# A Comparison of the Nedelsky and Angoff Cutting Score Procedures Using Generalizability Theory

**Robert L. Brennan and Robert E. Lockwood**
**The American College Testing Program**

Nedelsky (1954) and Angoff (1971) have suggested procedures for establishing a cutting score based on raters' judgments about the likely performance of minimally competent examinees on each item in a test. In this paper generalizability theory is used to characterize and quantify expected variance in cutting scores resulting from each procedure. Experimental test data are used to illustrate this approach and to compare the two procedures. Consideration is also given to the impact of rater disagreement on some issues of measurement reliability or dependability. Results suggest that the differences between the Nedelsky and Angoff procedures may be of greater consequence than their apparent similarities. In particular, the restricted nature of the Nedelsky (inferred) probability scale may constitute a basis for seriously questioning the applicability of this procedure in certain contexts.

Currently there is considerable debate concerning procedures for setting passing standards, or cutting scores, when scores on tests are used to make certain types of decisions (see, for example, National Council on Measurement in Education, 1978). Meskauskas (1976), Buck (1977), and Zieky and Livingston (1977), among others, have reviewed some current procedures for establishing cutting scores. For the most part these procedures can be grouped into two categories—procedures that use raters' subjective judgments and procedures that use examinee scores on the test itself and/or some criterion measure. The latter category of procedures is not discussed in this paper; rather, this paper examines two procedures suggested by Nedelsky (1954) and Angoff (1971) for establishing cutting scores based upon raters' judgments.

## Nedelsky and Angoff Procedures

Both of these procedures require judgments by raters concerning the performance of hypothetical "minimally competent" examinees on *each* item of a test. Using Nedelsky's procedure, raters are asked to identify, for each item, those distractors that a minimally competent examinee would eliminate as incorrect. The reciprocal of the number of remaining alternatives (including the correct answer) serves as an estimate of the probability that a "minimally competent" examinee would get the item correct. In Angoff's procedure, raters simply provide an estimate of the item probabilities with-

*APPLIED PSYCHOLOGICAL MEASUREMENT*
*Vol. 4, No. 2 Spring 1980 pp. 219–240*
*© Copyright 1980 West Publishing Co.*

out specifically identifying which distractors a "minimally competent" examinee would eliminate. For both procedures the mean of the item probabilities, over items and raters, is defined as the cutting score for the test (in terms of proportion of items correct). Notationally, throughout this paper, $\overline{X}$ is used to denote the cutting score, or more specifically, the mean cutting score that results from a particular study. For a particular rater, *r,* the mean of that rater's item probabilities will be denoted $\overline{X}_r$, which can be interpreted as the cutting score that would be assigned by that particular rater.

## Issues and Approach

The Nedelsky and Angoff procedures are appealing in many contexts because they are understandable to raters and test users; and these procedures force raters to give detailed consideration to the specific content of a test, rather than to its general characteristics. However, the validity and practical utility of these approaches, and similar approaches, for practical decision making may rest heavily upon the extent to which raters agree in their judgments. This concern for rater agreement has received very little attention in the context of establishing cutting scores, although Andrew and Hecht (1976) do address some aspects of this issue.

The principal purposes of this paper are (1) to consider a specific psychometric approach for characterizing and quantifying the magnitude of error variances (in either cutting score procedure) attributable to disagreement evident in rater judgments; (2) to illustrate this approach with experimental data; (3) to compare the Angoff and Nedelsky procedures; and (4) to examine the impact of rater disagreement on some issues relating to the reliability or dependability of measurement.

The principal psychometric approach employed here to address these issues is based upon generalizability theory, which is most completely explicated by Cronbach, Gleser, Nanda, and Rajaratnam (1972). In this paper concepts and equations from generalizability theory are used and explained as needed, but most results are not proven. Readers desiring more detail are referred to Cronbach et al. (1972) and/or Brennan (1977). It should be noted that there are many aspects of generalizability theory that are of little concern in this paper. For example, generalizability coefficients per se are not discussed. Indeed, the approach used here is essentially variance components analysis viewed from the perspective of generalizability theory. Nevertheless, for the purposes of this paper, generalizability theory is especially appropriate because it allows differentiation among multiple sources of error relevant to the cutting score procedures under consideration, and because generalizability theory considerably facilitates consideration of the impact of rater disagreement on measurement dependability. However, some of the issues treated here could be addressed using results from multiple matrix sampling theory (see Sirotnik & Wellington, 1977, p. 354) or results discussed in texts on mathematical statistics (e.g., Wilks, 1962, secs. 8.6 and 10.8).

Both the Nedelsky and Angoff procedures necessitate judgments about "minimum competence." In one section of this paper, consideration is given to aspects of how the two procedures allow a rater to operationalize some conception of minimum competence; otherwise, however, this paper is not intended to treat educational, philosophical, or psychological issues associated with defining minimum competence. The authors also recognize that in realistic settings evaluators sometimes use more than one cutting score procedure or a variant of one of the procedures discussed here, but this paper does not address such issues in detail.

## Angoff Procedure

For the Angoff procedure the probability assigned by rater *r* to item *i* can be represented as

$$X_{ri} = \lambda + \lambda_{r}\sim + \lambda_{i}\sim + \lambda_{ri}\sim \; ; \qquad [1]$$

where

$\lambda$ = grand mean for the population of raters and the universe of items,
$\lambda_{r}{\sim}$ = effect for rater *r*,
$\lambda_{i}{\sim}$ = effect for item *i*, and
$\lambda_{ri}{\sim}$ = effect for the interaction of rater *r* and item *i*.

(Technically, since there is only one observation for each rater-item combination, the effect $\lambda_{ri}{\sim}$ is completely confounded with any other sources of variation, sometimes called "residual" error.) Here, unless otherwise noted, it will be assumed that the actual raters in the study can be considered a random sample from an essentially infinite population of raters and that the actual items can be considered a random sample from an essentially infinite universe of items. Under this assumption, and assuming independent effects that sum to zero, Equation 1 represents what is usually called a random effects model for the *r*×*i* design.

Given this model, for rater *r* the expected probability over the universe of items is

$$\lambda_r = \lambda + \lambda_r{\sim} ; \qquad [2]$$

whereas the average probability over the sample of $n_i$ items is $\bar{X}_r$. Similarly, for item *i*, the expected probability over the population of raters is

$$\lambda_i = \lambda + \lambda_i{\sim} ; \qquad [3]$$

and the average probability over the sample of $n_r$ raters is $\bar{X}_i$.

## Sample Statistics

Table 1 reports means, standard deviations, and intercorrelations for five raters who independently applied the Angoff procedure to 126 four-alternative items in a health-related area. Also in-

Table 1
Means, Standard Deviations, and Intercorrelations
Among Raters for the Angoff Procedure

| | Raters | | | | | Mean over items | S.D. over items |
|---|---|---|---|---|---|---|---|
| Raters | 2 | 3 | 4 | 5 | c[a] | $\bar{X}_r$ | $\hat{\sigma}_i(X_{ri})$ |
| 1 | .525 | .053 | .046 | .150 | .731 | .671 | .203 |
| 2 | | .171 | .206 | .382 | .744 | .719 | .161 |
| 3 | | | .161 | -.036 | .237 | .656 | .119 |
| 4 | | | | .209 | .217 | .617 | .187 |
| 5 | | | | | .432 | .653 | .218 |
| c[a] | | | | | | .698 | .154 |

$\bar{X}$ = .663 (83.56)[b] $\hat{\sigma}(\bar{X}_r)$ = .037 (4.70)[b]

[a] c is reconciled rating arrived at by the raters themselves.
[b] Numbers within parentheses are expressed in terms of number of items.

cluded in Table 1 are sample statistics for a reconciled rating, or consensus judgment for each item, agreed upon by the raters after they independently employed the Angoff procedure. Each of the five raters was a practitioner or teacher in the appropriate field. However, the actual test was not a minimum competency test. Also, these data, and all data discussed subsequently, were collected only to study cutting score procedures per se, not to obtain cutting scores for operational use. Therefore, data reported in this paper are properly viewed as resulting from experimental use of cutting score procedures.

In Table 1 and subsequent tables, all results except those within parentheses are in terms of probabilities or proportions. Results within parentheses are in terms of number of items. For example, for these Angoff-type data the mean probability over $n_r = 5$ raters and $n_i = 126$ items is $\overline{X} = .663$, which is the (mean) cutting score, in terms of proportion of items correct. In terms of number of items correct, the (mean) cutting score is $n_i\overline{X}$, or 83.56.

The sample statistics reported in Table 1 will be examined in more detail later. Here, simply note that Table 1 suggests that there is some degree of variability among rater means, as reflected by $\hat{\sigma}(\overline{X}_r)$; there is some degree of variability within each rater, as reflected by $\hat{\sigma}_i(\overline{X}_{ri})$; and there is some degree of variability in the rater intercorrelations. The sample statistics in Table 1, however, do not indicate clearly the variability in the mean cutting score, $\overline{X}$, which is a principal concern of this paper. In other words, it is desired to estimate the variance (or standard deviation) of $\overline{X}$ if the entire study were replicated with different samples of raters and/or items.

## Estimates of Variability for Cutting Scores

The usual estimates of the variances associated with each of the random effects in Equation 1 are

$$\hat{\sigma}^2(r) = [MS(r) - MS(ri)]/n_i \;; \qquad [4]$$

$$\hat{\sigma}^2(i) = [MS(i) - MS(ri)]/n_r \;; \text{ and} \qquad [5]$$

$$\hat{\sigma}^2(ri) = MS(ri) \;. \qquad [6]$$

These are called estimated random effects variance components. For, example, $\hat{\sigma}^2(r)$ is an unbiased estimate of the variance of $\lambda_r$ (or $\lambda_{r\sim}$) over the population of raters. Similarly, $\hat{\sigma}^2(i)$ is an unbiased estimate of the variance of $\lambda_i$ (or $\lambda_{i\sim}$) over the universe of items.

It is important that $\hat{\sigma}^2\,(r)$ be differentiated from $\hat{\sigma}^2(\overline{X}_r)$. The former is an estimate of the variance, over the population of raters, of the scores (or probabilities) $\lambda_r$; the latter is the variance, over the sample of raters, of the scores (or probabilities) $\overline{X}_r$.

In terms of the random effects variance components

$$\hat{\sigma}^2(\overline{X}_r) = \hat{\sigma}^2(r) + \hat{\sigma}^2(ri)/n_i \;. \qquad [7]$$

In other words, the *observed* variance of rater means can be decomposed into two parts—one part that is uniquely associated with raters and another part that is associated with the interaction of raters and items.

In terms of the estimated variance components in Equations 4 to 6, there are three possible estimates for the variance of the mean cutting score, $\overline{X}$. These estimates differ in terms of the intended universe of generalization.

First, the expected value of the variance of $\bar{X}$ for generalizing over samples of $n_r$ raters *and* $n_i$ items is

$$\hat{\sigma}^2(\bar{X}) = \hat{\sigma}^2(r)/n_r + \hat{\sigma}^2(i)/n_i + \hat{\sigma}^2(ri)/n_r n_i \ . \quad [8]$$

Consider the possibility of determining $X$ a "very large" number of times, each time using a different sample of $n_r$ raters and $n_i$ items. Equation 8 estimates the variance of the distribution of this "very large" number of means. It is in this sense that $\hat{\sigma}^2(\bar{X})$ is an unbiased estimate of the variance of the mean for generalizing over both samples of raters and samples of items.

Second, the expected variance of $\bar{X}$ for generalizing over samples of $n_i$ items, for a fixed set of $n_r$ raters, is

$$\hat{\sigma}^2(\bar{X}|R) = \hat{\sigma}^2(i)/n_i + \hat{\sigma}^2(ri)/n_r n_i \ . \quad [9]$$

This variance is denoted $\hat{\sigma}^2(\bar{X}|R)$ to emphasize that raters are considered fixed. Again, consider the possibility of determining $X$ a "very large" number of times, each time using a different sample of $n_i$ items but the same $n_r$ raters. $\hat{\sigma}^2(\bar{X}|R)$ is an unbiased estimate of the variance of this distribution of means.

Third, when generalization is intended over samples of $n_r$ raters for a fixed set of $n_i$ items, the expected variance of $\bar{X}$ is

$$\hat{\sigma}^2(\bar{X}|I) = \hat{\sigma}^2(r)/n_r + \hat{\sigma}^2(ri)/n_r n_i \ . \quad [10]$$

Equations 8 to 10 provide three different estimates of error variance in the mean cutting score. Which of these estimates is appropriate can be determined only in the context of a specific study, i.e., it is the decision maker who must determine whether it is appropriate to generalize over samples of raters, items, or both. It is evident from Equations 8 to 10, however, that $\hat{\sigma}^2(\bar{X})$ must be at least as large as $\hat{\sigma}^2(\bar{X}|R)$ and $\hat{\sigma}^2(\bar{X}|I)$. This follows from the fact that $\hat{\sigma}^2(\bar{X}|R)$ does not involve variability due to raters, $\hat{\sigma}^2(r)$, and $\hat{\sigma}^2(\bar{X}|I)$ does not involve variability due to items, $\hat{\sigma}^2(i)$.

## Generalizability Results for Angoff Procedure

For the Angoff procedure, Table 2 reports the usual ANOVA results, estimated random effects variance components, and estimates of mean cutting score variability. (Brennan, 1979a, discusses a computer program that can be used to obtain all of these estimates using the observed data matrix as input.) It is usual in generalizability theory to report results in terms of variances; however, Table 2 also reports the three estimates of mean score variability in terms of standard deviations to facilitate interpretation. Note, for example, that in terms of proportion of items, the standard deviation of $\bar{X}$, for generalizing over raters and items, is .018; and in terms of number of items, it is 2.29. Furthermore, $\hat{\sigma}(\bar{X})$ and $\hat{\sigma}(\bar{X}|I)$ have approximately the same magnitude; both of them are almost twice as large as $\hat{\sigma}(\bar{X}|R)$. Clearly, for these data, the decision concerning whether or not to generalize over raters is an important determiner of the magnitude of the standard deviation of $\bar{X}$.

Readers familiar with generalizability theory will note that the above discussion does not differentiate between sample sizes for a G study (or generalizability study) and a D study (or decision study). That is, it has been assumed that the number of items and/or raters for a specific decision (e.g., calculating the expected variability of a cutting score) is identical to the number of items and/or raters characterizing the generalizability study. Also, the above results depend upon the assumption that the population or universe size for each facet (raters and items) is essentially infinite. Sometimes,

Table 2
ANOVA, Variance Components, and the Variability of Mean Scores for the Angoff Procedure

| Effect ($\alpha$) | df | SS | MS | $\hat{\sigma}^2(\alpha)$ |
|---|---|---|---|---|
| r | 4 | .700 | .175 | .0012 |
| i | 125 | 7.144 | .057 | .0061 |
| ri | 500 | 13.353 | .027 | .0267 |

| | |
|---|---|
| $\hat{\sigma}^2(\overline{X}_r) = .0014$ | $\hat{\sigma}(\overline{X}_r) = .037$ (4.70) |
| $\hat{\sigma}^2(\overline{X}) = .0003$ | $\hat{\sigma}(\overline{X}) = .018$ (2.29) |
| $\hat{\sigma}^2(\overline{X}\|R) = .0001$ | $\hat{\sigma}(\overline{X}\|R) = .010$ (1.20) |
| $\hat{\sigma}^2(\overline{X}\|I) = .0003$ | $\hat{\sigma}(\overline{X}\|I) = .017$ (2.10) |

Note. The terms $\hat{\sigma}^2(\alpha)$ are, more specifically, $\hat{\sigma}^2(r)$, $\hat{\sigma}^2(i)$, $\hat{\sigma}^2(ri)$. Results in the second half of this table for the variability of mean scores assume that $n_r = 5$ and $n_i = 126$. Results within parentheses are expressed in terms of number of items.

however, evaluators may wish to generalize to a finite population of raters and/or a finite universe of items. Cronbach et al. (1972) and Brennan (1977) discussed considerations relevant to differing G study and D study sample sizes; Brennan (1977) considered sampling from finite universes and/or populations; and Brennan (1979b) incorporated both considerations in equations for estimating the expected variability of a mean score. Equations 8 to 10 are special cases of these equations, as are certain equations resulting from multiple matrix sampling theory.

## Nedelsky Procedure

The Nedelsky and Angoff procedures are similar in that for each item and rater, both procedures result in a probability that a minimally competent examinee will get an item correct. However, the Angoff procedure directly elicits this probability from each rater, whereas the Nedelsky procedure involves inferring this probability from the number of distractors that a rater believes would be eliminated by a minimally competent examinee.

### Probabilities of Correct Response

Table 3 reports sample statistics, in terms of probability of correct response, that resulted from applying the Nedelsky procedure with the same raters and items discussed previously. Note that the mean cutting score, $\overline{X}$, was .556 (70.09) for the Nedelsky procedure; whereas for the Angoff procedure, $\overline{X}$ was .663 (83.56), as indicated in Table 1. Clearly, there is a substantial difference in mean scores for the two procedures. Furthermore, Tables 1 and 3 indicate that the standard deviation of the rater means for the Nedelsky procedure was approximately double the corresponding standard deviation for the Angoff procedure.

Table 4 reports a generalizability analysis of the Nedelsky probabilities based upon the same model and assumptions used to examine the corresponding results for the Angoff procedure in Table

Table 3
Means, Standard Deviations, and Intercorrelations Among Raters for Probability of Correct Response from Nedelsky Procedure

| | Raters | | | | Mean over items | S.D. over items |
|---|---|---|---|---|---|---|
| Raters | 2 | 3 | 4 | 5 | $\bar{X}_r$ | $\hat{\sigma}_i(X_{ri})$ |
| 1 | .307 | .118 | .196 | .377 | .644 | .283 |
| 2 | | .065 | .204 | .350 | .534 | .232 |
| 3 | | | .161 | .195 | .450 | .183 |
| 4 | | | | .242 | .570 | .238 |
| 5 | | | | | .584 | .231 |
| | $\bar{X}$ = .556 | (70.09)[a] | | $\hat{\sigma}(\bar{X}_r)$ = .072 | (9.03)[a] | |

[a] Results within parentheses are expressed in terms of number of items.

2. In comparing the Nedelsky results in Table 4 with the Angoff results in Table 2, note that each of the random effects variance components [$\hat{\sigma}^2(r)$, $\hat{\sigma}^2(i)$, and $\hat{\sigma}^2(ri)$] for the Nedelsky procedure is considerably larger than the corresponding variance component for the Angoff procedure. This directly results in larger estimates of $\hat{\sigma}(\bar{X})$, $\hat{\sigma}(\bar{X}|R)$, and $\hat{\sigma}(\bar{X}|I)$, for the Nedelsky procedure. For the two procedures $\hat{\sigma}(\bar{X}|R)$, for generalizing over items, is approximately the same. However, $\hat{\sigma}(\bar{X})$, for generalizing over both raters and items, is about twice as large for the Nedelsky procedure; and a similar statement holds for $\hat{\sigma}(\bar{X}|I)$, with generalization over raters only. In a later section these and other differences between the two procedures are examined in more detail.

Table 4
ANOVA, Variance Components, and Variability of Mean Scores for Probability of Correct Response from Nedelsky Procedure

| Effect (α) | df | SS | MS | $\hat{\sigma}^2(\alpha)$ |
|---|---|---|---|---|
| r | 4 | 2.589 | .647 | .0048 |
| i | 125 | 13.182 | .106 | .0125 |
| ri | 500 | 21.428 | .043 | .0429 |
| | $\hat{\sigma}^2(\bar{X}_r)$ = ,0051 | | $\hat{\sigma}(\bar{X}_r)$ = .072 | (9.03) |
| | $\hat{\sigma}^2(\bar{X})$ = .0011 | | $\hat{\sigma}(\bar{X})$ = .034 | (4.24) |
| | $\hat{\sigma}^2(\bar{X}\|R)$ = .0002 | | $\hat{\sigma}(\bar{X}\|R)$ = .013 | (1.64) |
| | $\hat{\sigma}^2(\bar{X}\|I)$ = .0010 | | $\hat{\sigma}(\bar{X}\|I)$ = .032 | (4.04) |

Note. The terms $\hat{\sigma}^2(\alpha)$ are more specifically $\hat{\sigma}^2(r)$, $\hat{\sigma}^2(i)$, and $\hat{\sigma}^2(ri)$. Results in the second half of this table, for the variability of mean scores, assume that $n_r$ = 5 and $n_i$ = 126.

## Eliminated Alternatives

One way of viewing the results presented thus far is that in terms of setting a single cutting score with the Nedelsky or Angoff procedure $\hat{\sigma}(ri)$ is always a source of error, $\hat{\sigma}^2(r)$ is a source of error if generalization is over raters, and $\hat{\sigma}^2(i)$ is a source of error if generalization is over items. This statement is based upon the linear model in Equation 1 for the probability assigned by a rater to an item. In the Nedelsky procedure, however, the data that are actually collected are eliminated distractors, not probabilities, even though the cutting score resulting from the Nedelsky procedure is based directly upon probabilities. (Technically, the cutting score is a linear function of the inferred probabilities and a nonlinear function of the eliminated distractors.)

Several potentially confounding issues arise when the set of eliminated distractors for raters and items is considered. For example, for a given item, if two raters indicate that the same number of distractors would be eliminated by a minimally competent examinee, then the (inferred) probability assigned to the item by these two raters is the same, whether or not the raters agree on *which* distractors would be eliminated. Technically, in terms of the way Nedelsky formulated his procedure, such disagreement among raters has no bearing upon the cutting score that results from the procedure. However, it seems reasonable to postulate that confidence in the Nedelsky procedure, in a specific context, might be influenced by the extent to which raters agree, not only with respect to the number of distractors that would be eliminated, but also with respect to which distractors would be eliminated.

To examine this issue, variance components can be estimated for a design in which raters are crossed with items, and distractors ($d$) are nested within items. This design is denoted $r\times(d{:}i)$.[1] Formulas for estimating variance components for this design are presented in Table 5, along with the estimated variance components for the data. It is usual in many applications of generalizability theory to report random effects variance components, based on the assumption that the population (or universe) size for each facet is essentially infinite. In this case, however, it seems unreasonable to consider the $n_d = 3$ distractors associated with each item as a sample from an essentially infinite universe of possible distractors for the item. Therefore, in Table 5, the variance components are reported under the assumption that distractors are fixed and this assumption is indicated by the notation $\hat{\sigma}^2(\alpha|D)$, where $\alpha$ is any one of the five effects in the design.

Consider the two variance components in Table 5 that involve variability attributable to distractors. The variance component $\hat{\sigma}^2(d{:}i|D)$ reflects the average, over items, of the variance attributable to the proportion of raters who eliminate each distractor. The magnitude of $\hat{\sigma}^2(d{:}i|D)$ will be large when, on the average, raters judge an item's distractors to vary in their difficulty, or attractiveness, to examinees. By contrast, the magnitude of $\hat{\sigma}^2(rd{:}i|D)$ reflects disagreement or variability among raters in their judgments of distractor attractiveness for an item. To put it another way, the magnitude of $\hat{\sigma}^2(rd{:}i|D)$ reflects the extent to which raters disagree in their judgments about *which* distractors would be eliminated by a minimally competent examinee.

If $\hat{\sigma}^2(d{:}i|D) = .063$ is considered as an estimate of "true" variability among distractors, then the estimate of "error" for $n_r = 5$ raters is $\hat{\sigma}^2\,(rd{:}i|D)\,/n_r = .181/5 = .036$. Evidently, the error variance (attributable to the differential attractiveness of distractors for different raters) is almost 50% as large as the true variance among distractors. This suggests that for these data, even when raters agree on the number of distractors that would be eliminated, there are substantial differences among raters concerning *which* distractors would be eliminated.

[1] For the $r \times (d{:}i)$ design it can be argued that a rater may not examine a given distractor independent of the other distractors for an item. If so, then one independence assumption associated with the linear model for this design becomes suspect, at least to some extent. This issue, however, is relatively unimportant here because the analysis is intended only to summarize data that have an indirect bearing on the principal analyses in this paper.

Table 5
ANOVA and Variance Components for Eliminated Distractors Using Nedelsky Procedure

| Effect ($\alpha$) | df | SS | MS | $\hat{\sigma}^2(\alpha \mid D)$ |
|---|---|---|---|---|
| r | 4 | 9.342 | 2.335 | .006 |
| i | 125 | 42.625 | .341 | .012 |
| d:i | 252 | 124.925 | .496 | .063 |
| ri | 500 | 79.984 | .160 | .053 |
| rd:i | 1008 | 182.885 | .181 | .181 |

$$\hat{\sigma}^2(r \mid D) = [MS(r) - MS(ri)]/n_i n_d \qquad \hat{\sigma}^2(ri \mid D) = MS(ri)/n_d$$

$$\hat{\sigma}^2(i \mid D) = [MS(i) - MS(ri)]/n_r n_d \qquad \hat{\sigma}^2(rd{:}i \mid D) = MS(rd{:}i)$$

$$\hat{\sigma}^2(d{:}i \mid D) = [MS(d{:}i) - MS(rd{:}i)]/n_r$$

Mean over items of proportion of distractors eliminated for raters 1 to 5:

$$\bar{X}_r = .698,\ .609,\ .505,\ .656,\ .688$$

$$\bar{X} = .631 \qquad \hat{\sigma}(\bar{X}_r) = .077$$

Note. In this table $\bar{X}_r$ and $\bar{X}$ refer to proportions of eliminated distractors--not probabilities of correct response.

In conducting a study with the Nedelsky procedure, it is usual to provide raters with complete items, including the correct alternatives. If the correct alternatives for all items are specified a priori for the raters, then it might be expected that no rater will indicate that a minimally competent examinee would eliminate a correct alternative. On the other hand, if correct alternatives are not specified a priori (and they were not in this study), then it might be expected that some raters will indicate that a minimally competent examinee would eliminate the correct alternative for some items. Indeed, this did occasionally occur in this study. No evidence of clerical error or miskeyed items was found to explain these results, and there is no reason to believe that raters did not take their task seriously. It is likely, however, that individual raters had differing levels of familiarity with the content tested by specific items; and it could be that some raters truly believed that a correct answer would be eliminated by a minimally competent examinee.

When a rater indicates that the correct answer would be eliminated by a minimally competent examinee, it might be argued that the (inferred) probability assigned by the rater to the item should be zero, no matter how many distractors are eliminated by the rater. However, for the purposes of this study, this argument was not adopted. Rather, Nedelsky's procedure, as he described it, was followed, and probabilities were assigned on the basis of eliminated distractors only. This approach was chosen for two reasons. First, since the principal purpose was to compare the Angoff and Nedelsky procedures as they are currently stated, it was not desired to alter the Nedelsky probabilities without a corresponding alteration of the Angoff probabilities; and there was no objective basis for altering the Angoff probabilities. Second, if a probability of zero had been assigned whenever a rater indicated that a minimally competent examinee would eliminate the correct alternative, then $\bar{X}$ would decrease and the estimates of variability would increase. The approach taken is a conservative one, for the purposes of this paper, in that the results reported here for the two procedures may be somewhat more similar than might be the case, otherwise.

## A Comparison of the Two Procedures

Since the Nedelsky and Angoff procedures were both applied to the same items by the same raters, the data from these two procedures can be analyzed jointly in a single design. Specifically, the appropriate analysis involves the *p* × *r* × *i* design, in which the two procedures (*p*) are crossed with both raters and items. Table 6 provides equations for estimating the variance components for this design, and Table 7 provides the numerical values of these estimated variance components for the data.

The variance components, identified as $\hat{\sigma}^2(\alpha)$ in Table 7 are obtained by letting $N_p$ approach infinity for the equations in Table 6; these are called random effects variance components. The variance

Table 6
Equations for Estimating Variance Components and the Expected Variance of the Mean Score for the p x r x i Design

| Effect | Estimated Variance Component |
|---|---|
| p | $[MS(p) - MS(pr) - MS(pi) + MS(pri)]/n_r n_i$ |
| r | $\{MS(r) - MS(ri) - (1 - n_p/N_p)[MS(pr) - MS(pri)]\}/n_p n_i$ |
| i | $\{MS(i) - MS(ri) - (1 - n_p/N_p)[MS(pi) - MS(pri)]\}/n_p n_r$ |
| pr | $[MS(pr) - MS(pri)]/n_i$ |
| pi | $[MS(pi) - MS(pri)]/n_r$ |
| ri | $[MS(ri) - (1 - n_p/N_p)MS(pri)]/n_p$ |
| pri | $MS(pri)$ |

When $n_p = N_p$, the variance components are identified as $\hat{\sigma}^2(\alpha|P)$. In terms of these variance components:

$$\hat{\sigma}^2(\overline{X}_r|P) = \hat{\sigma}^2(r|P) + \hat{\sigma}^2(ri|P)/n_i$$

$$\hat{\sigma}^2(\overline{X}|P) = \hat{\sigma}^2(r|P)/n_r + \hat{\sigma}^2(i|P)/n_i + \hat{\sigma}^2(ri|P)/n_r n_i$$

$$\hat{\sigma}^2(\overline{X}|P,R) = \hat{\sigma}^2(i|P)/n_i + \hat{\sigma}^2(ri|P)/n_r n_i$$

$$\hat{\sigma}^2(\overline{X}|P,I) = \hat{\sigma}^2(r|P)/n_r + \hat{\sigma}^2(ri|P)/n_r n_i$$

Note. These estimates are based on the assumptions that the model is fully restricted; and that, for finite values of $N_p$, all variance components associated with the p facet are defined using a factor of $N_p - 1$. For example, if $\mu_{pr}{\sim}$ is the effect for the interaction of p and r, then $\sigma^2(pr|P)$ is defined as $\sum_p \sum_r (\mu_{pr}{\sim})^2/(N_p - 1)$.

Table 7
ANOVA and Variance Components for Probability of Correct Response with Both Procedures

| Effect ($\alpha$) | df | SS | MS | $\hat{\sigma}^2(\alpha)$[a] | $\hat{\sigma}^2(\alpha\|P)$[b] |
|---|---|---|---|---|---|
| p | 1 | 3.599 | 3.599 | .0050 | .0050[c] |
| r | 4 | 1.554 | .388 | -.0002 | .0014 |
| i | 125 | 15.660 | .125 | .0074 | .0083 |
| pr | 4 | 1.735 | .434 | .0032 | .0032 |
| pi | 125 | 4.665 | .037 | .0019 | .0019 |
| ri | 500 | 20.952 | .042 | .0071 | .0210 |
| pri | 500 | 13.829 | .028 | .0277 | .0277 |

Means over procedures and items for raters 1 to 5:

$$\bar{X}_r = .658,\ .627,\ .553,\ .593,\ .619$$

$\bar{X} = .610\ (78.82)$ $\qquad$ $\hat{\sigma}^2(\bar{X}_r) = \hat{\sigma}^2(\bar{X}_r|P) = .040\ (4.99)$

| | |
|---|---|
| $\hat{\sigma}^2(\bar{X}\|P) = .0004$ | $\hat{\sigma}(\bar{X}\|P) = .020\ (2.46)$ |
| $\hat{\sigma}^2(\bar{X}\|P,R) = .0001$ | $\hat{\sigma}(\bar{X}\|P,R) = .010\ (1.26)$ |
| $\hat{\sigma}^2(\bar{X}\|P,I) = .0003$ | $\hat{\sigma}(\bar{X}\|P,I) = .018\ (2.22)$ |

[a] Values of $\hat{\sigma}^2(\alpha)$ are for $N_p \to \infty$ in the equations in Table 6.

[b] Values of $\hat{\sigma}^2(\alpha|P)$ are for $n_p = N_p = 2$ in the equations in Table 6.

[c] Some writers would refer to .0050 as an estimate of a quadratic form, rather than as an estimated variance component.

components identified as $\hat{\sigma}^2(\alpha|P)$ in Table 7 are obtained by letting $n_p = N_p = 2$ in Table 6; and these variance components are based on the assumption that procedures are fixed. The variance components $\sigma^2(\alpha|P)$ are appropriate when interest is restricted to the actual procedures in this study. Strictly speaking, the variance components $\hat{\sigma}^2(\alpha|P)$ seem more appropriate here than the random effects variance components $\hat{\sigma}^2(\alpha)$ because it seems difficult to consider these two procedures as a sample from some very large set of similar cutting score procedures.

Tables 6 and 7 also provide equations and numerical values for estimates of the variability of $\bar{X}$, where $\bar{X}$ is, in this case, the mean over raters, items, and procedures. For example, Table 7 reports that $\bar{X}$ (over procedures) is .610 (78.82), which is the mean of the $\bar{X}$'s reported in Tables 1 and 3.

The reader should note, however, that the estimates of the *variability* of $\bar{X}$ in Table 7 are *not* averages of the corresponding estimates in Tables 2 and 4. For example, $\hat{\sigma}(\bar{X}|P) = .020\ (2.46)$, which is similar to $\hat{\sigma}(\bar{X}) = .018\ (2.29)$ in Table 2 for the Angoff procedure but quite different from $\hat{\sigma}(\bar{X}) = .034\ (4.24)$ in Table 4 for the Nedelsky procedure. This pattern of results also holds for $\hat{\sigma}(\bar{X}|P,R)$ and $\hat{\sigma}(\bar{X}|P,I)$. For these data, therefore, one inference that might be drawn is that there would be no particular advantage in actually setting a cutting score by averaging $\bar{X}$ from procedures, assuming that interest is primarily in minimizing the variability of $\bar{X}$.

Perhaps the most outstanding result in Table 7 is that variance components that contain $p$ are relatively large, indicating that there are substantial differences between the two procedures and the probabilities that result from them. For example, the variance components suggest that there is considerably more variability attributable to differences in procedure means than to differences in rater means (over procedures). From another perspective, it can be shown that the observed variance in the two procedure means is

$$\hat{\sigma}^2(\bar{X}_p) = \hat{\sigma}^2(p) + \frac{\hat{\sigma}^2(pr)}{n_r} + \frac{\hat{\sigma}^2(pi)}{n_i} + \frac{\hat{\sigma}^2(pri)}{n_r n_i} . \qquad [11]$$

The relationship expressed by Equation 11 also holds using the "procedures fixed" variance components. In other words, the variance components that contain $p$ contribute directly to the disparity identified in the procedure means. In effect, Table 7 crystallizes many of the differences between the two procedures evident in comparing Table 2 with Table 4.

## Differences in Sample Statistics for Raters

Differences between the two procedures can also be examined using the sample statistics reported in Tables 1 and 3. In examining these differences, relationships between the results in Tables 1 and 3 and the analysis of variance results in Tables 2, 4, and 7 will occasionally be pointed out (without proof).

*Correlations and covariances among raters, within procedures.* Using Tables 1 and 3, the reader can verify that the average of the rater intercorrelations for the Angoff procedure is .187 and the corresponding result for the Nedelsky procedure is .222. In terms of covariances, these averages are .006 and .013 for the Angoff and Nedelsky procedures, respectively. The magnitude of these average covariances is influenced by the degree to which similar probabilities are assigned to items. Evidently, there is more variability over items in the probabilities assigned using the Nedelsky procedure. Further evidence of this fact will be seen below.

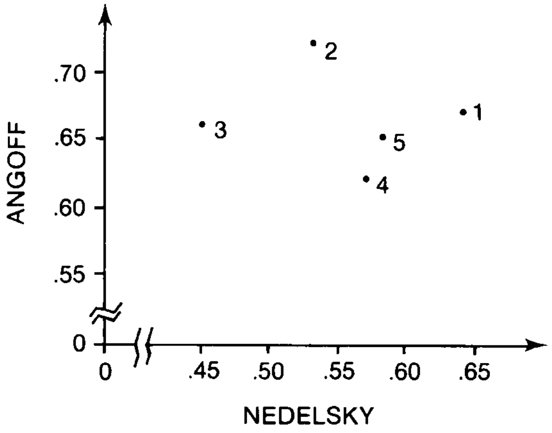
*Rater means.* Figure 1 provides a scatterplot of the rater means (over items) for the Angoff procedure (see Table 1) and the Nedelsky procedure (see Table 3). The reader can verify that the correlation in Figure 1 is −.052; and it can be shown that the covariance (in terms of the random effects variance components in Table 7) is $\hat{\sigma}^2(r) + \hat{\sigma}^2(ri)/n_i = -.0002 + .0071/126 \doteq -.0001$. Clearly, there is little, if any, linear relationship between the two procedures in terms of the five rater means.[2] Note that this result is not influenced by the difference in the grand means ($\bar{X}$'s) for the two procedures.

It appears from Figure 1, however, that there are two clusters of raters—Raters 2 and 3 and Raters 1, 4, and 5. Given the small numbers of raters involved, it cannot be said that there is a strong correlation among raters within clusters, but Figure 1 certainly does not preclude this possibility. In any case, Raters 2 and 3 are outstanding in that they assign relatively low probabilities using the Nedelsky procedure and relatively high probabilities using the Angoff procedure.

*Rater standard deviations.* Figure 2 provides a scatterplot of the statistics $\hat{\sigma}_i(\bar{X}_{ri})$ for each rater, by both procedures. Recall that for a given rater and procedure, $\hat{\sigma}_i(\bar{X}_{ri})$ is the standard deviation of the

[2] By definition, a variance component must be positive; however, *estimates* of variance components are occasionally negative. When a negative estimate occurs, sometimes it is advisable to treat it as zero (see Cronbach et al., 1972, and Brennan, 1977), and at other times it is best to leave the estimate unchanged (see Sirotnik, 1970). Here, $\hat{\sigma}^2(r)$ is *not* set to zero because it is a mathematical fact that an *observed* covariance of the type in Figure 1 is exactly $\hat{\sigma}^2(r) + \hat{\sigma}^2(ri)/n_i$, as shown by Cronbach et al. (1972, chap. 8).

**Figure 1**
Rater Means (Over Items) for
Probability of a Correct Response
using Nedelsky and Angoff Procedures



probabilities assigned to items. The standard deviations for the Nedelsky procedure are somewhat higher than those for the Angoff procedure, which is consistent with the variance components for items and interaction being higher for the Nedelsky procedure. Again, however, Rater 3 and, to some extent, Rater 2 appear to be different from the other three raters. Specifically, for both procedures, Raters 2 and 3 exhibit less variability in the probabilities they assign to items.

**Figure 2**
Standard Deviations, for each
Rater, of the Probabilities Assigned
to Items Using Nedelsky and Angoff Procedures

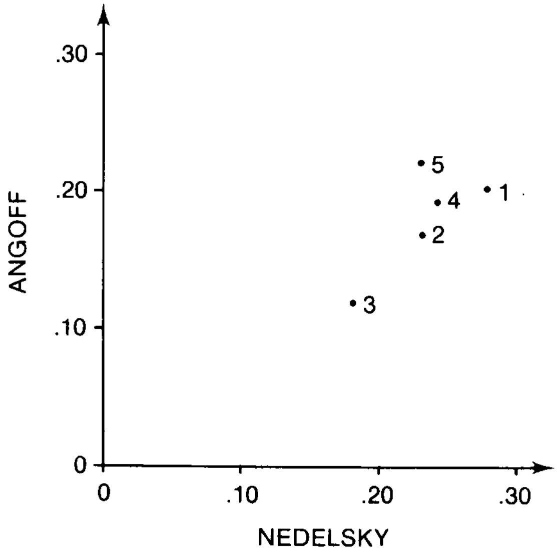


Figure 3 provides a frequency polygon for the average (over raters) of the probabilities assigned to items by both procedures; and Figure 4 provides a frequency polygon of the standard deviation of the probabilities assigned to items. Consistent with previously discussed results, Figure 3 indicates that the modal probability (interval) is considerably higher for the Angoff procedure. Also, consistent with previous results, Figure 4 indicates that there is somewhat more variability in the probabilities assigned to items using the Nedelsky procedure. Most importantly, however, the Nedelsky standard deviations in Figure 4 are bimodal. As discussed below, this bimodality is not an artifact of these data; it is a result that is virtually guaranteed by the Nedelsky procedure.

Recall that for each rater the probability assigned to an item by the Nedelsky procedure is the inverse of the number of noneliminated distractors. For the four-alternative items used in this study, this method of assigning probabilities implies that the only (inferred) probabilities that can be assigned to an item using the Nedelsky procedure are .25, .33, .50, and 1.00. In particular, note that there can be no probability between .50 and 1.00. Now, consider the probabilities assigned by raters to an item. If all raters assign probabilities in the range .25 to .50, the standard deviation will be relatively small; and, of course, if they all assign probabilities of 1.00, the standard deviation will be zero. However, the standard deviation will be relatively large when some raters assign a probability of 1.00 and other raters assign probabilities of .50 or lower.

The bimodality in Figure 4, then, seems almost certainly a direct result of having only a small number of *un*equally spaced probabilities with the Nedelsky procedure. Furthermore, this peculiar characteristic of the probability scale is a plausible explanation for the estimates of the variability of $\overline{X}$ being higher for the Nedelsky procedure than for the Angoff procedure (see Tables 2 and 4). Also, the restricted nature of the Nedelsky probability scale may account for the differences in the means for the two procedures, at least to some extent. To examine these issues in more detail, consider Tables 8 and 9.

**Figure 3**
Frequency Polygon of the Means
(Over Raters) of the Probabilities
Assigned to Items

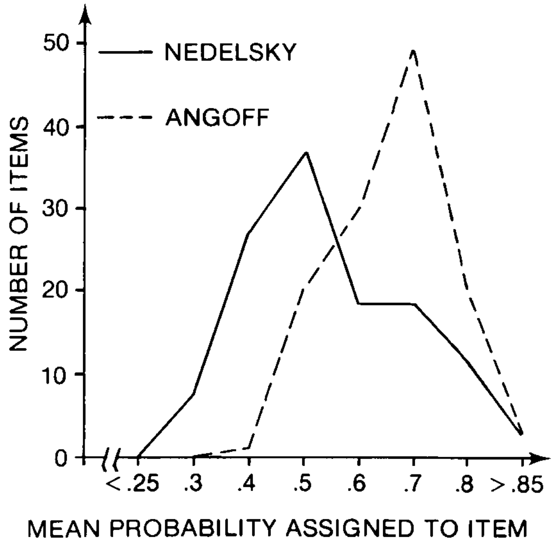
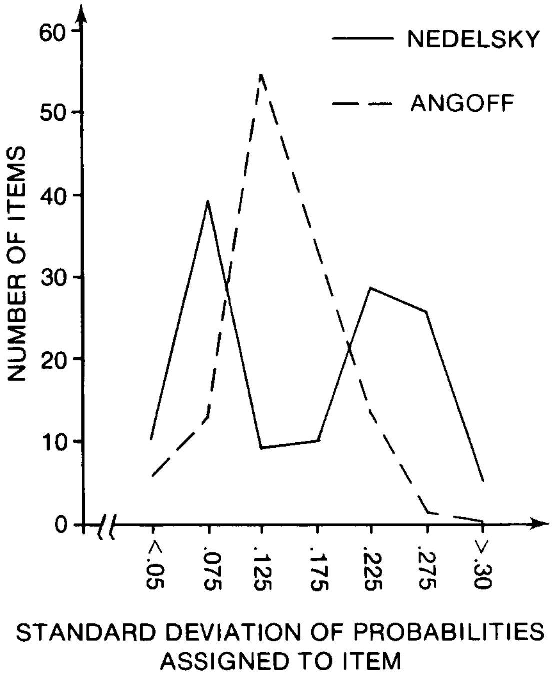


**Figure 4**
Frequency Polygon of the Standard Deviations (Over Raters) of the Probabilities Assigned to Items



Tables 8 and 9 provide relative frequency distributions, over items, for the probabilities assigned using the Angoff and Nedelsky procedures, respectively. Inspection of these tables reveals several points of interest. First, no rater assigned probabilities below .20 using the Angoff procedure. This implies that the range of probabilities for the two procedures is about the same; consequently, differential restriction in range is not a factor of importance in the data. Second, for the Nedelsky procedure, on the average, probabilities below .50 were used for 28% of the items, whereas for the Angoff procedure, they were used for only 7% of the items. Third, for the Angoff procedure, on the average, probabilities in the range .60 to .95 were used with 53% of the items, whereas the Nedelsky procedure precluded use of such probabilities.

These points, and visual inspection of Tables 8 and 9, reveal a consistent tendency for raters to assign more homogeneous probabilities using the Angoff procedure. Furthermore, it appears that a rater who uses a probability of .33 or .50 with the Nedelsky procedure is very likely to use a somewhat higher probability when given the opportunity to do so with the Angoff procedure.

## Operationalizing Conceptions of Minimum Competence

There are many ways in which the Nedelsky and Angoff procedures appear to be very similar. For example, they both involve raters' judgments about individual items; they both yield, directly or indirectly, a matrix of rater-by-item probabilities; and, given this matrix, the computational process for

Table 8
Raters' Frequency Distributions of Probability of Correct Response Using Angoff Procedure

| Probability of Correct Response[a] | Raters | | | | | Average |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | |
| <.20 | .00 | .00 | .00 | .00 | .00 | .00 |
| (.20, .25) | .03 | .02 | .00 | .07 | .06 | .04 |
| (.30, .35) | .02 | .00 | .00 | .02 | .06 | .02 |
| (.40, .45) | .00 | .00 | .06 | .00 | .00 | .01 |
| (.50, .55) | .42 | .24 | .19 | .44 | .36 | .33 |
| (.60, .65) | .01 | .04 | .31 | .00 | .01 | .07 |
| (.70, .75) | .26 | .33 | .23 | .37 | .31 | .30 |
| (.80, .85) | .01 | .23 | .20 | .00 | .02 | .09 |
| (.90, .95) | .12 | .07 | .01 | .10 | .04 | .07 |
| >.95 | .13 | .08 | .00 | .00 | .15 | .07 |
| $\bar{X}_r$ | .67 | .72 | .66 | .62 | .65 | .66 |

[a] Raters were constrained to report their probabilities in units of .05.

arriving at a cutting score is the same for both procedures. The procedures obviously differ in that probabilities are directly elicited in the Angoff procedure, whereas probabilities are inferred from eliminated distractors in the Nedelsky procedure.

It is also possible that the two procedures differ, to some extent, in the way they technically allow a rater to operationalize a conception of minimum competence. In the Angoff procedure, to arrive at a probability, a rater might conceptualize a *group* of minimally competent persons and reflect upon what proportion would get the item correct. Alternatively, for the Angoff procedure, a rater might

Table 9
Raters' Frequency Distributions of Probability of Correct Response Using Nedelsky Procedure

| Probability of Correct Response | Raters | | | | | Average |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | |
| .25 | .06 | .06 | .07 | .05 | .00 | .05 |
| .33 | .16 | .25 | .42 | .16 | .17 | .23 |
| .50 | .40 | .52 | .43 | .57 | .60 | .50 |
| 1.00[a] | .38 | .18 | .08 | .22 | .23 | .22 |
| $\bar{X}_r$ | .64 | .53 | .45 | .57 | .58 | .56 |

[a] Analyses of these data used a probability of 0.99, rather than 1.00, for coding convenience.

conceptualize a *single* minimally competent person and reflect upon what proportion of the time this person would correctly respond to the item if it were administered a large number of times.

For the Nedelsky procedure, however, there are only as many distinct probabilities that can be assigned (indirectly) as there are alternatives to the item, and these probabilities are not equally spaced. Logic suggests, therefore, that neither of the above two conceptualizations works very well with the Nedelsky procedure. For example, if a rater believes that 75% of a group of minimally competent persons would get an item correct, the rater cannot eliminate some number of alternatives that will yield a probability of .75. Technically, the rater cannot even report the average number of alternatives that a group of minimally competent persons would eliminate, unless this number is an integer.

It seems, then, that the Nedelsky procedure constrains a rater to conceptualize minimum competency in terms of the performance of a single person on a single administration of an item, with the additional constraint that this person will respond based upon a process of eliminating distractors. There is no compelling empirical evidence to suggest that examinees (specifically, minimally competent examinees) generally respond to an item based upon a process of eliminating distractors, even though this process is frequently recommended to potential examinees. However, even if examinees do respond in this manner, there still seem to be relatively clear differences in the conceptualization of minimal competence implicit in the Angoff and Nedelsky procedures.

This study cannot directly address the extent to which different conceptualizations of minimum competency may have influenced the study's results; and it is judged unlikely that raters gave this matter a great deal of conscious consideration. Nevertheless, any cutting score procedure necessitates some conceptualization of minimum competence; it seems likely that the conceptualizations are different for the Angoff and Nedelsky procedures; and evaluators are probably well advised to consider such differences in choosing a cutting score procedure in a given context.

### Cutting scores other than $\overline{X}$

It is important to note that, throughout this paper, it has been assumed that the cutting score, $\overline{X}$, resulting from either procedure is the mean of $\overline{X}_r$, for all raters who participated in the study. For example, it was pointed out that Raters 2 and 3 in this study appeared to be different from the other three raters. However, it was not suggested that they be eliminated from the study for the purposes of calculating a cutting score. In the opinion of the authors, unless there is clear evidence that a rater did not adhere to the intended procedure, it is probably not generally advisable to eliminate atypical raters in determining the cutting score. (It is assumed, of course, that raters were chosen carefully in the first place). However, if an atypical rater were eliminated, it would be best to redo analyses using only the remaining raters. This suggestion is made because the elimination of an atypical rater, after the study is completed, probably implies a change in the conceptualization of the intended population of raters.

*Reconciliation process.* Sometimes, rather than using $\overline{X}$ as the cutting score, it is suggested that a cutting score be determined by a reconciliation process. For example, after the five raters in this study completed the Angoff procedure, they were instructed, as a group, to reconcile their differences on each item (see Table 1). One typical result of using a reconciliation process is that certain raters tend to dominate, or to influence unequally, the reconciled ratings. This is indeed what happened in the study of the Angoff procedure, as indicated by the high correlations between the actual and reconciled ratings for Raters 1 and 2. The effect of this dominance by Raters 1 and 2 is that the reconciled cutting score (.70) is quite a bit different from $\overline{X}$ (.66).

There is a certain logic in using a reconciliation process that appears to be compelling. It might be argued that the ideal of using either the Nedelsky or the Angoff procedure is for raters to agree on every item. Therefore, why not force them to concur? One argument against this logic is that forced consensus is not agreement, although forced consensus may effectively hide disagreement. Also, a reconciliation process does not guarantee that the same cutting score will result each time a study is replicated. If a study is replicated a large number of times with different raters, the average reconciled cutting score might be considerably different from the average $\bar{X}$ or $\lambda$ in Equation 1; however, there could be as much, or even more, variability in the distribution of reconciled cutting scores as there is in the distribution of $\bar{X}$'s. This does not imply, however, that a reconciliation procedure should necessarily be avoided; rather, use of a reconciliation procedure involves complexities over and above those encompassed by either the Nedelsky or the Angoff procedure.

*Nedelsky's cutting score.* When Nedelsky originally described his procedure, he did not actually suggest using $\bar{X}$ (or $n_i\bar{X}$) as a cutting score. Instead, the cutting score he suggested using is $M_{FD} + k\sigma_{FD}$. Nedelsky's discussion of $M_{FD}$, $k$, and $\sigma_{FD}$ is somewhat confusing. However, it appears that $M_{FD}$ is intended to be the mean test score for a group of "border-line" examinees, only (Nedelsky, 1954, p. 5); $\sigma_{FD}$ is the standard deviation of this distribution; and $k$ is an a priori defined constant used to classify these "border-line" examinees into passing and failing examinees. Since Nedelsky suggests using $n_i\bar{X}$ as an estimate of $M_{FD}$, it is clear that his cutting score will equal $n_i\bar{X}$ only if $k$ is defined as zero or $\sigma_{FD}$ is zero.

It is not clear to these authors why $M_{FD} + k\sigma_{FD}$ would be used as a cutting score if there actually were test scores for a known group of "borderline" examinees. In such a case, the test data themselves would likely provide a reasonably sound basis for defining a cutting score independent of raters' judgments. It can be inferred, therefore, that Nedelsky probably wants a *hypothetical* group of borderline examinees to be considered. It has already been argued that there may be a logical inconsistency in conceptualizing a group of minimally competent examinees when using the Nedelsky procedure. However, even if this issue is overlooked, the problem still exists of estimating $\sigma^2_{FD}$ (a parameter for a test score distribution) using only the raters' probabilities.

It can be shown that the formula suggested by Nedelsky (1954, p. 12) for estimating $\hat{\sigma}^2_{FD}$ is

$$\hat{\sigma}^2_{FD} = \sum_r \sum_i X_{ri}(1 - X_{ri})/n_r \qquad [12a]$$

$$\doteq n_i[\bar{X}(1 - \bar{X}) - \hat{\sigma}^2(r) - \hat{\sigma}^2(i) - \hat{\sigma}^2(ri)] \;; \qquad [12b]$$

where $X_{ri}$ is the (inferred) probability assigned to item $i$ by rater $r$. Nedelsky provides a rationale for his estimate of $\hat{\sigma}^2_{FD}$; but in the authors' opinion, his rationale is weak in that it confounds considerations of parameters and estimates. Even if his formula for estimating $\hat{\sigma}^2_{FD}$ is accepted, the very process of defining a cutting score as $M_{FD} + k\sigma_{FD}$ requires fairly strong assumptions and a substantial degree of subjective judgment over and above that required to estimate the cutting score $n_i\bar{X}$. Whether or not such complexity is advisable depends upon the specific context of the cutting score decision process; however, there are probably not many contexts in which this complexity is warranted, and the procedure is easily defended.

## Measurement Reliability or Dependability

The numerical results reported in this paper are for a specific experimental study only; and, as such, these results are illustrative, rather than definitive. Nevertheless, there appear to be noticeable differences in the means (or cutting scores) for the two procedures. Also, for each procedure, there is evidence of error, as reflected in the expected variances of the distributions of means over replications; and these variances frequently have considerably different magnitudes for the two procedures. Given these results, it seems reasonable to consider their potential impact on issues of reliability, or measurement dependability. A complete discussion of these issues is beyond the intended scope of this paper. One relatively straightforward approach to this issue will, be considered however.

Brennan and Kane (1977a, 1977b), Kane and Brennan (1980), and Brennan (1980) discussed the following index of dependability for a domain-referenced test:

$$\Phi(\lambda) = \frac{\sigma^2(p) + (\mu - \lambda)^2}{\sigma^2(p) + (\mu - \lambda)^2 + \sigma^2(\Delta)} \quad . \qquad [13]$$

$\lambda$, in Equation 13 is identical to $\lambda$ in Equation 1. The other terms in Equation 13, however, are not evident from Equation 1. Rather, the other terms in Equation 13 result from a consideration of the following linear model for the observed response of person $p$ to item $j$:

$$Y_{pj} = \mu + \mu_{p}{\sim} + \mu_{j}{\sim} + \mu_{pj}{\sim} \ . \qquad [14]$$

Technically, the linear models in Equations 1 and 14 are formally identical. However, different notation is used in each of them for the purpose of emphasizing that Equation 1 is applied to a rater-by-item matrix of probabilities, whereas Equation 14 is applied to a person-by-item matrix of observed scores.

For any meaningful joint use of Equations 1 and 14, the item universe must be the same for both model equations, although the effect for items in Equation 1, $\lambda_{i}{\sim}$, is different from the effect for items in Equation 14, $\mu_{j}{\sim}$. Most importantly, $\lambda$ and $\mu$ in Equations 1 and 14 are very different. The parameter $\lambda$ is the cutting score (or grand mean of the probabilities) for the population of raters and universe of items; whereas the parameter $\mu$ is the grand mean of the observed scores, $Y_{pj}$, for the population of persons and the universe of items.

Using generalizability theory and the linear model in Equation 14, Brennan and Kane (1977a)[3] derived the following equation as an estimate of their index of dependability:

$$\hat{\Phi}(\lambda) = \frac{\hat{\sigma}^2(p) + (\overline{Y} - \lambda)^2 - \hat{\sigma}^2(\overline{Y})}{\hat{\sigma}^2(p) + (\overline{Y} - \lambda)^2 - \hat{\sigma}^2(\overline{Y}) + \hat{\sigma}^2(\Delta)} \quad ; \qquad [15]$$

In Equation 15,

$$\hat{\sigma}^2(p) = [MS(p) - MS(pj)]/n_j \quad ; \qquad [16a]$$

$$\hat{\sigma}^2(j) = [MS(j) - MS(pj)]/n_p \quad ; \qquad [16b]$$

[3] Brennan and Kane (1977a) also provide an easily calculated formula for $\hat{\phi}(\lambda)$ in terms of commonly used sample statistics.

$$\hat{\sigma}^2(pj) = MS(pj) \quad ; \qquad [16c]$$

$$\hat{\sigma}^2(\Delta) = \hat{\sigma}^2(j)/n_j + \hat{\sigma}^2(pj)/n_j \quad ; \text{ and} \qquad [16d]$$

$$\hat{\sigma}^2(\overline{Y}) = \hat{\sigma}^2(p)/n_p + \hat{\sigma}^2(j)/n_j + \hat{\sigma}^2(pj)/n_p n_j \quad . \qquad [16e]$$

The estimate in Equation 15 is identified as $\hat{\phi}(\lambda)$ to ɜmphasize that it is based on the assumption that $\lambda$ is somehow known, without error. This assumptıon is reflected in the term $(\overline{Y} - \lambda)^2$ in the numerator and denominator of Equation 15. When $\lambda$ is not known. however, and $\overline{X}$ from a particular study is used as an estimate of $\lambda$, then this term is no longer appropriate.

Furthermore, $\lambda$ may not simply be replaced with $\overline{X}$ in the term $(\overline{Y} - \lambda)^2$ because the expected value of a squared quantity is not equal to the square of the expected value. Rather, the expected value of $(\overline{Y} - \overline{X})^2$ is

$$\mathcal{E}_R \mathcal{E}_I (\overline{Y} - \overline{X})^2 = (\overline{Y} - \lambda)^2 + \hat{\sigma}^2(\overline{X}) \quad , \qquad [17]$$

if it is desired to generalize over samples of raters ($R$) *and* samples of items ($I$). If it is desired to generalize over samples of items only, then

$$\mathcal{E}_I (\overline{Y} - \overline{X})^2 = (\overline{Y} - \lambda)^2 + \hat{\sigma}^2(X|R) \quad . \qquad [18]$$

It follows from Equations 17 and 18 that when $\overline{X}$ is used as an estimate of $\lambda$, $\hat{\sigma}^2(\overline{X})$ or $\hat{\sigma}^2(\overline{X}|R)$, as appropriate, should also be subtracted from both the numerator and the denominator of Equation 15. The two resulting (modified) estimates of the index of dependability, $\hat{\phi}(\lambda)$, are as follows:

$$\hat{\Phi}(\overline{X}) = \frac{\hat{\sigma}^2(p) + (\overline{Y} - \overline{X})^2 - \hat{\sigma}^2(\overline{Y}) - \hat{\sigma}^2(\overline{X})}{\hat{\sigma}^2(p) + (\overline{Y} - \overline{X})^2 - \hat{\sigma}^2(\overline{Y}) - \hat{\sigma}^2(\overline{X}) + \hat{\sigma}^2(\Delta)} \qquad [19]$$

for sampling over both raters and items; and

$$\hat{\Phi}(\overline{X}|R) = \frac{\hat{\sigma}^2(p) + (\overline{Y} - \overline{X})^2 - \hat{\sigma}^2(\overline{Y}) - \hat{\sigma}^2(\overline{X}|R)}{\hat{\sigma}^2(p) + (\overline{Y} - \overline{X})^2 - \hat{\sigma}^2(\overline{Y}) - \hat{\sigma}^2(\overline{X}|R) + \hat{\sigma}^2(\Delta)} \qquad [20]$$

for sampling over items only.

Consider now the original question that motivated the development of Equations 19 and 20, namely, for the Nedelsky and Angoff procedures what effect do different values for $\overline{X}$ and its expected variability have on reliability or measurement dependability? Without loss of generality, consideration can be restricted to $\hat{\phi}(\overline{X})$ in Equation 19 for generalizing over raters and items. Since $\hat{\phi}(\lambda)$ can be no greater than one, decreasing the numerator and denominator in Equation 15 by $\hat{\sigma}^2(\overline{X})$ results in decreasing the magnitude of the estimate of the Brennan-Kane index. This is to be exᵞected, because additional sources of error attributable to the procedure used to establish a cutting score have been introduced.

Furthermore, all other things being equal, the larger the magnitude of $\hat{\sigma}^2(\bar{X})$, the smaller the magnitude of $\hat{\phi}(\bar{X})$. Since the results suggest that $\hat{\sigma}^2(\bar{X})$ is larger for the Nedelsky procedure, it might be expected that $\hat{\phi}(\bar{X})$ is smaller for the Nedelsky procedure. However, all things are *not* equal unless the cutting scores for the procedures are equal. When they are *un*equal the magnitude of $(\bar{Y}-\bar{X})^2$ will be different; and this difference, in turn, will affect the magnitude of $\hat{\phi}(\bar{X})$. Moreover, whether or not higher values of $\bar{X}$ will result in higher values of $(\bar{Y}-\bar{X})^2$ depends upon the magnitude of $Y$. In brief, it is not *necessarily* the case that lower values of $\hat{\sigma}^2(\bar{X})$ are always associated with higher values for estimates of measurement dependability.

Note that it has *not* been suggested that $\hat{\sigma}^2(\bar{X}|I)$ be considered in the context of modifying the Brennan-Kane index. Of course, there is an equation analogous to Equations 17 and 18, namely,

$$\mathcal{E}_R(\bar{Y} - \bar{X})^2 = (\bar{Y} - \lambda)^2 + \hat{\sigma}^2(\bar{X}|I) \quad , \qquad [21]$$

in which generalization is over samples of raters only. However, in this equation, items are considered fixed; and if $\hat{\sigma}^2(\bar{X}|I)$ is incorporated into an estimate of $\hat{\phi}(\lambda)$, items must then be considered fixed in estimating the other variance components, too. To do so means that there is no larger universe of items (or tests) to which it is desired to generalize; and, under such circumstances, estimates of reliability, generalizability, or dependability for the model in Equation 14 are usually undefined.

## Summary and Conclusions

Based upon an application of generalizability theory to a rater-by-item matrix of probabilities, equations have been provided and discussed for estimating the expected variability in a cutting score determined by the Nedelsky or Angoff procedure. The development assumes that the cutting score in a particular study is the observed mean (probability) over raters and items and that this mean may be viewed as an estimate of an "idealized" cutting score, defined as the mean for a population of raters and a universe of items. In this sense, the expected variability of the observed mean is error variance attributable to a particular application of the procedure used to define a cutting score.

This approach has been applied to data resulting from an experimental application of the Nedelsky and Angoff procedures by 5 raters to a 126-item test. Also, these results have been examined for each procedure separately, and results have been compared over procedures. The results indicate that both the cutting scores and their expected variances are considerably different for the two procedures. It has been postulated that these differences may be explained, in whole or in part, by differences in the ways probabilities are assigned using the two procedures or by differences in the ways minimum competency is conceptualized.

Finally, the influence of different values of $\bar{X}$, and the expected variance in the distribution of $\bar{X}$, on reliability, or measurement dependability, have been examined. To do so, a modification of the Brennan-Kane index of dependability, $\phi(\lambda)$ was developed. It was found that for a given value of $\bar{X}$, an increase in expected variance of $\bar{X}$ results in a decrease in the estimate of $\phi(\lambda)$. However, if both $\bar{X}$ and its variance change, then the estimate of $\phi(\lambda)$ could increase, decrease, or even remain unchanged.

The numerical results reported in this paper are for a single experimental study only. As such, these results clearly do not form a sufficient basis for a full evaluation of either the Nedelsky or the Angoff procedure. Even so, this study does suggest that differences between these procedures may be of greater consequence than their apparent similarities. In particular, the restricted nature of the Nedelsky (inferred) probability scale may constitute a basis for seriously questioning the applicability of this procedure in certain contexts.

## References

Andrews, 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, DC: American Council on Education, 1971.

Brennan, R. L. *Generalizability analyses: Principles and procedures* (ACT Technical Bulletin No. 26). Iowa City, IA: The American College Testing Program, September 1977.

Brennan, R. L. *GAPID: A FORTRAN IV computer program for generalizability analyses with single-facet designs* (ACT Technical Bulletin No. 34). Iowa City, IA: American College Testing Program, October 1979. (a)

Brennan, R. L. *Some issues involving the estimation of variance components with finite population and/or universe sizes* (ACT Technical Bulletin No. 35). Iowa City, IA: The American College Testing Program, December 1979. (b)

Brennan, R. L. Applications of generalizability theory. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art*. Baltimore, MD: The Johns Hopkins University Press, 1980.

Brennan, R. L., & Kane, M. T. An index of dependability for mastery tests. *Journal of Educational Measurement*, 1977, *14*, 277-289. (a)

Brennan, R. L., & Kane, M. T. Signal/noise ratios for domain-referenced tests. *Psychometrika*, 1977, *42*, 609-625. (Errata. *Psychometrika*, 1978, *43*, 289.) (b)

Buck, L. A. *Guide to the setting of appropriate cutting scores for written tests: A summary of the concerns and procedures* (U.S. Civil Service Commission Technical Memorandum 77-4). Washington, DC: U.S. Government Printing Office, 1977.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley, 1972.

Kane, M. T., & Brennan, R. L. Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement*, 1980, *4*, 105-126.

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

National Council on Measurement in Education. Special issue on standard setting. *Journal of Educational Measurement*, 1978, *15*, 237-327.

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

Sirotnik, K. An analysis of variance framework for matrix sampling. *Educational and Psychological Measurement*, 1970, *30*, 891-908.

Sirotnik, K., & Wellington, R. Incidence sampling: An integrated theory for "matrix sampling." *Journal of Educational Measurement*, 1977, *14*, 343-399.

Wilks, S. S. *Mathematical statistics*. New York: Wiley, 1962.

Zieky, M. J., & Livingston, S. A. *Basic skills assessment manual for setting standards*. Princeton, NJ: Educational Testing Service, 1977.

## Acknowledgments

*The second author is currently Director, Southeastern Regional Test Development Center, Educational Testing Service, Atlanta, Georgia.*

## Author's Address

Send requests for reprints or further information to Robert L. Brennan, Director of Measurement Research, American College Testing Program, P. O. Box 168, Iowa City, IA 52243.