PSYCHOMETRIKA—VOL. 48, NO. 2.
JUNE, 1983

UNBIASED ESTIMATORS OF ABILITY PARAMETERS, OF THEIR VARIANCE, AND OF THEIR PARALLEL-FORMS RELIABILITY

FREDERIC M. LORD

EDUCATIONAL TESTING SERVICE

Given known item parameters, unbiased estimators are derived i) for an examinee's ability parameter θ and for his proportion-correct true score ζ, ii) for the variances of θ and ζ across examinees in the group tested, and iii) for the parallel-forms reliability of the maximum likelihood estimator $\hat{\theta}$.

Key words: item response theory, true score, latent trait theory, maximum likelihood.

This paper is primarily concerned with determining the statistical bias in the maximum likelihood estimate $\hat{\theta}$ of the examinee ability parameter $\hat{\theta}$ in item response theory (IRT) [Lord, 1980]; also of certain functions of such parameters. It will deal only with unidimensional tests composed of dichotomously scored items. The item response function is assumed to be three-parameter logistic [see (2) below].

Formulas for the standard error of $\hat{\theta}$ are currently limited to the case where the item parameters are known; the present derivations are limited to this case also. This limitation is tolerable in situations where the item parameters are predetermined, as in item banking and tailored testing.

In the absence of a prior distribution for θ, it is well known that examinees with perfect scores have $\hat{\theta} = \infty$; also that examinees who perform near or below the chance level on multiple-choice items may be given large negative values of $\hat{\theta}$. This (correctly) suggests that $\hat{\theta}$ is positively biased for high-ability examinees and negatively biased for low-ability examinees. Will a correction of $\hat{\theta}$ for bias be helpful in such cases?

It is also 'well known' that for any ordinary group of examinees, the variance ($s_{\hat{\theta}}^2$) of $\hat{\theta}$ across examinees is larger than the variance (s_θ^2) of the true θ. The ratio $s_{\hat{\theta}}^2/s_\theta^2$ is closely related to the classical-test-theory reliability of $\hat{\theta}$ considered as the examinee's test score. Thus it is not enough for us to know that $s_{\hat{\theta}}^2 \to s_\theta^2$ as the number n of test items becomes large; we need to know how the relation of $s_{\hat{\theta}}^2$ to s_θ^2 varies as a function of n. We also need a better estimate of s_θ^2 than its maximum-likelihood estimator $s_{\hat{\theta}}^2$. These objectives can be achieved by correcting $s_{\hat{\theta}}^2$ for bias.

The methods used to derive formulas for correction for bias are presented here in detail for at least two reasons: i) Experience with similar derivations has shown that it is easy to reach erroneous results if details are not spelled out. ii) The general methods used here are easily transferred to solve other problems, such as a) correction of item parameters for bias, b) obtaining higher-order approximations to the sampling variance of $\hat{\theta}$.

1. Statistical Bias in $\hat{\theta}$ and ζ

The method used here to find the bias of $\hat{\theta}$ is adapted from the 'adjusted order of magnitude' procedure detailed by Shenton and Bowman [1977]. They assume their data

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. Reproduction in whole or in part is permitted for any purpose of the United States Government.

Requests for reprints should be addressed to: Frederic M. Lord, Educational Testing Service, Princeton, New Jersey, 08541.

0033-3123/83/0600-4123$00.75/0

to be a sample from a population divided into a denumerable number of subsets. For them, the population proportion of observations in a given subset is a known function of the parameter θ whose value they wish to estimate. Their sample estimate of θ is therefore a function of observed sample proportions in the various subsets. Since our data do not readily fit this picture, we cannot use their final published formulas but must instead derive our own.

Throughout Section 1, we deal with a single fixed examinee whose ability θ is the parameter to be estimated. All item parameters are assumed known.

1.1 Preliminaries

The maximum likelihood estimate $\hat{\theta}$ is obtained by solving the likelihood equation

$$\sum_{i=1}^{n} \frac{(u_i - \hat{P}_i)\hat{P}'_i}{\hat{P}_i\hat{Q}_i} = 0 \tag{1}$$

where $u_i = 0$ or 1 is the examinee's response to item i ($i = 1, 2, \ldots, n$), $P_i \equiv P_i(\theta)$ is the response function for item i, $Q_i \equiv 1 - P_i$, P'_i is the derivative of P_i with respect to θ, and a caret indicates that the function is to be evaluated at $\hat{\theta}$. We deal with the case where P_i is the three-parameter logistic function

$$P_i \equiv c_i + \frac{1 - c_i}{1 + e^{-A_i(\theta - b_i)}} \tag{2}$$

where A_i, b_i, and c_i are item parameters describing item i.

We will assume

i. θ is a bounded variable,

ii. the item parameters A_i and b_i are bounded,

iii. c_i is bounded away from 1, (thus P_i and Q_i are bounded away from 0 and 1);

iv. as n becomes large, the statistical characteristics of the test stabilize.

Rather than trying to define this last assumption formally, the reader may substitute the more restrictive assumption usually made in mental test theory: that a test is lengthened by adding strictly parallel forms.

With these assumptions, the conditions of Bradley and Gart [1962] are satisfied. It follows from their theorems that $\hat{\theta}$ is a consistent estimator of θ and that $(n)^{1/2}(\hat{\theta} - \theta)$ is asymptotically normally distributed with mean zero and variance $\lim_{n\to\infty} 1/n \sum_i^n P_i'^2/P_iQ_i$. The existence of this limit is guaranteed by assumption 4.

For compactness, we will rewrite (1) as

$$\hat{L}_1 \equiv \sum_{i=1}^{n} \hat{\Gamma}_{1i} = 0, \tag{3}$$

where by definition

$$\Gamma_{1i} \equiv \frac{(u_i - P_i)P'_i}{P_iQ_i}. \tag{4}$$

Now $\hat{L}_1$ considered as a function of $\hat{\theta}$ can be expanded formally in powers of $\hat{\theta} - \theta$, as follows:

$$\hat{L}_1 \equiv \sum_i \Gamma_{1i} + (\hat{\theta} - \theta) \sum_i \Gamma_{2i} + \frac{1}{2}(\hat{\theta} - \theta)^2 \sum_i \Gamma_{3i} + \ldots$$

where we define

$$\Gamma_{si} \equiv \frac{d^s}{d\theta^s} \log P_i^{u_i} Q_i^{1-u_i} \qquad (s = 1, 2, \ldots). \tag{5}$$

This definition is consistent with (3).

Let $x \equiv \hat{\theta} - \theta$, $\Gamma_s \equiv \sum_i \Gamma_{si}$. Rather than proving the convergence of the power series, let us use a closed form that is always valid:

$$\hat{L}_1 \equiv \Gamma_1 + x\Gamma_2 + \frac{1}{2}x^2\Gamma_3 + \frac{1}{6}x^3\Gamma_4 + \frac{\delta}{24}x^4\bar{\Gamma}_5 \tag{6}$$

where $\bar{\Gamma}_5 \equiv \text{Max}_\theta\, \Gamma_5$ and $|\delta| < 1$.

1.2 Derivatives and Expectations

To proceed further, it is necessary to evaluate the Γ_{ki}. It is found that

$$\Gamma_{ki} = (-A_i)^{k-1}\frac{P_i'}{Q_i}\sum_{s=1}^{k-1} s!\mathfrak{S}_{k-1}^s\left(\frac{-Q_i}{1-c_i}\right)^s\left(-1 + \frac{u_i c_i^s}{P_i^{s+1}}\right) \tag{7}$$

$$(k = 1, 2, \ldots)$$

where $\mathfrak{S}_{k-1}^s$ is a Stirling number of the second kind [Jordan, 1947, pp. 31–32, 168].

Define

$$\gamma_{si} \equiv \mathscr{E}\Gamma_{si}, \tag{8}$$

$$\varepsilon_{si} \equiv \Gamma_{si} - \mathscr{E}\Gamma_{si}. \tag{9}$$

Since $\mathscr{E}u_i = P_i$, we find that

$$\gamma_{1i} = 0, \tag{10}$$

$$\gamma_{2i} = -\frac{P_i'^2}{P_iQ_i}, \tag{11}$$

$$\gamma_{3i} = \frac{A_i^2}{(1-c_i)^2}\frac{P_i'}{P_i^2}(P_i - c_i)[2(P_i^2 - c_i) - P_i(1-c_i)], \tag{12}$$

$$\varepsilon_{1i} = \Gamma_{1i} = \frac{(u_i - P_i)P_i'}{P_iQ_i}, \tag{13}$$

$$\varepsilon_{2i} = \frac{A_ic_i}{(1-c_i)}\frac{P_i'(u_i - P_i)}{P_i^2}. \tag{14}$$

Let

$$\gamma_s = \sum_i \frac{\gamma_{si}}{n}, \qquad \varepsilon_s \equiv \sum_i \frac{\varepsilon_{si}}{n}. \tag{15}$$

We will denote the Fisher information by

$$I \equiv -\mathscr{E}\left(\frac{dL_1}{d\theta}\right) = -n\gamma_2 = \sum_i \frac{P_i'^2}{P_iQ_i}. \tag{16}$$

Setting (6) equal to zero, the likelihood equation can now be written in terms of the γ_s and the ε_s as

$$-\varepsilon_1 = x(\gamma_2 + \varepsilon_2) + \frac{1}{2}x^2(\gamma_3 + \varepsilon_3) + \frac{1}{6}x^3(\gamma_4 + \varepsilon_4) + \frac{\delta}{24}x^4\bar{\Gamma}_5. \tag{17}$$

We will need some information about the order of magnitude of the terms such as those in (17). It may be seen from (7) that each ε_s has the form

$$\varepsilon_s = \frac{1}{n}\sum_i K_{si}(u_i - P_i)$$

where K_{si} does not depend on n or on u_i. Since P_i, Q_i and $1 - c_i$ are bounded, the K_{si} and thus ε_s is bounded. By assumption (4), the bound of $\mathscr{E}\varepsilon_s$ does not depend on n. The same conclusion holds for γ_s.

Since $(n)^{1/2}x$ is asymptotically normally distributed with zero mean and finite variance, it follows that $\mathscr{E}x^r$ $(r = 1, 2, \ldots)$ is of order $n^{-r/2}$. A similar statement is true of $(n)^{1/2}\varepsilon_s$. Thus finally $\mathscr{E}x^r\varepsilon_s^t \leq (\mathscr{E}x^{2r}\mathscr{E}\varepsilon_s^{2t})^{1/2}$ so that $\mathscr{E}x^r\varepsilon_s^t$ is of order $n^{-(r+t)/2}$ $(r, t = 1, 2, \ldots)$.

1.3 First-Order Standard Error of $\hat{\theta}$

To clarify the procedure, let us derive from (17) the familiar formula for the asymptotic standard error of $\hat{\theta}$. Square (17) and take expectations to obtain

$$\mathscr{E}\varepsilon_1^2 = \gamma_2^2\mathscr{E}x^2 + 2\gamma_2\mathscr{E}x^2\varepsilon_2 + \mathscr{E}x^2\varepsilon_2^2 + \gamma_2\gamma_3\mathscr{E}x^3 + \gamma_2\mathscr{E}\varepsilon_3 x^3 + \ldots . \tag{18}$$

If we wish to neglect terms $o(n^{-1})$ (of higher order than n^{-1}), (18) becomes

$$\mathscr{E}x^2 = \frac{1}{\gamma_2^2}\mathscr{E}\varepsilon_1^2 + o(n^{-1}). \tag{19}$$

By (13) and (16), because of local independence,

$$\begin{aligned}
\mathscr{E}\varepsilon_1^2 &= \mathscr{E}\frac{1}{n^2}\sum_i \frac{P_i'}{P_iQ_i}(u_i - P_i)\sum_j \frac{P_j'}{P_jQ_j}(u_j - P_j) \\
&= \frac{1}{n^2}\sum_i\sum_j \frac{P_i'P_j'}{P_iQ_iP_jQ_j}\mathscr{E}(u_i - P_i)(u_j - P_j) \\
&= \frac{1}{n^2}\sum_i \frac{P_i'^2}{P_i^2Q_i^2}\operatorname{Var} u_i \\
&= \frac{1}{n^2}\sum_i \frac{P_i'^2}{P_iQ_i} \\
&= \frac{I}{n^2}.
\end{aligned} \tag{20}$$

Thus, finally

$$\text{S.E.}^2(\hat{\theta}) = \frac{1}{I} + o(n^{-1}), \tag{21}$$

a well-known result. It is derived here to clarify the reasoning to be used subsequently. If $\hat{\theta}$ is substituted for θ on the right side of (21), the formula will still be correct to the specified order of approximation.

1.4 Statistical Bias of $\hat{\theta}$

Take the expectation of (17) to obtain

$$-\mathscr{E}_1\varepsilon_1 = \gamma_2\mathscr{E}_1x + \mathscr{E}_1x\varepsilon_2 + \frac{1}{2}\gamma_3\mathscr{E}_1x^2, \tag{22}$$

where $\mathscr{E}_1$ indicates an expectation in which only terms of order n^{-1} are to be retained. Also multiply (17) by ε_2 and take expectations to obtain

$$-\mathscr{E}_1\varepsilon_1\varepsilon_2 = \gamma_2\mathscr{E}_1x\varepsilon_2. \tag{23}$$

By (9)

$$\mathscr{E}\varepsilon_r = 0 \qquad r = 1, 2, \ldots . \tag{24}$$

From (13) and (14)

$$\mathscr{E}\varepsilon_1\varepsilon_2 = \frac{1}{n^2}\sum_i \frac{A_i c_i}{1-c_i}\frac{P_i'^2}{P_i^3 Q_i}\mathscr{E}(u_i - P_i)^2$$

$$= \frac{1}{n^2}\sum_i \frac{A_i c_i}{(1-c_i)}\frac{P_i'^2}{P_i^2}. \tag{25}$$

Substituting (16) and (25) into (23), we have the sampling covariance

$$\mathscr{E}_1 x\varepsilon_2 = \frac{1}{nI}\sum_i \frac{A_i c_i}{(1-c_i)}\frac{P_i'^2}{P_i^2}. \tag{26}$$

Finally, substituting (16), (21), (24), and (26) into (22) and solving for $\mathscr{E}_1 x_1$, we have the bias

$$B_1(\hat{\theta}) \equiv \mathscr{E}_1(\hat{\theta} - \theta) = \frac{1}{I^2}\left(\sum_i \frac{A_i c_i}{1-c_i}\frac{P_i'^2}{P_i^2} + \frac{1}{2}n\gamma_3\right). \tag{27}$$

This may be rewritten as

$$B_1(\hat{\theta}) = \frac{1}{I^2}\sum_{i=1}^{n} A_i I_i\left(\phi_i - \frac{1}{2}\right) \tag{28}$$

where

$$\phi_i \equiv \frac{P_i - c_i}{1-c_i} \text{ and } I_i \equiv \frac{P_i'^2}{P_i Q_i}. \tag{29}$$

Since I is of order n, $B_1(\hat{\theta})$ is of order n^{-1}.

It may be of interest to note that in the special case where all items are equivalent (all P_i are the same), the bias simplifies to

$$B_1(\hat{\theta}) = \frac{APQ}{2nP'^2}\frac{P - Q - c}{1-c}.$$

In this special case, the bias is zero when $P = (1 + c)/2$.

1.5 Numerical Results

A hypothetical test was designed to approximate the College Entrance Examination Board's Scholastic Aptitude Test, Verbal Section. This test is composed of $n = 90$ five-choice items. Some information about the distributions of the parameters of the 90 hypothetical items is given in Table 1.

The standard error and bias of $\hat{\theta}$ were computed from (21) and from (27) respectively for various values of θ. The results are shown in Table 2. It appears that the bias in $\hat{\theta}$ is negligible for moderate values of θ, but is sizable for extreme values. Note that the bias is positively correlated with θ. Because of guessing, zero bias does not occurr at $\theta = 0$ but at $\theta = .34$ approximately.

1.6 Variance and Bias of Estimated True Score

Since the ability scale is not unique, any monotonic transformation of θ can serve as a measure of ability. Two transformations are particularly useful: e^{θ} and

$$\zeta \equiv \frac{1}{n}\sum_{i=1}^{n} P_i(\theta), \tag{30}$$

the proportion-correct true score (the number-right true score divided by the number of items). One important reason for using the latter transformation is the following.

TABLE 1

Range and Quartiles of the Item Parameters in 90-Item Hypothetical Test

	$a_i \equiv A_i/1.7$	b_i	c_i
Highest value	1.88	2.32	.47
Q_1	1.07	1.15	.20
Median	.83	.38	.15
Q_3	.69	-.41	.13
Lowest value	.41	-3.94	.01

Ordinarily, as in Table 2, we find large standard errors of $\hat{\theta}$ where θ is extreme. Usually these large standard errors are no more harmful to the user than are the smaller standard errors found when θ is near the level aimed at by the test. There is a reason why this is so: If it were not, the user should have designed his test so as to reduce those standard errors that were troublesome to him.

We see that from this point of view the size of a difference on the θ scale does not correspond to its importance. The discrepancy is greatly reduced, however, if we measure ability on the ζ scale instead of on the θ scale. This is one reason, among several, why we are interested in the variance and bias of

$$\hat{\zeta} \equiv \frac{\sum_i P_i(\hat{\theta})}{n}. \tag{31}$$

In most cases, as will be assumed in subsequent paragraphs, the sum in (31) will probably be taken over the n items used to estimate $\hat{\theta}$. The sum can be taken over any meaningful item set, however. If six different tests are being compared, for example, the sum could be over all items in all six tests. Another possibility would be to sum over a hypothetical "standard" test composed of n equivalent items having $a_i = 1$, $b_i = 0$, and $c_i = 0$, for example.

Although the proportion-correct score

$$z \equiv \frac{\sum_i u_i}{n} \tag{32}$$

is an unbiased estimator of ζ, z is never a fully efficient estimator of ζ unless $c_i = 0$ and $a_i = a_j$ $(i, j = 1, 2, \ldots, n)$: the squared standard error

$$\text{S.E.}^2(z) = \frac{1}{n^2} \sum_{i=1}^{n} P_i Q_i \tag{33}$$

is not so small as the squared standard error of $\hat{\zeta}$, which we must now derive.

TABLE 2

Standard Error and Statistical Bias in $\hat{\theta}$

$\hat{\theta}$	S.E.$(\hat{\theta})$	B$(\hat{\theta})$
3.5	.60	.24
3.0	.43	.12
2.5	.31	.06
2.0	.23	.032
1.5	.19	.011
1.0	.19	.0032
0.5	.20	.0012
0	.22	-.0028
-0.5	.25	-.010
-1.0	.31	-.025
-1.5	.41	-.05
-2.0	.54	-.09
-2.5	.70	-.14
-3.0	.89	-.22
-3.5	1.09	-.31

By (31)

$$d\hat{\zeta} \equiv \frac{1}{n}\sum_{i=1}^{n} \hat{P}_i' d\hat{\theta}. \tag{34}$$

Using the 'delta' method

$$\text{S.E.}^2(\hat{\zeta}) = \frac{1}{n^2}\left(\sum_{i=1}^{n} P_i'\right)^2 \text{S.E.}^2(\hat{\theta}).$$

By (21) and (16)

$$\text{S.E.}^2(\hat{\zeta}) = \frac{\left(\sum_i P'_i\right)^2}{n^2 \sum_i \frac{P'^2_i}{P_i Q_i}}. \tag{35}$$

To find the bias of $\hat{\zeta}$, we expand it in powers of $x \equiv \theta - \hat{\theta}$:

$$\hat{\zeta} - \zeta \equiv \frac{x}{n} \sum P'_i + \frac{x^2}{2n} \sum P''_i + \ldots \tag{36}$$

where

$$P''_i \equiv \frac{d^2 P_i}{d\theta^2}.$$

Taking expectations, and neglecting higher-order terms, we have for the bias

$$B_1(\hat{\zeta}) \equiv \mathscr{E}(\hat{\zeta} - \zeta) = \frac{1}{n}\left[B(\hat{\theta}) \sum P'_i + \frac{1}{2}\left(\sum P''_i\right)\text{S.E.}^2(\hat{\theta})\right]. \tag{37}$$

This can be rewritten as

$$B_1(\hat{\zeta}) = \frac{\zeta'}{I^2}\left(\sum \frac{A_i c_i P'^2_i}{(1 - c_i)P^2_i} + \frac{1}{2}\sum_i \gamma_{3i}\right) + \frac{\zeta''}{2I} \tag{38}$$

where $\zeta' \equiv \sum_i P'_i/n$ and $\zeta'' \equiv \sum_i P''_i/n$. Note in passing that when all items are equivalent (all $P_i(\theta)$ are the same), $\hat{\zeta} = z$ and its bias (38) is zero.

1.7 Numerical Results

Table 3 shows the bias in $\hat{\zeta}$ for the same hypothetical test considered in Section 1.5. The biases are all positive. However, they are negligible at all except the lowest ability levels. This tends to confirm our choice of the ζ scale of ability rather than the θ scale for many purposes.

As a matter of incidental interest, for selected values of true score Table 3 compares the standard error (35) of the maximum-likelihood estimator $\hat{\zeta}$ with the standard error (33) of the unbiased estimator z (proportion-correct score). There is little difference in accuracy between the two estimators for $\zeta \geq .5$. At low true-score levels, the maximum-likelihood estimator is much better than the proportion of correct answers.

2. Unbiased Estimation of s^2_θ, of s^2_ζ; Test Reliability

The symbols s^2_θ and s^2_ζ are used for the sample variance of θ and of ζ across the N examinees in the sample:

$$s^2_\theta = \frac{1}{N}\sum_{a=1}^{N} \theta^2_a - \left(\frac{1}{N}\sum_{a=1}^{N} \theta_a\right)^2. \tag{39}$$

The maximum-likelihood estimators of s^2_θ and s^2_ζ are $s^2_{\hat{\theta}}$ and $s^2_{\hat{\zeta}}$, the sample variances across examinees of $\hat{\theta}$ and of $\hat{\zeta}$.

2.1 Asymptotically Unbiased Estimator of σ^2_θ

Assume that our examinees are a random sample of N from some population, with N much greater than n. Denote by σ^2_θ the population variance of θ. Then $Ns^2_\theta/(N-1)$ is an

TABLE 3

Standard Error of z and of $\hat{\zeta}$, and Statistical Bias of $\hat{\zeta}$

$\hat{\theta}$	$\hat{\zeta}$	S.E.(z)	S.E.($\hat{\zeta}$)	B($\hat{\zeta}$)
3.5	.981	.014	.014	.00045
3.0	.966	.019	.018	.00052
2.5	.937	.024	.023	.00064
2.0	.891	.031	.029	.00059
1.5	.812	.037	.035	.00021
1.0	.715	.042	.040	.00026
0.5	.608	.045	.042	.00061
0	.506	.046	.043	.00061
-0.5	.416	.047	.042	.00062
-1.0	.344	.046	.038	.00061
-1.5	.291	.045	.037	.00085
-2.0	.254	.044	.033	.0014
-2.5	.227	.042	.029	.0020
-3.0	.211	.042	.025	.0024
-3.5	.199	.041	.021	.0026

unbiased estimator of σ_θ^2. Since s_θ^2 is unobservable, our first task is to find a function of $\hat{\theta}$ that is asymptotically unbiased estimator of σ_θ^2.

By the formula for the variance of a sum we have

$$\sigma_\theta^2 \equiv \sigma_{\theta+x}^2 \equiv \sigma_\theta^2 + \sigma_x^2 + 2\sigma_{\theta x}, \tag{40}$$

where σ^2 denotes a variance across all examinees in the population and $\sigma_{\theta x}$ is the corresponding population covariance. By a well-known identity from the analysis of variance

$$\sigma_x^2 \equiv \mathscr{E}_\theta \sigma_{x|\theta}^2 + \sigma_{\mathscr{E}(x|\theta)}^2 \tag{41}$$

where $\mathscr{E}_\theta$ denotes an expectation across all examinees in the population. Similarly,

$$\sigma_{\theta x} \equiv \sigma_{\theta, \mathscr{E}(x|\theta)}. \tag{42}$$

Substituting (41) and (42) into (40), transposing, writing $B_1 \equiv \mathscr{E}_1(x \mid \theta)$ as in (27), and dropping the subscript from B_1 for convenience, we have

$$\sigma_\theta^2 \equiv \sigma_{\hat{\theta}}^2 - 2\sigma_{\theta B} - \mathscr{E}_\theta \sigma_{x|\theta}^2 - \sigma_B^2. \tag{43}$$

Since by (28) B is of order n^{-1}, its variance is of order n^{-2}, so σ_B^2 can be neglected in (43). Since Section 1 deals with a single fixed examinee, the symbol $\sigma_{x|\theta}^2$ in (43) has the same meaning as S.E.$^2(\hat{\theta})$ in (21):

$$\sigma_{x|\theta}^2 = \frac{1}{I(\theta)} + o(n^{-1})$$

where $I \equiv I(\theta)$ is given by (16). Since $\sigma_{x|\theta}^2$ is of order n^{-1}, the effect of replacing θ by $\hat{\theta}$ on the right is negligible:

$$\mathscr{E}_\theta \sigma_{x|\theta}^2 = \mathscr{E}_\theta \frac{1}{I(\hat{\theta})} + o(n^{-1}).$$

By similar reasoning, we may replace $\sigma_{\theta B}$ in (43) by $\sigma_{\hat{\theta}\hat{B}}$ where $\hat{B}$ is defined by (27) with θ replaced by $\hat{\theta}$. The result of these approximations is that

$$\sigma_\theta^2 = \sigma_{\hat{\theta}}^2 - 2\sigma_{\hat{\theta}\hat{B}} - \mathscr{E}_\theta \frac{1}{I(\hat{\theta})} + o(n^{-1}). \tag{44}$$

A useful estimator of σ_θ^2 can be calculated from

$$\hat{\sigma}_\theta^2 \equiv \frac{N}{N-1} s_{\hat{\theta}}^2 - \frac{2N}{N-1} s_{\hat{\theta}\hat{B}} - \frac{1}{N} \sum_{a=1}^{N} \frac{1}{I(\hat{\theta}_a)}, \tag{45}$$

where

$$s_{\hat{\theta}\hat{B}} \equiv \frac{1}{N} \sum_{a=1}^{N} \hat{\theta}_a \hat{B}_a - \left(\frac{1}{N} \sum_{a=1}^{N} \hat{\theta}_a\right)\left(\frac{1}{N} \sum_{a=1}^{N} \hat{B}_a\right)$$

and $\hat{B}_a$ is given by (27) with θ replaced by $\hat{\theta}_a$. Equation (45) is unbiased over repeated random samples of examinees through terms of order n^{-1}. If we wish to estimate the sample variance of ability s_θ^2 rather than the population variance σ_θ^2, we can use

$$\hat{s}_\theta^2 \equiv s_{\hat{\theta}}^2 - 2s_{\hat{\theta}\hat{B}} - \frac{N-1}{N^2} \sum_{a=1}^{N} \frac{1}{I(\hat{\theta}_a)}. \tag{46}$$

The second and third terms of (44) are of order n^{-1}, an order of magnitude smaller than the first term but larger than the neglected terms. The covariance of $\hat{\theta}$ and $\hat{B}$ is usually positive, as can be readily seen from Table 2. Since $I(\hat{\theta})$ is necessarily positive, it appears that usually $\sigma_\theta^2 < \sigma_{\hat{\theta}}^2$, an inequality that is frequently assumed without proof. It is not clear whether this inequality is necessarily true.

2.2 The Reliability of $\hat{\theta}$

Consider the parallel-forms reliability coefficient $\rho_{\hat{\theta}\hat{\theta}'}$, the correlation between scores $\hat{\theta}$ and $\hat{\theta}'$ on two parallel tests. For present purposes, two tests are parallel when for each item in one test there is an item in the other test with the same item response function. Let us estimate

$$\rho_{\hat{\theta}\hat{\theta}'} = \frac{\sigma_{\hat{\theta}\hat{\theta}'}}{\sigma_{\hat{\theta}}\sigma_{\hat{\theta}}'} = \frac{\sigma_{\hat{\theta}\hat{\theta}'}}{\sigma_{\hat{\theta}}^2} \tag{47}$$

from a single test administration by substituting asymptotically unbiased estimators of the numerator and of the denominator into (47).

As in (41),

$$\sigma_{\hat{\theta}\hat{\theta}'} \equiv \mathscr{E}_\theta \sigma_{\hat{\theta}\hat{\theta}'|\theta} + \sigma_{\mathscr{E}(\hat{\theta}|\theta),\, \mathscr{E}(\hat{\theta}'|\theta)}. \tag{48}$$

Because of local independence, the first term on the right vanishes. Because of parallelism, the two expectations in the last term are identical, so this term is a variance. We thus have

$$\sigma_{\hat{\theta}\hat{\theta}'} = \sigma^2_{\mathscr{E}(\hat{\theta}|\theta)} \equiv \sigma^2_{B+\theta} \equiv \sigma^2_B + \sigma^2_\theta + 2\sigma_{\theta B}. \tag{49}$$

From (49) and (43),

$$\sigma_{\hat{\theta}\hat{\theta}'} = \sigma^2_{\hat{\theta}} - \mathscr{E}_\theta \sigma^2_{x|\theta}. \tag{50}$$

We see that the parallel-forms reliability of $\hat{\theta}$ is

$$\rho_{\hat{\theta}\hat{\theta}'} = 1 - \frac{1}{\sigma^2_{\hat{\theta}}} \mathscr{E}_\theta \frac{1}{I(\hat{\theta})} + o(n^{-1}). \tag{51}$$

Priority in obtaining this result belongs to Sympson [Note 1]. Replacing population values on the right by the corresponding sample statistics, we have a sample estimator of the parallel-forms reliability coefficient of $\hat{\theta}$:

$$\hat{\rho}_{\hat{\theta}\hat{\theta}'} \equiv 1 - \frac{N-1}{N^2 s^2_{\hat{\theta}}} \sum_{a=1}^{N} \frac{1}{I(\hat{\theta}_a)} + o(n^{-1}). \tag{52}$$

Since $\hat{\theta}$ is neither unbiased nor uncorrelated with θ, we should not expect the usual reliability formulas of classical test theory to apply. A similar but not identical case is discussed in Lord and Novick [1968, Section 9.8]. Thus $\rho_{\hat{\theta}\hat{\theta}'}$, $\rho^2_{\hat{\theta}\theta}$, and $\sigma^2_\theta/\sigma^2_{\hat{\theta}}$ are not interchangeable definitions of reliability. This difficulty is overcome if we seek the reliability of the unbiased estimator $\tilde{\theta} \equiv \hat{\theta} - B$, since by classical test theory $\rho_{\tilde{\theta}\tilde{\theta}} = \rho^2_{\tilde{\theta}\theta} = \sigma^2_\theta/\sigma^2_{\tilde{\theta}}$.

We now have

$$\rho_{\tilde{\theta}\tilde{\theta}'} = \frac{\sigma^2_\theta}{\sigma^2_{\tilde{\theta}}} = \frac{\sigma^2_\theta}{\sigma^2_{\hat{\theta}} + \sigma^2_B - 2\sigma_{\hat{\theta}B}}.$$

Dropping σ^2_B, substituting $\sigma_{\hat{\theta}B}$ for $\sigma_{\hat{\theta}B}$, and using (44), we find

$$\rho_{\tilde{\theta}\tilde{\theta}'} = 1 - \frac{\mathscr{E}_\theta \dfrac{1}{I(\hat{\theta})}}{\sigma^2_{\hat{\theta}} - 2\sigma_{\hat{\theta}B}} + o(n^{-1}).$$

It can be seen from (51) that $\tilde{\theta}$ and $\hat{\theta}$ only differ in reliability by $o(n^{-1})$: by terms that are negligible compared to n^{-1}. Since correlational measures are hard to interpret in the absence of homoscedasticity, and we will not now push this investigation of reliability further.

2.3 Corresponding Results for True Score

By the same reasoning used to obtain (44) we have

$$\begin{aligned}\sigma^2_\zeta &\equiv \sigma^2_\xi - 2\sigma_{\zeta,\, B(\xi)} - \mathscr{E}_\theta \sigma^2_{\xi|\zeta} - \sigma^2_{B(\xi)} \\ &= \sigma^2_\xi - 2\sigma_{\xi,\, B(\xi)} - \mathscr{E}_\theta \frac{\xi'^2}{I(\hat{\theta})} + o(n^{-1}).\end{aligned} \tag{54}$$

A useful estimator of σ^2_ζ can be calculated from

$$\hat{\sigma}^2_\zeta \equiv \frac{N}{N-1} s^2_\xi - \frac{2N}{N-1} s_{\xi,\, B(\xi)} - \frac{1}{N} \sum_{a=1}^{N} \frac{\xi'^2_a}{I(\hat{\theta}_a)}. \tag{54}$$

To estimate s_ζ^2, we can use

$$\hat{s}_\zeta^2 \equiv s_\xi^2 - 2s_{\xi,\,B(\xi)} - \frac{N-1}{N^2}\sum_{a=1}^{N}\frac{\zeta_a'^2}{I(\hat{\theta}_a)}. \tag{55}$$

As in (50)–(52) we have

$$\sigma_{\xi\xi'} = \sigma_\xi^2 - \mathscr{E}_\theta\,\sigma_{\xi|\zeta}^2, \tag{56}$$

$$\rho_{\xi\xi'} = 1 - \frac{1}{\sigma_\xi^2}\,\mathscr{E}_\theta\,\frac{\zeta'^2}{I(\hat{\theta})} + o(n^{-1}), \tag{57}$$

$$\hat{\rho}_{\xi\xi'} \equiv 1 - \frac{N-1}{N^2 s_\xi^2}\sum_{a=1}^{N}\frac{\zeta'^2}{I(\hat{\theta}_a)} + o(n^{-1}). \tag{58}$$

2.4 Numerical Results for True Scores

At moderate ability levels, (28) provides adequate but usually negligible corrections for bias in $\hat{\theta}$. Experience shows that at very low ability levels, the usual test length (n) of 50 or 100 items is not long enough for the asymptotic results of (28) to apply. For example, an examinee whose true θ is -3 may easily obtain an estimated ability $\hat{\theta}$ of -30 or of $-\infty$. For sufficiently long tests, such extreme values of $\hat{\theta}$ would have negligible probability, but with the usual values of n, (28) is totally inadequate for correcting $\hat{\theta}$ for bias at low ability levels.

This same difficulty carries over to the unbiased estimation of σ_θ^2 using (46). Since all ability levels are involved in (46), the formula is useless in practice for any group that contains even a few low-ability examinees. Fortunately, this difficulty does not carry over to the estimation of ability on the true-score (ζ) scale.

A simulation was carried out to administer the hypothetical SAT Verbal Test of Tables 1–3 to a typical group of 2995 hypothetical examinees. The bias in $\hat{\zeta}$ was estimated for each examinee and a corrected $\hat{\zeta}$ obtained from (51):

$$\tilde{\zeta} \equiv \text{corrected } \hat{\zeta} \equiv \hat{\zeta} - B_1(\hat{\zeta}).$$

In a few cases where the $\tilde{\zeta}$ would have been below the chance level $\sum_i^n c_i$, $\tilde{\zeta}$ was set equal to $\sum_i^n c_i$.

The mean of the 2995 true ζ used to generate the data was .5280, the mean of the uncorrected $\hat{\zeta}$ was .5294, the mean of the $\tilde{\zeta}$ was .5288. Thus the correction was in the right direction, but not large enough. The uncorrected mean $\hat{\zeta}$ was already so accurate as to leave little room for improvement.

Next, (55) was used to estimate s_ζ. The true value was $s_\zeta = .1610$, the standard deviation of $\hat{\zeta}$ was $s_{\hat{\zeta}} = .1660$, the corrected estimate from (55) was $\hat{s}_\zeta = .1614$. The correction worked very well here.

The parallel-forms reliability of $\hat{\zeta}$ was estimated from (58) to be $\hat{\rho}_{\hat{\zeta}\hat{\zeta}'} = .9420$. We have no 'true' value against which this can be compared, but the estimate seems a reasonable one. The Kuder-Richardson formula-20 reliability of number-right scores for these data is .9275.

It should be remembered that both the formulas and the numerical results in this report apply in situations where the item parameters are known. These formulas may be satisfactory for situations where the item parameters have been estimated from large groups not containing the examinees whose ability estimates are to be corrected for bias. These formulas will not be adequate for situations where the item parameters and ability parameters are estimated simultaneously from a single data set.

REFERENCE NOTE

1. Sympson, J. B. *Estimating the reliability of adaptive tests from a single test administration.* Paper presented at the meeting of the American Educational Research Association, Boston, April 1980.

REFERENCES

Bradley, R. A. & Gart, J. J. The asymptotic properties of ML estimators when sampling from associated populations. *Biometrika,* 1962, *49,* 205–214.

Jordan, C. *Calculus of finite differences* (2nd ed.). New York: Chelsea, 1947.

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

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

Shenton, L. R. & Bowman, K. O. *Maximum likelihood estimation in small samples.* (Griffin's Statistical Monograph No. 38.) New York: Macmillan, 1977.

Manuscript received 6/4/82
Final version received 10/12/82