An Estimator of Examinee-Level Measurement Error Variance That Considers Test Form Difficulty Adjustments

David Jarjoura
Northeastern Ohio Universities College of Medicine

A model and estimator for examinee-level measurement error variance are developed. Although the binomial distribution is basic to the modeling, the proposed error model provides some insights into problems associated with simple binomial error, and yields estimates of error that are quite distinct from binomial error. By taking into consideration test form difficulty adjustments often used in standardized tests, the model is linked also to indices designed for identifying unusual item response patterns. In addition, average error variance under the model is approximately that which would be obtained through a KR-20 estimate of reliability, thus providing a unique justification for this popular index. Empirical results using odd-even and alternate-forms measures of error variance tend to favor the proposed model over the binomial.

When drawing inferences from a particular examinee's observed score, it is critical that the score not be interpreted literally. The question of how much to allow for measurement error for a particular examinee has not been studied to the same degree as average measurement error variance (taken across some population of examinees). Often the average error variance is used in interpreting all examinees' observed scores, even in light of empirical evidence that error variance varies across examinees. Part of the problem is that, in contrast to measures of average error variance, there are no well-established procedures for estimating examinee-level error variance. Straightforward methods for estimating error variance at different performance levels have been suggested (e.g., fitting a polynomial regression to squared differences between scores on alternate forms as a function of performance level). Also, observed number-correct error models such as the binomial and the compound binomial provide means for estimating performance-level error variance. These types of procedures and error models will be compared and related to the estimator of examinee-level error variance that is derived here.

Within the framework of item response theory (IRT), practical applications of the individualized error that can be obtained from the information function of the maximum likelihood ability estimator have been described (e.g., Samejima, 1977; Weiss & Kingsbury, 1985). The modeling used here is outside the framework of IRT; rather, it provides examinee-level error within an item sampling framework.

Recently, several indices have been suggested for identifying examinees with unusual item response patterns, that is, patterns that indicate potential difficulty in interpreting observed scores (for a review see Tatsuoka & Linn, 1983). As measures of examinees' response consistency, these indices can also be related to the estimator of examinee-level error variance derived here.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 10, No. 2, June 1986, pp. 175–186

0146-6216/86/020175-12$1.85

Because the binomial error model serves as the foundation of the modeling, the present derivation provides insights into the nature of past criticisms of this model. As an example, under special circumstances the binomial error model can be said to hold by design (Lord, 1957; Lord & Novick, 1968, Chap. 11, Chap. 23, p. 254). If test forms are constructed by random sampling of items and the proportion-correct true score of interest is defined by the domain (of infinite size) from which items are sampled (rather than for a particular set of items), the binomial error model holds for a particular examinee as long as item responses are not made dependent by context and other similar effects.

The binomial error model is often criticized because items are not the same difficulty in a population of examinees. This issue is irrelevant when the focus is on a particular examinee for whom items are randomly sampled from some domain, that is, when the distribution of observable scores for a particular examinee follows the binomial. However, even if random item sampling were an acceptable description of test construction practice, the binomial model ignores the fact that adjustments for test form differences in difficulty are often made in standardized test scores.

That a sample of examinees is administered the same sample of items is an important aspect of score adjustments and is a major consideration in the modeling here. As will be shown, the average covariance of error across examinees administered the same sample of items is a component of error for adjusted scores. Feldt (1984) provided further discussion and development on this issue. Still, the binomial error model has been shown to be a good approximation in many cases (Huynh & Saunders, 1976; Keats, 1957; Keats & Lord, 1962; Subkoviak, 1978).

On a different issue, Lord (1957) defended the binomial error model against the intuitive notion that an examinee-level error model should result in larger error variance for an examinee who does much guessing than for one who does not (given both have the same true score). In the present paper it is shown how the proposed error model, which uses binomial model arguments, tends to result in larger error variances for examinees who guess substantially. An example provides evidence that the binomial error model yields error variance estimates that are larger for middle scoring examinees than alternate forms and odd-even estimates of error variance, as well as the model-based estimates developed here.

Development of the Error Model

Assume that interest is isolated to examinee a, and envision a process of examination whereby he/she could be administered any random sample of items from some infinite domain. It can then be argued that the binomial error variance is appropriate, that is, the observable proportion-correct scores for examinee a have variance $\mu_a(1 - \mu_a)/I$, where μ_a is his/her proportion-correct true score (over the item domain) and I is the number of items administered. Of course, this holds even if a group of examinees is administered the same set of items.

When a group of examinees is administered the same set of items, information about the difficulty of these items can be used for adjusting scores. Given that the group represents a random sample from some population of interest, the mean difficulty of the items (mean proportion correct) could be used to make a relative difficulty adjustment in this group's observed scores as long as there are data pertaining to the difficulty of at least one other set of items in the domain. Certainly, many types of adjustments could be made, but attention here is focused on this simple difficulty adjustment because it is fundamental to the kinds of score adjustments that are made in practice, and because it provides a starting point for studying other more complex adjustments.

The observable proportion-correct score for examinee a is

$$X_{aI} = \sum_i X_{ai} / I \quad , \tag{1}$$

where X_{ai} is a correct/incorrect (1/0) response to item i. The mean proportion-correct score across all examinees administered the same form is denoted as

$$X_{AI} = \sum_{a} X_{aI} / A \quad , \tag{2}$$

where there are A examinees contributing to the mean. It is assumed that this quantity represents a simple random sample from an infinite population of examinees.

Consider scores that are adjusted for the estimated mean difficulty of a particular test form. For examinee a the adjusted score is

$$X'_{aI} = X_{aI} - X_{AI} + c \quad , \tag{3}$$

where c is a constant applied to scores of all examinees across all possible test forms. Because examinees' scores from different forms of the same test are compared, it is important to have some adjustment (equating) for the variability of the difficulty of these forms. Equation 3 provides only a basic type of adjustment for test form difficulty. Other more complex considerations can provide test form difficulty adjustments that vary along the score scale.

The expected proportion-correct score across all items in the domain and the population of examinees is denoted as

$$\mu \equiv \mathrm{E}X_{ai} \quad . \tag{4}$$

If $c = \mu$, X'_{aI} is an unbiased estimator of the proportion-correct true score, which is denoted as

$$\mu_a \equiv \mathrm{E}(X_{ai} \mid a) \quad , \tag{5}$$

for any examinee a. Under many circumstances it does not seem important to have $c = \mu$ as long as c is constant across all examinees' adjusted scores, for then relative comparisons are not biased.

The measurement error variance of X'_{aI} for examinee a is

$$\begin{aligned} \sigma^2(e)_a \equiv \mathrm{VAR}(X'_{aI} \mid a) &= \mathrm{VAR}(X_{aI} \mid a) + \mathrm{VAR}(X_{AI} \mid a) - 2\mathrm{COV}(X_{aI}, X_{AI} \mid a) \\ &\approx \frac{\mu_a(1 - \mu_a)}{I} + \frac{\sigma^2(i)}{I} - \frac{2\mathrm{COV}(X_{ai}, \mu_i \mid a)}{I} \quad , \end{aligned} \tag{6}$$

where the variance (VAR) is conditional on a, but is taken across all items in the domain and, where appropriate (e.g., X_{AI}), across all other examinees in the population; similarly for the covariance (COV); and where

$$\mu_i = \mathrm{E}(X_{ai} \mid i) \quad , \tag{7}$$

and

$$\sigma^2(i) = \mathrm{VAR}(\mu_i) \quad . \tag{8}$$

Terms with A in the denominator have not been included in Equation 6 because A is expected to be large and because these terms are not basic components of error variance. (All terms are given in the Appendix.) Note that in contrast to the binomial error variance, $\mu_a(1 - \mu_a)/I$, examinees with the same true score can have quite different error variances depending on the covariance term.

In order to understand the covariance term better, define the residual effect as

$$r_{ai} = X_{ai} - \mu_a - \mu_i + \mu \quad , \tag{9}$$

and make the usual assumptions that

$$\underset{i \neq i'}{\mathrm{COV}}(r_{ai}, r_{ai'} \mid a) = \underset{a \neq a'}{\mathrm{COV}}(r_{ai}, r_{a'i} \mid i) = \underset{i \neq i',\ a \neq a'}{\mathrm{COV}}(r_{ai}, r_{a'i'}) = 0 \quad , \tag{10}$$

where the covariance is taken across all possible pairs of items and/or examinees. It is not assumed that the variance of residuals is a constant for each examinee, and therefore the examinee-level residual variance is defined as

$$\sigma^2(r)_a \equiv \mathrm{VAR}(r_{ai} \mid a) \quad . \tag{11}$$

Note that the usual residual variance component for an examinees × items sampling design, taken across both examinees and items, is defined as

$$\sigma^2(r) \equiv \mathrm{VAR}(r_{ai}) = \mathrm{E}\sigma^2(r)_a \quad . \tag{12}$$

The covariance of item difficulty (μ_i) and residual, across examinees and items, is zero by definition of the residual; but for a particular examinee this does not hold necessarily. For examinee a, define

$$\sigma(i,r)_a \equiv \mathrm{COV}(\mu_i, r_{ai} \mid a) \quad , \tag{13}$$

but note that

$$\mathrm{E}\sigma(i,r)_a = 0 \quad , \tag{14}$$

because $\mathrm{E}(r_{ai} \mid i) = 0$ for all i.

This digression allows the covariance term in Equation 6 to be written as

$$\mathrm{COV}(X_{ai}, \mu_i \mid a) = \sigma^2(i) + \sigma(i,r)_a \quad , \tag{15}$$

yielding the following reexpression of Equation 6:

$$\sigma^2(e)_a \approx \frac{\mu_a(1 - \mu_a)}{I} - \frac{\sigma^2(i)}{I} - \frac{2\sigma(i,r)_a}{I} \quad . \tag{16}$$

By Equation 14, the average error variance across examinees is

$$\mathrm{E}\sigma^2(e)_a \approx \frac{\mathrm{E}\mu_a(1 - \mu_a)}{I} - \frac{\sigma^2(i)}{I} \quad . \tag{17}$$

Therefore a mean difficulty adjustment makes *average* error variance smaller by $\sigma^2(i)/I$ as compared with binomial error variance. When A is not large this need not hold (see the Appendix). Note that $\sigma^2(i)/I$ is the average covariance of errors of unadjusted scores taken across all potential pairs of examinees administered the same items, that is,

$$\underset{a \neq a'}{\mathrm{COV}}(X_{aI} - \mu_a, X_{a'I} - \mu_{a'}) = \sigma^2(i)/I \quad . \tag{18}$$

The covariance $\sigma(i,r)_a$ can make $\sigma^2(e)_a$ for an examinee larger than the corresponding binomial variance. As an example, an examinee who randomly guesses for all items has $\mathrm{COV}(X_{ai}, \mu_i \mid a \text{ guesses}) = 0$ which means that $\sigma(i,r)_a = -\sigma^2(i)$. Therefore, given a guesses,

$$\sigma^2(e)_a = \mu_a(1 - \mu_a)/I + \sigma^2(i)/I \quad , \tag{19}$$

that is, random responses to items result in larger error variance for mean difficulty adjusted scores than for unadjusted scores.

Estimates of $\sigma^2(e)_a$ and Other Components

The error variance for mean adjusted scores can be expressed in a simpler form by noting that

$$\sigma^2(e)_a = \mathrm{VAR}(X_{aI} - X_{AI} \mid a) = \mathrm{VAR}\left[\sum_i r_{ai}/I - \sum_a \sum_i r_{ai}/(AI) \mid a\right] \quad . \tag{20}$$

By ignoring $1/A$ terms,

$$\sigma^2(e)_a \approx \sigma^2(r)_a/I \quad . \tag{21}$$

This result suggests a procedure for estimating $\sigma^2(e)_a$. The well-known unbiased estimator of average residual variance from the residual mean-square in an examinees × items design has the form

$$\hat{\sigma}^2(r) = \mathrm{MS}(r) = \sum_a \sum_i (X_{ai} - X_{aI} - X_{Ai} + X_{AI})^2/[(A - 1)(I - 1)] \quad . \tag{22}$$

For examinee a,

$$\mathrm{MS}(r)_a/I = \sum_i (X_{ai} - X_{aI} - X_{Ai} + X_{AI})^2/[I(I - 1)] \tag{23}$$

might then be used as an estimator of $\sigma^2(e)_a$. In fact, it can be shown that

$$\hat{\sigma}^2(e)_a = \mathrm{MS}(r)_a/I + \hat{\sigma}^2(a)/A \tag{24}$$

is an unbiased estimator of $\sigma^2(e)_a$ where $\hat{\sigma}^2(a)$ is the usual unbiased estimator of $\sigma^2(a) \equiv \mathrm{VAR}(\mu_a)$, that is,

$$\hat{\sigma}^2(a) = \sum_a (X_{aI} - X_{AI})^2/(A - 1) - \hat{\sigma}^2(r)/I \quad . \tag{25}$$

(All terms are included in Equation 24.)

Note that $\mathrm{MS}(r)/I$ is the error variance that is obtained from a KR-20 estimate of reliability, that is, it is identical to the observed score variance × (1 − KR-20). Because

$$\mathrm{MS}(r) = \sum_a \mathrm{MS}(r)_a/(A - 1) \quad , \tag{26}$$

these results provide an interesting and novel partial justification for KR-20 as a measure of reliability with large A. To complete the justification, note that the denominator of KR-20 $[\Sigma_a (X_{aI} - X_{AI})^2/(A - 1)]$ is a consistent (with increasing A) estimate of $\mathrm{VAR}(X_{aI})$, the relevant observed score variance for adjusted scores. Feldt (1984) provided a related motivation for KR-20.

Recall that in Equation 16, $\sigma(i,r)_a$ individualizes error variance for examinees with the same true score. Thus estimates of this covariance would be informative about differences in error variances. The residual mean square for examinee a can be broken down as follows:

$$\mathrm{MS}(r)_a = \frac{1}{I - 1}\left[\sum_i (X_{ai} - X_{aI})^2 - 2\sum_i (X_{ai} - X_{aI})(X_{Ai} - X_{AI}) + \sum_i (X_{Ai} - X_{AI})^2\right] \quad . \tag{27}$$

It is easy to show that the expected value of the cross-product term for a given a is

$$\mathrm{E}\left[\frac{\sum_i (X_{ai} - X_{aI})\,(X_{Ai} - X_{AI})}{I - 1} \,\Big|\, a\right] = \sigma^2(i) + \frac{(A - 1)\sigma(i,r)_a}{A} + \frac{\sigma^2(r)_a}{A} \quad , \tag{28}$$

and this term can therefore be used for unbiased estimation of $\sigma(i,r)_a$. (See the Appendix for an unbiased estimator of $\sigma^2(r)_a$.) This cross-product term is of special interest because it appears in the numerator of certain caution indices reviewed in Tatsuoka and Linn (1983). Essentially, this term estimates the covariance between an examinee's item responses and item difficulties. If an examinee's estimate is in some sense small, then the responses are atypical as compared to the sample of examinees on which the item difficulties are based. This explains its use in caution indices. From Equation 16, systematic variation in this covariance is primarily a result of variation in $\sigma(i,r)_a$ for large samples.

The other two components of Equation 16 can be estimated from the two other terms of $\mathrm{MS}(r)_a$. An unbiased estimator of the binomial error variance is

$$\frac{\widehat{\mu_a(1 - \mu_a)}}{I} = \frac{X_{aI}(1 - X_{aI})}{I - 1} = \frac{\sum_i (X_{ai} - X_{aI})^2}{I\,(I - 1)} \quad , \tag{29}$$

and an unbiased estimator of item difficulty variance is

$$\hat{\sigma}^2(i) = \frac{\sum_i (X_{Ai} - X_{AI})^2}{I - 1} - \frac{\hat{\sigma}^2(r)}{A} \quad . \tag{30}$$

Error in Estimating $\sigma^2(e)_a$

Refer to Equations 23 and 24 and note that $\hat{\sigma}^2(e)_a$ is a sum across items of the interaction function (divided by a constant) plus a term that can be ignored with large A. With large A, each item's contribution to the sum is essentially independent of the others. Therefore, the variance of $\hat{\sigma}^2(e)_a$ for examinee a can be estimated by simply calculating the variance of the interaction function across items,

$$\widehat{\text{VAR}}\,[\hat{\sigma}^2(e)_a] = \frac{\sum_i(\text{INT}_i - \text{INT}\cdot)^2}{I(I - 1)^2} \quad , \tag{31}$$

where $\text{INT}_i = (X_{ai} - X_{aI} - X_{Ai} + X_{AI})^2$ and $\cdot$ signifies a mean across i. With a moderate number of items this variance estimate will have a large error of its own, and perhaps should not be interpreted for a particular examinee. However, if an average is taken across examinees, it can be used to determine how much of the observed variability *across examinees* in $\hat{\sigma}^2(e)_a$ is attributable to systematic variation in examinee-level error variance and how much is attributable to error in estimating each $\sigma^2(e)_a$.

Empirical Results

A 26–item test of logical reasoning was used to obtain estimates of $\sigma^2(e)_a$ for 5,776 examinees. Table 1 contains means and variances of three estimators of error variance as well as other statistics for

Table 1
Means and Standard Deviations of
Error Estimates and Other Statistics

Estimator	Mean	SD
$\hat{\sigma}^2(e)_a$	.0065	.0020
$\dfrac{X_{aI}(1-X_{aI})}{I-1}$	.0080	.0021
$\hat{\sigma}(i,r)_a/I$	zero	.0007
Odd-Even Difference	.0063	.0090

$I = 26$; $A = 5776$; $X_{AI} = .650$;
$\hat{\sigma}^2(X_{aI}) = .028$; $\hat{\sigma}^2(i) = .038$
$\hat{\sigma}^2(a) = .022$; $\hat{\sigma}^2(r) = .169$.

this test. Note that the difference between the mean $\hat{\sigma}^2(e)_a$ and the mean of binomial variance estimates is $\hat{\sigma}^2(i)/I$. The odd-even estimator is explained later.

A scatterplot of $\hat{\sigma}^2(e)_a$ by number-correct score is given in Figure 1. The convex shape of the points resembles that of the binomial, but there is much variability of the $\hat{\sigma}^2(e)_a$ for each observed score. Also, the shape is asymmetric, showing higher error variances at low scores than at high scores. (For the binomial, variances are symmetric around 13.) The distribution of observed scores is negatively skewed, as is expected for a mean of .65 (see Table 1.) Many of the observations are to the right of 13.

Variability of the $\hat{\sigma}^2(e)_a$ at each number-correct score is attributable solely to variability of the $\hat{\sigma}(i,r)_a$. Figure 2 gives estimates of $\sigma(i,r)_a/I$ at each score. (Dividing by I yields a plot on the scale of the actual contribution to $\hat{\sigma}^2(e)_a$.) As indicated in Table 1, the mean of $\hat{\sigma}(i,r)_a$ is zero for the full sample of examinees (by definition of the estimator). From Figure 2 there appears to be a trend of lower covariances for lower scores. Recall that $\sigma(i,r)_a = -\sigma^2(i)$ for an examinee who randomly responds to the items. In

Figure 1
Scatterplot of $\hat{\sigma}^2(e)_a$ by Number-Correct Score

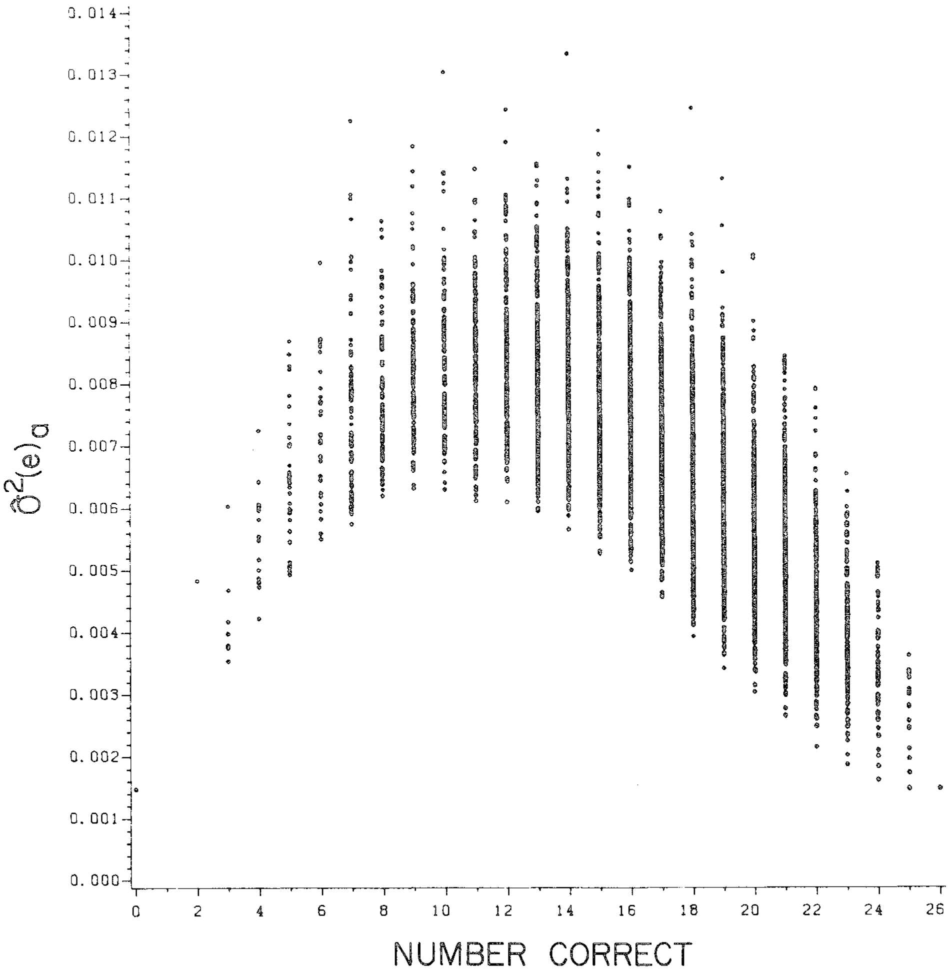
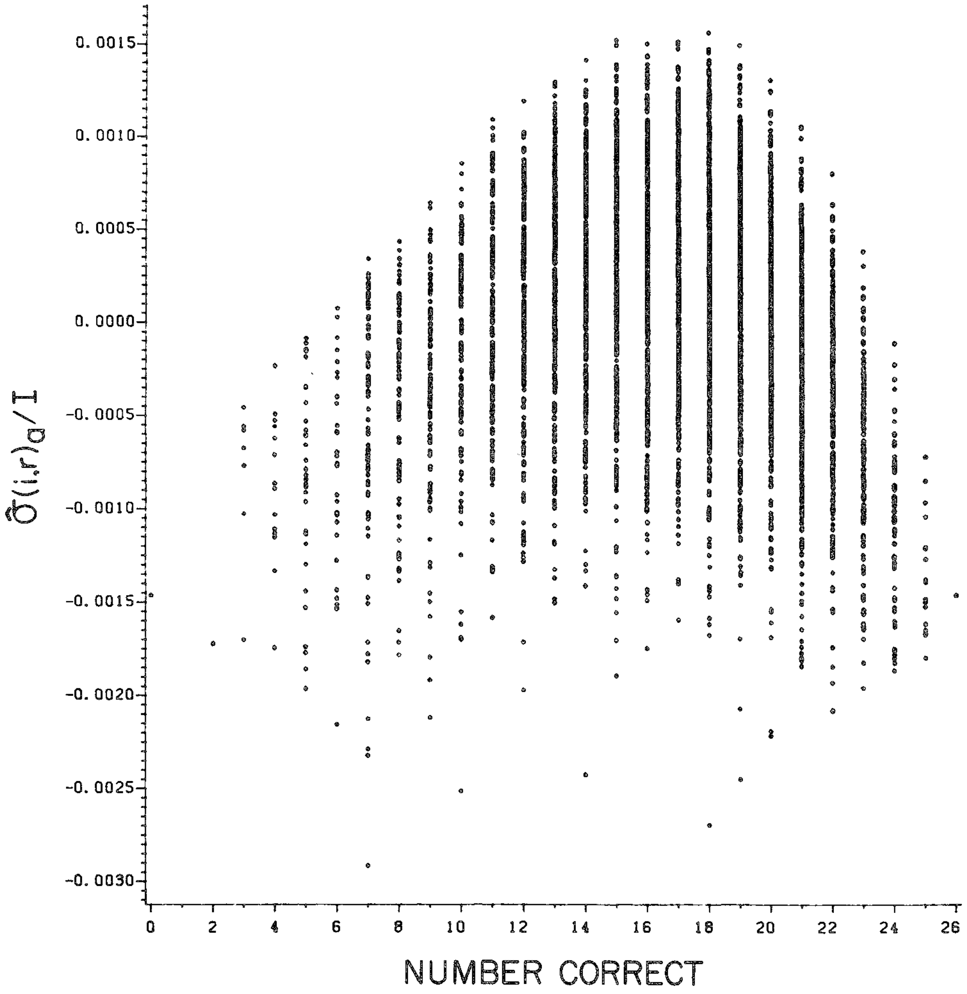


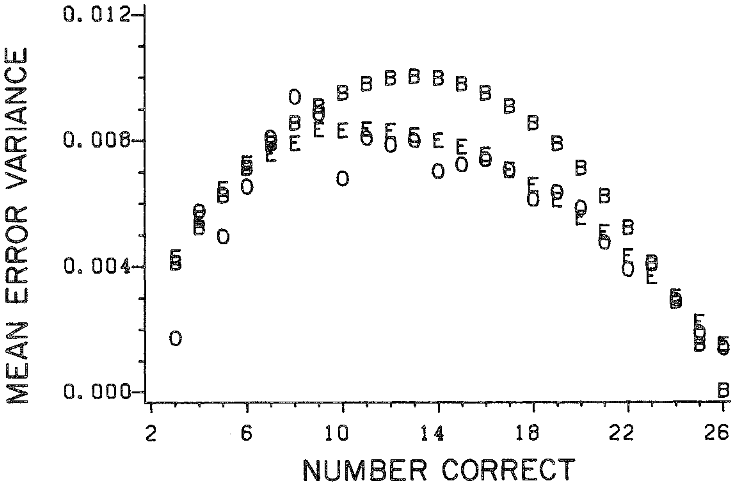
Figure 2
Scatterplot of $\hat{\sigma}(i,r)_a/I$ by Number-Correct Score



general, any response vector that is correlated zero with the item difficulty estimates has $\hat{\sigma}(i,r)_a = -\hat{\sigma}^2(i)$. In Figure 2, the point on the vertical axis corresponding to this is $-.0015$. It is of course possible for an examinee to show a negative covariance with item difficulties $[\hat{\sigma}(i,r)_a/I < -.0015]$, and there are several such examinees in this sample.

For scores of 12 through 20 the means of the $\hat{\sigma}(i,r)_a$ are above zero, and negative outside this range. The effect on the $\hat{\sigma}^2(e)_a$ is depicted in Figure 3. The mean of the $\hat{\sigma}^2(e)_a$ at each score point is plotted with an E, and the binomial variance estimate with a B. Notice that the points (B,E) are approximately equal for low-scoring examinees, indicating that adjusting for mean difficulty does not give greater measurement precision than unadjusted scores for these examinees. For the majority of examinees whose scores ranged from 8 (2.0 standard deviations below the mean) to 23 (1.1 standard deviations above the mean), the mean adjustment does appear to reduce measurement error.

Figure 3
Mean $\hat{\sigma}^2(e)_a$ (E), Mean Odd-Even (O), and Binomial (B) by Number-Correct Score



For an empirical comparison, odd versus even item scores on the 26 items were calculated for all examinees. In order to resemble a mean adjustment, the means for the two half-tests were set to the same constant. Differences between mean adjusted half-test scores yielded error variance estimates. Under the model, such estimation of error variance slightly underestimates $\sigma^2(e)_a$ by a constant of approximately .000004. This is because two means from the same examinees are used for the adjustment instead of one. The mean odd-even error variance estimate is plotted against each score using an O. These points are much closer to the means of the $\hat{\sigma}^2(e)_a$ than to the binomial points. The mean odd-even error variance across all examinees is given in Table 1. Note that it is also much closer to the overall mean of the $\hat{\sigma}^2(e)_a$ than to the overall mean for the binomial.

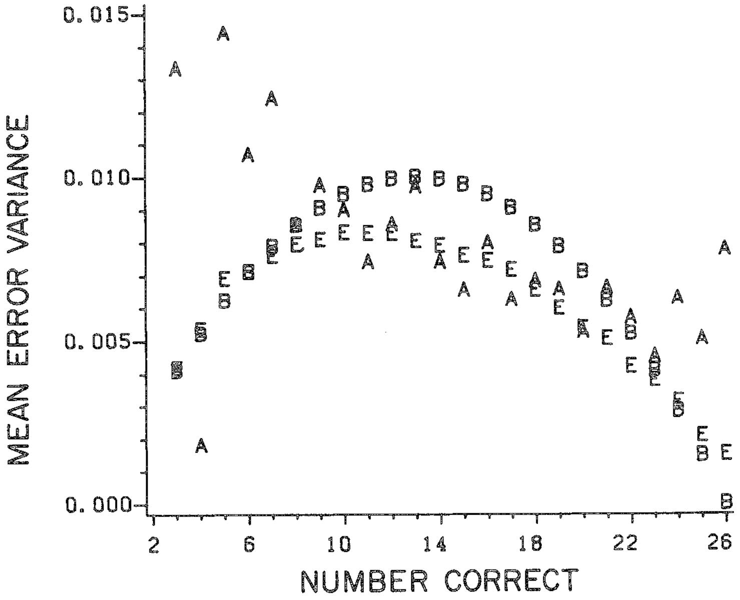
Because this logical reasoning test form was administered in a pre-equating design, alternate forms (pre-equating sections) of this test were also administered to some of the examinees. Thus, there were data for 1,915 examinees on two separately timed forms of the same test (one pre-equating, the other operational). After setting the means of the pair of alternate forms to the same constant, differences in scores on the forms were calculated. As for the odd-even differences, mean error variances (A) estimated from the alternate-forms differences by number-correct score were plotted in Figure 4. Means of $\hat{\sigma}^2(e)_a$ (E) for this sample were also plotted in Figure 4 as well as the binomial variances (B). As was the case for the odd-even error variances, the alternate forms means are closer to the means of the $\hat{\sigma}^2(e)_a$ than the binomial error variances for the middle scoring examinees. Under the model, the alternate forms error variances underestimate $\sigma^2(e)_a$ by .00001, which was discounted. Note that the means for low scores are based on few examinees.

Table 2 gives average error variances for both subsets of examinees. Notice that average error variance from alternate forms is closer to average $\hat{\sigma}^2(e)_a$ than to the binomial variance estimates. Data on 1,941 examinees that were administered a different alternate form yielded the same results.

Error in Estimating $\sigma^2(e)_a$

From Equation 31, the VAR$[\hat{\sigma}^2(e)_a]$ was estimated for each examinee. The average over all examinees

Figure 4
Mean $\hat{\sigma}^2(e)_a$ (E), Mean Alternate Forms (A), and Binomial (B) by Number-Correct Score



was 2.2×10^{-6}. The observed variance of the $\hat{\sigma}^2(e)_a$ across examinees was 3.84×10^{-6}. This indicates that 60% of the observed variance is attributable to error in estimation and 40% to systematic variation in $\sigma^2(e)_a$ across examinees. From the plot of means × score (Figure 3), there is a strong indication of systematic variation in the $\sigma^2(e)_a$, and this percentage breakdown simply serves to corroborate it. In fact, the variance of the means by score accounts for 65% of the observed variance in the $\hat{\sigma}^2(e)_a$. A comparison of the observed variance of $\hat{\sigma}^2(e)_a$ for each score with the average $\widehat{\text{VAR}}[\hat{\sigma}^2(e)_a]$ for each score indicated that variation in $\hat{\sigma}^2(e)_a$ at each score could be attributed solely to estimation error. For this particular test, then, it seems appropriate to provide a mean $\hat{\sigma}^2(e)_a$ for every examinee with the same score. Still, large and unlikely values of $\hat{\sigma}^2(e)_a$ should raise suspicion.

Discussion

Test equating typically involves more than a mean adjustment in scores. At the least, a standard deviation adjustment is also applied. Consideration for such linear equating is a necessary extension of the present results. Some further details on equating designs for even a simple mean adjustment are also needed. One particular design is discussed in the Appendix.

Test construction typically does not resemble simple random sampling. For this reason, some consideration of the usual categorization of items into cells of a table of specifications would be a logical extension of these results.

For the 26-item test studied here, error in estimating $\hat{\sigma}^2(e)_a$ proved to be substantial, which suggests that mean values by score be used for reporting of measurement error. With longer tests, the error could be reduced to an extent that would allow individual interpretation of $\hat{\sigma}^2(e)_a$. Along these lines, an empirical

Table 2
Means and Standard Deviations of Error Estimates Based on Alternate Forms (A=1915)

Estimator	Mean	SD
$\hat{\sigma}^2(e)_a$	.0065	.0019
$\frac{X_{aI}(1-X_{aI})}{I-1}$	.0080	.0021
Alternate Forms Difference	.0071	.0089

Bayes estimator of $\sigma^2(e)_a$ would reduce error across examinees and could avoid the degree of bias that is possible when using mean $\hat{\sigma}^2(e)_a$ for all examinees with the same score.

References

Feldt, L. S. (1984). Some relationships between the binomial error model and classical test theory. *Educational and Psychological Measurement, 44*, 883–891.

Huynh, H., & Saunders, J. C. (1976). Accuracy of two procedures for estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement, 13*, 265–276.

Keats, J. A. (1957). Estimation of error variances of test scores. *Psychometrika, 22*, 29–41.

Keats, J. A., & Lord, F. M. (1962). A theoretical distribution for mental test scores. *Psychometrika, 27*, 59–72.

Lord, F. M. (1957). Do tests of the same length have the same standard error of measurement? *Educational and Psychological Measurement, 17*, 510–521.

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

Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement, 1*, 233–247.

Subkoviak, M. J. (1978). Empirical investigation of procedures for estimating reliability of mastery tests. *Journal of Educational Measurement, 15*, 111–116.

Tatsuoka, K. K., & Linn, R. L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement, 7*, 81–96.

Weiss, D. J., & Kingsbury, G. G. (1985). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361–375.

Appendix

Further Equations for Error Variance

1. The full expression of $\sigma^2(e)_a$:

$$\sigma^2(e)_a = \frac{\mu_a(1-\mu_a)}{I} - \frac{\sigma^2(i)}{I} - \frac{2\sigma(i,r)_a}{I} - \frac{(2A-1)\sigma^2(r)_a}{A^2I} + \frac{(A-1)\sigma^2(r)}{A^2I} + \frac{\sigma^2(a)}{A}\ . \tag{A1}$$

This is obtained by deriving $\text{VAR}(X_{aI} - X_{AI} \mid a)$. Because this variance is conditional on a, some complications arise (note that a contributes to X_{AI}). Otherwise, it is a straightforward application of expectation rules.

Recall that $\text{MS}(r)_a$ (Equation 23) plus $\hat{\sigma}^2(a)/A$ provide an unbiased estimator of $\sigma^2(e)_a$; that is, E $[\text{MS}(r)_a]$ for a given a yields Equation A1 minus $\sigma^2(a)/A$.

2. Average error variance and sample size:

   The full expression of average error variance over examinees is

$$\mathrm{E}\,\sigma^2(e)_a = \frac{\mathrm{E}\,\mu_a(1-\mu_a)}{I} - \frac{\sigma^2(i)}{I} - \frac{\sigma^2(r)}{AI} + \frac{\sigma^2(a)}{A} \quad . \tag{A2}$$

   Therefore this equation can be used to determine for various values of A and I whether $\mathrm{E}\,\sigma^2(e)_a <$ $\mathrm{E}\,\mu_a(1 - \mu_a)/I$ (i.e., whether the mean adjustment reduces average error variance). For example, if A is small and I is large, the mean adjustment could result in larger average error variance than if no adjustment were performed.

3. An unbiased estimator of $\sigma^2(r)_a$:

$$\hat{\sigma}^2(r)_a = \frac{\mathrm{MS}(r)_a A^2}{(A-1)^2} - \frac{\hat{\sigma}^2(r)}{A-1} \quad . \tag{A3}$$

   (See Equations 22 and 27.)

4. Consideration of a mean adjustment in which c (Equation 3) has error.

   Suppose c is to be estimated. For example, say two forms are independently administered to two random samples of examinees. One form (old) brings with it a constant mean adjustment ($d_{I'}$) from a previous administration (error in $d_{I'}$ is ignored here). The other form is new (i.e., it has not been previously administered). A mean adjustment in the new form can then be written as

$$X''_{aI} = X_{aI} - X_{AI} + X_{A'I'} - d_{I'} \quad , \tag{A4}$$

   where $X_{A'I'} - d_{I'}$ takes the place of c in X'_{aI} (Equation 3), and where $X_{A'I'}$ is the mean of the old form on the current sample of A' examinees ($A \cap A' = 0$, $I \cap I' = 0$). It is easy to show that

$$\mathrm{VAR}(X''_{aI} \mid a) = \sigma^2(e)_a + \frac{\sigma^2(a)}{A'} + \frac{\sigma^2(r)}{I'A'} \quad , \tag{A5}$$

   where A' and I' also symbolize sample sizes. Note that the $d_{I'}$ adjusts for $\mu_{I'}$ and therefore $\sigma^2(i)$ for I' does not enter into Equation A5. Error variance for the case in which $A \cap A' \neq 0$ can be similarly obtained.

Acknowledgments

Much of this work was completed while the author was at the American College Testing Program.

Author's Address

Send requests for reprints or further information to David Jarjoura, Division of Community Health Sciences, Northeastern Ohio Universities College of Medicine, Rootstown OH 44272.