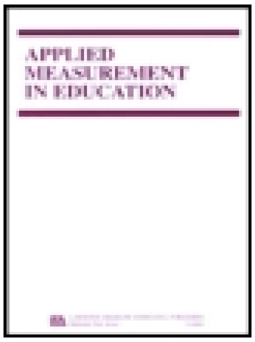

APPLIED MEASUREMENT IN EDUCATION

Applied Measurement in Education

Routledge
Taylor & Francis Group

ISSN: 0895-7347 (Print) 1532-4818 (Online) Journal homepage: https://www.tandfonline.com/loi/hame20

Judgmental Item Analysis of the Nedelsky and Angoff Standard-Setting Methods

Lei Chang

To cite this article: Lei Chang (1999) Judgmental Item Analysis of the Nedelsky and Angoff Standard-Setting Methods, Applied Measurement in Education, 12:2, 151-165, DOI: 10.1207/s15324818ame1202_3

To link to this article: https://doi.org/10.1207/s15324818ame1202_3

APPLIED MEASUREMENT IN EDUCATION, *12*(2), 151–165

Judgmental Item Analysis of the Nedelsky and Angoff Standard-Setting Methods

Lei Chang
Department of Educational Psychology
Chinese University of Hong Kong

It was hypothesized that the Nedelsky versus the Angoff methods would have (a) lower intrajudge inconsistency and (b) lower cutscores, especially for items presenting a challenge to the judges. These hypotheses were tested and supported in 3 standard-setting studies. These studies used 80 graduate students in education as judges to set standards for exams of a research method course they were taking. Lower intrajudge inconsistency of the Nedelsky method is attributed to focusing on response options and making multiple decisions. The strengths of the Nedelsky method, however, are limited by its discrete judgmental estimates. It is suggested that combining the strong features of both the Angoff and Nedelsky methods would make a stronger standard-setting procedure.

In reviewing the standard-setting literature, I identified 40 comparisons between the Nedelsky (1954) and Angoff (1971) methods reported in 10 studies (Baron, Rindone, & Prowda, 1981; Behuniak, Archambault, & Gable, 1982; Brennan & Lockwood, 1984; Cross, Impara, Frary, & Jaeger, 1984; Halpin, Sigmon, & Halpin, 1983; Harasym, 1981; Livingston & Zieky, 1989; Poggio, Glasnapp, & Eros, 1981; Rock, Davis, & Werts, 1980; Smith & Smith, 1988). In only 8 out of the 40 comparisons did the Nedelsky method produce a higher cutscore. In the other 80% of the comparisons, the Angoff cutscore was higher than the Nedelsky cutscore. Why is there such a consistent difference between the two methods? Which cutscore is more adequate? In this study, I sought to explore these questions

Requests for reprints should be sent to Lei Chang, Department of Educational Psychology, Chinese University of Hong Kong, Shatin, N. T., Hong Kong. E-mail: leichang@cuhk.edu.hk

by conducting a more conclusive comparison between these two most widely used standard-setting methods (Smith & Smith, 1988).

WHICH METHOD IS MORE ADEQUATE?

The adequacy of a standard-setting method depends on two processes—whether the judges adequately conceptualize the (minimum) competency of target examinees and whether judges adequately estimate item difficulty based on their conceptualized examinee competency. Jaeger (1991) called the latter process "judgmental item analysis" (p. 3). The Nedelsky and Angoff methods differ in the technique for judgmental item analysis, whereas the two methods have the same requirements in training judges to conceptualize examinee competency. Assuming a constant training effect on both methods, the superiority of one method over the other depends on which judgmental item analysis renders item difficulty estimates that are more consistent with the actual performance of target examinees. The lack of judgmental consistency is referred to as intrajudge inconsistency. One possible contributor to intrajudge inconsistency is the information judges use during judgmental item analysis.

Extending the research on decision making, Smith and Smith (1988) investigated the difference between the Nedelsky and Angoff methods in terms of the information judges used in making judgments. They tested and supported the hypothesis that Nedelsky judges made use of response options almost exclusively as salient information in making decisions whereas the Angoff judges used various other sources of information. This finding underscores the procedural difference between the two methods—although both methods require judges to evaluate an item including its response options, the mechanism of the Nedelsky method is such that an item difficulty estimate will not materialize without studying the response options, whereas there is no similar mechanism to ensure that response options are fully utilized in arriving at an Angoff estimate. In a multiple-choice item, similarity and plausibility of response options represent primary factors contributing to the difficulty or easiness of the item. The Nedelsky judges use response options in estimating item difficulty. The Angoff method has less control over the use of response options but more latitude in using other information, some of which may not be relevant (Smith & Smith, 1988).

The Nedelsky judgmental estimate of item difficulty involves multiple estimates, that is, making judgments regarding each of the response alternatives. The final Nedelsky estimate is, practically, an average of as many such error-prone judgments as there are distractors. Averaging serves to reduce error. Intrajudge inconsistency can thus be reduced in the Nedelsky method by averaging over the multiple estimates. In the Angoff method, one decision is made for each item, leaving no room to balance out the error associated with the decision. Thus, the

Nedelsky method provides another mechanism to reduce judgmental inconsistency. By using consistent information and by making multiple decisions, the Nedelsky method was hypothesized to have lower intrajudge inconsistency.

WHY IS THE ANGOFF CUTSCORE HIGHER?

The difference between the Nedelsky and Angoff cutscores could be due to a differential influence of judges' item-related knowledge during judgmental item analysis. The impact of judges' subject matter knowledge on standard setting has been shown in many studies (Busch & Jaeger, 1990; Chang, Dzuiban, Hynes, & Olson, 1996; Cross, Frary, Kelly, Small, & Impara, 1985; Jaeger, 1982; Pavia & Vu, 1979). For example, Chang et al. found that judges tended to set high standards for items they answered correctly and low standards for items they answered incorrectly. Pavia and Vu observed that Nedelsky judges had difficulties in separating their own difficulty with items from rendering judgments on these items.

As a Nedelsky judge evaluates each response alternative of an item, her or his knowledge underlying the item will be probed more than that of an Angoff judge who may or may not review each of the response options equally closely. Cognitively, an Angoff decision is driven primarily by confirming one correct answer whereas a Nedelsky decision is based on disproving multiple false alternatives. The judges' underlying knowledge has a higher chance of being tested and factored into a Nedelsky than an Angoff decision. If item-related knowledge indeed influences judges' ratings as has been shown in the literature, the Nedelsky method should produce lower cutscores than the Angoff method, especially for the items with which judges have difficulties. It was hypothesized that a Nedelsky cutscore was lower than an Angoff cutscore and the difference was larger for judges who answered the items incorrectly than judges who answered the items correctly.

These hypotheses were investigated in this study by making three standard-setting comparisons between the Nedelsky and Angoff methods. Because item difficulty estimates based on minimally competent examinees are rarely available (Plake, Melican, & Mills, 1991), this study had judges estimate item difficulty of average performance level (APL), instead of minimum performance level (MPL), so that item difficulty estimates for the entire population could be used as an adequate criterion to evaluate the APLs yielded by the two contrasting methods. Although setting an APL did not exactly represent the usual standard-setting objective, this approach did not change the procedural features of the two standard-setting methods that were being compared. More important, this approach ensures the availability of an adequate objective criterion the lack of which makes most of the existing standard-setting comparisons inconclusive (Kane, 1994).

METHOD

Judges

Judges in this study were 80 master and doctoral education majors from a metropolitan university in the United States. They were enrolled in four sections of a research method course. Sample 1 consisted of 21 students enrolled in one section of the course. Sample 2 contained 39 students enrolled in two other sections of the course. Sample 3 consisted of 20 students enrolled in the fourth section of the course. The great majority of these students were school teachers pursuing their degrees part-time. These students were taught the Angoff and Nedelsky standard-setting methods in this course. They practiced these two methods on 10 items taken from their midterm exam. In this exercise, the students were asked to provide judgmental item difficulty estimates for APL but not for MPL. Specifically, they were asked to estimate the probability that an average student in this course could get an item right for the Angoff and whether an average student could successfully eliminate each of the three alternatives for the Nedelsky method. APLs based on 167 past examinees on these 10 items as well as the mean item estimates of the class were later provided to the students. They were asked to adjust their original estimates based on the empirical item information and the class estimates. This exercise was intended to prepare the student judges for the actual standard setting reported in this study.

Items

Items were taken from the midterm and final exams of this course. All items were of the four-option multiple-choice format. None of the items had "none of the above" or "all of the above" as response options. Items taken from the midterm had performance data on 167 past examinees excluding the 80 students participating in this standard-setting study. Items from the final exam had performance data on 345 past students excluding participants in this study. Both the midterm and final exam items have high internal consistency reliability and adequate validity evidence (see Chang, 1996). Item difficulties estimated from the past examinees were used as empirical values to evaluate intrajudge inconsistency for both the Angoff and Nedelsky methods.

Procedure

Student judges in Sample 1 were asked to provide both the Angoff and Nedelsky ratings on nine items taken from their final exams. Right after turning in his or her final exam, a student was given the nine items from the exam on a separate sheet of paper with the correct answers marked. The student was asked to apply the Angoff procedure on these items first. Instead of estimating item difficulty for minimum competency or MPL, they were instructed to estimate the probability that an average student in this course would correctly answer each of the nine items (APL). After turning in the Angoff estimates, the student judge was given another sheet of paper containing the nine items with the correct responses marked. The student was instructed to cross out the false responses he or she thought an average student in this course would be able to eliminate. Students were not asked to calculate the Nedelsky estimate of item difficulty.

In Sample 2, the 39 students were first matched by their midterm scores in this course. Two students that were closest in their test scores were made into a pair, even though, in some cases, the pair could be several score points apart. Matched pairs were randomly assigned to one of the two standard-setting conditions, resulting in 20 and 19 students assigned to the Angoff and Nedelsky method, respectively. Twenty-five items taken from the midterm were distributed to the students in class 1 week after the exam on a separate sheet of paper with correct answers marked. The students were given the same standard-setting instructions described in Sample 1. Nineteen matched pairs were used in the analysis, eliminating one Angoff judge who did not have a matched Nedelsky counterpart.

In the same manner, the 20 students in Sample 3 were matched by their midterm scores before being randomly assigned to one of the two standard-setting methods. There were 10 matched pairs. They were asked to provide APL estimates on 18 items taken from their final exam using either the Angoff or Nedelsky method. As in Sample 1, each student rated the 18 items on a separate sheet of paper right after turning in the exam.

Intrajudge Inconsistency

Intrajudge inconsistency was defined as the average absolute deviation of a judge's item difficulty estimates from the items' empirical difficulty estimates:

$$\bar{d}_j = \sum\nolimits_i \left|P_{ij} - P_{ie}\right| / n_i,$$

where p_{ij} is a judge's item difficulty estimate, which in this study, was that of APL; p_{ie} is the item's empirical probability value, which in this study, was based on past examinees; and n_i is the number of items.

This operational definition of intrajudge inconsistency was the same as van der Linden's (1982) definition with two exceptions. First, van der Linden used item response theory-derived item probabilities whereas this study used empirical item probabilities computed from past examinees. Second, this definition produced an average absolute value of inconsistency whereas van der Linden computed a ratio-like index of intrajudge consistency that, ranging from 0 to 1, represents the degree to which the computed average absolute deviation differs from its maximum possible value.

The aforementioned definition of intrajudge inconsistency was modified for the Nedelsky procedure to account for the loss of consistency due to the discrete nature of the Nedelsky method (van der Linden, 1982). For the four-choice items that were used in this study, the Nedelsky method produces four fixed values of judgment—.25, .33, .50, and 1.0. Even though a Nedelsky judge has a perfect estimate of MPL (or APL used in this study) on this four-choice item, and thus, zero intrajudge inconsistency, there will be a discrepancy between the judge's rating and the actual performance level unless the actual performance level on this four-choice item falls on one of the four fixed values. For example, if the actual performance level (probability for answering the item correctly) is .30, then the minimum discrepancy between a Nedelsky estimate and this performance level is $.33 - .30 = .03$, which is due to the discreteness of the Nedelsky method. In this example, the perfect judge cannot achieve zero inconsistency on this item. The best this perfect judge can do is .03 inconsistency, which is not attributable to the judge's misconception (because the judge is perfect) but to the technical nature of the Nedelsky method. Thus, in calculating intrajudge inconsistency for the Nedelsky method, this minimum inconsistency due to the discreteness of the method is deducted from the "true" inconsistency attributable to misjudgment. In the aforementioned example, the intrajudge inconsistency for the perfect judge would be adjusted to zero as it should be. This adjustment was first introduced by van der Linden (1982).

RESULTS

Estimates of intrajudge inconsistency were computed for each Angoff and Nedelsky judge. To test the first hypothesis that the Nedelsky method produced smaller intrajudge inconsistency than the Angoff method, significance tests were conducted that compared the mean intrajudge inconsistency of the Angoff judges against that of the Nedelsky judges. In the first sample where the same 21 judges used both the Angoff and Nedelsky methods, a dependent t test showed that the Angoff method had a significantly higher mean intrajudge inconsistency ($M = .249$, $SD = .061$) than the Nedelsky method ($M = .129$, $SD = .045$), $t(20) = 8.31$, $p < .01$. Dependent t tests were also used for the other two samples where matched pairs were randomly assigned to the two methods. In Sample 2, the mean intrajudge inconsistency based on 19 Angoff judges ($M = .211$, $SD = .065$) was significantly higher than that ($M = .103$, $SD = .043$) from the 19 Nedelsky judges, $t(18) = 5.49$, $p < .01$. In Sample 3, where 10 matched pairs were used for the Angoff and Nedelsky method, respectively, mean intrajudge inconsistency was .20 ($SD = .043$) for the Angoff method and .139 ($SD = .048$) for the Nedelsky method. The t test was again significant, $t(9) = 3.04, p < .05$. These results supported the first hypothesis that the Nedelsky method had lower intrajudge inconsistency.

Further support of this hypothesis was obtained by comparing the Angoff and Nedelsky cutscores with actual performance levels based on the past examinees. The average of the item difficulty estimates on 345 past examinees was 0.58 for the 9 final-exam items used in Sample 1 and .60 for a different set of 18 items used in Sample 3. The Nedelsky cutscores in these two samples were .57 and .58, which were very close to the past performance levels. The Angoff cutscores (.71 for Sample 1 and .72 for Sample 3), however, deviated widely from the empirical values. For Sample 2, the actual performance level of the 25 midterm items based on 167 past examinees was .60. The Angoff and Nedelsky cutscores were .76 and .59, respectively. The Nedelsky cutscore was almost identical to the empirical value, whereas the Angoff cutscore had a large deviation from it.

The Nedelsky cutscores were also very similar to the averages of the judge-based item difficulty estimates in Sample 2 ($M = .61$; Nedelsky cutscore = .59) and Sample 3 ($M = .60$; Nedelsky cutscore = .58), but not in Sample 1 ($M = .70$, Nedelsky cutscore = .57). In Samples 2 and 3, these averages were also almost identical to those based on the past examinees. In Sample 1, however, the judge-based average was much higher than that derived from the past examinees. In this sample, the Nedelsky cutscore was much lower than this judge-based average whereas the Angoff cutscore was very close to it. This finding was somewhat unexpected.

The second hypothesis that the Angoff cutscore was higher than the Nedelsky cutscore was tested and supported in all three samples. In Sample 1, Angoff cutscore was .71, Nedelsky cutscore was .57, $t(20) = 5.44, p < .01$. In Sample 2, Angoff cutscore was .76, Nedelsky cutscore was .59; $t(18) = 5.64, p < .01$. In Sample 3, Angoff cutscore was .72, Nedelsky cutscore was .58; $t(9) = 2.65, p < .05$.

Item analysis also bore out the second hypothesis that the Angoff method produced higher item difficulty estimates than the Nedelsky method. Angoff APL estimates were higher than the Nedelsky estimates on 51 out of the 52 items used in the three samples. Table 1 contains the Angoff and Nedelsky item difficulty estimates and their standard deviations as well as two empirical item difficulty estimates, one based on the participating student judges and one based on past examinees.

A closer look at Table 1 shows larger differences between the Angoff and Nedelsky estimates for items having lower judge-based difficulty values. These results lend support to the third hypothesis that there were larger differences between the Angoff and Nedelsky methods (Nedelsky having lower probability estimates) for more difficult items. To further test this hypothesis, the APL estimates were divided into those from the Angoff and Nedelsky judges who answered the items correctly and those from the judges who answered the items incorrectly. These results are also contained in Table 1. The mean difference between the

TABLE 1
Comparisons of the Nedelsky and Angoff Item Difficulty Estimates

							Judges Who Answered							
			All Judges				*Correctly*				*Incorrectly*			
			Angoff		*Nedelsky*		*Angoff*		*Nedelsky*		*Angoff*		*Nedelsky*	
Item	*P1*	*P2*	*APL*	*SD*	*APL*	*SD*	*APL*	*SD*	*APL*	*SD*	*APL*	*SD*	*APL*	*SD*
Sample 1, 21 judges														
1	.76	.53	.73	.21	.67	.30	.77	.20	.77	.27	.60	.20	.37	.13
2	1.00	.77	.80	.15	.85	.26	.80	.15	.85	.26				
3	.81	.61	.68	.21	.58	.26	.66	.21	.65	.23	.76	.19	.29	.05
4	.81	.66	.67	.22	.54	.28	.72	.21	.60	.28	.49	.16	.31	.04
5	.48	.36	.65	.22	.44	.25	.67	.18	.62	.27	.62	.26	.29	.08
6	.29	.11	.74	.17	.33	.10	.62	.21	.43	.11	.79	.12	.29	.07
7	.95	.83	.73	.19	.57	.26	.73	.19	.58	.26	.80	.00	.25	.00
8	.57	.66	.69	.22	.56	.33	.78	.22	.76	.31	.57	.15	.29	.08
9	.67	.78	.68	.25	.61	.29	.72	.27	.75	.26	.61	.21	.34	.11
Sample 2, 39 judges														
1	.11	.49	.68	.14	.41	.10	.90	.00	.50	.00	.66	.12	.41	.11
2	.53	.30	.79	.16	.43	.11	.76	.19	.46	.09	.81	.12	.40	.12
3	.63	.81	.76	.21	.71	.29	.87	.11	.83	.24	.58	.23	.50	.24
4	.66	.62	.82	.14	.66	.27	.83	.13	.79	.25	.81	.16	.44	.11
5	.84	.57	.79	.18	.69	.27	.84	.15	.77	.25	.55	.15	.33	.00
6	.63	.64	.71	.19	.61	.24	.73	.16	.69	.25	.67	.23	.43	.11
7	.55	.56	.75	.21	.61	.28	.81	.16	.77	.26	.69	.25	.40	.12
8	.71	.64	.81	.18	.69	.27	.87	.12	.76	.27	.71	.23	.46	.09
9	.66	.57	.84	.09	.59	.26	.86	.06	.67	.25	.79	.12	.46	.25
10	.63	.61	.81	.13	.66	.30	.83	.11	.75	.29	.79	.19	.55	.29

11	.63	.58	.79	.14	.67	.26	.74	.13	.79	.26	.88	.08	.46	.09
12	.92	.56	.80	.16	.65	.27	.78	.16	.68	.27	.92	.11	.33	.00
13	.58	.60	.75	.19	.53	.27	.73	.19	.60	.27	.78	.20	.44	.26
14	.66	.46	.77	.20	.51	.19	.81	.15	.57	.18	.74	.27	.35	.09
15	.63	.66	.83	.19	.65	.25	.76	.22	.73	.25	.92	.11	.47	.07
16	.71	.66	.71	.18	.56	.20	.81	.12	.57	.18	.50	.10	.53	.28
17	.53	.72	.83	.13	.65	.23	.82	.12	.75	.26	.84	.15	.55	.15
18	.71	.74	.82	.15	.68	.28	.84	.13	.81	.25	.77	.21	.42	.13
19	.63	.66	.77	.14	.55	.23	.73	.09	.65	.24	.82	.17	.32	.10
20	.21	.42	.55	.23	.40	.23	.61	.21	.75	.29	.54	.24	.31	.09
21	.63	.57	.75	.16	.52	.23	.76	.13	.64	.23	.74	.23	.35	.10
22	.61	.67	.75	.15	.53	.22	.76	.14	.60	.25	.73	.17	.43	.12
23	.55	.54	.67	.17	.54	.22	.69	.19	.62	.25	.64	.15	.43	.10
24	.66	.66	.79	.14	.58	.19	.77	.12	.58	.19	.82	.17	.58	.20
25	.66	.66	.75	.20	.66	.30	.78	.16	.78	.29	.70	.24	.42	.09
Sample 3, 20 judges														
1	.70	.53	.72	.12	.59	.29	.70	.13	.71	.27	.77	.07	.30	.05
2	.80	.77	.77	.17	.59	.29	.81	.13	.67	.28	.58	.24	.29	.06
3	.65	.61	.77	.13	.67	.29	.83	.13	.76	.30	.69	.10	.44	.10
4	.45	.59	.78	.14	.49	.20	.84	.07	.63	.25	.72	.18	.40	.11
5	.50	.71	.76	.16	.56	.31	.85	.17	.77	.33	.68	.11	.35	.09
6	.25	.36	.84	.11	.65	.31	.83	.04	.61	.35	.84	.12	.67	.32
7	.65	.66	.72	.17	.62	.33	.79	.14	.78	.34	.57	.12	.40	.13
8	.45	.36	.55	.15	.51	.28	.58	.14	.75	.29	.52	.18	.35	.12
9	.30	.11	.52	.13	.47	.20	.63	.18	.54	.32	.49	.11	.42	.09
10	.70	.81	.63	.24	.44	.22	.66	.26	.50	.26	.50	.00	.35	.11
11	.75	.56	.59	.22	.49	.20	.59	.24	.52	.23	.58	.25	.42	.14
12	.70	.58	.69	.19	.60	.29	.76	.08	.67	.32	.52	.27	.44	.10
13	.90	.83	.79	.15	.69	.33	.84	.11	.69	.33	.58	.10		
14	.75	.66	.83	.11	.67	.29	.85	.11	.76	.30	.75	.07	.44	.10
15	.85	.78	.82	.17	.67	.29	.91	.07	.67	.29	.60	.10		

(continued)

TABLE 1 (Continued)

							Judges Who Answered							
			All Judges				*Correctly*				*Incorrectly*			
			Angoff		*Nedelsky*		*Angoff*		*Nedelsky*		*Angoff*		*Nedelsky*	
Item	*P1*	*P2*	*APL*	*SD*	*APL*	*SD*	*APL*	*SD*	*APL*	*SD*	*APL*	*SD*	*APL*	*SD*
16	.60	.70	.72	.09	.61	.28	.80	.04	.67	.28	.66	.06	.38	.18
17	.50	.43	.75	.14	.57	.24	.84	.11	.70	.27	.65	.09	.45	.11
18	.60	.72	.74	.12	.58	.30	.81	.08	.75	.27	.64	.09	.33	.12

Note. P1 = item difficulty based on student judges who participated in the standard-setting study, 21 for Sample 1, 38 for Sample 2, and 20 for Sample 3; P2 = item difficulty based on 345 past examinees for Samples 1 and 3 and 169 past examinees for Sample 2; APL = Nedelsky or Angoff judgmental estimates of item difficulty of average performance level.

Angoff and Nedelsky estimates was .09 from judges answering the items correctly and .27 from judges answering the items incorrectly. The Angoff–Nedelsky difference was significantly larger from judges who answered the items incorrectly, $t = 8.5, p < .01$.

Table 1 also shows that the Nedelsky item estimates were associated with larger standard deviations than were the Angoff estimates. In other words, the Nedelsky method produced larger interjudge inconsistency than the Angoff method.

DISCUSSION

The Nedelsky cutscores were significantly lower than the Angoff cutscores. This finding is consistent with the literature. A more important finding is that the difference between the two methods changed as a function of judges' item-related knowledge; there was a much larger difference between the two kinds of judges when they failed to answer the items correctly. This finding confirms the influence of judges' content knowledge in standard setting. Logically, judges set higher standards for items for which they possess the underlying knowledge and set lower standards for items when they lack the underlying knowledge. This influence seems to be more pronounced in the Nedelsky procedure in which going through response alternatives is likely to subject judges' item-related knowledge to a more direct test. Although Angoff judges are instructed to evaluate an item including its response alternatives, there is no mechanism in this method to enforce this instruction. In their global estimation of item difficulty, the Angoff judges may not pay as close attention to response alternatives as do the Nedelsky judges who have to make a decision regarding each alternative. Thus, when an item presents a potential challenge to the judges, the experienced difficulty has a higher chance of being factored into a Nedelsky than an Angoff decision.

When judges had no difficulty with an item, the differences between the Angoff and Nedelsky estimates were reduced. However, the Nedelsky estimates were still lower than the Angoff estimates from judges who answered the items correctly possibly because plausibility of the alternatives adds to the difficulty of an item. The Nedelsky method, which is designed to tune in to the similarities of alternatives, is more likely to factor in this added item difficulty. As Burton (1978) and Gross (1982) pointed out, the Nedelsky method correctly addresses the fact that multiple-choice item difficulty is a function not only of the complexity of the tested concept but also of the plausibility of the distractors.

The lower intrajudge inconsistency of the Nedelsky method might be attributed to its focused attention on response alternatives. Plausibility of response alternatives contributes to the difficulty of an item (Gross, 1982), whereas the stem of an item appeared to be unrelated to item difficulty (Smith & Smith, 1988). By focusing on the alternatives, the Nedelsky method provides a more deterministic judgmental item analysis. The Angoff judgmental item analysis, on the other hand, lacks a clearly defined frame of structure (Shepard, 1995). Without a mechanism to enforce the use of response alternatives, the Angoff judges might pay more attention to the stem and the correct answer in estimating item difficulty. Similar explanations were also made by Meskauskas (1976) of the Ebel method, which shares similar judgmental item analysis as the Angoff method.

The low intrajudge inconsistency of the Nedelsky method might also be explained by the counter-balancing effect of multiple decision making. For a four-option item, the item difficulty estimate derived from the Nedelsky procedure is the sum of three judgments. Intrajudge inconsistency associated with the final estimate may be ameliorated to the extent that not all three judgments are incorrect. For the same item, the Angoff difficulty estimate represents one decision with one error that cannot be adjusted.

The Nedelsky cutscores were very consistent with both the past performance levels and the judge-based item difficulty estimates in two of the three samples. In Sample 1, however, the Nedelsky cutscore was much lower than the judge-based item difficulty average; although it was comparable with that of the past examinees. Also, across three samples, the Nedelsky method had larger standard deviations than the Angoff method. Both of these findings in particular and the finding of lower Nedelsky estimates in general may also be explained by the discreetness of the Nedelsky item difficulty estimation.

The Nedelsky method produces a fixed number of probability estimates. For a four-choice item, they are 0.25, 0.33, 0.5, and 1.0. These fixed numbers are unequally spaced with a large gap between 0.5 and 1.0. Unless a judge believes that examinees are able to eliminate all three false alternatives of an item, the difficulty estimate of the item will be .5 or lower. Thus, there is a "depression" effect due to the discrete nature of the Nedelsky method that is independent from the soundness of judgment. The judge-based item difficulty average was .70 in Sample 1 in contrast to .61 and .60 in the other two samples. If judges were making estimates according to their own experience with the items, the depression effect would be more pronounced in this sample where judges were more successful with the items. To some degree, the depression effect could account for the lower Nedelsky item difficulty estimates in all three samples, especially by judges who answered the items correctly. However, there were larger differences between the Nedelsky and Angoff estimates (Nedelsky being lower) from judges who answered the items incorrectly than those who answered the items correctly. Thus independent from the depression effect, the explanation would still hold that the experienced difficulty with an item has a higher chance of being factored into a Nedelsky than an Angoff estimation. These findings suggest that the Nedelsky method may be more appropriate for difficult tests ($\overline{p} < .5$) than easy tests ($\overline{p} > .5$).

The discreteness of the Nedelsky estimates may also contribute to the large standard deviations associated with the Nedelsky cutscores. In this study, discreteness was adjusted in the computation of intrajudge inconsistency but not from the calculation of standard deviation or interjudge inconsistency. Brennan and Lockwood (1980) and Melican, Mills, and Plake (1989) noted this technical influence on standard deviations. Large Nedelsky standard deviations are also observed in a large number of the studies reviewed earlier. Without such technical inconsistency, the Nedelsky method was expected to have more satisfactory interjudge inconsistency. On the other hand, the Nedelsky method was, in a way, given a comparative advantage in this study by the adjustment of its intrajudge inconsistency.

The assumption for the adjustment of intrajudge inconsistency is that judgmental inconsistency should be distinguished from technical inconsistency (van der Linden, 1982). In practice, however, the two inconsistencies are inseparable. Consequently, the superior judgmental consistency of the Nedelsky method, as was found in this study, is dissipated by the discreteness of its estimates. A solution is to modify the discreteness of the method while retaining its judgmental strengths—focusing on response options and making multiple decisions. Probably, the best way to change the discreteness is to adopt the Angoff continuous judgmental procedure that has been widely accepted in practice. Specifically, the Nedelsky method's dichotomous decision regarding an examinee's ability to eliminate a distractor can be changed into a continuous Angoff-like probability judgment. The final judgmental estimate of an item will be the sum of the probabilities of successfully eliminating each distractor (plus one for guessing) divided by the number of response options. This modification of the Nedelsky method will produce evenly spaced continuous item difficulty estimates. Similar modifications have been proposed by Reilly and Zink (1984). With continuous estimates, intrajudge and interjudge consistency are expected to improve. More important, the mechanism of focusing on response options is kept and the counter-balancing power of multiple decisions is maximized.

There are several limitations of this study. First, the reported standard setting was a classroom exercise. The training of the judges was much more limited than what is normally done in a formal standard-setting context. The lack of training is expected to have the same effect on the Nedelsky and Angoff judgments. However, in current Angoff studies, judges often are given two opportunities to make item estimates rather than one as was done in this study. Thus, the Angoff results of this study could likely be improved. Second, judges in this study estimated item difficulty for average performance rather than MPL. The findings may not be directly generalizable to formal standard setting of minimum competency. Finally, the lower Nedelsky cutscores and intrajudge inconsistency were attributed to focusing on response alternatives and multiple decision making. However, this study did not empirically manipulate these variables to test their causal contributions to the observed differences between the Nedelsky and Angoff methods. Like any nonexperimental research, this study suffers from the weakness of inferring, inversely, from the effect to its possible causes. Such inverse inference exists in almost all existing studies on standard setting that tried to explain preexisting differences between the contrasting methods. Future research should aim at experimentally testing the independent variables suggested by this study to determine their causal relations to judgmental estimation of minimum competency.

ACKNOWLEDGMENTS

Preparation of this article was supported by a Mainline Research Grant from the Chinese University of Hong Kong. Special thanks go to Wim van der Linden for his explanations on the Nedelsky adjustment of intrajudge inconsistency and to James Impara and four anonymous reviewers for their helpful comments on previous drafts of this article.

REFERENCES

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

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.

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.

Burton, N. W. (1978). Societal standards. *Journal of Educational Measurement, 15*, 263–271.

Busch, J. C., & Jaeger, R. M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations. *Journal of Educational Measurement, 27*, 145–163.

Chang, L. (1996). Quantitative Attitudes Questionnaire: Instrument development and validation. *Educational and Psychological Measurement, 56*, 1037–1042.

Chang, L., Dzuiban, C., Hynes, M., & Olson, A. (1996). Does a standard reflect minimal competency of examinees or judge competency? *Applied Measurement in Education, 9*, 161–173.

Cross, L. H., Frary, R. B., Kelly, P. P., Small, R. C., & Impara, J. C. (1985). Establishing minimum standards for essays: Blind versus informed reviews. *Journal of Educational Measurement, 22*, 137–146.

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.

Gross, L. J. (1982). Standards and criteria: A response to Glass' criticism of the Nedelsky technique. *Journal of Educational Measurement, 19*, 159–161.

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.

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. (1982). An interactive structured judgment process for establishing standards on competency tests: Theory and application. *Educational Evaluation and Policy Analysis, 4,* 461–475.

Jaeger, R. M. (1991). Selection of judges for standard-setting. *Educational Measurement: Issues and Practice, 10*(2), 3–6, 10.

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

Livingston, S. A., & Zieky, M. J. (1989). A comparative study of standard-setting methods. *Applied Measurement in Education, 2,* 121–141.

Melican, G. J., Mills, C. N., & Plake, B. S. (1989). Accuracy of item performance predictions based on the Nedelsky standard setting method. *Educational and Psychological Measurement, 49,* 467–478.

Meskauskas, J. A. (1976). Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. *Review of Educational Research, 46,* 133–158.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14,* 3–19.

Pavia, 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.

Plake, B. S., Melican, G. J., & Mills, C. N. (1991). Factors influencing intrajudge consistency during standard-setting. *Educational Measurement: Issues and Practice, 10*(2), 15–16, 22, 25.

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.

Reilly, R. R., & Zink, D. L. (1984). Comparison of direct and indirect methods for setting minimum passing scores. *Applied Psychological Measurement, 8,* 421–429.

Rock, D. A., Davis, E. L., & Werts, C. (1980, June). *An empirical comparison of judgmental approaches to standard setting procedures* (Research Rep. No. 80–7). Princeton, NJ: Educational Testing Service.

Shepard, L. A. (1995, October). *Implications for standard setting of the National Academy of Education Evaluation of the National Assessment of Educational Progress achievement levels.* Paper presented at the Joint Conference on Standard Setting for Large Scale Assessments, National Assessment Governing Board, National Center for Educational Statistics, Washington, DC.

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.

van der Linden, V. 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.