

A Comparison of Five Methods for Estimating the Standard Error of Measurement at Specific Score Levels

Leonard S. Feldt
The University of Iowa

Manfred Steffen
Mississippi State University

Naim C. Gupta
Ball State University

The *Standards for Educational and Psychological Testing* (1985) recommended that test publishers provide multiple estimates of the standard error of measurement—one estimate for each of a number of widely spaced score levels. The presumption is that the standard error varies across score levels, and that the interpretation of test scores should take into account the estimate applicable to the specific level of the examinee. This study compared five methods of estimating conditional standard errors. All five of the methods yielded a maximum value close to the middle of the score scale, with a sharp decline occurring near the extremes of the scale. These trends probably characterize the raw score standard error of most standardized achievement and ability tests. Other types of tests, constructed using alternative principles, might well exhibit different trends, however. Two methods of estimation were recommended: an approach based on polynomial smoothing of point estimates suggested by Thorndike (1951) for specific score levels and a modification proposed by Keats (1957) for the error variance derived under the binomial error model of Lord (1955).

To describe the accuracy of the measurement of person i , examiners use the standard error of measurement—the hypothetical standard deviation of a large number of repeated measurements on that individual. It is customary to use the same estimate

of the standard error for all persons, an estimate obtained by the equation

$$S_E = S_X(1 - r_{xx'})^{1/2} \quad (1)$$

where S_E is the estimated standard error of measurement,

S_X is the estimated standard deviation of observed scores for the population of interest, and

$r_{xx'}$ is the estimated test reliability coefficient for that population, preferably obtained from parallel forms administered on different days.

This formula is well established and easily derived from classical test theory (Lord & Novick, 1968, p. 59). However, as a descriptive index of the potential inaccuracy of observed scores, it has several weaknesses: (1) like any standard deviation, its interpretive value is limited when obtained for grossly non-normal distributions, and (2) although descriptively accurate as an *average* value for an entire population of examinees, it may fail to reflect the variability of the measurement errors of low-scoring examinees, average examinees, or high-scoring examinees. In effect, it may be an average that poorly describes measurement accuracy at any specific score level. For this reason the most recent edition of the *Standards for Educational and Psychological Testing* (Committee of AERA, APA, & NCME to Develop Standards, 1985)

encouraged the estimation of *conditional* standard errors—values specific to examinees at specified score levels. The recommendation is phrased as follows:

Standard 2.10 Standard errors of measurement should be reported at critical score levels Comment: Reporting standard errors of measurement at every score level may not be feasible in some circumstances, but they should be reported at appropriate, well-separated levels or intervals. (p. 22)

The purpose of the present study was to compare a number of methods of estimating these conditional standard errors. Although many test authors and users may be familiar only with Equation 1 as a means of estimating the standard error, there are, in fact, at least six ways of approximating S_E at specific levels. The primary goal of this investigation was to assess the numerical comparability of the estimates yielded by various approaches and, if possible, to arrive at a recommendation regarding which method(s) might be routinely used.

Methods of Estimating Conditional Standard Errors

As most textbook authors note, every estimate of reliability entails an implicit definition of measurement error. Some definitions are more comprehensive than others. The same is true of direct estimates of S_E . Different estimators allow different configurations of error sources to reflect their impact. The estimates considered in the present study do not recognize error that arises from day-to-day changes in the examinee and variation due to unequal means of “not-quite-parallel” test forms, observers, raters, or measurements. In effect, the study employed methods of estimating S_E that might be referred to as internal consistency methods, though this phrase is more often applied to reliability coefficients than to estimates of S_E . Ideally, both statistics should be based on approaches that encompass the most comprehensive definition of measurement error. However, for the purposes of this study, more restricted estimates of S_E seemed adequate. Since the definition of error was similarly restricted under all approaches, it seemed unlikely

that the comparisons among the approaches would be biased by the exclusion of day-to-day sources of error.

Thorndike Method

Thorndike (1951) was among the first to propose a technique for estimating S_E at selected score levels. Thorndike conceived of an individual's total score, X , as the sum of two parallel half-test scores, X_1 and X_2 . Following the classical test theory model, each half-test score and the total test score is regarded as the sum of a “true” component and an error component:

$$\begin{aligned} X_1 &= T_1 + E_1 \quad , \\ X_2 &= T_2 + E_2 \quad , \\ X &= X_1 + X_2 \\ &= (T_1 + T_2) + (E_1 + E_2) \\ &= T + E \quad . \end{aligned} \quad (2)$$

Consistent with classical test theory, the correlation between E_1 and E_2 is assumed equal to zero. This leads to

$$\sigma_E = (\sigma_{E_1}^2 + \sigma_{E_2}^2)^{1/2} \quad . \quad (3)$$

The difference between half-test scores, that is, $X_1 - X_2$, equals $(T_1 - T_2) + (E_1 - E_2)$. For parallel half-tests, $T_1 = T_2$ or $T_1 - T_2 = 0$. Hence,

$$\begin{aligned} \sigma_{(X_1 - X_2)} &= \sigma_{(E_1 - E_2)} \\ &= (\sigma_{E_1}^2 + \sigma_{E_2}^2)^{1/2} \\ &= \sigma_E \quad . \end{aligned} \quad (4)$$

Therefore, the total test standard error equals the standard deviation of half-test differences for any large group of examinees. If the examinees are restricted to those at a particular value of X , the standard deviation of half-test differences for this subgroup provides an estimate of S_E at this specific score level.

The approach is limited by the size of the group at each score level of interest. There is the strong possibility that relatively few examinees may be available at some score values, particularly at the extremes. This limitation may be partially overcome by grouping examinees into short intervals of total score rather than by subgroups defined by one particular value of total score.

Polynomial Method

As part of a larger study, Mollenkopf (1949) proposed a refinement of the Thorndike (1951) approach. In essence, it provided a technique to smooth and stabilize the estimates, which is more effective than grouping examinees into score intervals. Mollenkopf proposed that the squared difference between half-test scores for each individual be regarded as a value to be "predicted" by regression techniques from the total score. The prediction would be made using a quadratic, cubic, or fourth degree polynomial regression equation. In essence, $Y = (X_1 - X_2)^2$ would be regarded as a "criterion" to be predicted using a least squares regression line approach. Thus, if a fourth degree polynomial model were used,

$$\hat{Y} = a_0 + a_1(X) + a_2(X^2) + a_3(X^3) + a_4(X^4) \quad (5)$$

Under regression theory, \hat{Y} may be interpreted as an estimate of μ_Y for that restricted population of individuals who earned a given value of X . However, the average value of Y , as defined in this case, is the average squared half-test difference. With parallel half-tests, the average squared difference is the variance of differences. This quantity, as Thorndike (1951) showed, is the variance of errors of measurement for the full test. Therefore, $(\hat{Y})^{1/2}$ for a given value of X is an estimate of S_E at score point X .

To apply this method, the coefficients a_0, a_1, \dots , must be solved for just as if a multiple regression equation for predicting Y was being determined. This can be done quite easily with the statistical analysis programs available through almost all computer installations. Once the coefficients are determined, $(\hat{Y})^{1/2}$ may be obtained by substituting any value of X in Equation 5 for which S_E is desired.

**Lord's Binomial Approach:
 Keats' Modification**

A third estimation technique is based on Lord's (1955) binomial error model. Lord conceived of a universe of acceptable test items from which a randomly selected set of size k constitutes a test form. Another independent set of k items constitutes a

second form, and so on. A given individual i is perceived to be able to answer a certain proportion, ϕ_i , of the entire population of items.

Under this conceptualization, the score of examinee i on one test form is analogous to the frequency count of the occurrence of phenomenon Q in a random sample of k units. The fundamental notion of the standard error of measurement is that of the standard deviation of scores for a given examinee on many parallel test forms. Thus, the concept of σ_E is directly comparable to the statistical concept of a standard error of a frequency. As noted in many statistics textbooks, the standard error of a frequency determined from a sample of size k is $[k(\phi)(1 - \phi)]^{1/2}$. For person i , the standard error of measurement equals this quantity, with ϕ_i inserted. This parameter of person i is unknown, of course. Lord (1955) proposed the use of the observed proportion correct as an estimate of ϕ_i . He also recommended correction for the known bias in the variance determined from finite samples. These modifications lead to

$$S_E = \left[\frac{X(k - X)}{k - 1} \right]^{1/2} \quad (6)$$

as the estimate of the standard error of measurement at the score level X for the total test.

Subsequent discussion of this equation in the measurement literature brought out an obvious weakness. It fails to take into account the careful matching of test forms in content, item difficulty, and other characteristics. Thus, Equation 6 seems destined to overestimate S_E at score value X . The fact that the average squared value of Equation 6 leads to Kuder-Richardson 21 as the reliability coefficient for the test (Lord, 1955) reinforced the belief that the equation overestimates the conditional S_E .

Keats (1957) proposed a modification of this equation to scale its average value down to a more appropriate level. This is done by multiplying the right-hand side of Equation 6 by a constant, which will result in an average value that is consistent with a more defensible reliability coefficient. This reasoning gives rise to

$$S_E = \left[\left(\frac{X(k - X)}{k - 1} \right) \left(\frac{1 - r_{xx'}}{1 - r_{21}} \right) \right]^{1/2} \quad (7)$$

where $r_{xx'}$ is the most defensible estimate of reliability for the test, and r_{21} is Kuder-Richardson 21 for the test. Keats recommended a parallel forms coefficient for $r_{xx'}$, but in practice it might be necessary to use a split-halves coefficient or Cronbach's (1951) alpha.

Lord's Binomial Approach: Compound Binomial

In a subsequent publication, Lord (1965) proposed that the standard error applicable to matched test forms be estimated in a somewhat different manner. Matching forms during test construction is essentially a process of selecting stratified samples of items rather than completely random samples from the population of items. Lord drew upon the statistical theory that provides an equation for the standard error of a frequency determined from a stratified sample. In the present context, this theory leads to the equation

$$S_{E(i)} = \left[\sum_{h=1}^c \frac{X_{ih}(k_h - X_{ih})}{k_h - 1} \right]^{1/2}, \quad (8)$$

where $S_{E(i)}$ is the standard error for person i ,

X_{ih} is the score of person i on the cluster of items corresponding to category h of the test specifications,

c is the number of item categories, and
 k_h is the number of items in category h .

Clearly, k_h must be greater than 1.

To use this equation, the test must be scored for every item category as if the categories were subtests of the overall instrument. Application of Equation 8 then provides an estimate of each examinee's personal standard error. Examinees are grouped into intervals on the basis of total test score, and the average of $S_{E(i)}^2$ is obtained for each interval. The square root of this average is taken as the estimate of S_E for the midpoint of the interval.

Variance Components Estimates

A fifth approach, not previously proposed in the measurement literature, draws upon analysis of

variance (ANOVA) methodology. As first noted by Hoyt (1941), the examinees by items score matrix for a test may be analyzed to obtain mean squares for examinees (MS_S), items (MS_I), and interaction ($MS_{S \times I}$), and the reliability of the test may be estimated from these mean squares. In particular, the error variance for a test of k items may be approximated by $k(MS_{S \times I})$. If individuals are grouped into intervals of total score, the standard error of measurement for the interval midpoint may be estimated by

$$S_E = [k(MS_{S \times I})]^{1/2} \quad (9)$$

This estimate is closely related to that of Thorndike (1951), and shares that estimate's principal weakness—instability arising from small numbers of examinees in many of the more extreme intervals. Some smoothing of the set of interval values will be needed if the estimates of conditional S_E are not to show erratic fluctuations at the extremes of X .

Estimates Based on Item Response Curve Theory

The final method studied in this investigation draws upon item response curve theory (IRT; Lord, 1980). This theory addresses the problem of estimating the probability that individual i with true ability score θ_i will answer item j correctly. This probability is approximated by a function, $P_j(\theta)$, which generally resembles a normal ogive. Several parameters must be estimated to define each item function. If these parameters are well approximated, the measurement error variance may be estimated for examinee i by a two-step process: (1) obtaining an estimate of the examinee's θ_i , and (2) evaluating the function

$$S_{E(i)}^2 = \left\{ \sum_j^k [P_j(\theta_i)] [1 - P_j(\theta_i)] \right\}^{1/2} \quad (10)$$

$P_j(\theta_i)$ is the value of the function for item j when evaluated at the ability level, θ_i , of subject i .

Like the compound binomial approach to which it is related, this method yields an estimate for each examinee. To use this method, examinees are grouped according to their level of total score, the

average value of $S_{E(i)}^2$ is computed for the examinees in each interval, and the square root of the average is determined for the interval. An alternative approach is to derive a table of one-to-one correspondences between values of X and values of θ . Each specific pair (X_0, θ_0) is derived by determining the value of θ_0 , which satisfies the relationship

$$X_0 = \sum_j^k P_j(\theta_0) \quad (11)$$

The squared standard error at value θ_0 , as computed by Equation 10, is then associated with the raw score value X_0 . This latter method was used in the present study. The modified two-parameter logistic model was adopted for the function $P_j(\theta)$. It was implemented using the LOGIST computer program of Wood, Wingersky, and Lord (1976).

The principal drawback of this method is that adequate estimation of the parameters of each item response function takes rather large examinee samples. Some authorities recommend samples of at least 500 cases. It also must be assumed that any test so analyzed is unidimensional in a factorial sense.

Method

Data

The six methods described above were applied to data from over 16,000 examinees in each of Grades 9 and 11 for three subtests of the Iowa Tests of Educational Development (ITED) battery: vocabulary (V), reading of literary materials (L), and use of sources of information (SI). Since the Grade 9 and Grade 11 batteries were not identical, six separate tests were involved. The examinees included a random half of the students in Grades 9 and 11 who were tested in the 1980–81 Iowa High School Testing Program.

Preliminary Analyses

Three preliminary analyses were undertaken. The first sought to test the tenability of the unidimensionality assumption of the IRT approach. The second involved a comparison of alternative models for the item characteristic curves (ICCs). The third

was carried out to establish the degree of the polynomial to be employed under the polynomial approach.

The first of these analyses involved a factor analysis of the six separate tests under study. These indicated the presence of one major factor and possibly one secondary factor of minor importance in each test. An examination of the potency of the first factor relative to the second led to the conclusion that the assumption of unidimensionality was adequately met.

The second preliminary analysis included a comparison of alternative two-parameter logistic models for the form of the ICCs. The study led to a choice of the model referred to in the literature as the modified two-parameter model. This model allows for variation among items in their difficulty and discrimination indices, and incorporates a constant nonzero value for the lower asymptote (c) of all ICCs. The preliminary analysis led to the adoption of $c = .10$ as the lower asymptote for each ICC.

The third preliminary study consisted of regression analyses to determine how high a degree was needed for the polynomial regression equation used to predict $(X_1 - X_2)^2$. It was found that a third degree polynomial was adequate. This suggests that the trend in S_E^2 over the range of total score may be satisfactorily described by a cubic equation in X . Such an equation allows for a curvilinear trend with one or two maximums and/or minimums. This form of polynomial accommodated to the intuitively attractive hypothesis that error variance peaks near the middle of the score range and declines fairly rapidly as it moves toward the lower or upper end of the score scale.

Because straightforward application of the Thorndike (1951) method resulted in the anticipated fluctuations in S_E , even with the large samples that were available, this method was explored only in conjunction with the polynomial method. The latter is, in essence, a technique that smoothes the Thorndike estimates using the least squares criterion of best fit. The precise number of answer sheets available for analysis varied from test to test and grade to grade. For Grade 9, the sample size per test ranged between 16,607 and 16,671. For Grade 11, N ranged from 17,053 to 17,119 per

test. These sample sizes more than satisfy the minimums suggested for factor analyses and the application of IRT methods.

The application of the compound binomial approach necessitates assignment of items to strata, reflecting the most salient features of the process of matching items in parallel test forms. In practice, it is customary to match forms on a number of item characteristics, including item difficulty, item discrimination, and content. Preliminary studies indicated that examinee score consistency across forms was most sensitive to the item difficulty factor. Therefore, the item strata in this study were defined on the basis of proportion of correct response. The first stratum included the five most difficult items. The second stratum included the next five items in difficulty, and so forth. Each test was thus partitioned into *c* strata of five items each, except for the last stratum of tests L and SI, which included the six easiest items.

Results and Discussion

The statistical characteristics of the six tests for the population under study are summarized in Table 1. It may be noted that each test score distribution is slightly skewed either positively or negatively. All of the distributions are platykurtic; they are "flatter" than normal distributions and have relatively heavy tails. This is characteristic of achievement tests composed of items with a wide range of difficulties and with reliabilities above .80 (Cook, 1959). The reliability coefficients for the tests are internal consistency coefficients computed using Flanagan's equation (Thorndike, 1951). This is a special case of coefficient alpha with part tests defined by halves. In the present case, the halves were balanced with respect to content and item difficulty. The distributions of item difficulties summarized in the lower half of Table 1 are typical of standardized tests, showing a broad range and

Table 1
Test Characteristics: Length, Mean, SD, Skewness (γ_1),
Kurtosis (γ_2), Reliability, and Item Difficulties (%)

	Grade 9			Grade 11		
	Test V	Test L	Test SI	Test V	Test L	Test SI
Items	40	46	46	40	46	46
Mean	20.95	23.91	27.59	18.36	25.95	26.43
SD	8.63	8.89	8.06	8.94	9.13	8.13
γ_1	+0.12	+0.08	-0.33	+0.43	-0.21	-0.11
γ_2	-0.90	-0.90	-0.53	-0.73	-0.95	-0.71
Rel.	.90	.89	.87	.90	.90	.87
Item Difficulty						
85-89			4			4
80-84	2	1	4		1	4
75-79	2	1	4	1	1	
70-74	2	4	5	3	6	3
65-69	5	6	6	2	7	3
60-64	4	4	4	2	6	3
55-59	4	4	2	4	5	5
50-54	4	4	3	4	7	6
45-49	4	7	3	5	4	4
40-44	5	7	4	4	3	4
35-39	1	4	2	5	2	4
30-34	3	2	2	6	3	3
25-29	2	2	3	3	1	
20-24	2			1		1

Table 2
 Conditional Standard Errors of Measurement, as a Function
 of Total Test Score, Estimated by Five Methods

Test and Score	Grade 9						Grade 11						
	Interval	N	IRT	F(X ²)	Keats	Comp Binom	ANOVA	N	IRT	F(X ²)	Keats	Comp Binom	ANOVA
Test V													
1-3	37	1.97	1.58	1.49	1.57	1.60	139	1.97	1.78	1.45	1.58	1.58	1.56
4-6	418	2.18	2.14	2.01	2.14	2.13	973	2.16	2.18	2.00	2.14	2.14	2.13
7-9	1175	2.46	2.51	2.38	2.52	2.50	1883	2.44	2.48	2.37	2.52	2.52	2.51
10-12	1638	2.68	2.77	2.65	2.75	2.74	2337	2.67	2.69	2.64	2.75	2.75	2.74
13-15	1801	2.81	2.92	2.83	2.88	2.87	2190	2.81	2.82	2.82	2.87	2.87	2.86
16-18	1893	2.86	3.00	2.93	2.92	2.91	1945	2.87	2.88	2.92	2.92	2.92	2.91
19-21	1900	2.85	3.00	2.96	2.91	2.90	1735	2.87	2.89	2.96	2.93	2.93	2.91
22-24	1851	2.79	2.94	2.93	2.83	2.82	1483	2.83	2.84	2.92	2.89	2.89	2.87
25-27	1688	2.67	2.82	2.83	2.71	2.70	1221	2.73	2.73	2.82	2.82	2.82	2.78
28-30	1450	2.49	2.64	2.65	2.56	2.52	1036	2.57	2.57	2.64	2.67	2.67	2.62
31-33	1267	2.23	2.39	2.37	2.32	2.26	901	2.33	2.33	2.36	2.42	2.42	2.36
34-36	910	1.87	2.07	1.97	1.96	1.89	699	1.98	1.73	1.96	2.05	2.05	1.99
37-39	579	1.35	1.65	1.35	1.40	1.34	511	1.37	1.53	1.32	1.40	1.40	1.36
Test L													
1-3	5	2.16	1.49	1.45	1.55	1.53	9	2.17	1.62	1.28	1.37	1.37	1.31
4-6	93	2.31	2.33	2.13	2.22	2.23	95	2.26	2.31	2.09	2.19	2.19	2.21
7-9	538	2.57	2.73	2.51	2.63	2.62	441	2.59	2.65	2.48	2.62	2.62	2.62
10-12	1235	2.84	3.00	2.79	2.91	2.91	943	2.89	2.91	2.77	2.91	2.91	2.92
13-15	1650	3.03	3.16	3.00	3.10	3.10	1340	3.08	3.08	2.98	3.13	3.13	3.12
16-18	1681	3.15	3.25	3.15	3.20	3.20	1420	3.20	3.20	3.13	3.25	3.25	3.24
19-21	1786	3.20	3.26	3.23	3.22	3.22	1449	3.25	3.26	3.21	3.30	3.30	3.30
22-24	1793	3.19	3.21	3.26	3.21	3.21	1580	3.24	3.26	3.24	3.28	3.28	3.28
25-27	1810	3.13	3.12	3.23	3.15	3.15	1712	3.17	3.21	3.21	3.20	3.20	3.20
28-30	1708	3.02	3.00	3.15	3.06	3.05	1846	3.04	3.10	3.12	3.07	3.07	3.07
31-33	1587	2.87	2.84	3.00	2.91	2.90	2052	2.86	2.94	2.97	2.90	2.90	2.89
34-36	1265	2.66	2.67	2.79	2.74	2.71	1886	2.63	2.71	2.77	2.69	2.69	2.66
37-39	927	2.38	2.49	2.49	2.48	2.43	1499	2.31	2.39	2.46	2.39	2.39	2.34
40-42	488	1.99	2.33	2.07	2.12	2.03	735	1.93	1.99	2.06	2.03	2.03	1.96
43-45	105	1.51	2.20	1.51	1.56	1.47	107	1.48	1.42	1.50	1.52	1.52	1.47
Test SI													
1-3	4	2.26	1.54	1.48	1.59	1.62	1	2.20	1.79	1.29	1.41	1.41	----
4-6	36	2.33	2.17	2.01	2.21	2.19	26	2.32	2.41	2.05	2.23	2.23	2.19
7-9	173	2.50	2.59	2.40	2.63	2.61	222	2.55	2.75	2.45	2.63	2.63	2.63
10-12	461	2.78	2.86	2.68	2.91	2.90	506	2.77	2.95	2.71	2.89	2.89	2.89
13-15	841	2.96	3.04	2.88	3.08	3.07	1059	2.93	3.09	2.92	3.04	3.04	3.04
16-18	1028	3.07	3.13	3.02	3.16	3.15	1400	3.02	3.17	3.06	3.10	3.10	3.09
19-21	1367	3.11	3.16	3.10	3.16	3.15	1750	3.06	3.19	3.14	3.10	3.10	3.08
22-24	1661	3.09	3.14	3.12	3.12	3.11	2060	3.06	3.16	3.16	3.09	3.09	3.09
25-27	2142	3.01	3.06	3.09	3.03	3.03	2166	3.02	3.08	3.14	3.05	3.05	3.04
28-30	2340	2.90	2.95	3.01	2.91	2.90	2093	2.93	2.97	3.06	2.97	2.97	2.96
31-33	2292	2.76	2.79	2.87	2.80	2.79	2105	2.80	2.81	2.92	2.84	2.84	2.83
34-36	1985	2.58	2.59	2.67	2.64	2.62	1729	2.61	2.62	2.71	2.68	2.68	2.65
37-39	1420	2.32	2.35	2.38	2.42	2.36	1199	2.35	2.39	2.42	2.44	2.44	2.40
40-42	744	1.96	2.08	1.98	2.06	1.99	643	1.99	2.12	2.00	2.07	2.07	2.02
43-45	166	1.48	1.79	1.41	1.50	1.43	160	1.55	1.83	1.43	1.53	1.53	1.49

relatively flat distribution of the percentage correct index.

For each test and grade, individuals were grouped into three-point intervals of total score. For test V, with 40 items, this resulted in 13 intervals; for tests

L and SI, with 46 items each, there were 15 intervals of total score. The values of S_E for the six combinations of grade and test are displayed in Table 2.

Perhaps the most striking feature of these data

is the similarity of the values and the similarity of the trends evidenced by the various methods. In all instances the general trends are parabolic, concave downward. The peaking of S_E occurs, as most researchers have suspected, in the middle of the score range. There is a steady decline in S_E with increases in the absolute deviation of X from the score point at which S_E is greatest. These data agree closely with those results reported by Feldt (1984).

The largest differences among the methods occur in the S_E values for the most extreme score intervals. Near the end points of the score scale the Keats (1957) modification of Lord's (1955) basic binomial equation almost always yields a smaller conditional standard error than do the other methods. The IRT approach tends to yield the highest value for S_E at the low end of the score scale, and the polynomial method tends to yield the highest value at the high end of the scale. Exclusive of these extreme intervals, the range of S_E values for any given interval is rarely greater than .2. With regard to the maximum value of S_E , the methods agree closely with respect to the interval at which the maximum occurs and the value of the maximum.

The fact that the IRT approach consistently yields the highest value for score interval 1 to 3 is not surprising. It undoubtedly is largely the result of the adoption of .10 as the lower asymptote for the two-parameter model for $P_i(\theta)$. With this value, the lowest scoring examinees were always assigned a probability of at least .10 of making a correct response to each item. This resulted in a minimum standard error (at the low end of the scale) of $[k(.1)(.9)]^{1/2}$. For Test V this yields a minimum value of 1.90 for S_E , and for Tests L and SI a minimum value of 2.03. No such constraints hold for the other methods. Had a two-parameter logistic model with $c_j = 0$ been adopted, the IRT estimate for interval 1 to 3 would probably have been more consistent with the other estimates for this interval. The discrepancy is of little practical consequence in the present case, however, since less than 1% of examinees had a score of 3 or lower.

The data leave little doubt that the standard error of measurement does, in fact, vary across score levels. Regardless of the method used, the maxi-

mum is often more than twice the minimum. This implies that the standard error of measurement computed by the traditional formula for the test as a whole does not adequately summarize the error propensity of many—perhaps most—examinees. If the test user adopts a confidence interval approach in the communication of test results to examinees or to their parents, it would be advisable to base these intervals on the conditional standard error applicable to the examinee's score level, not on the standard error for the test as a whole. Anticipation of results such as those reported here no doubt prompted the testing standards committee to make the recommendation quoted earlier.

Two qualifications to the foregoing inferences should be noted. First, the observed variation in S_E from interval to interval, as estimated by some methods, may well be related to the distribution of item difficulties. Even the trend in the standard error might be different with tests constructed using different principles than those used with achievement tests such as ITED. It is quite possible that tests with more homogeneous item difficulties would exhibit less variation in S_E , at least under the polynomial and ANOVA methods. Whether this would also be true under the IRT and compound binomial methods remains to be investigated. On the other hand, the Keats (1957) method might well produce greater variation in S_E for a test with homogeneous item difficulty, since $(1 - r_{xx})/(1 - r_{21})$ might well be closer to 1.0 for such a test.

The second qualification is that the trends reported here apply to the standard error quantified in raw score units. Conversion of raw scores to normative scales such as intelligence quotients (IQs), grade equivalents, age equivalents, or standard scores might alter the trends significantly. For example, the small raw score S_E at high score levels might well translate into a comparatively large value when raw scores are converted to grade equivalents or IQs.

The consistency of the present results suggests an underlying consistency in the logical foundations of the various methods—particularly IRT methods, the methods based on the binomial error model, the Thorndike (1951) approach, and the ANOVA method. Reconsideration verifies this.

The IRT methods lead to estimates of an examinee's true score, θ . Examinees with the same true score (θ_0) exhibit variability in their observed scores on item j , some obtaining a score of 0 and others a score of 1. The variance of these item scores equals $P_j(1 - P_j)$, and this variability must be due to error since all these persons have the same true ability. Hence, $P_j(1 - P_j)$ evaluated at θ_0 can be taken as the error variance for item j . On the assumption that the errors of measurement for individual items are independent, $\sum P_j(1 - P_j)$ evaluated at θ_0 is the error variance for the total test. To obtain the estimates of θ_0 it must be assumed that P_j increases in the manner represented by the corresponding logistic function.

The binomial models are an empirical version of the same idea. Instead of using a maximum likelihood approach to the estimation of θ for each individual, binomial methods rely more heavily on the observed score of each examinee, grouping together individuals with the same X_0 . Within each subgroup of examinees, each item has an estimated proportion correct, P_j , and a corresponding proportion incorrect, $(1 - P_j)$. The quantity $P_j(1 - P_j)$ is summed over items, on the same assumption of independence of error that is invoked under IRT. This is the essence of the binomial model approach. As a point of detail, the sum is multiplied by $k/(k - 1)$ to remove the bias in the variance due to sampling of items from the populations of items.

The binomial approach has the advantage that it is not tied to any assumed model for P_j as a function of overall true or observed score, but it is subject to the fallibility of empirical estimates based on small samples. The Lord (1965) and Keats (1957) modifications of the original binomial theory accommodate to the fact that parallel forms are typically stratified samples of items and hence differ only slightly in their means. The heart of the approach, however, is the idea that for homogeneous sets of items and homogeneous groups of examinees, the sum of $P_j(1 - P_j)$ times $k/(k - 1)$ approximates error variance for the full test.

The ANOVA approach, like the binomial approach, depends on empirical estimates of the P_j rather than on a prescribed model for P_j as a function of true score. For a group of examinees with

total score X_0 , the ANOVA approach estimates the error of measurement of examinee i on item j by the interaction effect:

$$E_{ij} = X_{ij} - P_j - (X_0/k) + M \quad (12)$$

However, since every examinee has score X_0 , $M = X_0/k$. Thus,

$$E_{ij} = X_{ij} - P_j \quad (13)$$

where P_j is the proportion correct for item j in this subgroup with $X = X_0$ and X_{ij} equals 0 or 1. Squaring this quantity, summing over examinees and items, and dividing by $(N - 1)(k - 1)$ produces an estimate of the error variance of a single item. Multiplication by k gives an estimate of the error variance for the total score based on k items. Thus, the ANOVA approach leads to the estimate

$$\hat{\sigma}_{E(\text{conditional})}^2 = \frac{k \sum_j \sum_i (X_{ij} - P_j)^2}{(N - 1)(k - 1)} \quad (14)$$

However, algebraic manipulation leads to the following equivalent expression:

$$\hat{\sigma}_{E(\text{conditional})}^2 = \left(\frac{k}{k - 1} \right) \left(\frac{N}{N - 1} \right) \sum_j P_j(1 - P_j) \quad (15)$$

With large numbers of examinees, $N/(N - 1)$ may be taken equal to 1. For a test of moderate length $k/(k - 1)$ will also be close to 1. Thus, this approach leads to an expression for conditional error variance that is comparable to that of IRT models, once the clustering of examinees has been accomplished.

The ANOVA approach, in turn, is closely related to the Thorndike (1951) approach. Suppose Thorndike's estimate, $\hat{\sigma}_{(x_1 - x_2)}^2$, were obtained for all possible pairs of parts of a test with k parts. Suppose further that these $k(k - 1)/2$ variance estimates were averaged. This mean of variances would approximate the error variance of a two-part test. A test of k parts, assuming error independence, could therefore be estimated to have an error variance $k/2$ times this average. However, it can be shown that this quantity exactly equals k times the mean square for interaction of parts by examinees. This is true for the total group of examinees, and it is true for any subgroup defined in terms of X . Thus, the Thorndike method and the polynomial

method are integrally related to the ANOVA method and through it to IRT and binomial methods.

In summary, it may be observed that all of the methods share a common theoretical foundation. Although the estimation methods appear to differ, they are all variations on the same concept: if examinees could be clustered on the basis of their true scores, error variance at a particular true score would be equal to $\sum P_j(1 - P_j)$, where P_j is the difficulty of item j for individuals at the true score being considered. It seems clear, then, that the similarity of the empirical data for the several methods reflects an underlying conceptual similarity among the methods.

What may be said regarding a preference among the estimation methods? Clearly, the various methods agree sufficiently closely to make the choice heavily dependent on practical considerations and on the user's preference for the logic underlying one approach or another. Test publishers that use IRT for various aspects of test equating, standardization, and scaling may well find support in the present data for application of that theory in the estimation of conditional standard errors.

Individuals without access to very large examinee samples are probably well-advised to use the polynomial approach if the test data may be analyzed using a large scale computer. This method entails undemanding assumptions and requires a minimum of additional test scoring. It involves the computation of only two additional scores for each examinee, those for the two half-tests. Input to the computer for each examinee is X , X_1 , and X_2 . Available computer packages can obtain $Y = (X_1 - X_2)^2$ for each examinee, determine the coefficients of the polynomial of best fit, substitute successive integer values of X into this polynomial, and print the estimated conditional standard error, $(\hat{Y})^{1/2}$, for each value of X . It should be noted, however, that the two half-tests must satisfy the condition which Lord and Novick (1968) call essential tau equivalence. This means that the half-tests must be parallel in content and close (though not necessarily equal) in their means and standard deviations.

If computer analysis is impractical, Keats' (1957) solution for conditional S_E is even less demanding computationally. The quantity $[X(k - X)/$

$(k - 1)]^{1/2}$ can be computed for successive integer values of X with no data whatsoever. The constant $[(1 - r_{xx})/(1 - r_{21})]^{1/2}$, which is used to reduce these values, entails preliminary computation of only the test score mean and variance, in addition to the parallel forms, split-halves, or coefficient alpha reliability coefficient.

The compound binomial and ANOVA techniques have attractive features from a theoretical perspective, but they offer no obvious advantages for the practical problem of estimation. The compound binomial approach might present difficulties if the numbers of items in the various item categories are small, say two or three, and the size of the examinee sample is modest. In such a case, the binomial estimate of error variance for an item category might well equal zero for a significant number of examinees. This would happen whenever an item category score equals zero or k_n . For this reason, the other methods are probably preferable with small samples.

References

- Cook, D. L. (1959). A replication of Lord's study of skewness and kurtosis of observed test-score distributions. *Educational and Psychological Measurement*, 19, 81-87.
- Committee of AERA, APA, & NCME to Develop Standards. (1985). *Standards for educational and psychological testing*. Washington DC: American Psychological Association.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.
- Feldt, L. S. (1984). Some relationships between the binomial error model and classical test theory. *Educational and Psychological Measurement*, 44, 883-891.
- Hoyt, C. (1941). Test reliability obtained by analysis of variance. *Psychometrika*, 6, 153-160.
- Keats, J. A. (1957). Estimation of error variances of test scores. *Psychometrika*, 22, 29-41.
- Lord, F. M. (1955). Estimating test reliability. *Educational and Psychological Measurement*, 15, 325-336.
- Lord, F. M. (1965). A strong true score theory, with applications. *Psychometrika*, 30, 239-270.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum Associates.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Wesley.

- Mollenkopf, W. G. (1949). Variation of the standard error of measurement. *Psychometrika*, 14, 189–229.
- Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 560–620). Washington DC: American Council on Education.
- Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976). *LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters* (Research Memorandum 76–6). Princeton NJ: Educational Testing Service.

Acknowledgments

The data analyzed in this study were gathered through the auspices of the Iowa Testing Programs.

Author's Address

Send requests for reprints or further information to Leonard S. Feldt, 334 Lindquist Center, University of Iowa, Iowa City IA 52242, U.S.A.