DOCUMENT RESUME

ED 218 345 TM 820 386

AUTHOR van der Linden, Wim J.
TITLE Assessing Inconsistencies in Standard Setting with the Angoff or Nedelsky Technique.
PUB DATE Mar 82
NOTE 20p.; Presentation of this paper supported by a grant from the Dutch Foundation of Educational Research.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Criterion Referenced Tests; *Cutting Scores; *Error of Measurement; *Latent Trait Theory; Mathematical Models; Probability; Secondary Education; Test Items; *Test Reliability
IDENTIFIERS *Angoff Methods; Inter Rater Reliability; *Nedelsky Method; Standard Setting

ABSTRACT

A latent trait method is presented to investigate the possibility that Angoff or Nedelsky judges specify inconsistent probabilities in standard setting techniques for objectives-based instructional programs. It is suggested that judges frequently specify a low probability of success for an easy item but a large probability for a hard item. The responses of 156 pupils to a 25-item test from a tenth grade physics course were inspected by eight Angoff and nine Nedelsky judges. The latent trait analysis produced 18 items showing a satisfactory fit to the Rasch model. Serious errors of specification were found and errors were considerably larger for the Nedelsky technique. Special difficulties with the Nedelsky judges are discussed. Applications of the latent trait method are discussed. (Author/CM)

Assessing Inconsistencies in Standard Setting with the Angoff or Nedelsky Technique

Wim J. van der Linden

Twente University of Technology

Enschede, The Netherlands

Paper presented at the Annual Meeting of the American Educational Research Association, New York, March 19-23, 1982. Presentation of this paper was made possible by a grant from the Dutch Foundation of Educational Research.

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, namely the possibility that Angoff or Nedelsky judges specify inconsistent probabilities, e.g., a low probability for an easy item but a large probability for a hard item. A latent trait method is proposed to estimate such misspecifications and to determine whether the judge has worked consistently. Results from an empirical study are given which indicate that serious errors of specification can be expected, and that these are considerably larger for the Nedelsky than for the Angoff technique.

Assessing Inconsistencies in Standard-Setting with the Angoff or Nedelsky Technique

This paper is concerned with the use of standard-setting techniques in objectives-based instructional programs. For such programs, a great variety of techniques has been proposed (for reviews, see Glass, 1978; Hambleton, 1980; Hambleton, Powell, & Eignor, 1979; Jaeger, 1979; Shepard, 1980a, 1980b). The emphasis in this paper will be on the Angoff (1971) and Nedelsky (1954) techniques. These two techniques, which are based on an item by item judgment of test content, are among the most popular techniques in use in objectives-based instruction.

It has been argued that all standard setting is arbitrary (Glass, 1978; Shepard, 1979, 1980a, 1980n). This is correct since standards ought to reflect learning objectives, and these ultimately rest on values and norms. In addition, the various standard-setting techniques available differ, more or less, in the conception of mastery underlying the way standards are obtained. Therefore, different results can be expected both for different techniques and for different persons using the same technique. This has been confirmed in many experiments (Andrew & Hecht, 1976; Brennan & Lockwood, 1980; Koffler, 1980; Saunders, Ryan, & Huynh, 1981; Skakun & Kling, 1980). In this paper we do not share the concern with inconsistent results due to differences between techniques or persons. Instead, the interest is in a related but hitherto ignored apsect of standard setting, namely the possibility of intrajudge inconsistency. Intrajudge inconsistency arises when the judge specifies probabilities of success on the items that are incompatible with

each other and imply different, conflicting standards. An example of intrajudge inconsistency is a judge specifying a low probability of success for an easy item but a large probability for a hard item. These two judgments are obviously inconsistent: the former implies a low standard, whereas the latter indicates that a high standard should be set. Another example is a judge specifying approximately equal probabilities for highly discriminating items (of differing difficulties). Generally, inconsistencies such as in these examples are due to a discrepancy between the actual properties of the items and the judge's perception of them.

Thus far, no attention has been paid to the possibility of intrajudge inconsistency, and results of the Angoff or Nedelsky technique are generally employed without checking the quality of the judge. This may be due to the fact that classical test theory does not provide satisfactory methods for analyzing such inconsistencies. It is the purpose of this paper to show how latent trait theory can be used to decide whether Angoff or Nedelsky standards have been set consistenly enough for use in practice and to assess for which items inconsistencies have occurred. The second purpose of the paper is to present empirical results showing how consistently the Angoff and Nedelsky techniques were used in a typical educational situation. In the following it is assumed that the reader is familiair with the elementary concepts from latent trait theory as well as the technical aspects of the Angoff and Nedelsky techniques (see the appendix). A fuller description of the method and the empirical results is given in van der Linden (1981a).

Method

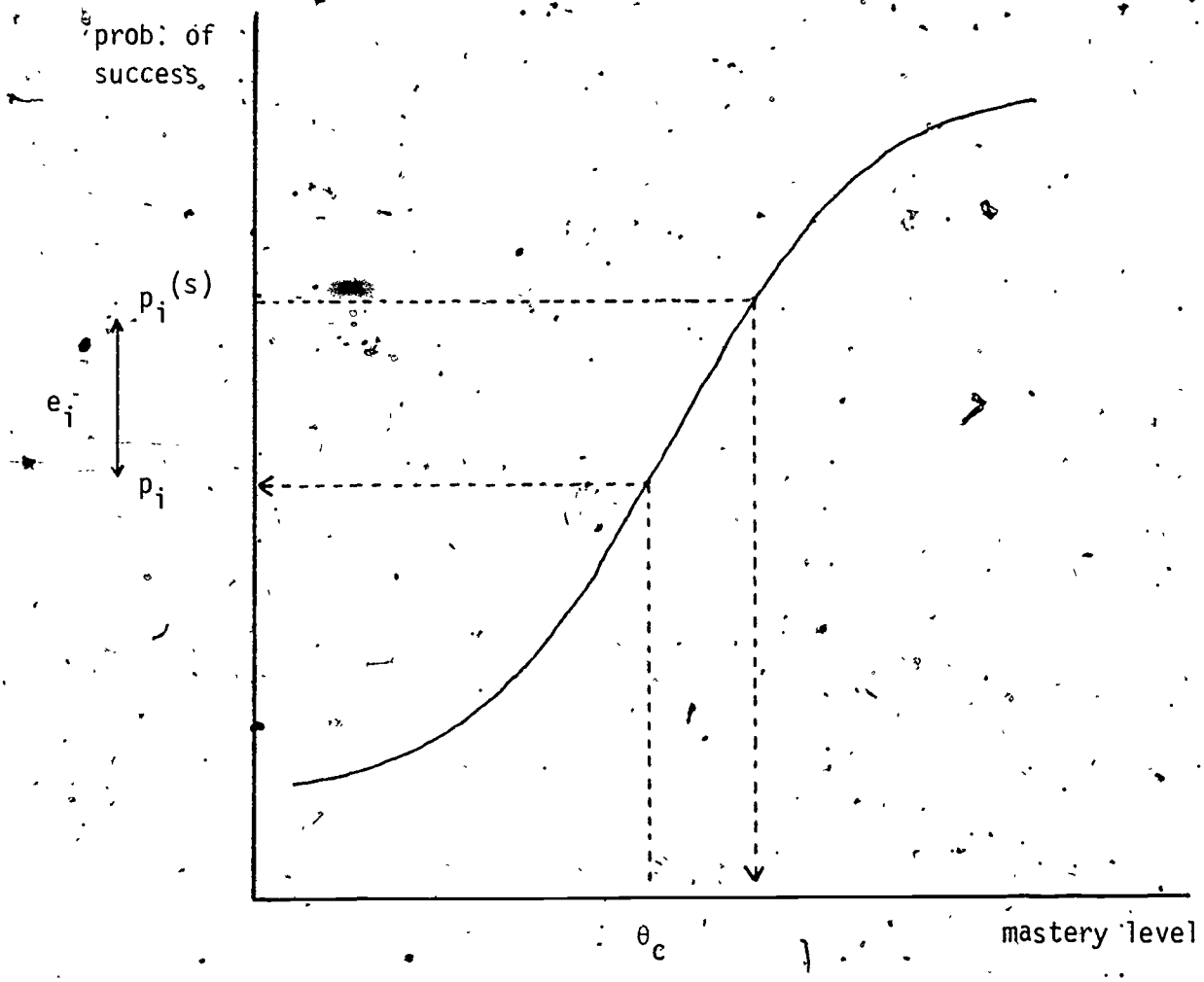
As mentioned earlier, intrajudge inconsistencies arise when probabilities are specified that are incompatible with each other and imply different, conflicting standards. Figure 1 shows how this can be viewed from latent trait theory. In this example, θ_c denotes the level of mastery belonging to the borderline student whom the judge has in mind. From the item characteristic curve it follows that this borderline student has a probability of success equal to p_i. However, the judge specifies a probability equal to $p_i^{(s)}$. Now, a misspecification occurs if

$$e_i = p_i^{(s)} - p_i$$

is unequal to zero. It is easy to see that a judge is only consistent if no misspecifications occur, that is, if $e_i = 0$ for all items. As soon as misspecifications are obtained for some items, the judge is inconsistent in the sense that his probabilities do not imply the same mastery level and therefore cannot belong to one person. Thus, in order to decide whether a judge has worked consistently, a method is needed to assess the misspecifications e_i.

The following steps summarize how latent trait theory can be used for this purpose:

1. A latent trait model is chosen, its parameters are estimated, and its fit is tested. Suppose that n items fit the model.


prob. of success
$p_i(s)$
e_i
p_i
θ_c
mastery level

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 equations 1 and 2 (appendix), the Angoff or Nedelsky standard, τ_c, is computed.
4. The hypothesis to be tested is that the judge has worked consistently, i.e., has specified correct probabilities of success. Note that under this hypothesis, the Angoff or Nedelsky standard technically is a true score (expected observed score). The true score standard τ_c is next transformed into a standard on the θ-scale of the latent trait model via the estimated test characteristic curve (appendix, equation 3). Since the latent trait standard θ_c is no explicit function of τ_c, trial values must be substituted for the former until the value of the latter is obtained. The task is simplified by the fact that θ is monotonically related to τ. However, some computer programs standardly produce 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 $\hat{p}_i$ are computed.
6. In order to determine whether the hypothesis of a consistent judge is tenable, a comparison between the subjective probabilities, provided by the judge, $p_i^{(s)}$, and the objective probabilities estimated under the model, $\hat{p}_i$, must be made. This can be done using the index of consistency C_1 (appendix, equation 5). C_1 is the degree to which the average absolute misspecification differs from its maximum possible value, measured on the standard interval [0, 1]. The closer the value of C_1 to zero, the less tenable the hypothesis is that the judge has worked consistently.

7. A special difficulty is associated with the use of the Nedelsky technique. This technique can provide only a limited number of possible probabilities of success, and inconsistencies may therefore be attributable to the discrete character of the technique rather than the judge's behavior. The index λ (appendix, equation 6) can be used to assess how large a reduction of consistency has occurred because of the discrete character of the technique.
8. Finally, the pattern of differences between $p_i^{(s)}$ and $\hat{p}_i$, is analyzed. Technically, these differences are the "residuals" left over after the hypothesis of a consistent judge has been fitted to the data. An analysis of this pattern can be used, for instance, to detect items with systematic errors across judges, or items for which the judge needs additional training.

Results

An empirical investigation was carried out to illustrate the above method and to compare results for the Nedelsky and Angoff techniques. Eight Angoff and nine Nedelsky judges were used who each inspected the same 25-item test belonging to an instructional unit from a physics course introducing grade ten pupils to elementary mechanical concepts. All items were of the three- or four-choice type. A latent trait analysis, based on the responses of 156 pupils, produced 18 items showing a satisfactory fit to the Rasch model (appendix, equation 4). A more detailed description of the items and the design of the study is given in van der Linden (1981a, 1981b).

Table 1

Results for Nine Judges Using the Nedelsky Technique

Judge	E	C_1	C_2	λ
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

	Judge 2				Judge 5			
Item	$p_i(s)$	$\hat{p}_i$	$\hat{e}_i(u)$	$\hat{e}_i(\ell)$	$p_i(s)$	$\hat{p}_i$	$\hat{e}_i(u)$	$\hat{e}_i(\ell)$
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

Estimated 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

Table 1 shows the results for the nine Nedelsky judges. The first column gives the average absolute errors of specification (E); the next columns show the values for the consistency index (C_1) and the reduction in consistency due to the discrete character of the Nedelsky technique (θ). The mean error of specification for all nine judges was no less than .25. The mean value of θ was equal to .09.

Table 2 contains the values of $p_i^{(s)}$ and $\hat{p}_i$ for the least consistent as well as the most consistent judge. The last two columns show the upper ($\hat{e}^{(u)}$) and lower ($\hat{e}^{(\ell)}$) bounds to the estimated misspecification ($p_i^{(s)} - \hat{p}_i$) for the individual items. Note the larger variability of these specification errors for the worst judge.

The results for the eight Angoff judges are given in Table 3. The mean absolute error for all eight judges was equal to .18 and thus less serious than for the Nedelsky technique. Correspondingly, the values for C_1 are higher than the ones in Table 1. Table 4 gives more detailed information about the results for the least consistent and most consistent Angoff judges.

The conclusion from the above findings is that when using the Angoff or Nedelsky technique one has to reckon with serious misspecifications of the probabilities of success from which the standards are computed. On the whole, these errors are however noticeably less unfavorable for the Angoff than for the Nedelsky technique, the explanation being the fact that the latter admits only discrete probabilities and thus always forces the judge to be inconsistent to some extent.

Discussion

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 consequences of modifying standard-setting techniques, or (4) item analysis to detect items with systematic errors across judges or techniques.

For all these applications of the method, it is necessary that the items fit the latent trait model. However, if some of the items do not satifactorily fit the model, the method can still be used for the other items in the test. The only modification necessary is the computation of a new standard skipping the items not fitting the model. The estimation of the errors of specification and the consistency index are based on the new standard, and these estimates, then, still give an impression of how consistently the judge has worked.

References

Andrew, B.J., & 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.

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. 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.

Koffler, S.L. A comparison of approaches for setting standards. Journal of Educational Measurement, 1980, 17, 167-187.

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

Saunders, J.C., Ryan, J.P., & Huynh, H. A comparison of two approaches to setting passing scores based on the Nedelsky procedure. Applied Psychological Measurement, 1981, 5, 209-217.

Skakun, E.N. & Kling, S. Comparability of methods for setting standards. Journal of Educational Measurement, 1980, 17, 229-235.

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 mimimum 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. A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard setting. Submitted for publication, 1981. (a)

van der Linden, W.J. A latent trait look at pretest-posttest validition of criterion-referenced test items. Review of Educational Research, 1981, 51, 379-402. (b)

Appendix: Main Formulas and Equations

Angoff standard

For a test of length n, the Angoff standard is equal to

$$(1) \qquad \sum_{i=1}^{n} p_i^{(s)},$$

where $p_i^{(s)}$ is the borderline student's probability of success as specified by the judge.

Nedelsky standard

The Nedelsky standard is equal to (1) with

$$(2) \qquad p_i^{(s)} = [q_i - k_i]^{-1},$$

where q_i is the number of alternatives of item i, and k_i is the number of alternatives of which the judge indicates that the borderline student knows they are incorrect.

Test characteristic curve

For a student with θ_c, it holds that

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

where $P_i(+|\upsilon_c)$ is the probability of a correct response to item i; E( ) is the expected value operator, u_i is the item response variable (1 = correct, 0 = incorrect), and υ_c is the classical test theory true score.

Rasch model

$$(4) \qquad P_i(+|\theta_c) = \{1 + \exp[-(\theta_c - b_i)]\}^{-1},$$

where b_i is the difficulty parameter of item i.

Index of consistency

$$(5) \qquad C_1 = \frac{M - E}{M},$$

where

$$E = \sum_{i=1}^{n} |p_i^{(s)} - p_i|/n;$$

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

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

Note that $e_i^{(u)}$ is the maximum value of $|p_i^{(s)} - p_i|$, and that M is the maximum value of E.

Reduction in consistency

For the Nedelsky technique the reduction in consistency due to the discrete character of its probabilities is equal to

$$\lambda = C_2 - C_1, \tag{6}$$

where

$$C_2 = \frac{M - E}{M - m},$$

$$m = \sum_{i=1}^{n} e_i^{(\ell)}/n;$$

$$e_i^{(\ell)} = \left|(q_i - k_i^*)^{-1} - p_i\right|,$$

and k_i^* is the value of k_i in (2) chosen such that $e_i^{(\ell)}$ is minimal. Note that $e_i^{(\ell)}$ is the minimum value of $\left|p_i^{(s)} - p_i\right|$, and that m is the minimum value of E.