Errors of Measurement and Standard Setting in Mastery Testing

Michael Kane
American College Testing Program, Iowa City, Iowa

Jennifer Wilson
National League for Nursing, New York, New York

A number of studies have estimated the dependability of domain-referenced mastery tests for a fixed cutoff score. Other studies have estimated the dependability of judgments about the cutoff score. Each of these two types of dependability introduces error. Brennan and Lockwood (1980) analyzed the two kinds of errors together but assumed that the two sources of error were uncorrelated. This paper extends that analysis of the total error in estimates of the difference between the domain score and the cutoff score to allow for covariance between the two types of error.

Glaser and Nitko (1971) have defined a criterion-referenced test as one that is designed "to yield measurements that are directly interpretable in terms of specified performance standards" (p. 653). A domain-referenced test is given a criterion-referenced interpretation in terms of each person's level of performance on some content domain. For a domain-referenced test, the parameter of interest for each person is the proportion of items in some domain of content that the person could answer correctly. The observed score, the proportion correct on a sample of items from the domain, provides an estimate of the proportion of items in the domain that the examinee could answer correctly. A domain-referenced test that is used to decide whether individuals have attained some particular level of performance is called a *mastery test*.

It is assumed that the domain consists of a large number of discrete tasks or items and that independent random samples can be drawn from the domain. The proportion of items that person p could answer correctly, if exposed to all the items in the domain, is the person's *domain score*, represented by μ_p. The domain score, μ_p, is a parameter defined for the person on the domain.

Mastery of the domain is defined by establishing a *cutoff score*, γ, on the domain. A person whose domain score, μ_p, is at or above the cutoff score, γ, is said to be a *master*, and a person whose domain score is below the cutoff score is said to be a *nonmaster*.

Since it is generally not practical to include the entire domain in a test, the domain score is not directly observable. Rather, decisions about mastery are based on the person's performance on a sample of items from the domain. The *observed score*, X_{pI}, of person p on the Ith sample of n items, in this case a mastery test, is the proportion of items that the person answers correctly on that mastery test. Although μ_p is assumed to be a constant for each person, X_{pI} will typically vary from one sample of items to another.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 8, No. 1, Winter 1984, pp. 107–115

0146-6216/84/010107-09$1.70

In using judgmental standard-setting procedures, the estimated cutoff score, Y_c, is generally based on the judgments of experts who review the items in the test and decide on an appropriate minimum passing level for each item. The estimated cutoff score for a mastery test is simply the sum of the minimum passing levels for the individual items. The value of the estimated cutoff score will depend on the items included in the test and on the raters.

Since the cutoff score, γ, for the domain is not defined in terms of any particular group of raters but rather in terms of a wider population of qualified raters, variability among raters is a source of error in estimating the cutoff score. Similarly, since the cutoff score is defined on the domain as a whole, variability due to the sampling of items is a source of error in estimating the cutoff.

For the pth person taking the Ith mastery test, a mastery decision is made by comparing a fallible estimate of the domain score, X_{pI}, to a fallible estimate of the cutoff score, Y_c. A *false positive* is said to occur when a nonmaster is incorrectly classified as a master (i.e., when μ_p is less than γ and X_{pI} is greater than or equal to Y_c). A *false negative* is said to occur when a master is incorrectly classified as a nonmaster (i.e., when μ_p is greater than or equal to γ, and X_{pI} is less than Y_c).

Almost all of the literature discussing the reliability of domain-referenced test scores assumes a fixed cutoff score, known a priori, and analyzes either the consistency of classification across tests (e.g., Hambleton & Novick, 1973; Huynh, 1976, 1978; Millman, 1973; Subkoviak, 1976), or the consistency in observed deviations from the fixed cutoff score (e.g., Brennan & Kane, 1977a, 1977b; Kane & Brennan, 1980; Livingston, 1972). In the first approach, the focus is on the accuracy of decisions, where accuracy is defined either in terms of the proportion of examinees who are consistently classified as masters or nonmasters, or in terms of Cohen's Kappa. The second approach analyzes the difference between the observed score and the cutoff score in terms of the sources of variance (including various sources of error variance) contributing to the difference. The analyses presented below follow the second approach and make it possible to identify a number of sources of error in estimating the difference between the observed score and the cutoff score.

The precision of estimates of the cutoff score based on judgmental standard-setting procedures has also been investigated (Andrew & Hecht, 1976; Brennan & Lockwood, 1980; Shepard, 1980). However, with one exception, the error of measurement and the errors in standard setting have not been addressed together. The one exception is the study by Brennan and Lockwood (1980) in which errors in the test scores and errors in the estimated cutoff score were analyzed together. However, Brennan and Lockwood assumed that these two kinds of errors are uncorrelated. As will be discussed later, this assumption is unrealistic because the observed score and the cutoff score are generally based on the same sample of items.

This paper evaluates the magnitude of the total error in estimates of the difference between an examinee's domain score and the cutoff score, assuming that the estimate of the universe score and the judgmental standard are based on the same sample of items. The work of Brennan and Lockwood (1980) is extended by explicitly considering the covariance across items between errors of measurement and errors in standard setting. The implications of these results for the probability of misclassification are also discussed briefly.

Domain Scores and Errors of Measurement

The linear model given below partitions the observed scores into a number of effects and uses the framework of generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972):

$$X_{pI} = \mu + \alpha_p + \alpha_I + \alpha_{pI}, \quad (1)$$

where

μ is the grand mean over persons and items,
α_p is the main effect for persons,
α_I is the main effect for samples of n items, and
α_{pI} represents the interaction between persons and samples of items.

(Note that the observed score for a single item would be stated in terms of effects, α_i and α_{pi} for single items). Because the effects are defined in terms of deviations from mean scores, the expected value of an effect over any of its subscripts is zero, and the correlation, over persons and items, of any two of the effects is zero.

The results of a random effects ANOVA with persons crossed with items would typically be stated in terms of the variance components, $\sigma^2(\alpha_p)$, $\sigma^2(\alpha_i)$, and $\sigma^2(\alpha_{pi})$, for sampling individual persons and items. The variance component for the average effect over a sample is inversely proportional to the size of the sample. For example, for a sample of n items:

$$\sigma^2(\alpha_I) = \sigma^2(\alpha_i)/n \tag{2a}$$

$$\sigma^2(\alpha_{pI}) = \sigma^2(\alpha_{pi})/n \ . \tag{2b}$$

The pI effect and the I effect are sources of error, and the relationships in Equation 2 are examples of the Spearman-Brown correction; as the number of items increases, the error variance decreases. In general, the equations in this paper are expressed in terms of the variance components for samples of items rather than for single items, but the impact of increasing sample sizes can always be made explicit (e.g., using Equation 2.)

The domain score for the pth person, μ_p is defined as the expected value of the observed score, X_{pI}, taken over the domain of items. Using the model in Equation 1, the domain score is

$$\mu_p = E_I(X_{pI}) = \mu + \alpha_p \tag{3}$$

and the expected value of the domain score over the population is equal to the grand mean, μ. Therefore, the domain score variance is given by

$$E_p(\mu_p - \mu)^2 = E_p(\alpha_p^2) = \sigma^2(\alpha_p) \ . \tag{4}$$

The domain score variance is the same for crossed or nested designs. The domain score variance depends on the domain definition and on the definition of the population, but it does not depend on the sampling designs used to estimate the domain scores.

Cronbach et al. (1972, p. 84) defined the error in estimates of domain scores as the difference between the observed score and the domain score:

$$\Delta_{pI} = X_{pI} - \mu_p \ . \tag{5}$$

The variance of Δ_{pI}, taken over the population and the domain, is the same for nested and crossed designs and is given by

$$\sigma^2(\Delta) = \sigma^2(\alpha_I) + \sigma^2(\alpha_{pI}) \ . \tag{6}$$

Therefore, the error variance for point estimates involves both the variance of the item main effect and the person-item interaction variance.

For a crossed design, the systematic error, α_I, affects all observed scores in the same way and therefore tends to bias observed scores in one direction or the other. For example, if the sample of items included in the test is particularly easy, α_I will be positive, and the observed scores will tend to be higher than they would be if the items were of average difficulty. This kind of bias will increase the probability of false positives and will decrease the probability of false negatives. It will also increase the size of the errors in estimates of the difference between the domain score and the cutoff score. Therefore, the systematic error, α_I, is included in the error variance for mastery tests, regardless of whether the sampling

of items is nested within persons or crossed with persons, and an appropriate error variance for estimates of the observed scores is $\sigma^2(\Delta)$.

The Cutoff Score and its Estimation

The three most commonly discussed procedures (Angoff, 1971; Ebel, 1972; Nedelsky, 1954) for setting a cutoff score are based on judgments about what constitutes minimally competent performance and therefore are all influenced by variability among and within raters. The Angoff procedure will be used as the basis for subsequent discussion, but the issues to be discussed also apply to the Nedelsky procedure, and in a modified form to the Ebel procedure. In the Angoff procedure, expert judges are asked to consider the expected level of performance on each item (the probability of answering the item correctly) of hypothetical "minimally competent candidates." The judges are instructed to assign a minimum passing level (MPL) for each item in terms of the probability that a minimally competent candidate could answer that item correctly. Since the cutoff score for the sample of items defining a test is simply the sum of the MPLs for the individual items, it will depend on the sample of items and on the sample of raters.

Note that unless a behavioral interpretation of the test scores is available, the results of a judgmental standard-setting procedure do not indicate the kind of behavior that distinguishes passing candidates from failing candidates. Although individual raters undoubtedly use some behavioral standards in setting the MPL for each item (e.g., their individual experiences with persons considered to be minimally competent), the judgmental standard-setting procedures do not provide a mechanism for making these behavioral standards explicit. Therefore, the interpretation of the standard depends on the criteria for selecting judges. Without an independently developed behavioral interpretation, the burden of interpretation falls on the new reference population, the population of raters.

As their name indicates, the standard-setting procedures are not designed to estimate the difficulty level of the items; rather, they are designed to provide a systematic approach to the task of establishing a standard. Nevertheless, it would be expected that the item MPLs would be positively correlated with item difficulty levels. The judges assigning MPLs to items are indicating how difficult they think each item would be for a hypothetical minimally competent candidate. The judges should assign relatively high MPLs to items that they perceive to be relatively easy and relatively low MPLs to items that they perceive to be relatively difficult. Assuming that the judges operate at better than the chance level in estimating item difficulty, the MPLs should be positively correlated with the item difficulty level. As shown later in this paper, a positive covariance between item MPLs and item difficulties will reduce the overall error in evaluating performance relative to the standard set by the judges.

The cutoff score estimate for raters, R, and items, I, can be represented by a linear model that parallels that given in Equation 1 for observed scores,

$$Y_{RI} = \gamma + \beta_R + \beta_I + \beta_{RI} \ , \tag{7}$$

where

γ is the grand mean over samples of raters and items,

β_I is the main effect for items,

β_R is the main effect for raters, and

β_{RI} is the item-rater interaction.

The "true" value of the cutoff score can be defined as the expected value of Y_{RI} over the domain of items and over the population of raters, and is equal to the constant, γ. Therefore, given a domain of items and a population of raters, there are three identifiable sources of error in estimating the cutoff score:

the item effect, the rater effect, and the item-rater interaction. The contribution of the different sources of error to the variance in estimated cutoff scores is given by

$$\sigma^2(\Delta_{RI}) = \sigma^2(\beta_R) + \sigma^2(\beta_I) + \sigma^2(\beta_{RI}) \ . \quad (8)$$

Each of the variance components in Equation 8 represents the variance of the average value of an effect over a sample of items and a sample of raters.

The assumption, made in this paper, that items are randomly sampled from a domain forms the basis for most discussions of the reliability of mastery tests. An alternative approach that could enhance the reliability of mastery decisions would involve selecting items with an MPL close to .5. The discriminating power of an item for a population of examinees tends to be highest when the item has a difficulty level close to .5; items that either are very easy or very difficult are not very effective in making differential decisions. Items with MPLs near .5 would tend to be maximally discriminating for the marginal candidates because, by definition, the probability that a marginal candidate will be able to answer a question is given by its MPL. By selecting items with MPLs near .5, the precision of a mastery test would be maximized near the cutoff score where it is most critical. However, if this procedure is adopted, the scores on the examination would not provide good estimates of the examinees' domain scores, since items with MPLs near .5 cannot be considered a random (or representative) sample from the domain.

The Total Error

The question that is of central interest for mastery decisions is whether the difference between an examinee's domain score and the cutoff score, given by $\mu_p - \gamma$, is positive or negative. The magnitude of this difference indicates the strength of the signal to be detected in making decisions about mastery. From Equations 6 and 8, it can be seen that there are five distinct sources of error in estimating this signal. Brennan and Lockwood (1980) included all five of these sources of error in their analysis but assumed that these five sources of error are uncorrelated. In general, however, α_I, the item effect for observed scores, and β_I, the item effect for ratings of the average MPL, will be positively correlated; that is, the raters' estimates of how difficult an item would be for a minimally competent examinee will be positively related to the average difficulty of the item over the population of examinees.

The total error in estimating the difference score, $\mu_p - \gamma$, is given by

$$(X_{pI} - Y_{RI}) - (\mu_p - \gamma) = \alpha_I + \alpha_{pI} - \beta_R - \beta_I - \beta_{RI} \ . \quad (9)$$

Since the covariance between any two terms that do not have identical subscripts is zero, the total error variance is given by

$$\sigma^2(\Delta_T) = \sigma^2(\alpha_{pI}) + \sigma^2(\beta_R) + \sigma^2(\beta_{RI}) + \sigma^2(\alpha_I) + \sigma^2(\beta_I) - 2 \text{ cov } (\alpha_I, \beta_I) \ , \quad (10)$$

where cov (α_I, β_I) is the covariance between the item effect in observed scores and the item effect in the estimated cutoff score.

If the covariance is zero, the last term in Equation 10 disappears, and the total error is that reported by Brennan and Lockwood (1980). At the other extreme, if α_I is equal to β_I, the last three terms in Equation 10 cancel out, and the total error will involve only the first three terms in Equation 10. Therefore, to ignore a positive covariance would yield an inflated estimate of the error variance.

The item effect for observed scores indicates how easy the items are, and the item effect for the ratings represents the average score expected of a hypothetical minimally competent examinee. As discussed earlier, the estimates of the probability that a minimally competent examinee answers an item correctly should be higher for items that the raters consider relatively easy and lower for items that the raters consider relatively difficult. Assuming that the raters exhibit some accuracy in evaluating the relative difficulty of the items, the covariance in Equation 10 is likely to be positive, thus decreasing the total error variance.

To the extent that $\text{cov}(\alpha_i, \beta_i)$ is positive, the standard-setting procedure provides a way to correct for differences in difficulty from one set of items to another and thereby to control one source of systematic error. However, if the covariance is negative, judgmental standard setting would exacerbate the problems caused by unequal item difficulty. In this case, the cutoff score would tend to be low for sets of easy items, thus making it even easier to be classified as a master on these items, and would tend to be high for difficult items, thus making it even more difficult to be classified as a master on these items. More seriously perhaps, a negative covariance would suggest that the item characteristics being emphasized by the raters in determining MPLs are not the item characteristics that determine student performance. Therefore, a negative value for $\text{cov}(\alpha_i, \beta_i)$ would cast doubt on either the validity of the test or on the appropriateness of the criteria used in standard setting.

Note that Equation 10 indicates the contribution to the total error of various sources of variance and therefore provides guidance on how to control the magnitude of the total error. For example, if $\sigma^2(\beta_R)$ were particularly large, indicating that the judges varied considerably in the standards they were applying, it would be reasonable to take steps to reduce this component of the error. This might be accomplished by giving the judges more training, or perhaps more effective training. Of course, as indicated by Equation 2, the rater variance could also be reduced by increasing the number of raters used to set the standard.

Estimation Issues

As noted earlier, the variance components in Equation 10 can be estimated from two random effect ANOVAs (Brennan, 1983; Cronbach, et al., 1972). An ANOVA of the item responses with persons crossed with items yields unbiased estimates of $\sigma^2(\alpha_p)$, $\sigma^2(\alpha_i)$, and $\sigma^2(\alpha_{pi})$. An ANOVA of the ratings, with items crossed with raters, yields unbiased estimates of $\sigma^2(\beta_i)$, $\sigma^2(\beta_r)$, and $\sigma^2(\beta_{ri})$. These variance components can then be modified, using relationships like those in Equation 2, to reflect the sample sizes being used.

As shown below, an unbiased estimate of $\text{cov}(\alpha_i, \beta_i)$ in terms of sample statistics is given by

$$\widehat{\text{cov}}(\alpha_i,\beta_i) = \frac{1}{n-1}\sum_i (X_{Pi} - X_{PI})(Y_{Ri} - Y_{RI}) \ , \tag{11}$$

where

X_{Pi} is the average score over the sample of persons on the ith item,

Y_{Ri} is the average rating on the ith item,

X_{PI} is the average score over the sample of persons and items, and

Y_{RI} is the average rating over the sample of raters and items.

Expanding Equation 11 in terms of effects yields

$$\widehat{\text{cov}}(\alpha_i,\beta_i) = \frac{1}{n-1}\sum_i [(\alpha_i - \alpha_I) + (\alpha_{Pi} - \alpha_{PI})][(\beta_i - \beta_I) + (\beta_{Ri} - \beta_{RI})] \ . \tag{12}$$

Taking the expected value of Equation 12 over R and P, and recalling that the expected value of an effect over any of its subscripts is zero, gives

$$E_P E_R \widehat{\text{cov}}(\alpha_i,\beta_i) = \frac{1}{n-1}\sum_i (\alpha_i - \alpha_I)(\beta_i - \beta_I) \ . \tag{13}$$

Now, taking the expected value of Equation 13 over samples of n items gives

$$\begin{aligned} E_I[E_P E_R \widehat{\text{cov}}(\alpha_i,\beta_i)] &= \frac{1}{n-1}\sum_i E_I(\alpha_i - \alpha_I)(\beta_i - \beta_I) \\ &= \frac{1}{n-1}\sum_i [E_I \alpha_i\beta_i - E_I \alpha_i\beta_I - E_I \alpha_I\beta_i + E_I \alpha_I\beta_I] \ . \end{aligned} \tag{14}$$

Then,

$$E_I\,\alpha_i\beta_I = E_I\,\alpha_i\frac{1}{n}\sum_j \beta_j = \frac{1}{n}E_i\,\alpha_i\beta_i + \frac{1}{n}\sum_{j\neq i}E_j\,\alpha_i\beta_j \ . \tag{15}$$

Since taking the expectation over I involves taking n expectations (one for each of the n independently sampled items), and since the expectation of $\alpha_i\beta_j$, over i or j is zero for i not equal to j, this gives

$$E_I\,\alpha_i\beta_I = \frac{1}{n}E_i\,\alpha_i\beta_i \ . \tag{16}$$

Similarly,

$$E_I\,\alpha_I\,\beta_i = E_I\,\alpha_I\beta_I = \frac{1}{n}E_i\,\alpha_i\beta_i \ . \tag{17}$$

Substituting Equations 16 and 17 in Equation 14 yields

$$E_I\,E_P\,E_R\,\widehat{\mathrm{cov}}(\alpha_i,\beta_i) = \frac{1}{n-1}\sum_i[E_i\alpha_i\beta_i - \frac{1}{n}E_i\alpha_i\beta_i]$$

$$= E_i\,\alpha_i\beta_i \ , \tag{18}$$

which is, by definition the covariance of α_i with β_i. Therefore, the estimator defined in Equation 11 is an unbiased estimate of $\mathrm{cov}(\alpha_i,\beta_i)$. An estimate of the covariance between α_I and β_I, for a sample of n items, can then be obtained using the relationship

$$\mathrm{cov}(\alpha_I,\beta_I) = \frac{1}{n}\mathrm{cov}(\alpha_i,\beta_i) \ , \tag{19}$$

which is analagous to the relationship for variance components given in Equation 2.

Effect of Errors on the Probability of Misclassification

A detailed analysis of the effects of errors of measurement and errors in standard setting on the probability of misclassification is beyond the scope of this paper. The discussion below is intended to indicate the relevance of both types of errors to the probability of misclassification without providing the kind of detailed analysis that would require strong distributional assumptions.

Consider a person, p, whose universe score, μ_p, is above the cutoff score, γ; that is $\mu_p - \gamma$ is positive. What is the probability that this person will be misclassified as a nonmaster? Such a false negative result will occur for person p on test I if X_{pI} is less than Y_{RI}. It would be desirable to approximate the probability that $X_{pI} - Y_{RI}$ is less than zero given that $\mu_p - \gamma$ is greater than zero.

The expected value over R and I of the observed deviation score for person, p, is given by

$$E_{RI}\,(X_{pI} - Y_{RI}) = E_{RI}X_{pI} - E_{RI}Y_{RI} = \mu_p - \gamma \ . \tag{20}$$

Therefore, the expected value of the observed deviation score is positive whenever the universe deviation score is positive. In order for a false negative to occur for person p the observed deviation score, $X_{pI} - Y_{RI}$, must be less than zero. Since the distribution of observed deviation scores for person p has a mean of $\mu_p - \gamma$, the probability of a false negative for person p is equal to the probability of obtaining an observed deviation score that is below the mean deviation score for person p by a distance greater than $\mu_p - \gamma$. The exact value of this probability will, of course, depend on the shape of the distribution of the observed deviation scores for person p. In fact, the probability of a false negative is given by the value of the appropriate cumulative probability distribution at a point $(\mu_p - \gamma)$ below the mean.

Without making any specific distributional assumptions, however, it is possible to draw some general conclusions about the effect of the two types of errors on the probability of a false negative. In particular, the probability of obtaining an observed deviation score that is below the mean of the observed deviation score distribution for person p by a distance that is greater than $\mu_p - \gamma$ will be an increasing function of

the variance of the person's observed deviation score distribution over samples of items and raters. This variance is given by the total error variance for person p. Since the probability of a false negative is an increasing function of the total error variance, all of the components included in Equation 10 will contribute to the probability of a false negative. A similar analysis applies to the probability of false positives, which is also an increasing function of the total error variance.

Conclusions

This paper has extended the work of Brennan and Lockwood (1980) in analyzing the errors in criterion-referenced mastery tests in terms of both errors of measurement and errors in standard setting. The analysis presented here goes beyond that produced by Brennan and Lockwood (1980) in suggesting that the covariance between the item effect, α_i, involved in errors of measurement and the item effect, β_i, involved in errors in standard setting might be correlated.

An analysis of the total error in terms of variance components and covariances indicates the contribution of specific sources of error included in the analysis to the total error variance. It therefore provides a basis for controlling the total error variance by controlling those sources of error that make the largest contribution to the total error. Since the probability of misclassification depends on the total error, it also provides a way of decreasing the probability of misclassification. For example, if it is found that the person-item interaction variance, $\sigma^2(\alpha_{pI})$, is very large compared to all other terms in Equation 10, increasing the number of items would be more effective in decreasing the total error variance, and thereby the probability of misclassification, than would a proportional increase in the number of raters used to establish the cutoff score.

In the analyses presented here, the covariance between the item main effect in errors of measurement and the item main effect in errors in standard setting plays a significant role. It has been argued that this covariance should be positive because the item difficulty should, in general, be positively related to the MPL representing the judges' estimates of how difficult the item would be for a particular subgroup of examinees, namely, those who are minimally competent. Assuming that the covariance is positive, the total error will be decreased by the magnitude of the covariance term.

The covariance term is important because a positive covariance will decrease the total error variance and thereby the probability of misclassification, whereas a negative covariance will increase the total error variance and thereby the probability of misclassification. Of more importance, perhaps, the covariance term provides an empirical test of the reasonableness of the overall mastery decision process. A negative value for the covariance term suggests that the criteria being used by judges to set the cutoff score are not consistent with the attribute being measured by the items, as reflected in their relative difficulty levels. It therefore suggests either that the domain of items has not been defined appropriately or that the judges are using inappropriate criteria in setting the cutoff score. The covariance term thus provides a check on the appropriateness, or validity, of the interpretation that is applied to mastery decisions based on domain-referenced tests and judgmental standard-setting procedures.

References

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

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

Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa City IA: American College Testing Program.

Brennan, R. L., & Kane, M. T. (1977a). An index of

dependability for mastery tests. *Journal of Educational Measurement, 14,* 277–289.

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

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.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana: University of Illinois Press.

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

Ebel, R. L. (1972). *Essentials of educational measurement*. Englewood Cliffs NJ: Prentice Hall.

Glaser, R., & Nitko, A. J. (1971). Measurement in learning and instruction. In R. L. Thorndike (Ed.), *Educational measurement*. Washington DC: American Council on Education.

Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement, 10,* 159–170.

Huynh, H. (1976). Statistical consideration of mastery scores. *Psychometrika, 41,* 65–78.

Huynh, H. (1978). Reliability of multiple classifications. *Psychometrika, 43,* 317–325.

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

Livingston, S. A. (1972). A criterion-referenced application of classical test theory. *Journal of Educational Measurement, 9,* 13–26.

Millman, J. (1973). Passing scores and test lengths for domain-referenced measures. *Review of Educational Research, 43,* 205–216.

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

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

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement, 13,* 265–276.

Author's Address

Send requests for reprints or further information to Michael T. Kane, ACT, P.O. Box 168, Iowa City IA 52243, U.S.A.