaining alternatives:

$$(4) \qquad P_i(+|\theta_c) = \left[q_i - k_i^{(c)}\right]^{-1}.$$

The Nedelsky standard is defined as

$$(5) \qquad \tau_c \equiv \sum_{i=1}^{n} \left[q_i - k_i^{(c)}\right]^{-1}.$$

Note that this is again a true cutoff score since it is equal to the sum of probabilities used in (3).

Obviously, the Angoff and Nedelsky techniques are based on different conceptions of the behavior that a student exhibits when responding to test items. In the Angoff technique it is only assumed that this behavior is stochastic and that a student has different probabilities of success for different items. The Nedelsky technique supposes that a student proceeds by eliminating incorrect alternatives and then chooses at random between the remaining alternatives. The assumptions underlying the Angoff technique are extremely weak and consistent with the fact that behavioral measurements are liable to error. The Nedelsky technique, on the other hand, asserts that a specific behavior pattern can be expected, and this assertion can be

true or false. That the Nedelsky technique is much stronger is also clear from the range of values its probabilities of success can assume. In the Angoff technique, these probabilities can assume any value (between zero and one) but in the Nedelsky technique strong restrictions on the range of possible values are imposed and, in fact, only $q_i + 1$ different values are possible. Hence, it can be expected that there are many situations for which the Nedelsky technique does not hold but the Angoff technique still does.

Latent Trait Analysis

The probability of a successful item response, $P_i(+|\theta)$, varies as a function of the mastery level, θ. Generally, the higher the mastery level, the larger this probability. Latent trait theory is concerned with how $P_i(+|\theta)$ varies as a function of θ. It provides models for this function (usually called the item characteristic curve) and methods for analyzing their statistical properties and their fit to test data

A versatile model in latent trait theory is the three-parameter logistic model

$$(6) \qquad P_i(+|\theta) = c_i + (1 - c_i)\left\{1 + \exp[-a_i(\theta - b_i)]\right\}^{-1},$$

in which

a_i is the discriminating power of item i;

b_i is the difficulty of item i;

c_i is a lower asymptote representing the probability of guessing item i correct

(Birnbaum, 1968, pp. 399-405; Lord, 1980, pp. 12-14). The model is attractive because of its flexibility and the fact that it explicitely allows for such item properties as difficulty, discriminating power, and the possibility of guessing, all of which influence the probability of a successful response. For $a_i = 1$ and $c_i = 0$, the Rasch model (Rasch, 1960; Wright & Stone, 1979) is obtained. This model has unique, attractive statistical properties (Andersen, 1980, chap. 6) but is, due to the reduction of the number of parameters, less flexible than the model in (6). More latent trait models are available. For further particulars, the reader is recommended to consult the above references.

All latent trait models are approximations of the actual characteristic function of the items under consideration. If a model fits this function satisfactorily, it can be used for analyzing the item responses. For convenience, the model in (6) will be used to describe how latent trait models can be employed for determining inconsistencies in the use of the Angoff and Nedelsky techniques. The choice is not essential, however; any other latent trait model could be used as well. The only important point is that each item can be assumed to have a characteristic function showing how the probability of a successful response depends both on the level of the student and the properties of the items.

Method

Suppose that (6) holds for the n items for which the Angoff or

Nedelsky technique is used. The first thing to note is that if a probability of success is specified for a student on an item this in fact fixes his/her level of mastery. For example, if a = 1.25, b = .80, and c = .20, and if the probability of success is specified as .53, it follows from (6) that this specification only holds if the student has θ = .50. The value θ = .50 may point to another mastery level than that of the borderline student whom the judge has in mind. If so, and inconsistency arises because the mastery level of the borderline student and the properties of the item imply a probability of success differing from the one the judge has actually specified. Suppose that another item has a = .80, b = -.20, and c = .30, and that the probability of success is specified as .80. This implies θ = .95, a value contradicting the value θ = .50 obtained via the previous item. The judge is now inconsistent because he/she specifies probabilities of success for different items that can never belong to the same person.

Inconsistencies as above arise because judges do not use the item characteristic function correctly. This may be due to the fact that, for instance, they misjudge the difficulties or discriminating powers of the items or inadequately allow for the consequences of guessing. The general point, however, is that for each item there is specific relationship between mastery level and probability of success and that these relationships can be used inconsistently when the success probabilities for the borderline student are specified.

A special difficulty is associated with the use of the Nedelsky technique. As only $q_i + 1$ values can be specified for the probability of success on item i, it follows from (6) that no more than $q_i + 1$ values for θ_c are possible which each may differ from the mastery level that the

judge has in mind. Misspecifications are thus quite likely to occur.

For convenience, a somewhat different notation for the probabilities of success will be used. Let $p_i^{(s)}$ denote the borderline student's probability of success on item i as specified by an Angoff or Nedelsky judge. The superscript s is used to indicate that subjective probabilities are obtained. Further, the objective probabilities $P_i(+|\theta_c)$, which follow from the characteristic curve of item i for $\theta = \theta_c$, will be abbreviated as p_i. Now, a misspecification for item i occurs if

$$(7) \qquad e_i \equiv p_i^{(s)} - p_i$$

is unequal to zero. Note that the value of $p_i^{(s)}$ has been provided by the judge but that p_i is unknown. Hence it is important to be able to estimate p_i. This can be done using (6) provided that the item parameters have been estimated and the value of θ_c is known. However, the Angoff and Nedelsky techniques yield a cutoff score on the true score scale, τ_c, and this scale is related to the θ scale via the test characteristic curve. Thus, the value of θ_c can be determined by computing τ_c from (1) or (5) and using (3) from the left to the right. Once this has been done and one of the available computer programs has been used to estimate the item parameters, (7) can be estimated for the test items to determine how consistently the judge has worked.

It seems obvious to compute

$$(8) \qquad E \equiv \sum_{i=1}^{n} \left| p_i^{(s)} - p_i \right| / n,$$

when is the average absolute error of specification for the n items of the test. To obtain an index of consistency, however, the scale of (8) must be reversed. Moreover, due to the fact that the maximal size of e_i depends on p_i, (8) will mostly be restricted to the interval [0, c], $0 < c < 1$, while indices on [0, 1] are common in educational measurement.

Hence, a natural index of consistency is

$$(9) \qquad c_1 \equiv \frac{M - E}{M},$$

where

$$M \equiv \sum_{i=1}^{n} e_i^{(u)} / n;$$

$$e_i^{(u)} \equiv \max \{p_i, 1 - p_i\}.$$

Note that $e_i^{(u)}$ is the maximum absolute value of the error of specification which follows when either $p_i^{(s)} = 0$ or $p_i^{(s)} = 1$ is substituted into (7). C_1 is thus the degree to which the average absolute error of specification deviates from its maximum possible value, measured on the standard interval [0, 1].

For the Nedelsky technique a modification of (9) is needed because,

as a result of the discrete character of the Nedelsky probabilities, in most instances the minimum value of $|p_i^{(s)} - p_i|$ will be larger than zero. Generally, this minimum value is equal to

$$e^{(\ell)} \equiv \left|[q_i - k_i^*]^{-1} - p_i\right|,$$

where k^* is the value of $k_i^{(c)}$ in (4) chosen such that $e^{(\ell)}$ is minimal. Let

$$m \equiv \sum_{i=1}^{n} e^{(\ell)}/n,$$

then

$$C_2 \equiv \frac{M - E}{M - m} \tag{10}$$

is a modification of C_1 that allows for the fact that for the Nedelsky technique the smallest possible value of E is equal to m. Note that C_1 follows from C_2 for m = 0. Note also that if both C_1 and C_2 are computed for the Nedelsky technique, the difference $C_2 - C_1$ shows the reduction of consistency due to the discrete character of the technique. This difference can thus be used as a measure of the degree to which the model underlying the Nedelsky technique fits the situation.

Based on the above results, a method for determining inconsistencies in the Angoff or Nedelsky procedure can be used which consists of the following steps:

1. A latent trait model is chosen, its parameters are estimated, and its fit is tested. Suppose that n items fit the model.
2. For these n items the Angoff or Nedelsky technique is used to specify for each item the probability of success $p_i^{(s)}$.
3. Using (1) or (5), the Angoff or Nedelsky true cutoff score, τ_c, is computed.
4. The true cutoff score τ_c is transformed into the θ scale of the latent trait model via the estimated test characteristic curve (3). Note that in (3), θ_c is not an explicit function of τ_c so that trial values must be substituted for the former until the in step 3 computed value of the latter is obtained. This can be done, for example, using a short computer program. The task is simplified by the fact that θ is monotonically related to τ. However, some computer programs for latent trait analysis standardly output the estimated test characteristic curve and in that case $\hat{\theta}_c$ can simply be read off.
5. Next, substituting $\hat{\theta}_c$ and the estimated item parameters into the model, the estimated probabilities of success $\hat{p}_i$ are computed.
6. The differences between $p_i^{(s)}$ and $\hat{p}_i$ show how the judge has misspecified the probabilities of success. The latter are estimates of the probabilities that he/she should have specified if the item properties had been used consistently. An analysis of the differences reveals where large misspecifications have occurred and whether peculiarities in the judgments are present.
7. Finally, C_1 is computed to obtain an overall impression of how consistently the judge has worked. If the Nedelsky technique has been used, C_2 is computed and $C_2 - C_1$ shows how large a reduction of consistency has occurred because the Nedelsky model did not fit the situation.

Results

An empirical investigation was carried out to illustrate the above method and to compare results for the Nedelsky and Angoff techniques. The items and Nedelsky data were taken from a previous investigation in which the values of item information functions at the Nedelsky standard were compared with pretest-posttest indices of item validity (van der Linden, 1981). All items were from a test belonging to the unit "Forces and Motion" from a physics course introducing grade ten pupils to elementary mechanical concepts. The test was written by professional item writers of the National Institute of Educational Measurement, Arnhem, The Netherlands, in co-operation with the Project Team Curriculum Development Physics of the State University at Utrecht, The Netherlands. All items were of the three- and four-choice type. A latent trait analysis, based on the responses of 156 pupils to an end-of-unit administration of the test, yielded 18 items showing a satisfactory fit to the Rasch model (equation 5 with a=1 and c=0). For a futher description of the test and the items, the reader is referred to van der Linden (1981).

The Nedelsky data were obtained using nine judges who all were involved in the curriculum development project. The judges were asked to conform to the learning objectives of the instructional unit as formulated in the project. The Angoff data were obtained for the same 18 items one year after the Nedelsky study took place. In this part of the study eight judges were used.

Table 1 shows the Nedelsky results for the nine judges. The first

Insert Table 1 about here

column gives the average absolute errors of specification. The next columns show the values for the indices C_1 and C_2 and their difference. As the last row displays, the mean error of specification in the whole study was no less than .25 . The mean difference between the values of C_1 and C_2 in this study was equal to .09 . As indicated earlier, this difference has to be explained by the lack of fit of the model underlying the Nedelsky technique. The deviation of C_2 from its optimal value of 1.00 cannot be ascribed to this lack of fit but is due to inconsistencies in the judgments.

In Table 2 the probabilities of success on all items are given both

Insert Table 2 about here

for the least consistent and the most consistent judge. The first column contains the Nedelsky probabilities which should be approximately equal to the estimated objective probabilities in the second column. For Judge 2 the differences between the two columns show large variability about their average absolute value of .30 . These differences, albeit still considerably, are markedly smaller for Judge 5. The last two columns give the estimated values of $e^{(u)}$ and $e^{(\ell)}$ on which the computations of C_1 and C_2 are based. These values can also be used as benchmarks when inspecting the differences $p_i^{(s)} - \hat{p}_i$ for the individual items.

The results for the Angoff technique are given in Table 3. As this

Insert Table 3 about here

table demonstrates, the average absolute errors are less serious than for the Nedelsky technique. The mean error in the whole study was equal to .18 . Correspondingly, the values of C_1 are higher than these in Table 1.

Finally, Table 4 gives more detailed information for the most consistent and the least consistent Angoff judge. This information confirms

Insert Table 4 about here

the general findings from the investigation, namely that when using the Angoff or Nedelsky techniques one must reckon with serious misspecifications of the probabilities of success from which the standards are computed but that these are noticeably less unfavorable for the former than for the latter.

Discussion

Three possible sources of arbitrariness in standard setting using the Angoff or Nedelsky technique have been distinguished: (1) the interpretation of the learning objectives by the judge, (2) the conception of mastery underlying the technique, and (3) inconsistencies in specifying the probabilities of success for the borderline student. It has been argued that differences in outcomes as a result of the first two sources can be expected and do not necessarily lead to unusable standards. What should be

required is an explicit interpretation of the objectives as well as a conscious choice of the conception of mastery which both can be defended when asked for. The third source of arbritariness is serious, however. The results show that errors of .20 - .25 are typical but that, especially for the Nedelsky technique, errors larger than .50 are no exception.

The method proposed in this paper can be used for several purposes. An obvious possibility is a routine check of standard setting results before they are used in educational practice. Other possibilities are, for example: (1) selecting judges meeting predetermined criteria of consistency, (2) evaluating programs for training judges, (3) assessing the consequences of modifications of standard-setting techniques, or (4) item analysis to detect items yielding systematic errors across persons or techniques.

For all these applications, it is necessary that items are available fitting one of the latent trait models. Two situations can arise. First, latent trait models can be used for item analysis in which items not fitting the model at first are revised or replaced until a test of appropriate length is obtained that fits the model. This procedure, albeit not always possible for practical reasons, is recommended because experience with latent trait analysis shows that items having unwanted properties are often not detected until such an analysis indicates that something is wrong. Moreover, all items are calibrated before the standard is set and the method proposed in the paper can immediately be used for the full test. Second, the case can arise that it is decided to use the method for a test and a standard that are already in use. This is no ideal situation for the above reasons. However, even if not all items fit satisfactorily the method can still be used. In this case a new standard is computed skipping items not fitting the model.

The new standard is then used to estimate (7), (9), or (10) for the items that do fit the model and these estimates give an impression of how consistently the judge has worked.

References

Andersen, E.B. Discrete statistical models with social science applications. Amsterdam: North-Holland Publishing Company, 1980.

Andrew, J.A., & Hecht, J.T. A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 1976, 36, 45-50.

Angoff, W.H. Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.). Educational Measurement. Washington, D.C.: American Council on Education, 1971.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Brennan, R.L., & Lockwood, R.E. A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. Applied Psychological Measurement, 1980, 4, 219-240.

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

Hambleton, R.K. On the use of cut-off scores with criterion-referenced tests in instructional settings. Journal of Educational Measurement, 1978, 15, 277-290.

Hambleton, R.K. Test score validity and standard-setting methods. In R.A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore, MD: The Johns Hopkins University Press, 1980.

Hambleton, R.K., Powell, S., & Eignor, D.R. Issues and methods for standard-setting. Paper presented at the Annual Meeting of the National Council

on Measurement in Education, San Francisco, California, April 9-11, 1979.

Jaeger, R.M. Measurement consequences of selected standard-setting models. In M.A. Bunda and J.R. Sanders (Eds.), Practices and problems in competency-based measurement. Washington, D.C.: National Council on Measurement in Education, 1979.

Lord. F.M. Applications of item response theory to practical testing problems. Hillsdale, New Jersey; Erlbaum, 1980.

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.

Popham, J.W. As always, provocative. Journal of Educational Measurement, 1978, 15, 297-300.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmark Paedagogisk Institut, 1960.

Shepard, L.A. Setting standards. In M.A. Bunda and J.R. Sanders (Eds.), Practices and problems in competency-based measurement. Washington, D.C.: National Council on Measurement in Education, 1979.

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

Shepard L. Technical issues in minimum competency testing. In D.C. Berliner (Ed.), Review of research in education (Vol. 8). Itasca, Illinois: F.E. Peacock Publishers, 1980. (b)

van der Linden, W.J. Decision models for use with criterion-referenced tests. Applied Psychological Measurement, 1980, 4, 469-492.

van der Linden, W.J. A latent trait look at pretest-posttest validation of criterion-referenced test items. Review of Educational Research,

1981, 51, in press.

Wright, B.D., & Stone, M.H. Best test design: A handbook for Rasch measurement. Palo Alto, Cal.: The Scientific Press,

Table 1

Results for Nine Judges Using the Nedelsky Technique

Judge	E	C_1	C_2	$C_2 - C_1$
1	.25	.65	.74	.09
2	.30	.63	.71	.08
3	.25	.65	.76	.11
4	.25	.69	.77	.08
5	.20	.75	.84	.09
6	.25	.69	.77	.08
7	.23	.69	.78	.09
8	.23	.73	.78	.05
9	.25	.67	.76	.09
Mean	.25	.68	.77	.09

Table 2

Estimated Probabilities of Success for Two Nedelsky Judges

Item	Judge 2				Judge 5			
	$p_i(s)$	$\hat{p}_i$	$\hat{e}_i(u)$	$\hat{e}_i(\ell)$	$p_i(s)$	$\hat{p}_i$	$\hat{e}_i(u)$	$\hat{e}_i(\ell)$
1	.50	.73	.73	.08	.33	.66	.66	.01
2	1.00	.11	.89	.12	.33	.08	.92	.08
3	1.00	.93	.93	.07	1.00	.90	.90	.10
4	.50	.50	.50	.16	.50	.41	.59	.04
5	1.00	.94	.94	.05	1.00	.92	.92	.08
6	.50	.84	.84	.15	.50	.79	.79	.12
7	1.00	.87	.87	.12	.50	.83	.83	.16
8	.50	.92	.92	.07	1.00	.89	.89	.11
9	.50	.71	.71	.05	.33	.63	.63	.04
10	.50	.86	.86	.13	.50	.81	.81	.14
11	.50	.74	.74	.01	.50	.67	.67	.08
12	.50	.16	.84	.08	.50	.12	.88	.12
13	.33	.82	.82	.17	1.00	.76	.76	.01
14	1.00	.22	.78	.01	.33	.17	.83	.08
15	.50	.26	.74	.02	.33	.20	.80	.05
16	.25	.62	.62	.12	.50	.53	.53	.03
17	1.00	.94	.94	.06	1.00	.91	.91	.09
18	.25	.17	.83	.07	.25	.13	.87	.12

Table 3

Results for Eight Judges Using the Angoff Technique

Judge	E	C_1
1	.21	.73
2	.15	.81
3	.16	.81
4	.20	.75
5	.16	.80
6	.17	.78
7	.22	.71
8	.19	.76
Mean	.18	.77

Table 4

Estimating Probabilities of Success for Two Angoff Judges

	Judge 2			Judge 7		
Item	$p_i(s)$	$\hat{p}_i$	$\hat{e}(u)$	$p_i(s)$	$\hat{p}_i$	$\hat{e}(u)$
1	.70	.74	.74	.30	.57	.57
2	.50	.11	.89	.30	.06	.94
3	.80	.93	.93	.90	.87	.87
4	.30	.50	.50	.70.	.34	.66
5	.80	.94	.94	.70	.89	.89
6	.90	.84	.84	.80	.72	.72
7	1.00	.87	.87	.50	.76	.76
8	.60	.92	.92	.30	.86	.86
9	.70	.72	.72	.30	.56	.56
10	.90	.86	.86	.60	.76	.76
11	.60	.75	.75	.70	.60	.60
12	.40	.16	.84	.30	.09	.91
13	.80	.82	.82	.50	.70	.70
14	.40	.23	.77	.50	.13	.87
15	.50	.27	.73	.30	.15	.85
16	.50	.62	.62	.50	.46	.54
17	.70	.94	.94	.80	.88	.88
18	.30	.18	.82	.50	.10	.90

Acknowledgement

Thanks are due to Ronny F.A. Wierstra and the PLON - CITO team for participating in the empirical study and to Paula Achterberg for typing the manuscript.

Author's address

Wim J. van der Linden, Onderafdeling Toegepaste Onderwijskunde, Technische Hogeschool Twente, Postbus 217, 7500 AE Enschede, The Netherlands.