SETTING STANDARDS AND DETECTING INTRAJUDGE INCONSISTENCY USING INTERDEPENDENT EVALUATION OF RESPONSE ALTERNATIVES

LEI CHANG
Chinese University of Hong Kong

WIM J. VAN DER LINDEN AND HANS J. VOS
University of Twente

This article introduces a new test-centered standard-setting method as well as a procedure to detect intrajudge inconsistency of the method. The standard-setting method that is based on interdependent evaluations of alternative responses has judges closely evaluate the process that examinees use to solve multiple-choice items. The new method is analyzed against existing methods, particularly the Nedelsky and Angoff methods. Empirical results from three different experiments confirm the hypothesis that standards set by the new method are higher than those of the Nedelsky but lower than those of the Angoff method. The procedure for detecting intrajudge inconsistency is based on residual diagnosis of the judgments, which makes it possible to identify the sources of inconsistencies in the items, response alternatives, and/or judges. An empirical application of the procedure in an experiment with the new standard-setting method suggests that the method is internally consistent and has also revealed an interesting difference between residuals for the correct and incorrect alternatives.

Keywords: *standard setting; Angoff method; Nedelsky method; intrajudge inconsistency; judgmental item analysis; multiple-choice test; polytomous response models*

A thorny and recurring issue in the standard-setting literature has been how to determine the validity of a standard set on an achievement test. Since the discussion on standard setting in the 1978 special issue of the *Journal of Educational Measurement* (Glass, 1978; Hambleton, 1978; Popham, 1978), the dominant view has been that a standard cannot be justified by any inde-

Educational and Psychological Measurement, Vol. 64 No. 5, October 2004 781-801
DOI: 10.1177/0013164404264847
© 2004 Sage Publications

pendent criterion; rather, its validity should derive from the method used to set it (Kane, 1998).

The focus of this article is on judgmental, test-centered, standard-setting methods. These methods are based on one or more judges going through the process of solving the items in the test and specifying the behavior they expect a minimally competent examinee to display. The standard is derived from the judges' expectations about the performance of a minimally competent examinee.

We believe that a good method of judgmental standard setting should have at least the following two features. First, it should force the judges to carefully examine the process that examinees follow when solving the test items. Second, it should allow judges to specify their expectations about the behavior of a minimally competent examinee in a consistent way.

One purpose of this article is to present a new standard-setting method that has been developed explicitly to meet the first requirement. The method does so by capitalizing on the best characteristics of the well-known Angoff and Nedelsky methods. The second purpose is to present a method to assess the consistency with which judges specify their expectation about the behavior of a minimally competent examinee. The method not only quantifies possible inconsistencies but also allows the identification of the sources of inconsistencies in the judges, the items, and/or the response alternatives.

Study 1: Development of a New Standard-Setting Method

In judgmental standard setting for multiple-choice (MC) tests, three different steps can be distinguished (Chang, 1999). First, a definition of the minimum competency needed to pass the test is established. Second, a judgmental item analysis (Jaeger, 1989) is executed in which judges specify their expectations about the performances of an examinee with minimum competence on the test. Third, an algorithm is used to transform the specifications of the judges into a cutoff score on the test. Because judgmental standard-setting methods primarily differ in the second and third step, we will further ignore the role of the definition of minimum competency in this article.

The authors are equally responsible for the content of this article; the order of authorship is determined alphabetically. This article was completed while the second author was a fellow at the Center for Advanced Study, Stanford University. He is indebted to the Spencer Foundation for a grant awarded to the Center to support his fellowship. This article is supported by a Mainline research grant (44M2003) from the Chinese University of Hong Kong and partly supported by an earmarked grant (4047/98H) from the Research Grants Council of the Hong Kong SAR. Correspondence concerning this article should be addressed to Lei Chang at leichang@cuhk.edu.hk or to Wim van der Linden at w.j.vanderlinden@edte.utwente.nl.

Angoff and Nedelsky Methods

The Angoff and Nedelsky methods are the pioneering examples of judgmental standard setting. At a superficial level, seemingly the only distinction between the two methods is the way by which judges specify their expectations about the performances of an examinee with minimum competence. In the Angoff method, judges indicate the probabilities of success with which a minimally competent examinee is expected to choose the correct response alternatives of the items. In the Nedelsky method, they indicate which alternatives such an examinee is expected to identify as wrong.

CALCULATION OF CUTOFF SCORES

The Angoff and Nedelsky methods use different rules to calculate cutoff scores. Let p^s_{ij} denote the (subjective) probability of success for a minimally competent examinee on item i as specified by judge j in the Angoff standard-setting procedure. The Angoff cutoff score on a test of n items for judge j is defined as

$$\tau_{cj} = \sum_{i=1}^{n} p^s_{ij}. \tag{1}$$

Let q_i be the number of alternatives on item i and k^s_{ij} the number of alternatives that judge j believes a minimally competent examinee will identify as wrong in a Nedelsky standard-setting procedure. The Nedelsky cutoff score on a test of n items for judge j is defined as

$$\tau_{cj} = \sum_{i=1}^{n} (q_i - k^s_{ij})^{-1}. \tag{2}$$

DIFFERENCES BETWEEN METHODS

However, at a more fundamental level, the two methods differ in important assumptions about the way that examinees are believed to find a solution for the items on an MC test. These assumptions deal with the following two aspects of examinee behavior: (a) the information in the items on which examinees focus and (b) the item-solving strategy they adopt. We first analyze how the Angoff and Nedelsky methods succeed and fail to correctly address these two aspects of examinee behavior correctly, and then we propose and discuss a new method that combines the stronger properties of these existing methods.

Focus on information in the items. Both methods are typically implemented in a way in which judges are asked to solve the items on a test and then

to provide the data needed to calculate the cutoff score. However, in the Angoff procedure, the judges have to provide only a probability estimate for the correct response alternative, whereas in the Nedelsky method, judges have to make a decision on each of the distractors. Thus, in the Angoff method, it is not necessary for a judge to evaluate each response alternative to provide the data. Although Angoff judges are required to evaluate all the response alternatives before arriving at an item probability estimate, there is no built-in mechanism to ensure the full compliance to this requirement. On the contrary, it is possible and highly likely that the judges focus only on the stem of the item or on the stem and the correct alternative. On the other hand, in the Nedelsky method, judges must inspect both the stem and each of the distractors. Thus, with regard to the information available in the items, the Nedelsky method invokes a wider focus for the judges than the Angoff method does. However, the Nedelsky method does not have a mechanism to draw judges' attention to the correct alternative.

Because it is possible for the judges in the Angoff method not to be distracted by the distractors, they may underestimate the difficulty of the items. However, for judges using the Nedelsky method, it is possible to ignore the correct alternative while evaluating the distractors. Therefore, Nedelsky judges may fail to consider possible clues contained in the correct alternatives and overestimate the difficulty of the items. As a result of these differences in focus, cutoff scores set using the Nedelsky method are expected to be lower than those using the Angoff method.

The fact that the Nedelsky method has a tendency to set low cutoff scores has been observed, for instance, by Chang, Dziuban, Hynes, and Olson (1996); Cross, Impara, Frary, and Jaeger (1984); and Shepard (1995). The fact that the Angoff method has a tendency to set standards that are more difficult to meet has been observed by Paiva and Vu (1979) and Rock, Davis, and Werts (1980). Empirical findings confirming these tendencies can be found in Baron, Rindone, and Prowda (1981); Behuniak, Archambault, and Gable (1982); Brennan and Lockwood (1980); Cross et al. (1984); Halpin, Sigmon, and Halpin (1983); Harasym (1981); Poggio, Glasnapp, and Eros (1981); Rock et al. (1980); and Smith and Smith (1988). For a review of the empirical differences between Angoff and Nedelsky standard setting, see Chang (1999).

Item-solving strategy. The Nedelsky method assumes that minimally competent examinees follow an elimination process when solving MC items. They are expected to identify incorrect alternatives until they end up with a set of alternatives about which they are confused. On the other hand, the Angoff method seems to assume that examinees follow a selection process. Examinees are expected to go directly for the correct alternative, which they find with a probability that varies as a function of their level of competence.

In fact, the Nedelsky method is even more specific in its assumptions about the behavior of examinees with minimum competence in that it expects them to guess blindly among the remaining alternatives. This model of "knowledge or random guessing" that is assumed in the Nedelsky method is the same as the one underlying the well-known correction for guessing on MC tests (formula scoring) or the three-parameter logistic item response model (Birnbaum, 1968). The model has been criticized for two reasons (Lord & Novick, 1968, section 14.2). First, examinees may have misinformation or "negative knowledge" about an item. Second, they may have partial information about some of the alternatives. In either case, the assumption of random guessing with equal probabilities among the remaining alternatives is unrealistic. Empirical research suggests that low ability students do not rely on random guessing in choosing among the plausible distractors (Kassirer & Kopelman, 1989; Ramsden, Whelan, & Cooper, 1989). As Maguire, Skakun, and Harley (1992) stated, "There is almost always one alternative that is more attractive than the others" (p. 440).

The assumption of random guessing also introduces the unnecessary problem of having a small set of possible values for the probability estimates. For example, if a minimally competent examinee is judged able to eliminate two distracters of a four-choice item, the resulting Nedelsky probability estimate is .05. The next higher probability possible is 1.0. However, if examinees have partial knowledge, item probabilities between .5 and 1.0 are possible. Large gaps in the set of possible item probabilities are an important source of intrajudge inconsistency in the Nedelsky method (van der Linden, 1982). They also have a depression effect contributing to lower Nedelsky cutoff scores frequently reported in the literature (Chang, 1999).

Interdependent Evaluation of Alternatives (IDEA) Method

We propose a new standard-setting method that compensates for the weaknesses of the Nedelsky and Angoff methods explained above. In particular, the method assumes that examinees with minimum competence focus on the information in the entire item; that is, they consider both its stem and all of its response alternatives. Therefore, this method requires judges to do the same. Besides, the method assumes an item-solving strategy, coined here as IDEA, in which examinees do not make an absolute judgment about one response alternative but evaluate the plausibility of each alternative against all others. If an examinee views one alternative as more likely to be the correct answer, then the other alternatives are viewed as less likely to be correct. The new method therefore should invoke in its judgment the same weighing of the response alternatives as shown by the minimally competent examinees.

STEPWISE DESCRIPTION OF METHOD

The basic protocol of the proposed method contains the following steps:

1. The judges are asked to reconstruct the process that minimally competent examinees would follow when answering the items, that is, to evaluate the correctness of each of the alternatives against all others in relation to the problem formulated in the stem.
2. The judges are then asked to specify for each response alternative, including the correct alternative, the probability that a minimally competent examinee would choose the alternative as the correct answer. These probabilities are required to be specified such that their sum equals 1.
3. The cutoff score on the test is calculated as the sum of the subjective probabilities for the correct alternatives of the items specified by the judge. The probabilities for all other response alternatives of the items are ignored. If $p^s_{g_i j}$ denotes the probability specified by judge j on the correct alternative g_i of item $i = 1, \ldots, n$, the cutoff score is given by

$$\tau_{cj} = \sum_{i=1}^{n} p^s_{g_i j}. \tag{3}$$

Of course, it is possible to implement this protocol in different ways. For example, judges may be asked to discuss intermediate results with each other, to revise earlier probabilities based on feedback by the experimenter, to view actual answers by examinees, and so forth. However, the focus of this article is not to provide a panoply of possible implementations of the new method. Instead, the article focuses on the extent to which a standard-setting method attends to the item-solving strategies employed by the minimally competent examinees.

DISCUSSION

The proposed method is believed to improve both the Angoff method and the Nedelsky method. First, unlike the Angoff and Nedelsky methods, it requires judges to focus on all information in the item and not to ignore any of the alternatives. It requires judges to inspect the correct alternative and thus to evaluate the impact of possible clues about the item's answer. It also requires judges to inspect the incorrect alternatives and to evaluate the possibility that a minimally competent examinee is distracted by each of them.

Second, the method is not based on an item-solving strategy that consists of either an elimination or a selection process. An elimination process can take place only if, as in the Nedelsky method, the probabilities of an alternative being identified as incorrect are allowed to be equal only to 0 or 1. In fact, by imposing the restriction that the sum of the probabilities for all alternatives must be equal to 1, the IDEA method forces the judges to evaluate all response alternatives interdependently and allows for the processes of elimination and selection to happen simultaneously and in a more balanced way.

Third, the IDEA method does not necessarily assume random guessing on the part of the minimally competent examinees. The probabilities defined in Equation 3 allow judges to take into consideration the partial and negative knowledge of the examinees.

Fourth, the probability estimates can take any value on the standard interval [0, 1] and thus do not suffer from the intrajudge inconsistencies due to discrete estimates. It also reduces the tendency to set lower standards for higher performing minimally competent examinees, as seen in the Nedelsky method in which a restricted set of probabilities must be used. Finally, because judges are required to specify the probabilities for each of the distractors, the possibility of specifying unrealistic probabilities for the correct alternative equal to 1 and 0, as may happen among judges using the Angoff method, is absent.

Empirical Experiments

Three different standard-setting experiments were conducted to test predictions with respect to differences in the cutoff scores among the Nedelsky, Angoff, and IDEA methods. It was hypothesized that the cutoff score set by the IDEA method would be higher than those set by the Nedelsky method but lower than those set by the Angoff method. The rationale for this hypothesis follows from the earlier conceptual analysis of the three methods.

EXPERIMENT 1

The test consisted of 29 items from an exit exam of German as a second language administered nationally at the end of Dutch secondary education. All 29 items were of the multiple-choice format, with the number of alternatives ranging from three to five. The judges were 22 secondary school teachers from the Netherlands. There were 18 male and 4 female judges. Their average experience in teaching German was 8.5 years ($SD = 1.5$). These participants were thus qualified to serve as judges in the experiment.

The judges were randomly assigned to one of the three standard-setting methods, seven judges to each of the Nedelsky and Angoff methods and eight judges to the IDEA method. The judges first received training in a plenary session on how to conceptualize minimum competency. Subsequently, each judge received separate training on the standard-setting method he or she was asked to use. The actual standard-setting process started when judges indicated that they fully understood the method and deemed themselves confidant to use it to set a standard. On average, the actual standard-setting process lasted about an hour. The Angoff method took the shortest time, with most judges finishing in more than half an hour. The IDEA method took the longest time, with most judges exceeding 1 hour.

Table 1
Cutoff Scores and Standard Deviations of the Three Standard-Setting Methods

Method	N	Cutoff Score	SD
Experiment 1 (29 items)			
Nedelsky	7	13.68	2.74
IDEA	8	15.92	1.94
Angoff	7	17.76	2.09
Experiment 2 (20 items)			
Nedelsky	10	9.50	0.93
IDEA	10	11.05	0.84
Angoff	10	11.88	0.88
Experiment 3 (19 items)			
Nedelsky	4	9.11	2.33
IDEA	5	10.42	1.80
Angoff	4	12.40	1.45

Note. IDEA = Interdependent Evaluation of Alternatives.

EXPERIMENT 2

A 20-item geography test for fifth-grade elementary students was used. All 20 items were of the four-choice MC format. The judges were 30 teachers, of whom 23 were female, from elementary schools in the Netherlands. Ten judges were randomly assigned to each of the three standard-setting methods. The training and standard-setting procedures were similar to those in Experiment 1.

EXPERIMENT 3

The test consisted of 19 items from an exit exam of English as a second language administered nationally at the end of secondary education in the Netherlands. Each of these 19 MC items had three to five alternatives. The judges were 13 secondary school English teachers, of whom 7 were female. The average teaching experience of these judges was 18 years. Five judges were randomly assigned to the IDEA method and 4 to the Angoff and Nedelsky methods, respectively. These judges underwent similar training and standard-setting procedures as those of Experiment 1. However, this time, the training was conducted by e-mail.

Results and Conclusion

The cutoff scores and the standard deviations of the three methods from each of the three experiments are reported in Table 1. For each method, the cutoff scores on the tests were calculated as averages over the cutoff scores of the individual judges.

As hypothesized, in all three experiments, the IDEA cutoff score fell between the scores for the Nedelsky and Angoff methods. A one-way ANOVA showed the differences among the three cutoff scores to be significant in Experiment 1, $F(2, 19) = 5.68$, $p < .05$, and Experiment 2, $F(2, 27) = 14.63$, $p < .05$, but not in Experiment 3, $F(2, 10) = 3.09$, $p = .09$. The results from these three experiments are considered to confirm the hypothesis in a robust way, especially because the power of the statistical tests for Experiment 3 was low.

Post hoc comparisons using the Bonferroni procedure revealed the same pattern. In Experiment 1, there was a significant difference between the cutoff scores for the Angoff and Nedelsky methods ($p < .01$). In Experiment 2, significant differences were found between the cutoff scores for the IDEA and Nedelsky ($p < .05$) and between those for the Nedelsky and Angoff methods ($p < .01$).

We also conducted nonparametric binomial tests at the level of the judges' specifications for individual items. For each item, the probability estimates across judges within each of the three groups were averaged. The results were 29, 20, and 19 triplets of average probability estimates for Experiments 1, 2, and 3, respectively, each ordered from the lowest to the highest probability. Seventeen of the 29 triplets in Experiment 1 had the hypothesized order (Nedelsky method, IDEA method, and Angoff method). This ratio was highly significant under the null hypothesis of a uniform distribution of triplets ($p < .0001$). Similarly, in Experiment 2, 11 of the 20 triplets showed the hypothesized order ($p < .001$), and in Experiment 3, 9 of the 19 had the expected order ($p < .002$).

Consistent with the literature (Behuniak et al., 1982; Brennan & Lockwood, 1980; Chang, 1999; Cross et al., 1984), it was found that the Nedelsky method had a larger standard deviation than the Angoff method in all three experiments. The IDEA method had the lowest standard deviation in Experiments 1 and 2. In Experiment 3, the IDEA standard deviation was lower than the Nedelsky but a half-point higher than the Angoff standard deviation. This finding suggests that the new method produces more agreement between judges.

The previous results thus seem to confirm our predictions for the cutoff scores set by the IDEA method relative to those for the Nedelsky and Angoff methods. However, a more rigorous evaluation of the method can be conducted using a procedure to assess the intrajudge consistency of the subjective probabilities the method requires the judges to specify. Our next study deals with this issue.

Study 2: Detecting Intrajudge Inconsistency

van der Linden (1982) introduced both the concept of intrajudge consistency and a procedure to evaluate the intrajudge internal consistency of the Nedelsky and Angoff methods (see also Kane, 1998). The purpose of this study is to generalize the procedure to the case of standard setting with probabilities specified for each of the response alternatives, such as in the IDEA method. However, in principle, the procedure can also be applied to the evaluation of any standard-setting method that capitalizes on a polytomous response format for the items (e.g., items that require partial-credit scoring or grading by content experts).

The procedure is based on the statistical technique of residual analysis. It requires that a model for the probabilities on the response alternatives be fit to examinee response data and then analyzes the residuals in the judges' subjective probabilities under the hypothesis of consistent judgment.

Definitions and Notation

As before, the test items used in the IDEA method are denoted as $i = 1, \ldots, n$. To allow for items with different numbers of alternatives, as in the empirical experiments above, the response alternatives for item i are denoted as $k_i = 1, \ldots, m_i$. A separate notation is needed for the correct and incorrect alternatives of the items. The correct alternative of item i is still denoted as g_i, whereas an arbitrary incorrect alternative is denoted as w_i. The items are assumed to measure a (unidimensional) variable θ representing the performances of the examinees. Each of the judges $j = 1, \ldots, N$ is asked to choose a standard for the performance level required from the examinees. The standard for judge j is denoted as a cutoff score θ_{cj}. Observe that the standards are indexed by j because we evaluate the consistency of each individual judge. For each item, the judges are required to specify the probabilities of an examinee's operating at performance level θ_{cj} to produce response $X_i = k_i$ on item i. As before, these subjective probabilities are denoted as $p^s_{k_i j}$.

IRT Model

If the response data for the populations of examinees fit an IRT model for items with a polytomous response format, then we also have objective probabilities for response $X_i = k_i$ by an examinee at performance level θ_{cj}. In the empirical example below, Thissen and Steinberg's (1984, 1997) model for MC items was fitted to the data. The model defines the probability of an examinee at θ_{cj} producing response $X_i = k_i$ as

$$p_{k_i j} \equiv Prob\{X_i = k_i | \theta_{cj}\} \equiv \frac{\exp\{a_{k_i}(\theta_{cj} - b_{k_i})\} + d_{k_i}\exp\{a_{0_i}(\theta_{cj} - b_{o_i})\}}{\sum_{h_i=0}^{m_i}\exp\{a_{h_i}(\theta_{cj} - b_{h_i})\}}, \quad (4)$$

where b_{k_i} and a_{k_i} are the location and discriminating power of alternative k of item i, respectively. The model, which generalizes Bock's (1997) nominal response model, was chosen because of its flexibility to deal with guessing on MC items. It does so by assuming that among examinees that give response k_i to item i, a (priori unknown) proportion d_{k_i} guesses

$$(\sum_{k_i=1}^{m_i} d_{k_i} = 1).$$

The process of guessing is not assumed to be blind but to be dependent on θ with probabilities given by

$$\exp\{a_{0_i}(\theta_{cj} - b_{0_i})\} / \sum_{h_i=0}^{m_i}\exp\{a_{h_i}(\theta_{cj} - b_{h_i})\},$$

with a_{0_i} and b_{0_i} denoting the location and discriminating power of the response function for the examinees who guess.

When the model in Equation 4 is fitted to data from achievement tests, the response function for the correct alternative should be monotone in θ. In the application below, the validity of this assumption is tested against the alternative of a nonmonotone response function.

Error Definition

Observe that θ_{cj} should be calculated from the subjective probabilities provided by judge j under the hypothesis of consistent judgments. The assumption is typical of the technique of residual analysis used in this article. This technique consists of the following steps. First, a model for the probabilities on the alternatives is fitted to the response data from a representative set of examinees. In the application below, the model is the one specified in Equation 4. Second, under the null hypothesis of a consistent judge, a cutoff score is fitted to the judge's subjective probabilities. Third, the residuals, which are the differences between the objective probabilities from the model and the subjective probabilities from the judge, are calculated. Fourth, the residuals are analyzed for inconsistencies, and then potential explanations for the inconsistencies are developed.

For the current response model in Equation 4, the calculation of the cutoff score θ_{cj} for judge j in the second step is based on the following operations:

1. summing the probabilities $p^s_{g_i j}$ over the items in the test;
2. summing the objective probabilities for the correct alternatives over the items in the test; and
3. equating the two sums and calculating θ_{cj} as the root of the equation.

That is, θ_{cj} is calculated as the root of

$$\sum_{i=1}^{n} p^s_{g_i j} = \sum_{i=1}^{n} \frac{\exp\{a_{g_i}(\theta_{cj} - b_{g_i})\} + d_{k_i} \exp\{a_{0_i}(\theta_{cj} - b_{0_i})\}}{\sum_{k_i=0}^{m_i} \exp\{a_{k_i}(\theta_{cj} - b_{k_i})\}}. \tag{5}$$

The error by judge j on alternative k of item i is thus equal to the residual probability

$$e_{k_i j} \equiv p^s_{k_i j} - p_{k_i j}. \tag{6}$$

It is now possible to aggregate the error in Equation 6 over response alternatives, items, or judges. This aggregation results in inconsistency indices for (combinations of) judges and items. We first introduce a set of nonstandardized inconsistency indices and then indicate how to standardize these indices to take possible values only in the interval [0, 1].

ERRORS BY INDIVIDUAL JUDGES

The absolute errors by judge j on the correct and incorrect alternatives of item i are given by

$$\in_{g_i j} \equiv \left| p^s_{g_i j} - p_{g_i j} \right| \sum_{k=1;k\neq g}^{m_i} \left| p^s_{k_i j} - p_{k_i j} \right| \tag{7}$$

and

$$\in_{w_i j} \equiv (m_i - 1)^{-1} \sum_{k=1;k\neq g}^{m_i} \left| p^s_{k_i j} - p_{k_i j} \right|, \tag{8}$$

respectively.

Aggregating these errors over the items gives the following indices for the average errors by judge j on the correct alternative, incorrect alternatives, and across all alternatives at the level of the test:

$$\in_{gj} = n^{-1} \sum_{i=1}^{n} \left| p^s_{g_i j} - p_{g_i j} \right| \tag{9}$$

$$\in_{wj} \equiv \left(\sum_{i=1}^{n} m_i - n\right)^{-1} \sum_{i=1}^{n} \sum_{k=1;k\neq g}^{m_i} |p^s_{k_i j} - p_{k_i j}| \tag{10}$$

$$\in_{j} \equiv \left(\sum_{i=1}^{n} m_i\right)^{-1} \sum_{i=1}^{n} \sum_{k=1}^{m_i} |p^s_{k_i j} - p_{k_i j}|. \tag{11}$$

This choice for the absolute values of the errors is made to prevent them from compensating for each other when they are aggregated within or between items or judges.

Errors by a Panel of Judges

Subjective probabilities specified in standard-setting experiments differ in the likelihood of having an error. The reason for such differences may reside, for example, in sloppy behavior by a judge, the formulation of the items, the difficulty of the correct alternative, or the familiarity of the judge with specific topics in the domain tested. Item analysis based on the aggregate errors of a panel of judges can help to reveal the actual sources of such differences.

The following equations give the average errors for the correct alternative, the incorrect alternatives, and across all alternatives of item *i* across a panel of judges:

$$\in_{g_i} \equiv n^{-1} \sum_{j=1}^{N} |p^s_{g_i j} - p_{g_i j}| \tag{12}$$

$$\in_{wi} = N^{-1}(m_i - 1)^{-1} \sum_{j=1}^{N} \sum_{k=1;k\neq g}^{m_i} |p^s_{k_i j} - p_{k_i j}| \tag{13}$$

$$\in_{i} \equiv (Nm_i)^{-1} \sum_{j=1}^{N} \sum_{k=1}^{m_i} |p^s_{k_i j} - p_{k_i j}|. \tag{14}$$

Analogous to Equations 9 to 11, the errors by a panel of judges can be aggregated over all items in the test. These aggregates can be used, for example, to detect differences between the error levels for the correct and incorrect alternatives or the general error level for a panel of judges on the test. The equations are as follows:

$$\in_{g} \equiv (Nn)^{-1} \sum_{j=1}^{N} \sum_{i=1}^{n} |p^s_{g_i j} - p_{g_i j}| \tag{15}$$

$$\in_w \equiv N^{-1}\left(\sum_{i=1}^{n} m_i - n\right)^{-1} \sum_{j=1}^{N}\sum_{i=1}^{n}\sum_{k=1;k\neq g}^{m_i} |p^s_{k_i j} - p_{k_i j}| \tag{16}$$

$$\in \equiv \left(N\sum_{i=1}^{n} m_i)\right)^{-1} \sum_{j=1}^{N}\sum_{i=1}^{n}\sum_{k=1}^{m_i} |p^s_{k_i j} - p_{k_i j}|. \tag{17}$$

Standardized Consistency Indices

The above inconsistency indices should be used descriptively. The use of a statistical test for the hypothesis of consistent judgments is still hampered by the difficulty in formulating a statistical model for the distribution of the subjective probabilities $p^s_{k_i j}$ across replications. To support the comparison of errors among items, judges, or occasions, it is therefore important to have standardized versions of the indices that have a common range of possible values.

Standardization of the above indices that can take values in the full interval [0, 1] is achieved through the transformation

$$C \equiv \frac{M_\in - \in}{M_\in}, \tag{18}$$

where $\in$ is a generic symbol for the inconsistency indices and $M_\in$ is the maximum possible value of the index (van der Linden, 1982). The maximum is found if index $\in$ is calculated with the expression $p^s_{k_i j} - p_{k_i j}$ in Equation 6 replaced by

$$\max\{p_{k_i j}, 1 - p_{k_i j}\}. \tag{19}$$

Because the calculations are straightforward, no equations for the consistency indices are given.

The main purpose for standardizing the residuals is to make them independent of the objective probabilities of success at the performance level of the borderline examinee, θ_{cj}. The maximum residual in Equation 19 varies as a function of θ, whereas index C does not. Observe that the direction of C is also opposite to the direction of $\in$. C should therefore be considered as a consistency index; the closer its value to 1, the more consistent the judgments. The maximum $C = 1$ is obtained if at θ_{cj}, it holds that $p^s_{k_i j} = p_{k_i j}$ for all alternatives, items, and/or judges over which the index is defined.

Table 2
Summary of Residuals for Correct Alternative

	Judge								
Item	1	2	3	4	5	6	7	8	Average
1	0.28	0.19	0.34	0.19	0.15	0.22	0.09	0.24	0.21
2	0.24	0.50	0.37	0.43	0.22	0.41	0.35	0.30	0.35
3	0.16	0.06	0.07	0.09	0.21	0.04	0.07	0.01	0.09
4	0.31	0.13	0.01	0.13	0.01	0.40	0.27	0.34	0.20
5	0.08	0.00	0.10	0.11	0.14	0.11	0.09	0.07	0.09
6	0.04	0.13	0.20	0.24	0.02	0.17	0.07	0.02	0.11
7	0.17	0.09	0.13	0.12	0.32	0.35	0.21	0.16	0.19
8	0.18	0.20	0.19	0.14	0.26	0.34	0.15	0.17	0.20
9	0.02	0.05	0.01	0.29	0.12	0.12	0.01	0.09	0.09
10	0.23	0.21	0.13	0.40	0.00	0.20	0.39	0.18	0.22
11	0.06	0.14	0.10	0.18	0.11	0.25	0.18	0.39	0.18
12	0.42	0.41	0.45	0.59	0.37	0.45	0.47	0.45	0.45
13	0.04	0.03	0.03	0.14	0.07	0.21	0.01	0.03	0.07
14	0.11	0.14	0.18	0.32	0.22	0.13	0.06	0.30	0.18
15	0.14	0.08	0.09	0.05	0.09	0.18	0.22	0.22	0.13
16	0.09	0.05	0.03	0.37	0.06	0.01	0.13	0.06	0.10
17	0.18	0.40	0.40	0.02	0.20	0.19	0.12	0.23	0.22
18	0.03	0.02	0.18	0.38	0.26	0.14	0.10	0.15	0.16
19	0.23	0.22	0.00	0.17	0.15	0.24	0.22	0.14	0.17
Average	0.16	0.16	0.16	0.23	0.16	0.22	0.17	0.19	0.18

Empirical Example

Data reported here are from the eight IDEA judges in Experiment 1 from Study 1. The 29 MC items were previously calibrated under the model in Equation 4 using the response data from 161,648 examinees and the software program Multilog (Thissen, 1991). The goodness of fit of the model was assessed against both a less restrictive model and a more restrictive model fitted to the same data set. The direct likelihood-ratio test of the model against the general multinomial alternative in Multilog could not be used because the number of examinees was of a much smaller order than the number of possible response patterns (which was equal to $1{,}938 \times 10^{11}$). For the use of such alternative goodness-of-fit tests, see Thissen and Steinberg (1984). The less restrictive model was Mokken's (1997) nonparametric response model. This model was used to check the items for the unidimensionality of θ, as well as for the monotonicity of the response function for the correct alternative using the software program MSP 5 (Molenaar & Sijtsma, 2000). A set of 19 items yielded a scalability coefficient $H = .14$, which is to be considered a conservative value (Molenaar & Sijtsma, 2000). Because the Mokken model does not assume any parametric form for the response functions, it follows that the

Table 3
Summary of Residuals for Incorrect Alternatives

	Judge								
Item	1	2	3	4	5	6	7	8	Average
1	0.13	0.09	0.12	0.08	0.05	0.11	0.11	0.09	0.09
2	0.11	0.17	0.12	0.14	0.07	0.14	0.12	0.10	0.12
3	0.09	0.08	0.03	0.02	0.10	0.03	0.03	0.06	0.06
4	0.11	0.10	0.08	0.19	0.06	0.20	0.17	0.11	0.13
5	0.10	0.12	0.17	0.04	0.23	0.10	0.08	0.19	0.14
6	0.12	0.09	0.07	0.08	0.10	0.07	0.05	0.04	0.08
7	0.10	0.21	0.07	0.07	0.18	0.12	0.22	0.13	0.17
8	0.11	0.12	0.06	0.13	0.10	0.12	0.06	0.13	0.10
9	0.11	0.06	0.05	0.26	0.06	0.12	0.05	0.04	0.09
10	0.14	0.16	0.09	0.17	0.02	0.10	0.18	0.07	0.12
11	0.21	0.04	0.17	0.18	0.14	0.24	0.09	0.24	0.16
12	0.18	0.14	0.17	0.20	0.22	0.20	0.16	0.22	0.18
13	0.03	0.01	0.06	0.10	0.10	0.07	0.12	0.06	0.07
14	0.10	0.06	0.09	0.13	0.08	0.23	0.05	0.10	0.10
15	0.08	0.09	0.03	0.05	0.08	0.18	0.09	0.07	0.08
16	0.15	0.08	0.08	0.14	0.03	0.08	0.19	0.18	0.12
17	0.12	0.13	0.13	0.01	0.07	0.06	0.04	0.08	0.08
18	0.05	0.12	0.08	0.13	0.09	0.05	0.07	0.05	0.08
19	0.12	0.11	0.05	0.09	0.07	0.12	0.11	0.07	0.09
Average	0.13	0.11	0.09	0.11	0.10	0.12	0.10	0.11	0.11

data support these two critical assumptions. The assumption of monotonicity of the response functions for the correct alternatives is particularly important because the model in Equation 4 was applied to achievement test items. The more restrictive model was the nominal response model (Bock, 1997). For the same set of items, a likelihood-ratio test showed that this model had to be rejected in favor of the model in Equation 4 ($p < .001$). This set was therefore used in the experiment.

Summaries of the (aggregated) residuals for the correct and incorrect alternatives are given in Tables 2 and 3, respectively. A consistent trend in the two tables is the difference between the residuals for the correct and incorrect alternatives. The average residuals across all judges and items is 0.18 for the correct alternative and 0.11 for the incorrect alternatives. The ranges for the average residuals per judge are remarkably small: 0.16 to 0.23 for the correct alternatives and 0.10 to 0.13 for the incorrect alternatives. The difference in range can be explained by the fact that the results for the incorrect alternative are based on the extra step of averaging (the number of incorrect alternatives was 2 to 4).

The average residuals per item are also in a relatively small range for the incorrect alternatives (0.07-0.18). However, the average residuals per item

Table 4
Summary of Consistency Index for Correct Alternative

	Judge								
Item	1	2	3	4	5	6	7	8	Average
1	0.69	0.76	0.64	0.77	0.85	0.76	0.87	0.73	0.75
2	0.74	0.44	0.62	0.54	0.77	0.57	0.59	0.69	0.62
3	0.76	0.88	0.92	0.85	0.74	0.94	0.88	0.98	0.87
4	0.47	0.78	0.99	0.77	0.99	0.33	0.53	0.42	0.66
5	0.86	1.00	0.82	0.81	0.74	0.80	0.86	0.87	0.85
6	0.92	0.79	0.66	0.59	0.96	0.67	0.90	0.97	0.81
7	0.75	0.87	0.80	0.82	0.48	0.46	0.71	0.76	0.71
8	0.77	0.75	0.73	0.82	0.60	0.54	0.81	0.78	0.73
9	0.97	0.92	0.99	0.58	0.85	0.84	0.99	0.88	0.88
10	0.64	0.72	0.78	0.43	0.99	0.61	0.51	0.68	0.66
11	0.92	0.83	0.82	0.77	0.78	0.61	0.77	0.47	0.75
12	0.49	0.52	0.36	0.30	0.45	0.40	0.46	0.44	0.43
13	0.92	0.95	0.95	0.74	0.88	0.59	0.99	0.94	0.87
14	0.85	0.81	0.71	0.55	0.65	0.81	0.92	0.57	0.74
15	0.82	0.89	0.89	0.94	0.88	0.77	0.69	0.72	0.83
16	0.81	0.91	0.94	0.29	0.90	0.99	0.78	0.89	0.82
17	0.80	0.53	0.56	0.98	0.78	0.78	0.85	0.74	0.75
18	0.96	0.96	0.77	0.39	0.68	0.81	0.81	0.79	0.77
19	0.72	0.73	1.00	0.79	0.83	0.71	0.74	0.84	0.79
Average	0.78	0.77	0.78	0.68	0.78	0.69	0.76	0.74	0.75

for the correct alternatives showed two outlying results: 0.35 for Item 2 and 0.45 for Item 12. If these results are excluded, the range is 0.07 to 0.22.

A comparison between the residuals for Item 2 and 12 in Tables 2 and 3 shows that both are uniformly high across judges for the correct alternative. Item 12 also shows uniformly high residuals for the incorrect alternatives, whereas Item 2 shows results for the judges that are not systematically larger than those for other response items (not clear). There are two reasons why residuals can be large: (a) attributes specific to the item that can make it difficult to specify subjective probabilities for one or more of its alternatives and (b) the dependency of the residuals on θ_{cj}.

The latter explanation should be rejected if the residuals disappear in an analysis based on the standardized consistency indices. Tables 4 and 5 show the values of these indices for the same items and judges. A comparison between these two sets of tables seems to support the hypothesis that the results for Item 12 are due to the attributes of the item or, particularly, attributes of the correct alternatives (the values for the incorrect alternatives do not show any remarkable pattern). The results from Item 2 were more in line with those of other items (albeit the average consistency across judges was among the lowest in value). Referring to the response data for the examinees, we

Table 5
Summary of Consistency Indices for Incorrect Alternatives

	Judge								
Item	1	2	3	4	5	6	7	8	Average
1	0.89	0.91	0.88	0.92	0.95	0.88	0.88	0.91	0.90
2	0.92	0.83	0.88	0.85	0.92	0.86	0.88	0.90	0.88
3	0.83	0.91	0.96	0.98	0.90	0.97	0.97	0.94	0.93
4	0.88	0.89	0.91	0.78	0.93	0.76	0.80	0.87	0.85
5	0.72	0.85	0.80	0.95	0.72	0.87	0.89	0.77	0.82
6	0.86	0.89	0.91	0.90	0.88	0.92	0.93	0.95	0.90
7	0.57	0.72	0.92	0.91	0.77	0.84	0.71	0.83	0.78
8	0.86	0.85	0.92	0.84	0.89	0.85	0.92	0.84	0.87
9	0.87	0.93	0.95	0.69	0.93	0.86	0.93	0.95	0.89
10	0.83	0.79	0.90	0.78	0.97	0.88	0.76	0.91	0.85
11	0.72	0.94	0.79	0.76	0.83	0.69	0.87	0.69	0.79
12	0.79	0.83	0.79	0.76	0.73	0.76	0.80	0.74	0.78
13	0.96	0.99	0.93	0.88	0.88	0.92	0.85	0.93	0.92
14	0.88	0.92	0.89	0.83	0.90	0.71	0.93	0.87	0.87
15	0.91	0.90	0.97	0.94	0.92	0.81	0.90	0.92	0.91
16	0.82	0.90	0.90	0.82	0.96	0.90	0.77	0.79	0.86
17	0.87	0.86	0.86	0.99	0.93	0.93	0.95	0.92	0.91
18	0.94	0.86	0.91	0.85	0.91	0.95	0.92	0.95	0.91
19	0.87	0.88	0.95	0.91	0.92	0.87	0.88	0.92	0.90
Average	0.84	0.87	0.90	0.87	0.89	0.86	0.87	0.87	0.87

found that the *p* value for these examinees (.40) was lower than the *a* value for one of the incorrect alternatives (.49). This observation may suggest ambiguity with respect to the correct alternative. In a real-life application of this method, in which judges receive feedback, the next step would be to ask the judges to discuss this alternative. If the conclusions did not converge, or if they indicated a technical error in this alternative, the natural decision would be to remove the item from the test and to ask the judges to reconsider their subjective probabilities having excluded this item.

Discussion

The systematic trend in the results in Tables 2 through 5 is more consistent with behavior for the incorrect than for the correct alternatives of the items. This trend seems to hold for nearly each judge (the only clear exception is Judge 5). The fact that this trend holds for the standardized consistency indices as well as for the residuals seems to exclude explanations based on differences in response probabilities for examinees performing at the cutoff scores θ_{cj}. As a tentative explanation, it is suggested that correct alternatives are

more difficult to comprehend than incorrect alternatives and that therefore the judges were less capable of specifying probabilities of success on items.

It is not known if this trend generalizes to other content domains. If so, an interesting practical conclusion would be to calculate the standards θ_{cj} using probabilities on the incorrect rather than the correct alternatives. This can be done using a version of Equation 5 in which the left- and right-hand side sums are defined over the subset with the most consistent incorrect alternatives. The cutoff score on the number-correct scale would then follow from the one on the θ scale via the right-hand side of the current version of Equation 5. In fact, this calculation seems to imply a continuous version of the Nedelsky technique.

The IDEA method allows us to make the decision as to what probabilities to use in the definition of the sums in Equation 5 post hoc, that is, after all the probabilities have been obtained and it is known on which alternative the judges have operated most consistently. Another advantage of the IDEA method requirement is that judges specify probabilities for all alternatives on the items.

Concluding Summary

A new standard-setting method was derived from the operations of the Nedelsky and Angoff methods. The method was based on the assumption that examinees with a borderline competency evaluate all response options on an MC item against each other. Therefore, the method forces judges to focus on the process by which examinees choose among the response alternatives of MC items. It was hypothesized that the cutoff score set by the method should tend to be higher than that by the Nedelsky method because the judges are forced to inspect the correct alternative. At the same time, the cutoff score should tend to be lower than that by the Angoff method because the judges are required to evaluate the effectiveness of the distractors. The prediction was confirmed in a series of three standard-setting experiments with different judges and with tests from different domains.

In addition, an earlier procedure to assess intrajudge inconsistency was generalized to the polytomous response format used in the IDEA method. The method was used to check the data from one experiment for possible inconsistencies due to the behavior of the judge, the features of the items, or the features of the alternatives. It was found that the method led to judgments that were generally consistent, with the exception of two items. Another finding was that the probabilities were generally more consistent for the incorrect than for the correct alternatives. This finding suggests that the additional requirement to specify probabilities for incorrect alternatives does not make the method necessarily more inconsistent than the Angoff method. The finding also suggests postponing the decision to calculate the cutoff score from

the correct or the incorrect alternatives until it is known which choice would lead to the most consistent cutoff score.

One potential problem with the IDEA method is that it seems to take more time than the Nedelsky method, which has already been criticized for being time-consuming (Smith & Smith, 1988). However, for longer tests, it is possible to reduce the workload of judges by using a multiple-matrix sampling approach by assigning samples of judges from a well-defined population to samples of items. A variant that would allow us to profit maximally from the judges' knowledge is to block items according to their content specifications, group the judges according to their professional backgrounds, and assign random subsets of items from different blocks to groups of judges. If this implementation of the IDEA method is combined with the procedure to detect intrajudge inconsistency in this article, we can also aggregate the results over items with different content specifications and judges with different backgrounds and get more specific information about possible sources of inconsistency.

References

Baron, J. B., Rindone, D. A., & Prowda, P. (1981, April). *Will the "real" proficiency standard please stand up?* Paper presented at the annual meeting of the New England Educational Research Organization, Lenox, MA.

Behuniak, P., Jr., Archambault, F. X., & Gable, R. K. (1982). Angoff and Nedelsky standard setting procedures: Implications for the validity of proficiency test score interpretation. *Educational and Psychological Measurement*, *42*, 247-255.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bock, R. D. (1997). The nominal categories model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 33-49). New York: Springer.

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

Chang, L. (1999). Judgmental item analysis of the Nedelsky and Angoff standard-setting methods. *Applied Measurement in Education*, *12*, 151-165.

Chang, L., Dziuban, C. D., Hynes, M. C., & Olson, A. H. (1996). Does a standard reflect minimal competency of examinees or judge competency? *Applied Measurement in Education*, *9*, 151-160.

Cross, L. H., Impara, J. C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the national teacher examinations. *Journal of Educational Measurement*, *21*, 113-129.

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

Halpin, G., Sigmon, G., & Halpin, G. (1983). Minimum competency standards set by three divergent groups of raters using three judgmental procedures: Implications for validity. *Educational and Psychological Measurement*, *43*, 185-197.

Hambleton, R. K. (1978). On the use of cut-off scores with criterion-referenced tests in instructional settings. *Journal of Educational Measurement*, *15*, 227-290.

Harasym, P. H. (1981). A comparison of the Nedelsky and modified Angoff standard setting procedure on evaluation outcome. *Educational and Psychological Measurement*, *41*, 725-735.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485-514). New York: Macmillan.

Kane, M. (1998). Choosing between examinee-centered and test-centered standard-setting methods. *Educational Assessment*, *5*, 129-145.

Kassirer, J. P., & Kopelman, R. I. (1989). Cognitive errors in diagnosis: Instantiation, classification, and consequence. *American Journal of Medicine*, *86*, 433-441.

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

Maguire, T., Skakun, E., & Harley, C. (1992). Setting standards for multiple-choice items in clinical reasoning. *Evaluation and the Health Professions*, *15*, 434-452.

Mokken, R. L. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351-367). New York: Springer.

Molenaar, I. W., & Sijtsma, K. (2000). *MSP 5 for Windows: A program for Mokken scale analysis for polytomous items*. Groningen, the Netherlands: iecProGAMMA.

Paiva, R. E. A., & Vu, N. V. (1979, April). *Standards for acceptable level of performance in an objectives-based medical curriculum: A case study.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Poggio, J. P., Glasnapp, D. R., & Eros, D. S. (1981, April). *An empirical investigation of the Angoff, Ebel and Nedelsky standard setting methods*. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

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

Ramsden, P., Whelan, G., & Cooper, D. (1989). Some phenomena of medical students' diagnostic problem solving. *Medical Education*, *23*, 108-117.

Rock, D. A., Davis, E. L., & Werts, C. (1980, June). *An empirical comparison of judgmental approaches to standard setting procedures*. Research report of the Educational Testing Service, Princeton, NJ.

Shepard, L. A. (1995). *Implications for standard setting of the National Academy of Education Evaluation of the National Assessment of Educational Progress Achievement Levels*. Washington, DC: Proceedings of the Joint Conference on Standard Setting for Large Scale Assessments, National Assessment Governing Board, National Center for Educational Statistics.

Smith, R. L., & Smith, J. K. (1988). Differential use of item information by judges using Angoff and Nedelsky procedures. *Journal of Educational Measurement*, *25*, 259-274.

Thissen, D. (1991). *Multilog user's guide*. Chicago: Scientific Software.

Thissen, D., & Steinberg, L. (1984). A response model for multiple-choice items. *Psychomerika*, *49*, 501-519.

Thissen, D., & Steinberg, L. (1997). A response model for multiple-choice items. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 51-65). New York: Springer.

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