An Investigation of Methods for Reducing Sampling Error in Certain IRT Procedures

Marilyn S. Wingersky and Frederic M. Lord
Educational Testing Service

The sampling errors of maximum likelihood estimates of item response theory parameters are studied in the case when both people and item parameters are estimated simultaneously. A check on the validity of the standard error formulas is carried out. The effect of varying sample size, test length, and the shape of the ability distribution is investigated. Finally, the effect of anchor-test length on the standard error of item parameters is studied numerically for the situation, common in equating studies, when two groups of examinees each take a different test form together with the same anchor test. The results encourage the use of rectangular or bimodal ability distributions, and also the use of very short anchor tests.

Until recently, the asymptotic sampling variances and covariances for maximum likelihood estimates of item parameters in item response theory (IRT) have usually been computed by assuming abilities to be known. Conversely, the asymptotic sampling variances and covariances for ability estimates have been computed by assuming the item parameters to be known. In this paper, a suggested method for computing the asymptotic sampling variance-covariance matrix of joint maximum likelihood estimates when all parameters are unknown (Lord & Wingersky, in press) is used to try to answer various practical questions. (For many purposes, an alternative approach has recently become available: the use of marginal maximum likelihood estimation, exemplified by BILOG [Mislevy & Bock, 1981], which provides asymptotic sampling variances for the estimates obtained. This approach was not available to the authors at the time the investigation reported here was initiated. It is not discussed here.) Throughout this paper all sampling variances, covariances, and standard errors are asymptotic.

Section 2 presents needed additional, though not conclusive, evidence that the Lord-Wingersky method for computing the sampling variance-covariance matrix yields correct results. Section 3 investigates the effect of changing the number of items, the number of people, or the distribution of ability, on the standard errors of both the item parameters and the abilities. Section 4 presents a technique for displaying and understanding the standard errors and sampling covariances of estimates of item parameters.

Section 5 deals with the situation when there are two tests that contain a set of items in common and these tests are administered to two separate groups of examinees. An important problem in item banking or test equating is to put the parameter estimates for the two tests on a common scale. One way to do this is to estimate all of the parameters for both tests in one calibration run. When this is done,

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 8, No. 3, Summer 1984, pp. 347–364
© Copyright 1984 Applied Psychological Measurement Inc.
0146-6216/84/030347-18$2.15

how does the number and quality of the common items affect the standard errors of the parameter estimates for the unique (noncommon) items?

1. Preliminaries

The three-parameter Birnbaum logistic model is used throughout. The probability of examinee a answering item i correctly is

$$P_{ia} \equiv P_i(\theta_a) = c_i + (1 - c_i)/\{1. + \exp[-1.7a_i(\theta_a - b_i)]\} \tag{1}$$

where a_i is the discrimination of item i,

b_i is the difficulty for the item,

c_i is the lower asymptote of the item response function, and

θ_a is the ability for examinee a.

In a typical calibration run, poorly estimatable c_i are ordinarily fixed at some common value. In this paper, however, all c_i are considered unknown and must be estimated. Treating all of the c_i as unknown results in the "worst case" standard errors.

In IRT, the origin and unit of measurement of the ability scale is arbitrary. Until this scale is specified, all parameters except the c_i are *unidentifiable*. The origin and unit of the ability scale must be specified in terms of (as a function of) the true parameters. If the origin and unit of the ability scale were specified in terms of the parameter estimates, then the true parameters would be undefined. Since the true parameters are unknown but depend on the scale used, this means that the scale origin and the scale unit (each defined as a function of the true parameters) must be estimated from the data. The estimated origin and scale unit are obviously subject to sampling errors, which affect the accuracy of all parameter estimates. It is therefore important to define the origin and unit each by a function of parameters that can be estimated with good accuracy.

If the scale were fixed by setting $a_1 = 1$, $b_1 = 0$, for example, then the accuracy of all parameter estimates would depend on how accurately a_1 and b_1 can be estimated from the data. If Item 1 happened to be a nondiscriminating item, the pattern of observed responses would be independent of the value of b_1. Thus in this extreme case, the parameter scale could not be inferred from the data and all a_i and b_i would be unidentifiable—no a_i or b_i could be estimated from the data.

The scale recommended in Lord and Wingersky (in press) and used here requires that the mean of the difficulty parameters of certain selected items be 0 (the origin) and that the difference between two such means for two sets of selected items be 1 (the scale unit). This scale will be referred to as the "capital" scale: parameters on this scale are denoted by the capital letters A_i, B_i, C_i, Θ_a. The "small" scale or the "LOGIST" scale, referred to by lower-case letters, is the scale used by the LOGIST program (Wingersky, Barton, & Lord, 1982), the computer program used here for estimating the parameters of Equation 1 by maximum likelihood. LOGIST sets a truncated mean of the estimated abilities to 0 and a truncated standard deviation of the estimated abilities to 1. The following formulas convert the parameters from the LOGIST scale to the capital scale:

$$\Theta_a = (\theta_a - \bar{b}_0)/k \quad , \tag{2}$$

$$k = \bar{b}_1 - \bar{b}_0 \quad , \tag{3}$$

$$A_i = ka_i \quad , \tag{4}$$

$$B_i = (b_i - \bar{b}_0)/k \quad , \tag{5}$$

and

$$C_i = c_i \quad , \tag{6}$$

where $\bar{b}_0$ and $\bar{b}_1$ are means of the b_i for two selected subsets of items. The capital scale is a linear transformation of the LOGIST scale. The c_i are not affected by the scale.

2. Variance of p_i, the Proportion Correct

If it could be proven that the maximum likelihood parameter estimates for the Birnbaum model are consistent when all item and ability parameters are estimated simultaneously and when the number of examinees and the number of items both become large simultaneously, then the sampling variance-covariance matrix described in Lord and Wingersky (in press) would be the correct matrix to use. Since consistency has not yet been proven mathematically, any results that confirm the appropriateness of this variance-covariance matrix makes a researcher feel more comfortable about using it.

The sampling variance of p_i, the proportion of examinees in the sample who answer item i correctly, can be computed directly from familiar standard formulas; it can also be computed with some effort from the sampling variance-covariance matrix obtained by Lord and Wingersky (in press). These two methods should give the same results if the Lord-Wingersky matrix is correct.

The usual likelihood equations for $\hat{b}_i$ and for $\hat{c}_i$, obtained by setting the derivative of the likelihood function equal to zero, are (Lord, 1980, Eqs. 12.1 and 12.2)

$$\sum_{a=1}^{N} [u_{ia} - \hat{P}_i(\hat{\theta}_a)][\hat{P}_i(\hat{\theta}_a) - \hat{c}_i]/\hat{P}_i(\hat{\theta}_a) = 0 \quad , \tag{7}$$

$$\sum_{a=1}^{N} [u_{ia} - \hat{P}_i(\hat{\theta}_a)]/\hat{P}_i(\hat{\theta}_a) = 0 \quad , \tag{8}$$

where u_{ia} is the score (0 or 1) of examinee a on item i, N is the number of examinees, and a caret denotes substitution of parameter estimates for true parameter values. Multiplying Equation 8 by c_i, adding to Equation 7, and transposing gives

$$\sum_{a=1}^{N} \hat{P}_i(\hat{\theta}_a) = \sum_{a=1}^{N} u_{ia} \quad . \tag{9}$$

Since

$$p_i = \frac{1}{N} \sum_{a=1}^{N} u_{ia} \quad , \tag{10}$$

it follows that

$$p_i = \frac{1}{N} \sum_{a=1}^{N} \hat{P}_i(\hat{\theta}_a) \quad (i = 1, 2, \ldots, n) \quad . \tag{11}$$

From Equations 10 and 11, two separate formulas for the variance of p_i can be derived.

For some group of examinees whose abilities are specified by the vector $\boldsymbol{\theta} \equiv \{\theta_1, \theta_2, \ldots, \theta_N\}$, it follows from Equation 10 that

$$\begin{aligned} \mathrm{var}(p_i|\boldsymbol{\theta}) &= \frac{1}{N^2} \sum_{a=1}^{N} \sum_{a'=1}^{N} \mathrm{cov}(u_{ia}, u_{ia'}|\boldsymbol{\theta}) \quad , \\ &= \frac{1}{N^2} \sum_{a=1}^{N} \mathrm{var}(u_{ia}|\boldsymbol{\theta}) \quad , \\ &= \frac{1}{N^2} \sum_{a=1}^{N} P_i(\theta_a) Q_i(\theta_a) \quad , \end{aligned} \tag{12}$$

with

$$Q_i(\theta_a) = 1 - P_i(\theta_a) \quad , \tag{13}$$

since $\operatorname{cov}(u_{ia}, u_{ia'}|\boldsymbol{\theta}) = 0$ when $a \neq a'$. Similarly, $\operatorname{cov}(p_i, p_j|\boldsymbol{\theta}) = 0$.

By the formula for the covariance between two sums, it follows from Equation 11 for the same group of examinees that

$$\operatorname{var}(p_i|\boldsymbol{\theta}) = \frac{1}{N^2} \sum_{a=1}^{N} \sum_{b=1}^{N} \operatorname{cov}[\hat{P}_i(\hat{\theta}_a), \hat{P}_i(\hat{\theta}_b)|\boldsymbol{\theta}] \quad , \tag{14}$$

$$\operatorname{cov}(p_i, p_j|\boldsymbol{\theta}) = \frac{1}{N^2} \sum_{a=1}^{N} \sum_{b=1}^{N} \operatorname{cov}\,[\hat{P}_i(\hat{\theta}_a), \hat{P}_j(\hat{\theta}_b)|\boldsymbol{\theta}] \quad . \tag{15}$$

If the parameter estimates are consistent, the $\operatorname{cov}[\hat{P}_i(\hat{\theta}_a), \hat{P}_j(\hat{\theta}_b)|\boldsymbol{\theta}]$ can be evaluated asymptotically by applying the delta method (Kelley, 1947, pp. 524–526; Kendall & Stuart, 1969, Section 10.6) to Equation 1. For fixed $\boldsymbol{\theta}$ (for simplicity, the notation "$|\boldsymbol{\theta}$" is omitted from the following formula):

$$\begin{aligned}\operatorname{cov}[\hat{P}_i(\hat{\theta}_a), \hat{P}_j(\hat{\theta}_b)] \cong{}& w_{ia} w_{jb} \{t_{ia} t_{jb} [\operatorname{cov}(\hat{\theta}_a, \hat{\theta}_b) - \operatorname{cov}(\hat{b}_i, \hat{\theta}_b) - \operatorname{cov}(\hat{\theta}_a, \hat{b}_j) + \operatorname{cov}(\hat{b}_i, \hat{b}_j)] \\ &+ v_{ia} t_{jb} [\operatorname{cov}(\hat{a}_i, \hat{\theta}_b) - \operatorname{cov}(\hat{a}_i, \hat{b}_j)] + v_{jb} t_{ia} [\operatorname{cov}(\hat{\theta}_a, \hat{a}_j) - \operatorname{cov}(\hat{b}_i, \hat{a}_j)] \\ &+ v_{ia} v_{jb} \operatorname{cov}(\hat{a}_i, \hat{a}_j) + t_{jb} [\operatorname{cov}(\hat{c}_i, \hat{\theta}_b) - \operatorname{cov}(\hat{c}_i, \hat{b}_j)]/1.7 \\ &+ [v_{jb} \operatorname{cov}(\hat{c}_i, \hat{a}_j) + v_{ia} \operatorname{cov}(\hat{a}_i, \hat{c}_j)]/1.7 + t_{ia} [\operatorname{cov}(\hat{\theta}_a, \hat{c}_j) - \operatorname{cov}(\hat{b}_i, \hat{c}_j)]/1.7 \\ &+ \operatorname{cov}(\hat{c}_i, \hat{c}_j)/(1.7)^2\} \quad ,\end{aligned} \tag{16}$$

where

$$w_{ia} = [1.7 Q_i(\theta_a)]/(1 - c_i) \quad , \tag{17}$$

$$t_{ia} = a_i [P_i(\theta_a) - c_i] \quad , \tag{18}$$

and

$$v_{ia} = (\theta_a - b_i) [P_i(\theta_a) - c_i] \quad . \tag{19}$$

The standard errors for p_i were calculated from Equation 12 and again (asymptotically) from Equations 14 and 16 for each of the 45 items in the test described in Section 3. The results from two different approaches agree to at least three significant digits for each item. The $\operatorname{cov}(p_i, p_j|\boldsymbol{\theta})$ obtained from Equations 15 and 16 were all of order 10^{-7} or less. This gives increased confidence in the Lord-Wingersky sampling covariance matrix.

3. Effects of Changing Number of Items, Number of Examinees, or the Frequency Distribution of Ability

To investigate the effect of changing the number of items, the number of examinees, or the distribution of abilities on the sampling errors of parameter estimates, various sets of parameters were specified. The simplest set of parameters represents the administration of a 45-item test to 1,500 examinees. The numerical values used as the true θ_a were a spaced sample of 1,500 $\hat{\theta}_a$ drawn from the ability estimates obtained by LOGIST for a regular administration of the Test of English as a Foreign Language (TOEFL). A spaced sample of 15 items were drawn from the 60 TOEFL items whose parameters were estimated in the same run as the abilities. The estimated parameters for these 15 items were used as the true parameters. These 15 items were then replicated twice to obtain a total of 45 items, of which Items 16–30 and Items 31–45 had the same item parameters as Items 1–15. Note that various parameters were specified, but no sets of artificial data were generated for this study, since sampling variances and covariances depend only on the true parameters, not on sample observations.

To investigate the effect of increasing the number of examinees, each of 1,500 θ_a was repeated four times to represent the θ_a of 6,000 examinees. To study the effect of increasing the number of items, another 45 items were added exactly like the first 45 to create a 90-item test. For a different distribution of abilities, a random sample of 1,500 θ_a was drawn from the rectangular distribution in the interval $[-3,3]$.

Tables 1 through 4 give the standard errors of the parameter estimates that would be obtained from actual data in the various situations investigated. Only the standard errors for the 15 unique items (Items 1–15) are given in the tables of the standard errors for the item parameters. The abilities are grouped into 16 intervals between -4 and 3. Two of the intervals had no examinees. N is the number of examinees and n is the number of items. The values of both the small and capital parameters are given. The constants to convert from the small scale to the capital scale are $\bar{b}_0 = -.305$ and $k = .976$.

Figure 1 contains plots corresponding to these tables. Gaps in the curve for the standard error of $\hat{B}_i$ are due to some points located out of the range of the plot. The standard error for $\hat{C}_i$ is not plotted against C_i, since most of the C_i are equal, but against $B_i - 2/A_i$ instead. $B_i - 2/A_i$ is an indicator of the ability level at which the item response curve becomes asymptotic. The higher $B_i - 2/A_i$, the better C_i should be estimated.

As expected, quadrupling the number of examinees halved the standard errors of the estimated item parameters; doubling the number of items decreased the standard errors of the estimated abilities by a factor of $2^{1/2}$. Quadrupling the number of examinees reduced the largest standard error for $\hat{\theta}_a$ sharply, but had little effect on the smaller standard errors; doubling the number of items had only a moderate or small effect on the standard errors of item parameter estimates. Note that these effects cannot be investigated at all using the usual standard error formulas, which do not deal with the situation when item parameters and ability parameters are both estimated simultaneously.

The rectangular distribution of abilities definitely gave better estimates of the item parameters than the bell-shaped distribution of abilities. For C_i where $B_i - 2/A_i$ is low, the rectangular distribution gave standard errors nearly as low as the standard errors with quadruple the number of examinees.

Table 1
Standard Errors for $\hat{A}_i$

			Standard Errors for $\hat{A}_i$			
			Bell-shaped distribution			Rectangular
Item			n=45	n=90	n=45	n=45
No.	a_i	A_i	N=1500	N=1500	N=6000	N=1500
1	0.99	0.96	0.234	0.192	0.117	0.178
2	0.35	0.34	0.134	0.131	0.067	0.072
3	1.38	1.34	0.318	0.243	0.159	0.235
4	0.78	0.76	0.147	0.126	0.073	0.099
5	0.42	0.41	0.100	0.106	0.050	0.055
6	0.92	0.90	0.178	0.145	0.089	0.120
7	0.92	0.90	0.179	0.147	0.089	0.119
8	1.06	1.04	0.209	0.168	0.104	0.141
9	1.34	1.31	0.262	0.205	0.131	0.180
10	1.50	1.46	0.317	0.259	0.158	0.231
11	0.87	0.85	0.180	0.151	0.090	0.117
12	0.62	0.60	0.142	0.128	0.071	0.086
13	1.09	1.06	0.234	0.197	0.117	0.153
14	1.39	1.36	0.311	0.265	0.156	0.204
15	1.50	1.46	0.333	0.283	0.166	0.209

Table 2
Standard Errors for $\hat{B}_i$

			Standard Errors for $\hat{B}_i$			
			Bell-shaped distribution			Rectangular
Item			n=45	n=90	n=45	n=45
No.	b_i	B_i	N=1500	N=1500	N=6000	N=1500
1	-2.01	-1.75	0.516	0.466	0.258	0.339
2	-1.61	-1.33	2.544	2.344	1.272	1.470
3	-1.09	-0.80	0.353	0.259	0.177	0.242
4	-0.77	-0.48	0.257	0.240	0.128	0.177
5	-0.67	-0.38	0.965	0.929	0.483	0.591
6	-0.34	-0.04	0.191	0.161	0.095	0.141
7	-0.15	0.16	0.165	0.141	0.082	0.128
8	0.00	0.31	0.143	0.117	0.071	0.113
9	0.11	0.42	0.124	0.096	0.062	0.096
10	0.26	0.58	0.110	0.092	0.055	0.097
11	0.46	0.79	0.103	0.101	0.051	0.098
12	0.57	0.90	0.178	0.179	0.089	0.148
13	0.68	1.01	0.085	0.086	0.043	0.086
14	0.90	1.23	0.082	0.080	0.041	0.076
15	1.16	1.50	0.103	0.089	0.052	0.077

4. Displaying Standard Errors and Sampling Covariances

In looking at tables of standard errors, it is difficult to see how the standard errors for $\hat{A}_i$, $\hat{B}_i$, and $\hat{C}_i$ interrelate and how the standard errors relate to the magnitude of the parameters. A plot of the three-dimensional asymptotic joint normal distribution of $\hat{A}$, $\hat{B}$, and $\hat{C}$ would be useful but difficult to read.

Table 3
Standard Errors for $\hat{C}_i$

			Standard Errors for $\hat{C}_i$			
			Bell-shaped distribution			Rectangular
Item			n=45	n=90	n=45	n=45
No.	c_i	C_i	N=1500	N=1500	N=6000	N=1500
1	0.17	0.17	0.598	0.469	0.299	0.316
2	0.17	0.17	0.715	0.628	0.358	0.409
3	0.17	0.17	0.096	0.083	0.048	0.045
4	0.17	0.17	0.144	0.123	0.072	0.080
5	0.17	0.17	0.318	0.280	0.159	0.183
6	0.17	0.17	0.071	0.064	0.035	0.039
7	0.17	0.17	0.059	0.054	0.029	0.033
8	0.17	0.17	0.041	0.039	0.021	0.025
9	0.13	0.13	0.026	0.025	0.013	0.018
10	0.34	0.34	0.026	0.026	0.013	0.021
11	0.17	0.17	0.039	0.038	0.020	0.025
12	0.17	0.17	0.068	0.064	0.034	0.039
13	0.25	0.25	0.027	0.027	0.014	0.021
14	0.29	0.29	0.020	0.020	0.010	0.018
15	0.18	0.18	0.015	0.015	0.007	0.015

Table 4
Standard Errors for $\hat{\Theta}_a$

		Standard Errors for $\hat{\Theta}_a$			
		Bell-shaped distribution			Rectangular
θ_a	Θ_a	n=45 N=1500	n=90 N=1500	n=45 N=6000	n=45 N=1500
-2.75	-2.51	2.090	1.478	1.331	1.453
-2.25	-1.99	1.296	0.917	0.879	0.955
-1.75	-1.48	0.861	0.609	0.621	0.669
-1.25	-0.97	0.607	0.429	0.460	0.491
-0.75	-0.46	0.456	0.322	0.373	0.390
-0.25	0.06	0.349	0.247	0.309	0.317
0.25	0.57	0.278	0.196	0.266	0.268
0.75	1.08	0.261	0.185	0.260	0.261
1.25	1.59	0.303	0.214	0.292	0.295
1.75	2.11	0.422	0.298	0.394	0.401
2.25	2.62	0.628	0.444	0.589	0.599
2.75	3.13	0.931	0.658	0.888	0.900

However, projections of the contours of this distribution onto the 3 two-dimensional planes gives a graphical representation not only of the magnitude of the standard errors but also of the sampling correlations between the parameter estimates. The projected contours are two-dimensional ellipses. These plots are a refinement of a suggestion by T. Warm (personal communication, 1981).

For convenience, the subscript i is dropped. To plot the projection of the three-dimensional contour onto the $(\hat{A},\hat{B})$-plane, only var$(\hat{A})$, var$(\hat{B})$, and cov$(\hat{A},\hat{B})$ are needed. The exponent of the asymptotic bivariate normal distribution of $\hat{A}$ and $\hat{B}$ is given by the right side of Equation 20. The quadratic in brackets is asymptotically distributed as chi square with 2 degrees of freedom. The 95th percentile for a χ^2 with 2 degrees of freedom is 5.99. Thus 95% of the time, the obtained $(\hat{A},\hat{B})$ will lie within the ellipse given by the equation

$$5.99 = \frac{1}{1-\rho^2}\left\{\frac{(\hat{A}-A)^2}{\mathrm{var}(\hat{A})} - \frac{2\rho\,(\hat{A}-A)(\hat{B}-B)}{[\mathrm{var}(\hat{A})\,\mathrm{var}(\hat{B})]^{1/2}} + \frac{(\hat{B}-B)^2}{\mathrm{var}(\hat{B})}\right\} , \tag{20}$$

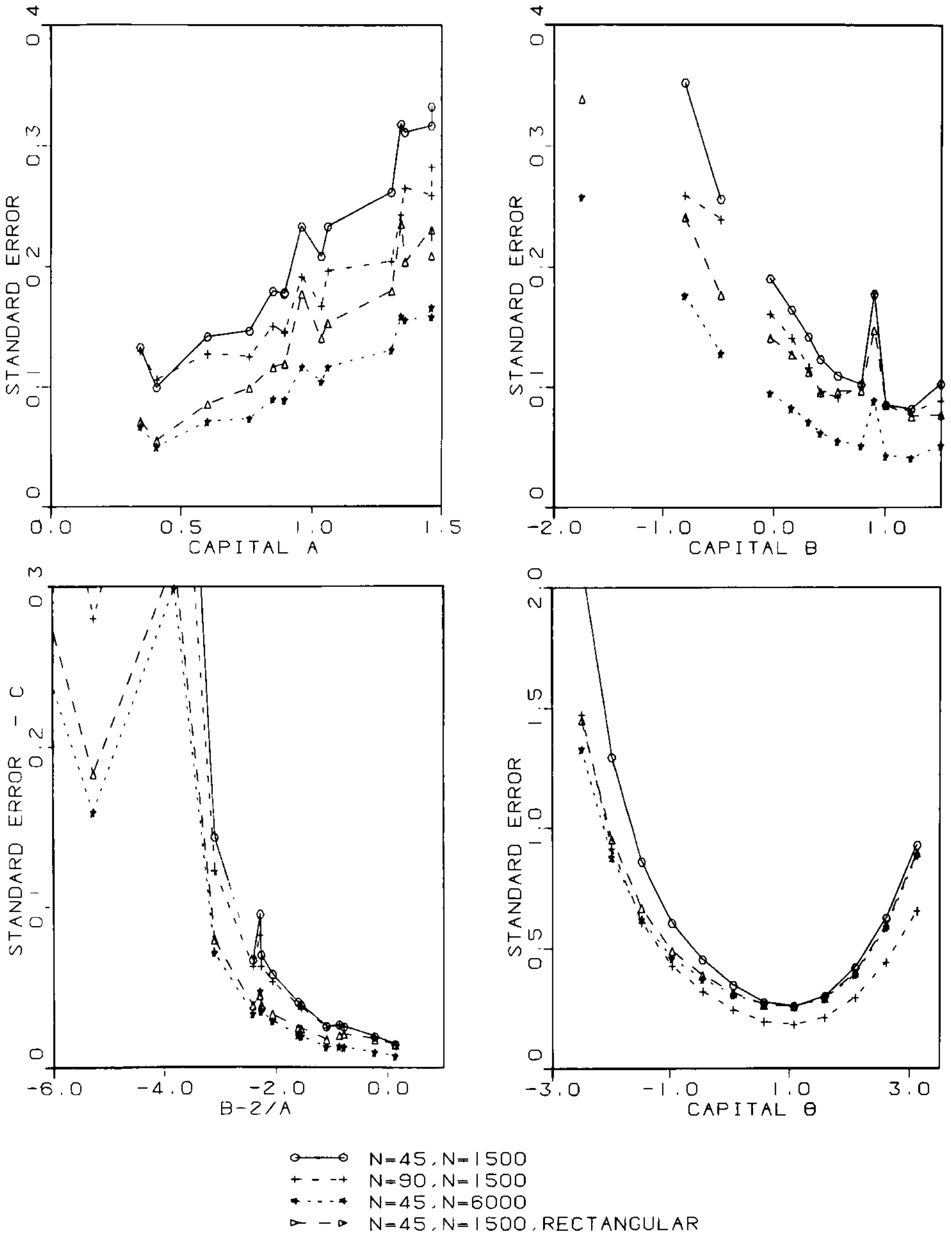
where

$$\rho = \frac{\mathrm{cov}(\hat{A},\hat{B})}{[\mathrm{var}(\hat{A})\,\mathrm{var}(\hat{B})]^{1/2}} . \tag{21}$$

Similar equations apply for the projections onto the $(\hat{A},\hat{C})$- and $(\hat{B},\hat{C})$-planes. The ellipse plotted from Equation 20 for a given N is identical to the 53% ellipse that would be plotted for a sample size of $N/4$.

The following procedure was used to plot a representative set of ellipses. A hypothetical test of 60 items was created by selecting 60 items from an operational Scholastic Aptitude Test (SAT) mathematics test and treating these item parameter estimates as the true parameters. A standard normal distribution of 1,000 abilities was generated. Fifteen new items were then generated with all combinations of the parameters $a = .5, 1.0, 1.5$; $b = -2, -1, 0, 1, 2$; and $c = .15$. Using these new items, fifteen 61-item tests were created, each containing the 60 original items and one of the new items. The sampling variance-covariance matrix for each of the fifteen 61-item tests was obtained. These matrices differ only because the 61st item differed for each matrix. Only the variances and covariances for the 61st item were used in Equation 20 to compute the ellipses.

Figure 1
Comparison of the Standard Errors for $\hat{A}_i$, $\hat{B}_i$, $\hat{C}_i$, and $\hat{\theta}_a$
for Different Numbers of Items, Different Numbers of Examinees,
and for a Different Distribution of Examinees



The plots were made for an N of 16,000 to avoid a confusing overlap of the ellipses. These ellipses are also the 53% confidence ellipses for an N of 4,000. The left and bottom axes are labeled with the small scale; the right and top axes are labeled with the capital scale. The standard errors used are for parameter estimates on the capital scale. The transformation parameters to transform from the small to the capital scale are $\overline{b}_0$ = .001, k = 1.336. The center of the ellipse is marked by a "+."

Figure 2 shows the ellipses on the $(\hat{A},\hat{B})$-plane. The plot shows that the standard error of $\hat{A}$ increases with A. The standard error of $\hat{B}$ increases as B approaches the extremes. The sampling correlation between $\hat{A}$ and $\hat{B}$ is moderately or strongly positive for easy items and moderately or strongly negative for difficult items.

Figure 3 shows the projections onto the $(\hat{B},\hat{C})$-plane. At each value of B there are three ellipses, which are concentric because $c = C = .15$ for all items. The longest ellipse along the C axis is for $a = .5$, the middle ellipse is for $a = 1.0$, and the shortest is for $a = 1.5$. The other triples of ellipses are similarly ordered on a. The standard error of $\hat{C}$ is large for easy items and moderately small for difficult items; the standard error of $\hat{C}$ decreases as a increases. As a decreases, the sampling correlation between $\hat{B}$ and $\hat{C}$ becomes strongly positive except for difficult items where $\hat{C}$ is well determined.

Figure 4 shows the projections onto the $(\hat{A}, \hat{C})$-plane. There are five concentric ellipses for each value of a. The ellipse with the longest c-axis is for $b = -2.0$, the ellipse with the shortest c-

Figure 2

Projections onto the $(\hat{A},\hat{B})$-Plane of the 95% Ellipses for an N of 16,000

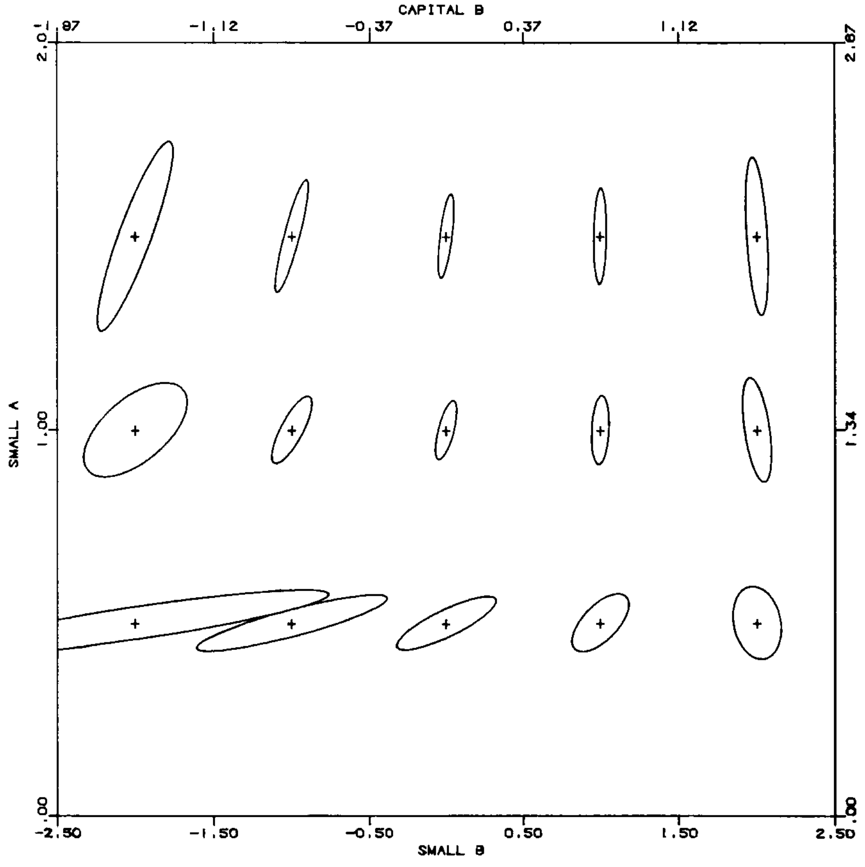
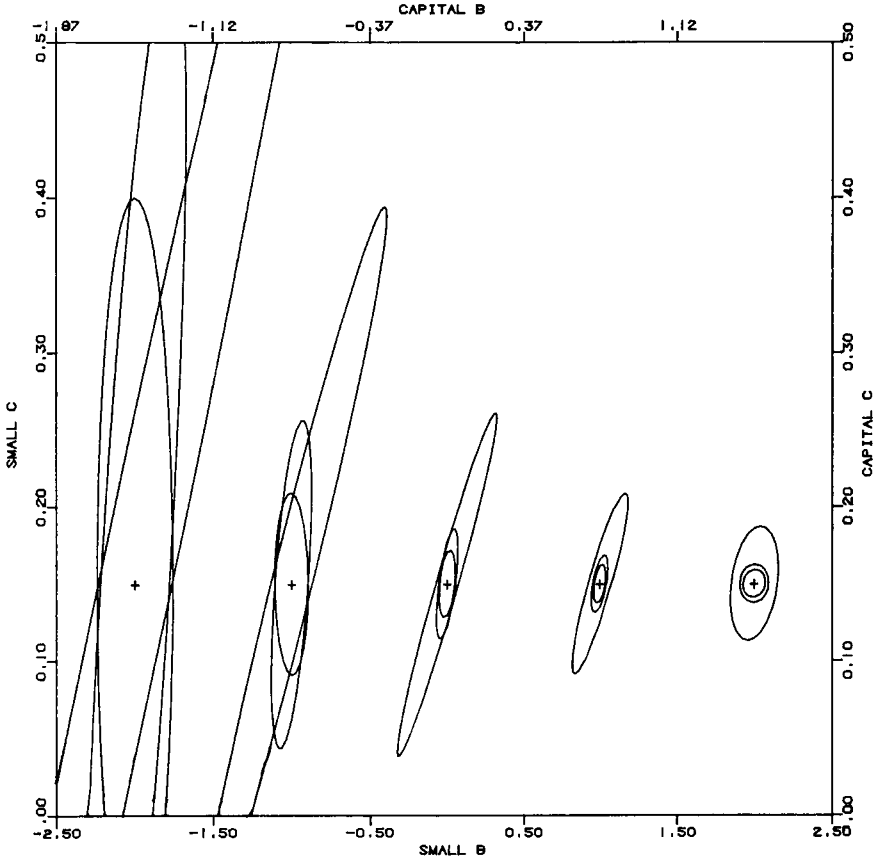


Figure 3

Projections onto the $(\hat{B},\hat{C})$-Plane of the 95% Ellipses for an N of 16,000



axis is for $b = 2.0$. Again $\hat{C}$ has large standard errors for easy items and for items with low as. For difficult items, the sampling correlation between $\hat{A}$ and $\hat{C}$ is positive and sometimes high; for easy items, the correlation is negative.

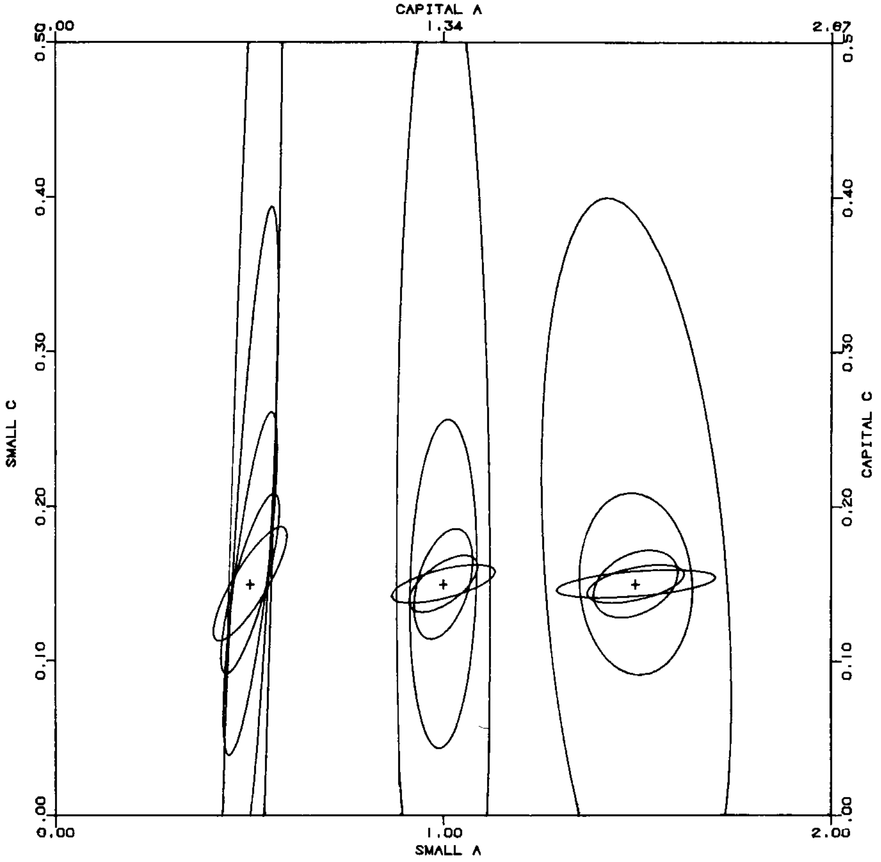
5. Standard Errors for Two Tests with Common Items

Suppose that each of two tests measuring the same ability is administered to a different group of examinees. Item response theory can then be used either to put the items for both tests into a common item pool or to equate the two tests. For either purpose, it is necessary that all the estimated parameters be on the same scale.

Unless equivalent groups of examinees are used, methods for doing this usually require a subset of items that are common to both tests. The unique items are the items in each test that are not common to the other test. The item parameters for each test can then be estimated, either separately in two calibration runs or together in one calibration run. If the parameters are estimated in two separate runs, there are two different parameter estimates for each common item. These should be the same except for sampling error and the arbitrary origin and unit of measurement of the ability scale. Several methods exist for determining the linear transformation necessary to transform the item parameter estimates for both tests

Figure 4

Projections onto the $(\hat{A},\hat{C})$-Plane of the 95% Ellipses for an N of 16,000



to the same scale. These methods are not described here (see Stocking & Lord, 1983). However, if all of the items for both tests are calibrated in one run, called a concurrent calibration, then the parameters for both tests are automatically put on the same scale and no linear transformation is necessary. This concurrent procedure is more efficient; it provides smaller standard errors and involves fewer assumptions than other procedures. The concurrent procedure is the procedure studied here.

One question that arises when applying the common item method for putting the parameters for both tests on a common scale is: How many common items are necessary? Vale, Maurelli, Gialluca, Weiss, and Ree (1981) investigated this problem using simulated data with 5, 15, and 25 common items and three different shapes of the common item section test information curve: peaked, normal, and rectangular. They also investigated many other linking methods. For the common item method, Vale et al. (1981) assumed that good estimates of the parameters for the common items were already known, and they required that there be enough common and unique items to get good estimates of the abilities. They used two estimates of the abilities—one obtained from the common items, the other from the unique items—to determine the transformation to put the unique items onto the common scale. They found that 15 to 25 items were necessary and that the common item sections with a rectangular or normal information function were better than those with a peaked information function.

Another study to determine the number of common items necessary was done by McKinley and Reckase (1981). They worked with real data from a multidimensional achievement test covering seven different areas of achievement. McKinley and Reckase concluded that 5 items were not adequate, 25 items were better than 15, but 15 were adequate for linking with the concurrent method. Since their data clearly violated the unidimensionality assumption of their model, there is little reason to consider their study in detail here.

Given the sampling variance-covariance matrix for all parameter estimates in a single concurrent run when all parameters are treated as unknown, what effect the number of common items has on the sampling standard errors of the unique items in both tests can be investigated. Note that this problem cannot be investigated at all with the limited sampling-error formulas that assume that item and ability parameters are not estimated simultaneously.

Numerical Procedures

Suppose Test 1 has a section of unique items labeled V4, and Test 2 has a section of unique items labeled Z5. Both tests have the same set of common items labeled C0. One group of examinees, Group X, took Test 1, another group of examinees, Group Y, took Test 2. The information matrix $\|\mathbf{I}_{pq}\|$, which must be inverted to get the variance-covariance matrix, has the following structure (Lord & Wingersky, in press):

$$\|\mathbf{I}_{pq}\| = \begin{array}{ccc|cc} \multicolumn{3}{c}{\text{Items}} & \multicolumn{2}{c}{\text{Examinees}} \\ \text{V4} & \text{C0} & \text{Z5} & \text{Group X} & \text{Group Y} \\ \hline \boldsymbol{S}_{11} & \mathbf{0} & \mathbf{0} & \boldsymbol{F}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}_{22} & \mathbf{0} & \boldsymbol{F}_{21} & \boldsymbol{F}_{22} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{S}_{33} & \mathbf{0} & \boldsymbol{F}_{32} \\ \hline \boldsymbol{F}_{11} & \boldsymbol{F}_{21} & \mathbf{0} & \boldsymbol{T}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{F}_{22} & \boldsymbol{F}_{32} & \mathbf{0} & \boldsymbol{T}_{22} \end{array}$$

Each **S** submatrix ($\mathbf{S}_{11}$ for the V4 items; $\mathbf{S}_{22}$ for the common items; $\mathbf{S}_{33}$ for the Z5 items) contains 3 × 3 Fisher information matrices for a_i, b_i, c_i on the diagonal and zeros elsewhere. Each **T** submatrix is a diagonal information matrix for examinees: $\mathbf{T}_{11}$ for those that took Test 1; $\mathbf{T}_{22}$ for those that took Test 2. Each **F** submatrix contains the vectors $\mathbf{f}_{ia}$, 3 × 1 Fisher information vectors for item i and examinee a. Note that for Group Y, $\mathbf{f}_{ia}$ is $\mathbf{0}$ for the V4 items; for Group X, $\mathbf{f}_{ia}$ is $\mathbf{0}$ for Z5.

The matrix $\|\mathbf{I}_{pq}\|$ is inverted by grouping the abilities for Group X into 16 groups and by grouping the abilities for Group Y into another set of 16 groups. Then the formulas for inverting a partitioned matrix using the method described in Lord and Wingersky (in press) are successively applied.

Data and Results

To study the effect of the number of common items on the standard errors of the parameter estimates for the unique items, two 60-item SAT Mathematics tests with an additional 25-item common-item section were selected. The 60 unique items in the first test are referred to as V4 and the 60 unique items in the second test are referred to as Z5. Estimates of all of the parameters were obtained in one concurrent

LOGIST run from real data. These estimates were treated as true parameter values in computing the standard errors for all 145 items.

Note that this is real data, not artificial data. Substituting parameter estimates for true parameter values is, of course, standard procedure in obtaining estimated standard errors, since true parameters are never known for real data.

The length of the common item section was then doubled by simply replicating the parameters for the 25 common items. Surprisingly, the standard errors for the 120 unique items in V4 and Z5 computed with 50 common items agreed with the standard errors computed with only 25 common items to two decimal places. If doubling the number of common items makes so little difference, what is the effect of halving the number of common items? Or at the extreme, reducing the number of common items to two?

To study the effect of two common items on the standard errors of the unique items, 2 "good" items and 2 "poor" items were selected from the 25 common items. The item parameters and their standard errors (SE) for the 2 good items were

a	SE($\hat{A}$)	b	SE($\hat{B}$)	c	SE($\hat{C}$)
.98	.09	$-.10$	.02	.06	.02
.96	.10	.21	.02	.15	.02

The item parameters and their standard errors for the 2 poor common items were

a	SE($\hat{A}$)	b	SE($\hat{B}$)	c	SE($\hat{C}$)
.32	.10	-1.51	.47	.07	.24
.53	.07	-1.19	.12	.07	.10

These standard errors were computed for the situations in which all 25 common items are included in the parameter estimation run.

The variance-covariance matrix was then obtained for the V4 and Z5 items when only the 2 good common items were included in the estimation run; the variance-covariance matrix was also obtained when only the 2 poor common items were used. The constants to transform from the small scale to the capital scale are $\bar{b}_0 = -.261$ and $k = 1.914$. Only V4 and Z5 items were used to compute $\bar{b}_0$ and k so that the same transformation would apply to all four variance-covariance matrices.

Table 5 gives the medians, and the bottom and top quartiles of the standard errors for $\hat{A}$, $\hat{B}$, and $\hat{C}$, for the V4 and Z5 unique items computed for four different situations: (1) using 50 common items, (2) using 25 common items, (3) using 2 good common items, and (4) using 2 poor common items. Using 2 good common items gave smaller standard errors for the unique items than using 2 poor common items. The standard errors using the 2 good items were not much larger than the standard errors using 25 common items. Even reliance on just 2 poor common items gave surprisingly good results. Since the purpose of the common items is to determine a common scale, it is not surprising that the number of common items has a negligible effect on the standard error of $\hat{C}$, since c is independent of the ability scale.

Table 6 gives the standard errors for the abilities computed with the four different sets of common items. Not surprisingly, if the number of common items is increased to 50, the standard error of the abilities is reduced, although not uniformly as shown by the ratio column. The standard error for the abilities at -2 was lower when computed using the 2 poor common items, which were easy items, than when computed using the 2 good common items.

Even though there is little difference between the standard errors when there are 2 common items and when there are 25 common items, the parameter estimates for the V4 and Z5 items would not have been adequately put on the same scale if all of the parameter estimates for V4 items err in one direction

Table 5
Comparison of the Standard Errors of Estimated Item Parameters across the Four Sets of Common Items

	50 Common Items	25 Common Items	2 Good Common Items	2 Poor Common Items
Standard Errors for $\hat{A}$				
First Quartile	0.114	0.115	0.123	0.131
Median	0.140	0.141	0.151	0.163
Third Quartile	0.224	0.226	0.236	0.243
Standard Errors for $\hat{B}$				
First Quartile	0.029	0.030	0.034	0.041
Median	0.042	0.042	0.048	0.056
Third Quartile	0.066	0.067	0.072	0.076
Standard Errors for $\hat{C}$				
First Quartile	0.013	0.013	0.013	0.013
Median	0.027	0.027	0.028	0.027
Third Quartile	0.055	0.055	0.058	0.056

and all of the parameter estimates for Z5 items err in the opposite direction. Is this what will happen in practice? To determine how well an anchor test of only two common items puts Tests V4 and Z5 on the same scale, the parameters were reestimated twice: once in a LOGIST run with the items for Z5 and V4 and the 2 good common items, and the other in a LOGIST run with the items for Z5 and V4 and the 2 poor common items.

The estimated parameters for Z5 and V4 computed with the 25 common items are used as the criterion for evaluating the calibrations with 2 common items. The 2 good common items do fairly well at putting the parameters on this scale. The 2 poor items do not do so well. The left plot in Figure 5 compares the *b*s for the 60 unique V4 items estimated with 2 good items with the *b*s estimated with 25 common items. Similarly, the plot on the right compares the $\hat{b}$s for the unique Z5 items. If the parameters were on the same metric, the $\hat{b}$s in both plots should fall on a 45° line. The difference from the 45° line is difficult to distinguish. The two points for Z5 that are far away from the 45° line had the $\hat{c}$s fixed by LOGIST at the common $\hat{c}$ value in one calibration but not in the other.

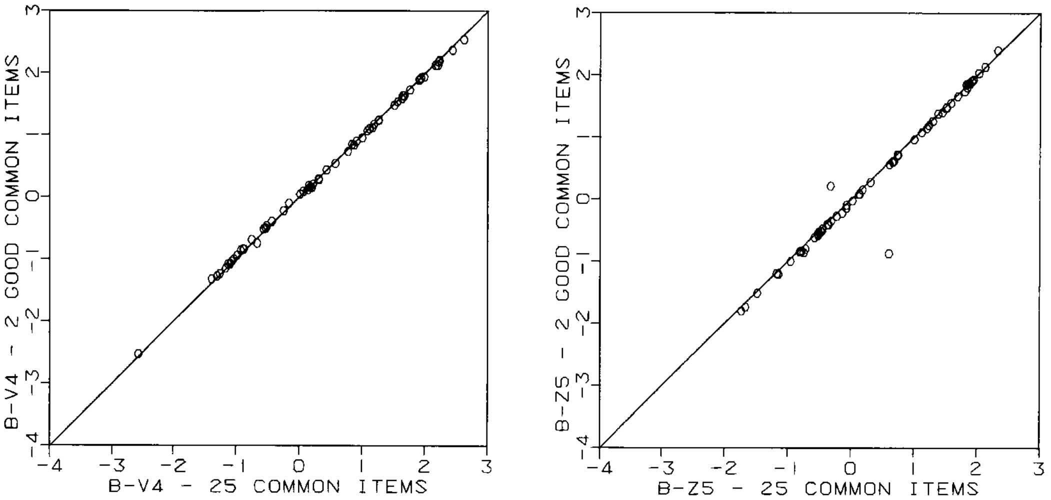
Figure 6 shows the plots for the $\hat{a}$s for V4 and Z5, respectively. Here, it definitely looks as if the $\hat{a}$s are not on the same scale. The $\hat{a}$s for the V4 items have a slope greater than 45°.

Figure 7 compares the *b*s estimated with the 2 poor common items with the *b*s estimated with 25 common items. Here, the points for the V4 items are above the 45° line and points for the Z5 items are below the line. The plots comparing the $\hat{a}$s in Figure 8 confirm that the 2 poor common items do not put

Table 6
Comparison of the Standard Errors of Estimated Abilities across the Four Sets of Common Items

θ_a	Θ_a	50 Common Items S.E	25 Common Items S.E.	Ratio	2 Good Common Items S.E.	2 Poor Common Items S.E.
2.00	1.18	0.097	0.109	0.894	0.127	0.132
1.00	0.66	0.089	0.102	0.870	0.122	0.126
0.0	0.14	0.100	0.115	0.874	0.134	0.138
-1.00	-0.39	0.129	0.145	0.892	0.165	0.167
-2.00	-0.91	0.221	0.248	0.891	0.288	0.281

Figure 5
Comparison of the *b*s Estimated with 2 Good Common Items and the *b*s Estimated with 25 Common Items, Separately for V4 and Z5



the parameters for Z5 and V4 on the same metric. As suspected, with the 2 poor items, the parameters for one set of the unique items err in one direction and for the other set, in the opposite direction.

The reason for putting Z5 and V4 on the same scale was to equate Z5 to V4 using true-score equating. What effect does using only two common items to put the two forms on the same scale have on the true-score equating between the two forms? Figure 9 shows three true-score equating lines: (1) the solid line is the equating line when the parameters are estimated with 25 common items, (2) the dotted line is the

Figure 6
Comparison of the *a*s Estimated with 2 Good Common Items and the *a*s Estimated with 25 Common Items, Separately for V4 and Z5

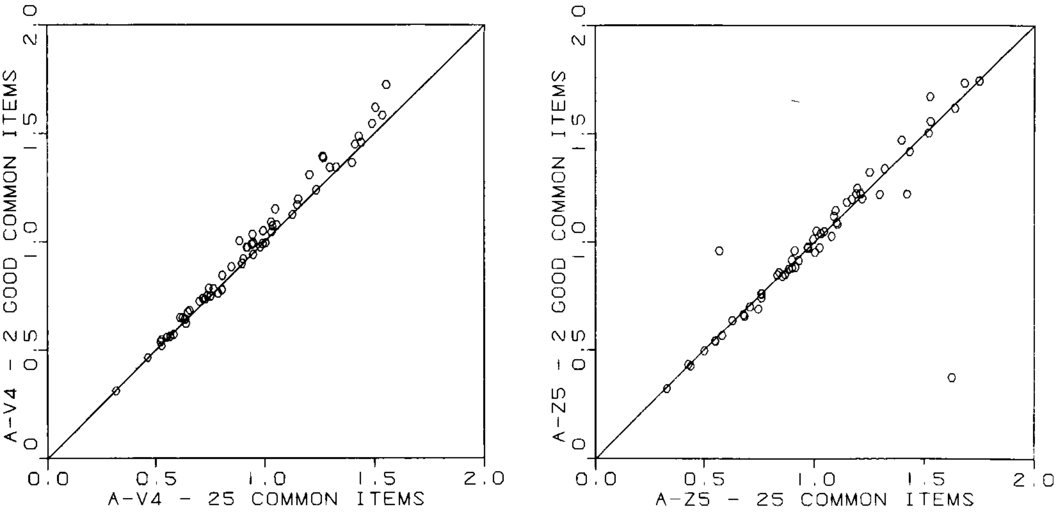
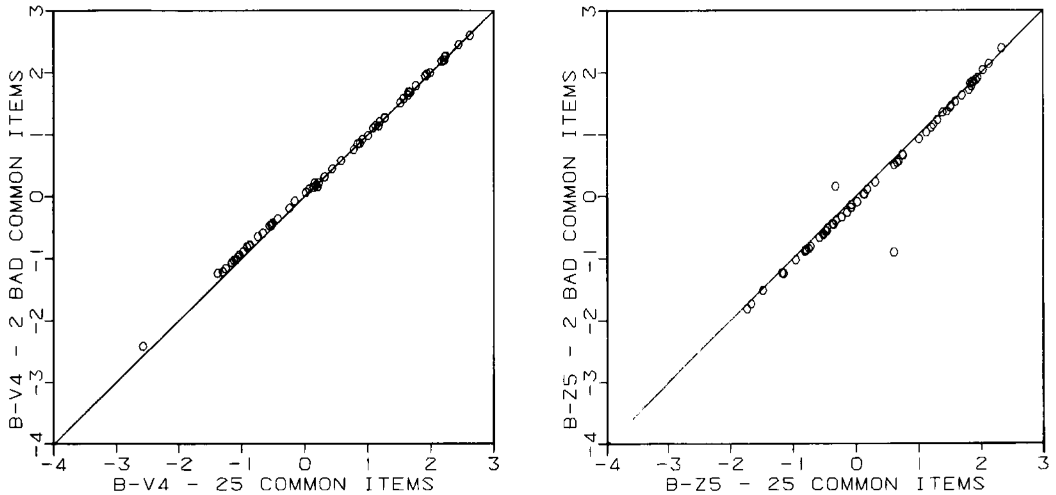


Figure 7
Comparison of the *b*s Estimated with 2 Poor Common Items and the *b*s Estimated with 25 Common Items, Separately for V4 and Z5



equating line when the parameters are estimated with the 2 good common items, and (3) the dashed line is the line when the parameters are estimated with the 2 poor common items. For this equating, true scores on form Z5 were first equated to true scores on V4. Then the true scores on V4 were converted to scaled scores between 100 and 800 by a linear transformation. Using the equating line with the 25 items as a criterion, the equating using 2 poor common items is worse than the equating using 2 good common items. The equating using the 2 good common items is close to the equating with 25 common items; the maximum scaled score difference is eight points.

Figure 8
Comparison of the *a*s estimated with 2 Poor Common Items and the *a*s Estimated with 25 Common Items, Separately for V4 and Z5

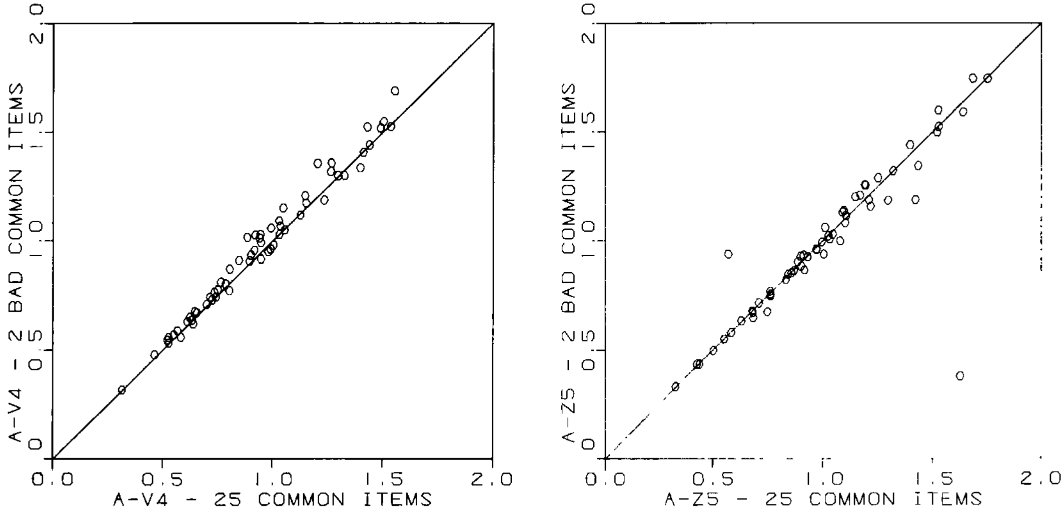
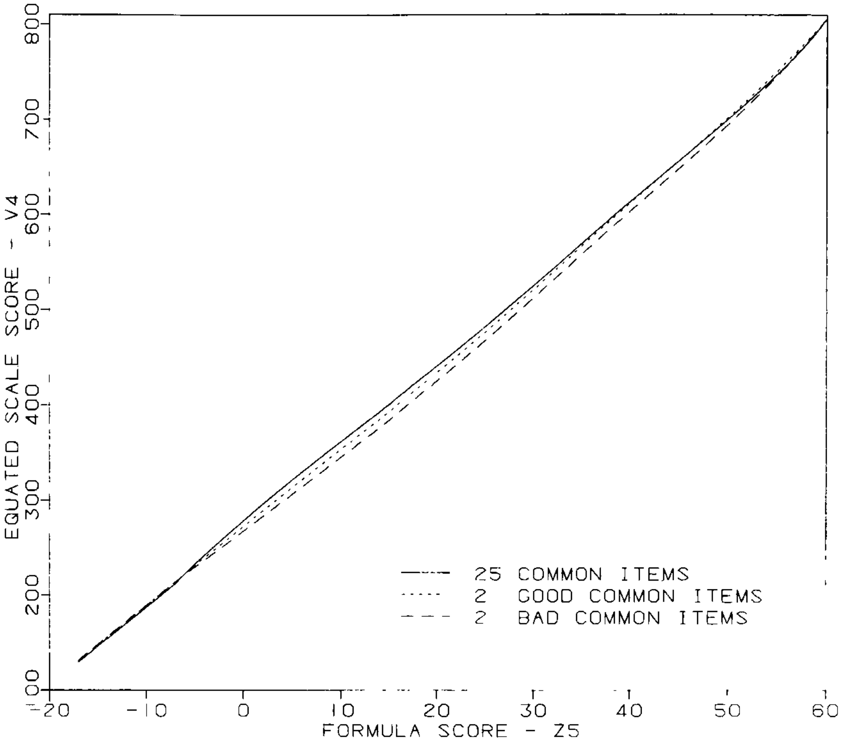


Figure 9
Comparisons of the Three True-Score Equatings of Test Z5 to Test V4:
Using 25 Common Items, Using 2 Good Common Items, and Using 2 Poor Common Items



All of these results assume that the item parameters estimated using 25 common items are on the same scale. This analysis should be repeated in a situation when a researcher knows that all of the parameters used as a criterion are on a common scale. From the results so far, it appears that good linking may be obtained with as few as five common items or less. However, these results only apply when the item parameters for the two forms are put on a common scale by estimating all of the item parameters in one calibration run. These results do not apply when the two tests are calibrated in two separate runs and the parameters are put on a common scale using some linear transformation determined from the common items.

The conclusion that good linking may be obtained with as few as five common items is more optimistic than the conclusions reached by Vale et al. (1981) and by McKinley and Reckase (1981). The differences between the results of Vale et al. and the present results may be because (1) their scaling was based on estimated θs, and (2) they used three estimation runs instead of one concurrent run. The differences from the results of McKinley and Reckase are probably because in their study (1) the responses of some examinees to some items apparently often appeared twice in the same concurrent LOGIST run, violating the assumption of local independence; and, more importantly, (2) they pooled the Iowa Tests of Educational Development covering seven different achievement areas, and analyzed the resulting multidimensional pool of items as if it were unidimensional.

6. Summary

The asymptotic sampling variance-covariance matrix of maximum likelihood estimators when both abilities and item parameters are unknown was used to study several problems in IRT, such as the extent to which more items, more examinees, or a different distribution of abilities will provide better estimates of parameters. It was found for the values of n and N studied that the standard error of $\hat{\theta}$ varies inversely as $n^{1/2}$, but is only moderately affected by changes in N; the standard error of the estimated item parameters varies inversely as $N^{1/2}$, but is only slightly affected by changes in n.

A rectangular distribution of abilities gives smaller standard errors for the item parameters than does doubling the number of items. In fact, for low As, also for Cs for items with $B - 2/A$ less than -1, the standard errors computed with a rectangular distribution of ability are nearly as low as the standard errors computed with a bell-shaped distribution and quadruple the number of people.

With the variance-covariance matrix computed when all parameters are treated as unknown, a researcher can study the effect of the number of common items on the standard errors of the unique items when each of two tests containing common items is administered to a different group of examinees and the parameters for both tests are calibrated in one LOGIST run. This problem cannot be dealt with at all by previously available sampling error formulas, which assume that item and ability parameters are not estimated simultaneously. The number of common items has little effect on the standard errors of the parameters for the unique items. The standard errors indicate that as few as two items may be sufficient providing the parameter estimates for these two items are well determined. However, when two tests were actually calibrated in one LOGIST run using two common items that had parameter estimates with low standard errors, the parameters are not quite on the same scale as the parameters estimated with 25 common items. The $\hat{b}$s are very close to the same scale, but the $\hat{a}$s for one of the tests are on a slightly different scale. Although two items are not quite enough, adequate linking may be possible with as few as five items.

References

Kelley, K. L. (1947). *Fundamentals of statistics*. Cambridge MA: Harvard University Press.

Kendall, M. G., & Stuart, A. (1969). *The advanced theory of statistics* (Vol. 1, 3rd ed.). New York: Hafner.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Wingersky, M. S. (in press). Sampling variances and covariances of parameter estimates in item response theory. In D. J. Weiss (Ed.), *Proceedings of the Item Response Theory and Computerized Adaptive Testing Conference*. Minneapolis MN: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

McKinley, R. L., & Reckase, M. D. (1981). *A comparison of procedures for constructing large item pools* (Research Report 81-3). Columbia MO: University of Missouri, Department of Educational Psychology.

Mislevy, R. J., & Bock, R. D. (1981). *BILOG—Maximum likelihood item analysis and test scoring: Logistic model*. Chicago: International Educational Services.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210.

Vale, C. D., Maurelli, V. A., Gialluca, K. A., Weiss, D. J., & Ree, M. J. (1981). *Methods for linking item parameters* (AFHRL-TR-81-10). Brooks Air Force Base TX: Air Force Human Resources Laboratory.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide*. Princeton NJ: Educational Testing Service.

Acknowledgment

This work was supported in part by contract N00014-80-C-0402, project designation NR 150-453, between the Office of Naval Research and Educational Testing Service.

Author's Address

Send requests for reprints or further information to Marilyn S. Wingersky, Educational Testing Service, Princeton NJ 08541, U.S.A.