PSYCHOMETRIKA—VOL. 54, NO. 3, 427–450
SEPTEMBER 1989

WEIGHTED LIKELIHOOD ESTIMATION OF ABILITY IN ITEM RESPONSE THEORY

THOMAS A. WARM

FAA ACADEMY

Applications of item response theory, which depend upon its parameter invariance property, require that parameter estimates be unbiased. A new method, weighted likelihood estimation (WLE), is derived, and proved to be less biased than maximum likelihood estimation (MLE) with the same asymptotic variance and normal distribution. WLE removes the first order bias term from MLE. Two Monte Carlo studies compare WLE with MLE and Bayesian modal estimation (BME) of ability in conventional tests and tailored tests, assuming the item parameters are known constants. The Monte Carlo studies favor WLE over MLE and BME on several criteria over a wide range of the ability scale.

Key words: maximum likelihood estimation, unbiased estimation, statistical bias, Bayesian modal estimation, item response theory, tailored testing, adaptive testing.

Introduction

Item response theory (IRT) is an elegant model of examinee behavior on multiple-choice tests in terms of item and person parameters. The parameters are invariant from test to test within a linear transformation. This invariance property makes possible the placement of scores and item parameters from different tests onto a common scale. With the parameters on a common scale, scores on different tests may be equated, and items from different tests may be accumulated into item banks with comparable item parameters. Since the true values of the parameters are never known and must be estimated, applications of IRT which use the invariance property, depend upon an assumption of parameter estimate invariance. Parameter estimates are not invariant, because of estimation error. Hence, in practice the invariance property must be phrased in terms of expectations of parameter estimates.

For the expected values of parameter estimates to be invariant the estimates must be unbiased. Unless unbiased estimates can be obtained, test score equatings will be erroneous and item banks will contain items with noncomparable statistics. Goldstein (1980) has already made this point concerning the Rasch model. The same is true of other IRT models, especially the three parameter model.

Linking the scales of tests can be divided into two broad categories: horizontal linking and vertical linking (Baker, 1984). Horizontal linking refers to linking tests of approximately equal difficulty, while vertical linking means that the tests are of deliberately very different difficulty. In both cases it is presumed that the tests to-be-linked are administered to examinees for whom the tests are of appropriate difficulty. The mechanics of linking in both situations are identical. All that is necessary for linking is to have the scale values for two different abilities on each of the two scales. When estimates are used for this transformation in place of the true values, the statistical bias of those estimates can have serious consequences. Lord (1983a) found that maximum likelihood estimates of θ are biased outward, and that the magnitude of bias is greater

Requests for reprints should be sent to Thomas A. Warm, FAA Academy AAC-934, Mike Monroney Aeronautical Center, PO Box 25082, Oklahoma City, OK 73125.

0033-3123/89/0900-9046$00.75/0

at negative values of θ than at positive values. This asymmetry of bias causes the slope of transformation to be underestimated. This is especially true in vertical linking, where the effects become exaggerated, because positive values from the easier test are equated to the more biased negative values from the harder test. When several tests of increasing difficulty are vertically linked, the effect is compounded at each step.

When scales are linked/equated with parameter estimates, the transformation for linking scales is also an estimate. The greater the error of the parameter estimates, the greater will be the error of the linear transformation. In order to minimize the error of transformation, averages of parameter estimates are used in place of the parameters themselves when solving for the linking transformation. Averages have reduced variance, a valuable property which should reduce the variability of the linking transformation. However, if the estimates are statistically biased, then the averages will also be biased, and so will the linking transformation. Usually linking is accomplished with the item location parameter (the *b*-parameter) and perhaps slope parameter (the *a*-parameter). But in practice item parameters are estimated using estimates of examinees' θs. Thus, no matter how unbiased one's item parameter estimation method may be, item parameter estimates will still be biased, if the θ estimates are biased. Therefore, unbiased estimation of θ must be the first order of business.

This paper mathematically derives and tests by Monte Carlo a new estimation method, called Weighted Likelihood Estimation (WLE). The new method is applied to estimation of θ, the ability parameter, assuming the true item parameters are known. The new estimator has small bias and is computationally efficient. The mathematical form of IRT used in this paper is the three parameter logistic model. Conclusions, reached for this model, apply immediately to the one and two parameter models as well.

The model gives the probability, P, that a scored item response, u_i, to item i is corrected ($u_i = 1$), is a function of the ability parameter, θ, and three item parameters, a_i, b_i, and c_i.

$$P(u_i = 1|\theta; a_i, b_i, c_i) = c_i + \frac{1 - c_i}{1 + \exp(-1.7a_i(\theta - b_i))} \tag{1}$$

The left hand side (LHS) of (1) is often abbreviated $P_i(\theta)$, $P(\theta)$, or just P, when the context excludes ambiguity.

Estimation Methods and Bias

There are five basic estimation methods that are used in IRT for parameter estimation: (a) maximum likelihood estimation (MLE; Lord, 1980), (b) Bayesian modal estimation (BME; Samejima, 1980; also called modal a-posteriori, MAP; Bock, 1983), (c) Owen's sequential Bayesian (OSB; Owen, 1975), (d) expected a-posteriori (EAP; Bock), and (e) marginal maximum likelihood (MML; Bock & Aitkin, 1981). In addition to these, there are variations such as the robustified jackknife (Wainer & Wright, 1980), *h*-estimators (Jones, 1982), and biweight estimates (Bock & Mislevy, 1981).

All of these estimation methods produce estimates that are biased to some degree. MLE (Lord, 1983a), and BME (Lord, 1983b, 1984) were shown to have bias of $O(n^{-1})$; that is to say the bias is inversely proportional to n, the number of items in the test, other things being equal. OSB, EAP, and MML are all Bayesian procedures (in spite of MMLs title), and, therefore, also have bias of $O(n^{-1})$. The biweight and *h*-estimators are robust *M*-estimators (modified MLE; Andrews, et al., 1972) designed to reduce the influence of misfitting outliers rather than to reduce bias.

The bias of jackknifed estimators, which were introduced by Quenouille (1956), is of one order less than the order of bias of the estimator jackknifed (Kendall & Stuart,

1973). Therefore, the bias of the Robustified Jackknife, which jackknifes an M-estimator, should be $o(n^{-1})$ except for any bias caused by the robustification. Unfortunately, jackknifing is computationally inefficient, because the reduced bias is achieved by increasing the required computations by an order of magnitude. That is to say, for a test of n items, the computational time for the jackknifed MLE is n times the computational time for the MLE itself. Even on large computers this computational intensity increases CPU time from minutes to hours; this increase is unacceptable in many settings.

Other methods of bias reduction have been used with some success in other contexts. For MLE, Cox and Hinkley (1974) suggest evaluating the bias at the value of the estimate, and then subtracting the estimated bias from the estimate to produce an improved estimate. Anderson and Richardson (1979) and Schaefer (1983) used this technique successfully on discrimination and location parameters, respectively, of linear logistic models (not IRT). One difficulty with this approach is that for models in which bias is a cubic-shaped function of the parameter being estimated (as is the case for MLE(θ) in IRT), error can actually be increased rather than reduced.

Lord (personal communication, January 1983) has suggested using as an estimate of θ that value of the parameter, which when added to the bias evaluated at the value, is equal to the maximum likelihood estimate. There are two difficulties with this estimator: (a) it is not necessarily unique, and (b) if the maximum likelihood estimate is infinite, so is this estimator.

Weighted Likelihood Estimation

For a test of n items the MLE of θ, MLE(θ), is the value of θ that maximizes the likelihood function, $L(\mathbf{u}|\theta)$, where

$$L(\mathbf{u}|\theta) = \prod_{i=1}^{n} P_i(\theta)^{u_i} \cdot Q_i(\theta)^{1-u_i}, \tag{2}$$

Π is the product operator, $\mathbf{u}$ is the vector of n scored item responses ($u_i = 1$ if item i is correctly answered, and $u_i = 0$ if item i is incorrectly answered; $i = 1, 2, \ldots, n$), and $Q_i(\theta) = 1 - P_i(\theta)$. Hereafter, the subscript i will be dropped for convenience, unless it is needed for clarity. MLE(θ) is found at the zero of the likelihood equation,

$$l_1 \equiv \frac{\partial \ln L(\mathbf{u}|\theta)}{\partial \theta} = \frac{\sum_{i=1}^{n} (u - P)P'}{PQ} = 0, \tag{3}$$

where $P' = \partial P/\partial\theta$.

A class of estimators, θ^*, may be defined as the value of θ that maximizes (4), for a suitably chosen function $f(\cdot)$,

$$f(\theta) \cdot L(\mathbf{u}|\theta) = f(\theta) \prod_{i=1}^{n} P(\theta)^{u_i} \cdot Q(\theta)^{1-u_i}. \tag{4}$$

θ^* is found at the zero of the estimation equation,

$$\frac{\sum_{i=1}^{n}(u-P)P'}{PQ}+\frac{\partial \ln f(\theta)}{\partial\theta}=0. \tag{5}$$

If $f(\theta)$ is a positive constant, θ^* is a maximum likelihood estimate of θ, MLE(θ), and (5) reduces to (3). If $f(\theta)$ is an assumed prior density function of θ, then (4) is proportional to the posterior density function, and θ^* is a Bayesian modal estimate of θ, BME(θ).

Lord (1983a) gives the following asymptotic expression for the bias of MLE(θ), BIAS(MLE(θ)), which is $O(n^{-1})$:

$$\text{BIAS}(\text{MLE}(\theta))=\frac{-J}{2I^2}, \tag{6}$$

where I is test information, $I=\Sigma P'^2/PQ$, $J=\Sigma P'P''/PQ$, and $P''=\partial^2 P/\partial\theta^2$. Equation (6) is equivalent to the general expression of BIAS(MLE(θ)) for a multinomially distributed variable, given by Cox and Hinkley (1974, p. 310). Lord (1984) gives the bias of BME(θ) with a standard normal prior:

$$\text{BIAS}(\text{BME}(\theta))=\text{BIAS}(\text{MLE}(\theta))-\frac{\theta}{I} \tag{7}$$

BIAS(BME(θ)) is also $O(n^{-1})$. The last term on the right hand side (RHS) of (7) is the derivative, with respect to θ, of the log of the standard normal density divided by test information. From this observation, we can conjecture that the bias of the estimator defined by (5) is

$$\text{BIAS}(\theta^*)=\text{BIAS}(\text{MLE}(\theta))+\frac{\dfrac{\partial \ln f(\theta)}{\partial\theta}}{I}. \tag{8}$$

Thus, in order to find the estimator that is unbiased, we only need set the RHS of (8) equal to zero, and solve for $f(\theta)$. Setting the RHS of (8) = 0 and substituting (6), we obtain

$$\frac{\partial \ln f(\theta)}{\partial\theta}=\frac{J}{2I}. \tag{9}$$

Replacing $f(\cdot)$ with $w(\cdot)$ in (5) to emphasize that the function is now specifically defined, yields

$$\frac{\sum_{i=1}^{n}(u-P)P'}{PQ}+\frac{\partial \ln w(\theta)}{\partial\theta}=0,$$

and substituting (9) into (5), gives

$$\frac{\sum_{i=1}^{n} (u - P)P'}{PQ} + \frac{J}{2I} = 0, \tag{10}$$

where I and J are as defined in (6). An estimate satisfying (10) is called a weighted likelihood estimate (WLE). Note that WLE is not in any sense Bayesian, because no assumptions have been made about the distribution of θ, and $w(\theta)$ is a function of the parameters of the items in the test.

If the conjecture above is correct, then BIAS(WLE(θ)) should be only $o(n^{-1})$.

Theorem. WLE(θ) is unbiased to order n^{-1}, that is,

$$\text{BIAS}(\text{WLE}(\theta)) = 0 + o(n^{-1}).$$

The appendix gives the mathematical proof of the theorem for the rather restrictive conditions of IRT. Apparently, however, this ". . . method of removing the first bias term will work in complete generality, and can be extended to any form of consistent estimating equation where the mathematical form of bias is computable" (D. Hinkley, personal communication, July 22, 1985). The generality of the theorem is demonstrated by the fact that the weighted likelihood estimates of the mean (μ) and variance (σ^2) of a normal population are *both* unbiased, whereas it is well known that MLE(σ^2) is biased.

WLE(θ) is asymptotically normally distributed with variance equal (asymptotically) to the variance of MLE(θ), that is, $\text{VAR}(\text{WLE}(\theta)) \simeq \text{VAR}(\text{MLE}(\theta)) \simeq I^{-1}$.

In general there is no known, closed-form solution for the indefinite integral of $\partial \ln w(\theta)/\partial\theta$ in order to solve for $w(\theta)$. Fortunately, only $\partial \ln w(\theta)/\partial\theta$ is necessary in order to solve for WLE(θ). However, if the c-parameters are equal to zero for all items (as they are in the one and two parameter models of IRT), then $w(\theta) = I^{1/2}$, if $c_i = 0$, all i.

Method

WLE(θ) was compared to MLE(θ) and BME(θ) by the Monte Carlo method in the context of conventional tests and tailored tests, assuming the true item parameters were known.

Design of the conventional tests. For the comparison of the three estimators under a wide range of test lengths, 12 conventional tests were constructed. There were two tests for each of six test lengths, n = 10, 20, 30, 40, 50, and 60. For each test length, the a-parameters of one test were set to $a_i = 1.0$, for all $i = 1, 2, \ldots n$, and for the other test $a_i = 2.0$, all i. The b-parameters of all 12 tests were distributed "normally," using the inverse normal transformation, $\Phi^{-1}(\cdot)$. That is, for a test of length n, $b_i = \Phi^{-1}((i - .5)/n)$. The c-parameters for all items in all tests were set to 0.20. These item parameters produce a test information curve that is roughly "normally" shaped, and is a commonly used conventional test design.

At each of 17 values of θ (= $-4, -3.5, \ldots, 4$), 1000 simulated examinees were administered all 12 tests, and WLE(θ), MLE(θ), and BME(θ) were computed for each examinee and for each test. The same item responses were used for all three estimators. The mean, standard deviation, and mean squared error of the 1000 estimates of θ were computed at each of the 17 values of θ for each test and each estimator.

Design of tailored tests. Six tailored tests, two for each estimator, were administered to 100 simulated examinees at each of the 17 values of θ. For all tailored tests all $c_i = .20$. For each of the three estimators one tailored test had all $a_i = 2.0$. The other tailored test for each estimator had declining a-parameters to simulate the declining item information available from a finite item pool, specifically $a_i = (71 - i)/35$. Following Weiss and McBride (1984), the values of the b-parameters for all tailored tests were chosen so that the maximum of the item information for the item a- and c-parameter was at the current estimate of θ. That is, the b-parameter of the $(i + 1)$-th item was

$$b_{i+1} = \hat{\theta}_{(i)} - \frac{\ln (0.5 \cdot (1 + (1 + 8c)^{1/2}))}{1.7a},$$

where $\hat{\theta}_{(i)}$ is the current estimate of θ after the i-th item.

The stopping rules for administering items were: (a) stop if test information exceeds 20 at the current estimate of θ, or (b) stop if the number (n) of items administered is 50, whichever occurred first. The mean, standard deviation, and mean squared error of the 100 estimates were computed for each of the 17 values of θ for each test and each estimator. In addition, for each tailored test and estimator the average number of items administered, and the average iteration computation time per item were computed.

Estimating θ. For the conventional tests, θ was iteratively estimated by the interval bisection method with $r = 15$ iterations.

$$\hat{\theta}_{m,r} = \hat{\theta}_{m,r-1} + \Delta_{m,r}, \qquad r = 1, 2, \ldots 15,$$

$$m = \text{WLE, MLE, or BME, and } \hat{\theta}_{m,0} = \theta.$$

For the first four iterations, $r = 1, 2, 3, 4$, $|\Delta_{m,r}| = 1$ with the sign the same as the objective function. In the remaining iterations, $r = 5, 6, \ldots, 15$, $|\Delta_{m,r}| = |\Delta_{m,r-1}|/2$. This iteration method has several advantages: (a) $|\hat{\theta} - \theta| < 5$, (b) it will find the local maximum closest to true θ, (c) the magnitude of the difference between the true maximum and the final estimate < 0.001, and (d) convergence is guaranteed.

Iterative estimation of MLE(θ) for the tailored tests was accomplished by "Fisher scoring" (Lord, 1980) with $\hat{\theta}_{m,0} = 0$, and the magnitude of $\Delta_{m,r}$ limited to 2.0. The respective analogues of Fisher scoring were used for WLE and BME. For m = WLE,

$$\Delta_{m,r} = \frac{l_1 + \frac{J}{2I}}{\frac{I - IJ' - I'J}{2I^2}},$$

evaluated at $\hat{\theta}_{m,r-1}$. Iterations were continued until $|\Delta_{m,r}| < 0.001$, $r = 21$, or $|\hat{\theta}_{m,r+1}| > 5$, whichever occurred first.

Generating the scored item responses u_i. For each item response $P(\theta)$ was calculated, and a pseudo-random number, y, uniformly distributed over the interval (0, 1), was generated. Then $u_i = 0$, if $y > P(\theta)$; else $u_i = 1$.

Results

Conventional tests. The results for the 12 conventional tests are remarkably similar, and virtually identical with respect to the relative results among the three estima-

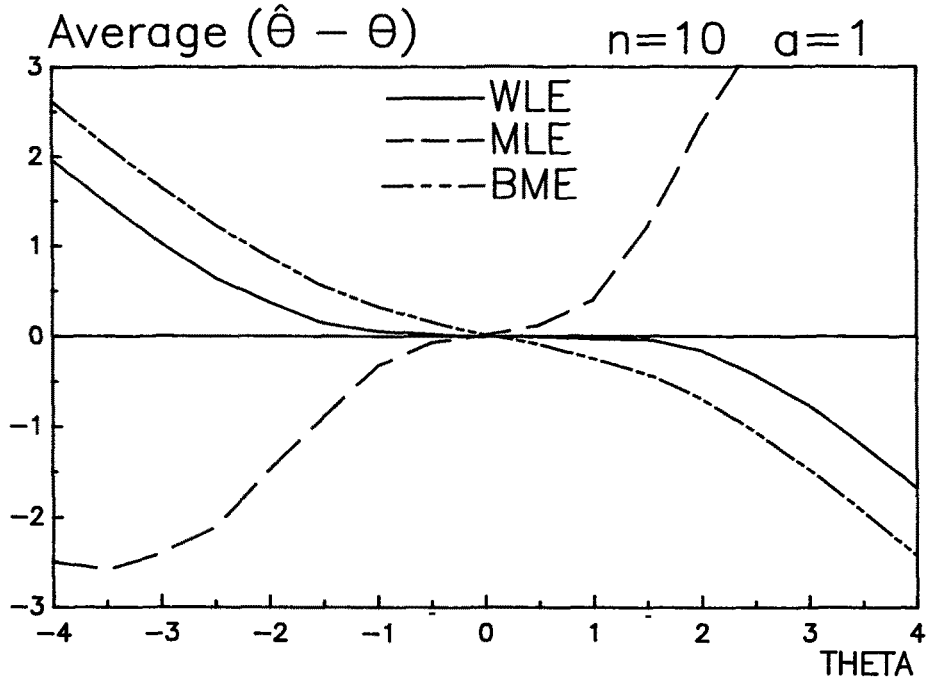

FIGURE 1
Average estimation error at $\hat{\theta}$ on conventional tests with 10 items, all $a = 1$, normally distributed b, and all $c = 0.20$. WLE(θ) is much less biased than MLE(θ) and BME(θ).

tors. Therefore, complete results of conventional tests will be presented here only for the test with 10 items and $a = 1$. Some results for the 30 and 60 item tests will also be shown to provide a range of values.

Figure 1 shows the average error of each estimator at each of the 17 values of θ. As predicted by (6) and (7) (and demonstrated by Lord, 1984), the bias of MLE(θ) is positively correlated with θ, while the bias of BME(θ) is negatively correlated. The bias of WLE(θ) is also negatively correlated with θ. It is very clear from the figure that WLE(θ) is considerably less biased than both MLE(θ) and BME(θ) over the entire range of θ; this holds for all test lengths and both values of the a-parameter. Note that the range of θ over which the bias of WLE(θ) is apparently negligible (i.e., indistinguishable from the zero reference line) is relatively broad, whereas the other two estimators have small bias essentially at a point.

Figure 2 shows the absolute bias of WLE(θ) for the 10 item test and the absolute bias of MLE(θ) for tests with 10, 30 and 60 items. Figure 2 indicates that WLE(θ) with 10 items is less biased than MLE(θ) with as many as six times as many items, depending upon the value of θ. Figure 3 gives the same comparisons between WLE(θ) and BME(θ). WLE(θ) is also less biased than BME(θ) in tests with four to six times as many items.

Comparison of the standard deviations of estimated θ in Figure 4 reveals a picture different from the comparisons on bias. BME(θ) has less variability than both WLE(θ) and MLE(θ) at all values of θ. (The decline of the standard deviations of MLE(θ) at high and low values of θ were caused by the artificial boundary of $|\text{MLE}(\theta) - \theta| < 5$. In some cases all 1000 values of MLE(θ) were equal to the artificial boundary.) It is not difficult to reduce variability at the expense of bias. This trade-off accounts for the low standard deviation of BME(θ), and will be expanded upon below. On the other hand

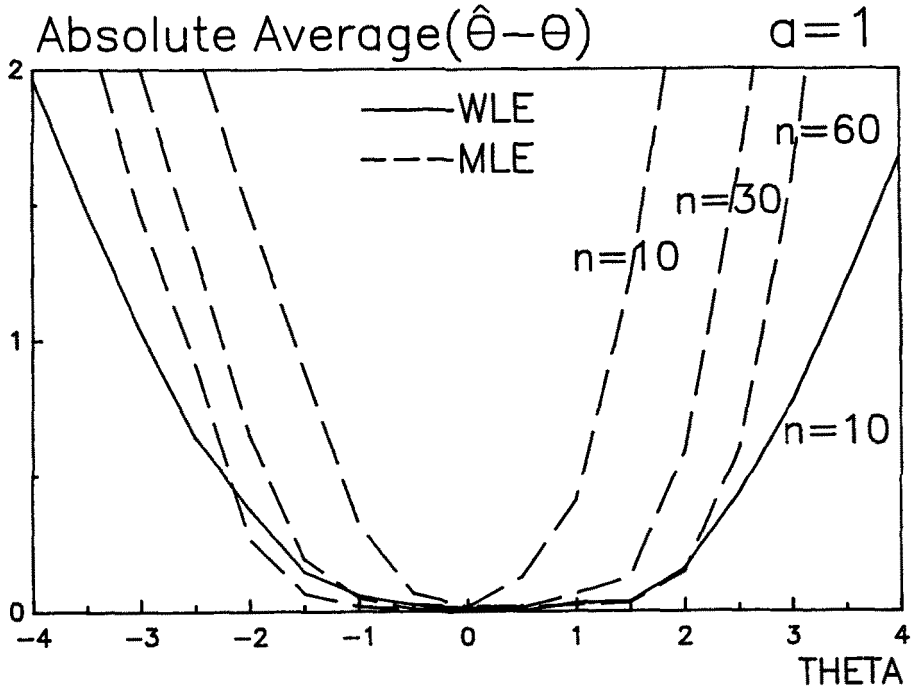

FIGURE 2
Absolute average estimation error of WLE(θ) with 10 items, and of MLE(θ) with 10, 30, and 60 items; all $a = 1$, normally distributed b, and all $c = 0.20$. WLE(θ) with only 10 items is less biased than MLE(θ) with 30 to 60 or more items.

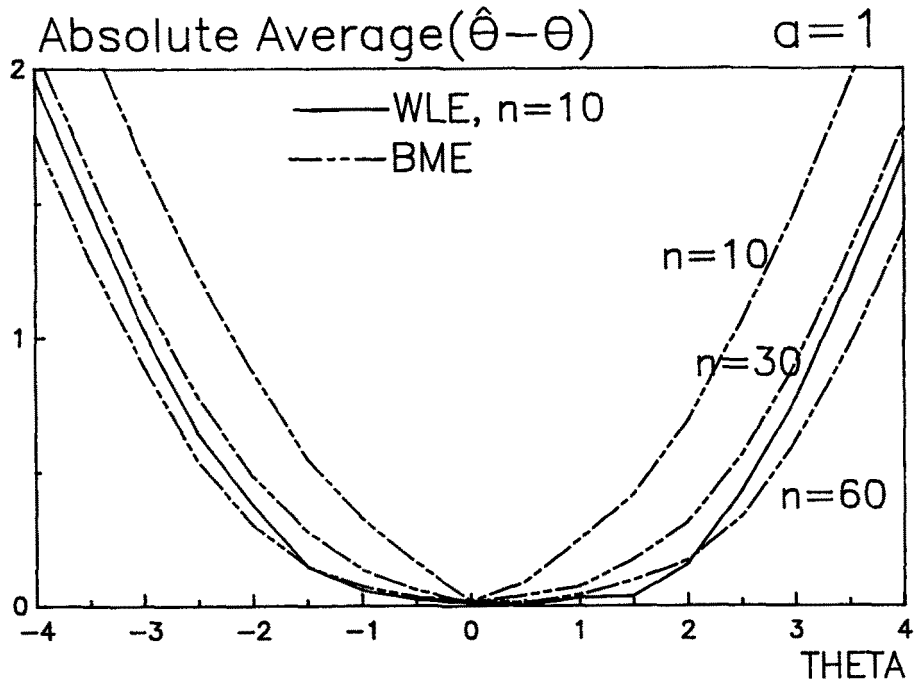

FIGURE 3
Absolute average estimation error of WLE(θ) with 10 items, and of BME(θ) with 10, 30, and 60 items; all $a = 1$, normally distributed b, and all $c = 0.20$. WLE(θ) with 10 items is less biased than BME(θ) with 30 items.

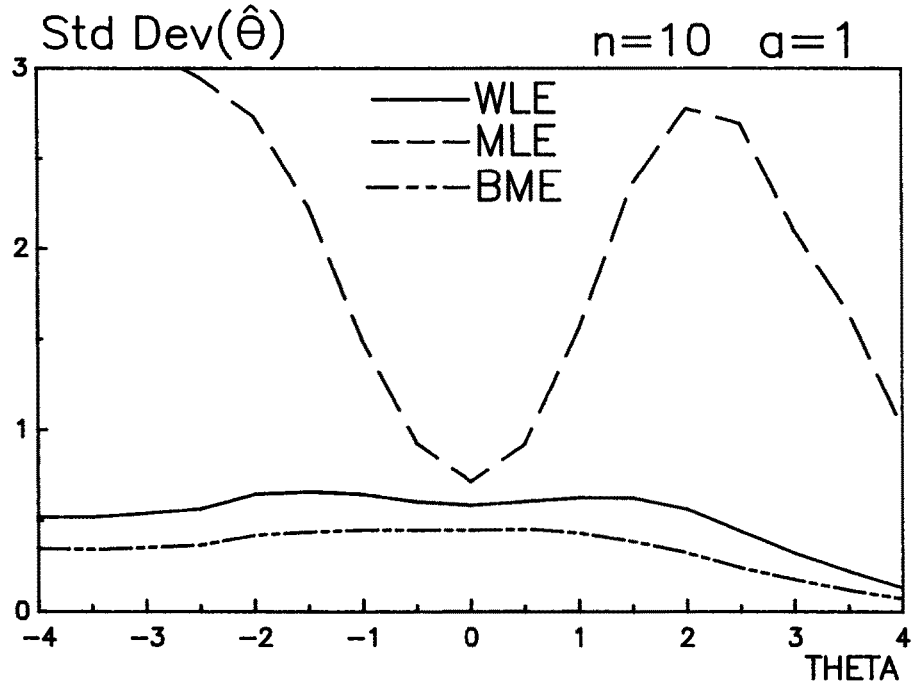

FIGURE 4
Standard deviation of $\hat{\theta}$ on conventional tests with 10 items, all $a = 1$, normally distributed b, and all $c = 0.20$. SD(WLE(θ)) is small despite WLE(θ) being relatively unbiased. BME(θ) trades small SD for large bias. (The decline of SD(MLE(θ)) at extreme θ is an artifact due to the artificial limit placed on MLE(θ).)

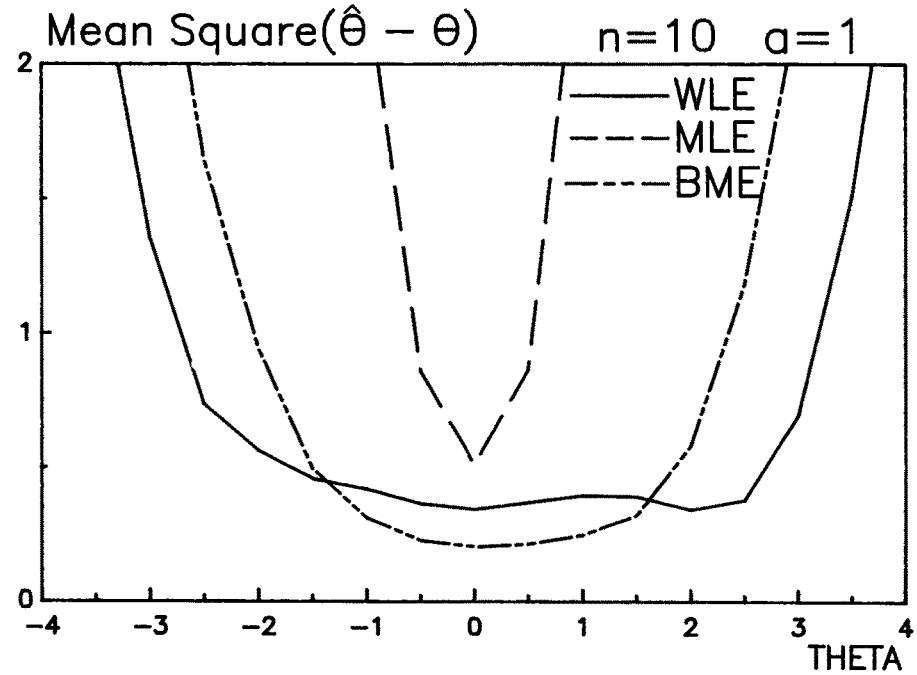

FIGURE 5
Mean squared error of $\hat{\theta}$ on conventional tests with 10 items, all $a = 1$, normally distributed b, and all $c = 0.20$.

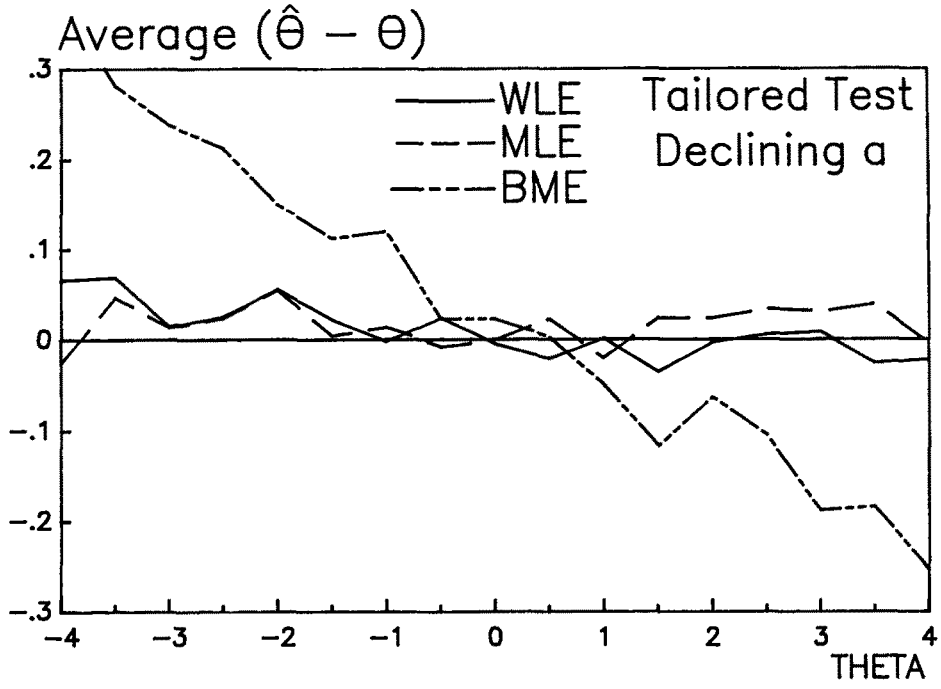

FIGURE 6
Average estimation error of $\hat{\theta}$ on tailored test with declining a-parameter, optimal b-parameter, and all c = 0.20. The declining a-parameter simulated the depletion of items with high information in a finite item bank. MLE(θ) and WLE(θ) are both unbiased. BME(θ) is very biased.

WLE(θ) also has small standard deviation, and is considerably less biased than BME(θ).

The mean squared error (MSE) of WLE(θ) (See Figure 5) is smaller than that of MLE(θ) at all values of θ for all 12 tests. BME(θ) has smaller MSE still at the more central values of θ. However, as the value of the a-parameter or the number of items increases, this advantage of BME(θ) over WLE(θ) shrinks rapidly. At all other values of θ, MSE(WLE(θ)) is considerably smaller than MSE(BME(θ)).

Tailored tests. Figures 6 through 10 give the average error of $\hat{\theta}$, SD($\hat{\theta}$), and MSE($\hat{\theta}$) for the tailored tests. BME(θ) is very biased at all non-zero values of θ on both tests. MLE(θ) is slightly positively biased at most values of θ on both tests. WLE(θ) is the least biased estimator at positive values of θ. At low values of θ, the biases for WLE(θ) and MLE(θ) are about equal and small. SD(WLE(θ)) and SD(BME(θ)) are about equal on both tailored tests, as are MSE(WLE(θ)) and MSE(BME(θ)) at most values of θ. SD(MLE(θ)) on the test with declining a-parameters is also about the same as SD(MLE(θ)) and SD(BME(θ)). However, SD(MLE(θ)) and MSE(MLE(θ)) are unexpectedly very large at central values of θ on the tailored test with $a = 2$.

The relative average number of items administered to reach the stopping rule for the three estimators on the tailored tests are shown in Figure 11. At central values of θ, WLE(θ) and BME(θ) used about the same number of items, and as few as half as many items as MLE(θ). Both WLE(θ) and BME(θ) required considerably fewer items at central values of θ than at more extreme values. The number of items required by MLE(θ) declined slightly on both tests as θ increased.

The average computation times for estimating θ between items divided by the number of items used in the estimation is displayed in Figure 12. For both tailored tests,

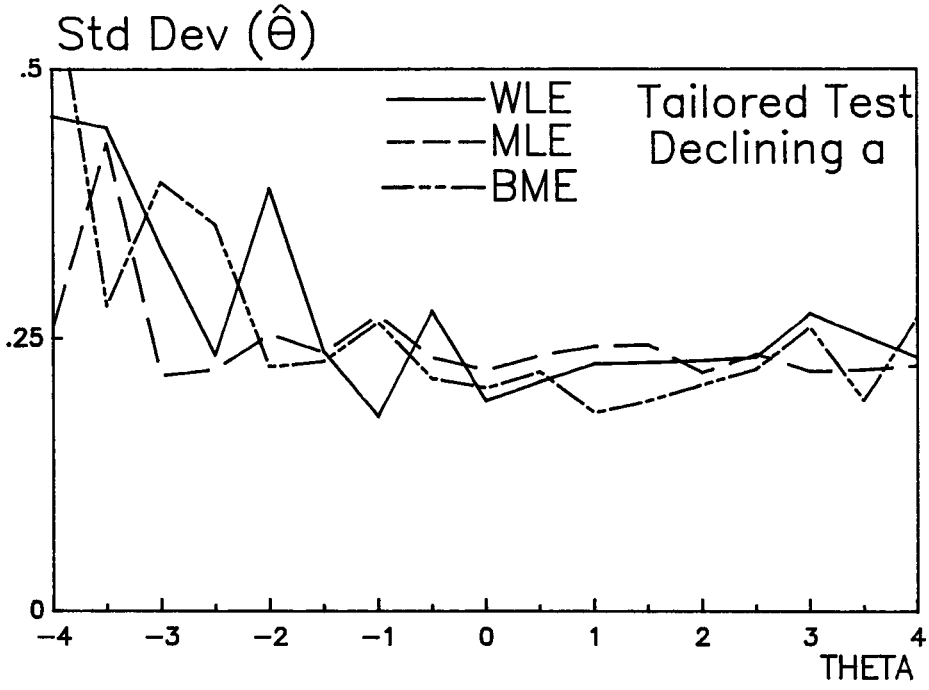

FIGURE 7
Standard deviation of $\hat{\theta}$ on tailored test with declining a-parameter, optimal b-parameter, and all $c = 0.20$. All methods had roughly the same SD due to the stopping rule based on test information.

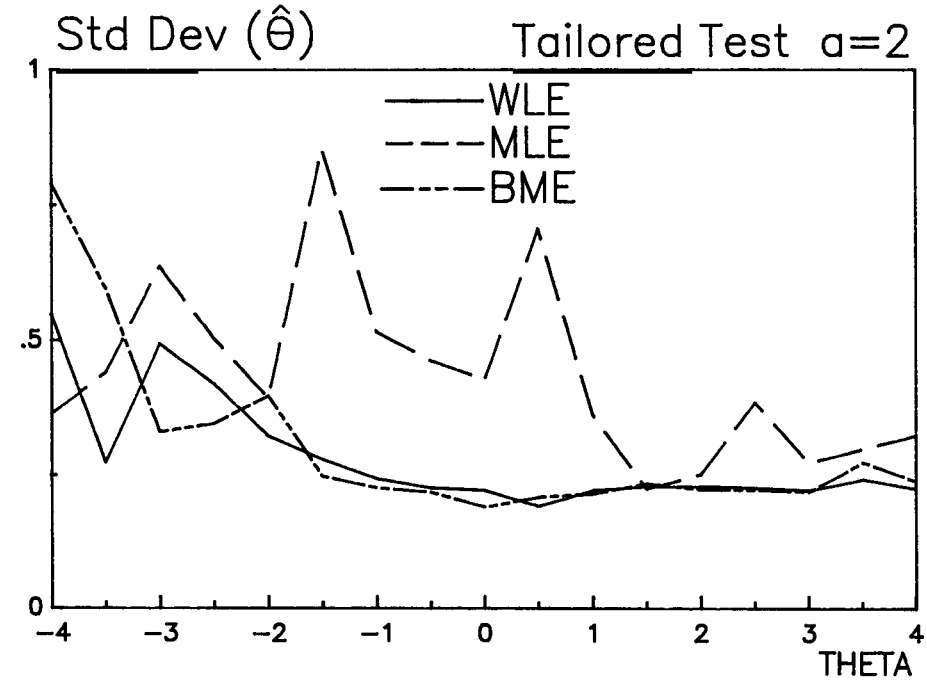

FIGURE 8
Standard deviation of $\hat{\theta}$ on tailored test with $a = 2$, optimal b-parameter, and all $c = 0.20$. SD(MLE(θ)) is large at central θ. This result was unexpected, and appears to be analogous to the attenuation paradox (conditional on θ).

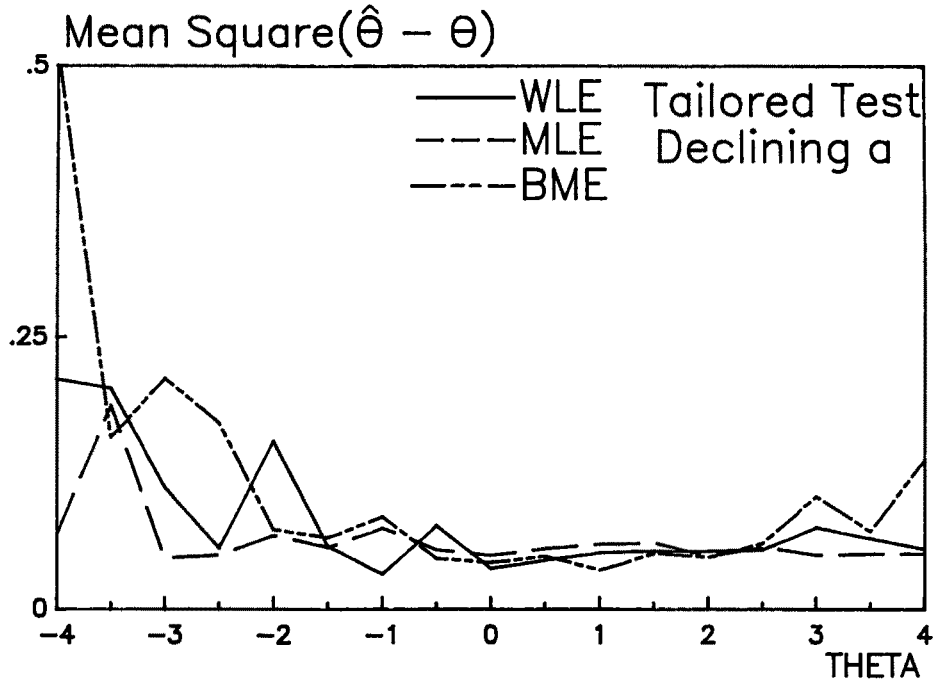

FIGURE 9
Mean squared error of $\hat{\theta}$ on tailored test with declining a-parameter, optimal b-parameter, and all $c = 0.20$.

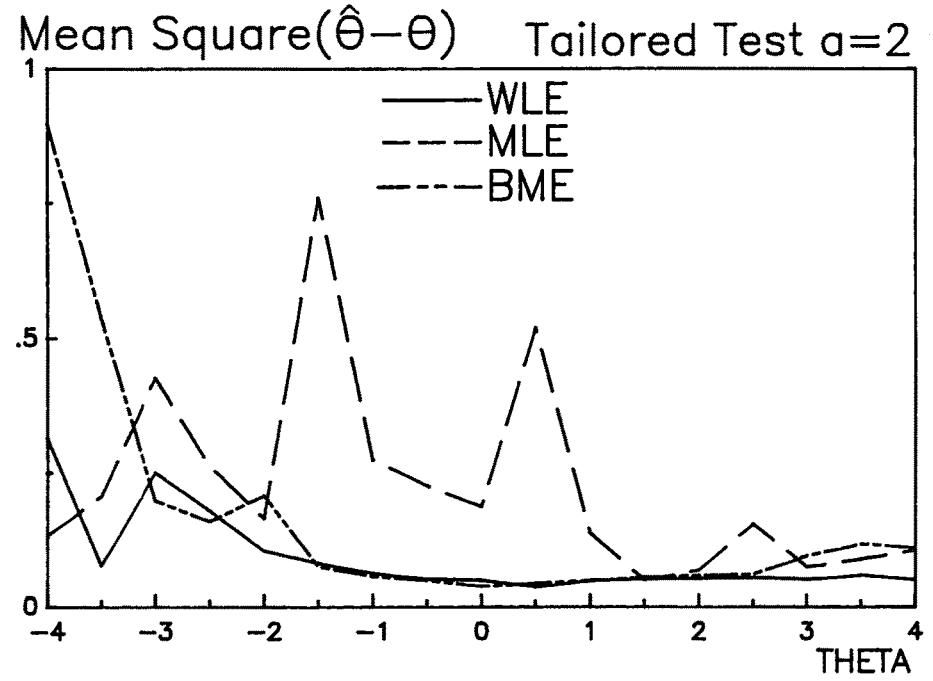

FIGURE 10
Mean squared error of $\hat{\theta}$ on tailored test with $a = 2$, optimal b-parameter, and all $c = 0.20$.

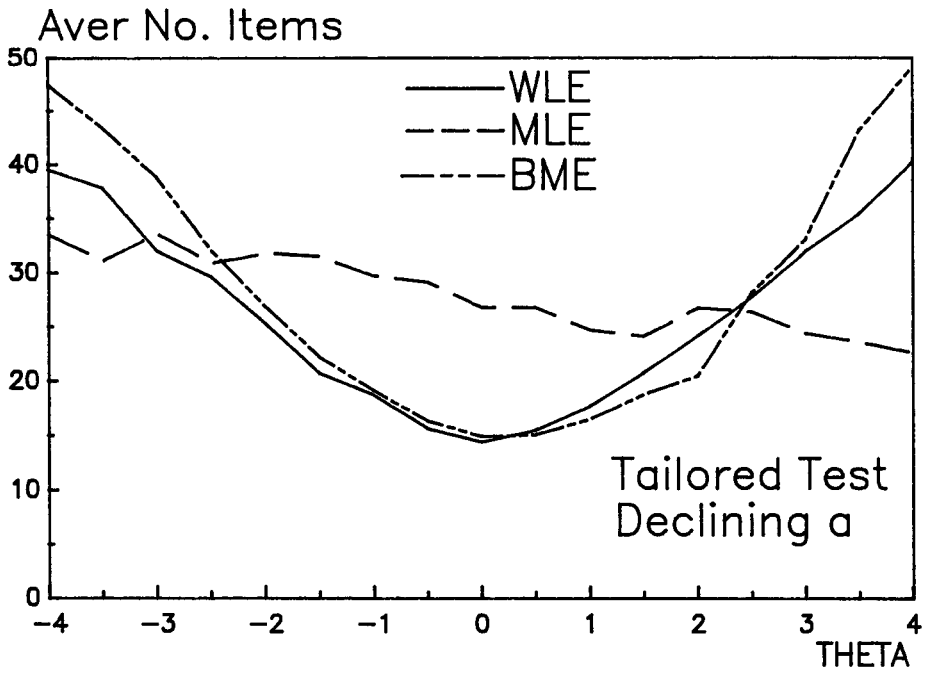

FIGURE 11
Average number of items administered on tailored test with declining a-parameter, optimal b-parameter, and all $c = 0.20$. WLE(θ) and BME(θ) used many fewer items than MLE(θ), but BME(θ) was very biased.

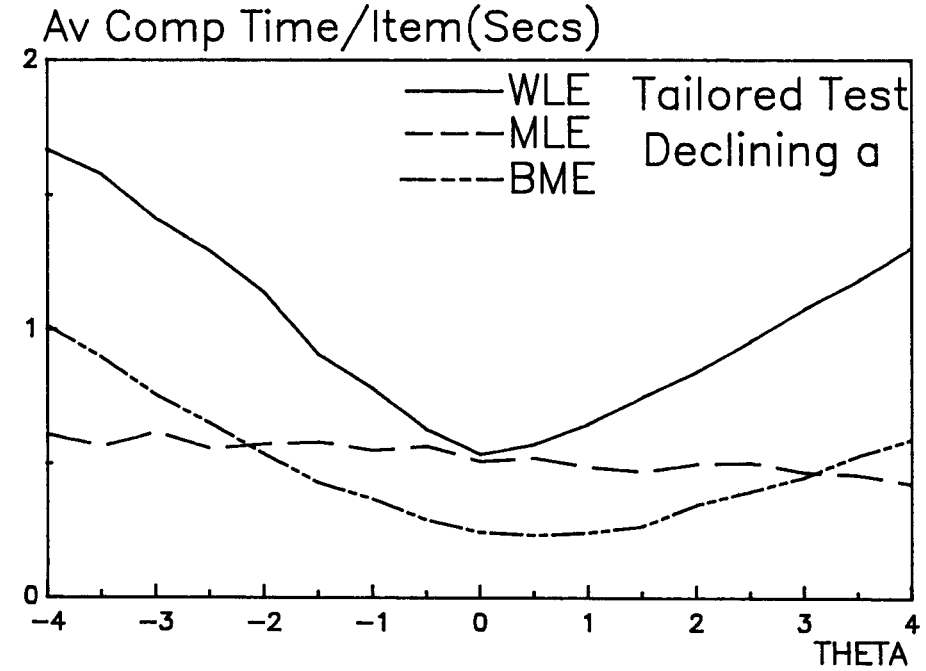

FIGURE 12
Relative average computation time between items on tailored test with declining a-parameter, optimal b-parameter, and all $c = 0.20$. WLE(θ) required about twice as much computation time per item as BME(θ).

WLE(θ) required more computation time than either of the other two estimators at all values of θ. Computation time for MLE was nearly constant on both tests and at all values of θ. BME used the least computation time of the three estimators on both tests at all except extreme values of θ.

Discussion

Conventional tests. WLE(θ) is clearly a less biased estimator of θ than both MLE(θ) or BME(θ). In light of the proof in the appendix, this outcome should not be surprising in some circumstances. What is surprising is that WLE(θ) was less biased in *every* condition investigated in this study: that is, for test lengths ranging from 10 to 60 items; for a-parameters of $a = 1$ and $a = 2$; for all values of θ, and for conventional tests as well as for tailored tests (utilizing both infinite item banks and simulated finite item banks). Moreover, WLE(θ) has small and roughly constant variance over a wide range of the θ-scale, as well as small MSE over a much wider interval than either MLE(θ) or BME(θ). For the conventional tests BME(θ) also fared better than MLE(θ) on all three criteria—bias, SD, and MSE. However, BME(θ) benefitted greatly by the design of the simulated tests. The tests were designed to have roughly "normally" distributed test information. The general effect of this design was to bias estimates outwardly (digress estimates away from the peak of test information). BME(θ) tends to regress estimates toward the peak of the prior probability distribution. Since the peaks of test information for these tests were located near $\theta = 0$ (which was the peak of the standard normal prior), the digression of test information and the regression of the prior tended to cancel out each other. For differently designed tests or a different prior, where the peak of the prior is not nearly coincident with the peak of test information, BME(θ) could be much worse.

Local unbiasedness and the Ω estimator. The width of the interval over which an estimator performs well is an important consideration. A comparison statistic of two estimators, R, may be defined as the ratio of the range of θ over which the magnitude of the bias is less than some small value β. For the test in Figure 1 with $\beta = 0.05$, R(WLE/MLE) = 4.4, and R(WLE/BME) = 5.0. Thus, WLE(θ) has small bias over 4.4 times the range of θ as does MLE(θ), and over 5 times the range as BME(θ).

It is trivially simple to create an estimator that has locally small bias, variance, and MSE (over a small interval of the θ-scale). For example, consider an estimator, Ω, which is defined as $\Omega = k$, where k is a constant. Then, used as an estimate of θ, BIAS(Ω) = $(k - \theta)$, VAR(Ω) = 0, and MSE(Ω) = $(k - \theta)^2$, but R(WLE/Ω) = 24.7. In the interval on the θ-scale where $|k - \theta|$ is sufficiently small, Ω may be superior to any other estimator by all three criteria. Note that a nonlinear transformation of Ω is equally superior over a small interval of the nonlinearly transformed θ-scale. The estimator Ω can be considered to be a Bayesian estimator for which the prior has non-zero density only where $\theta = k$. Ω compresses all estimates to the constant k, which has no functional relationship to the test. Other Bayesian estimators have the same characteristic—they "regress" estimates toward the locality of the peak of a prior distribution, which also has no functional relationship to the test.

This analysis of the estimator Ω makes it clear why Bayesian estimators sometimes work well; they sometimes "regress" estimates in a direction opposite to the bias of MLE(θ). Bayesian estimators may also compress the estimates into the locality of the θ-scale that is of interest; this is exactly the reason that BME(θ) had the smallest MSE for certain relatively small intervals. On the other hand, if the prior is not correct, then the "regression" of the Bayesian estimators will worsen estimation, rather than im-

prove it. In general, BME(θ) will work well if the prior approximates $w(\theta)$, the weighting function for WLE(θ). For the tests in this study a normal prior with standard deviation of two would have approximated $w(\theta)$ much better than the standard normal prior. Since the tests in this study have more peaked test information curves than many actual tests, in practice the priors for BME(θ) should have standard deviations greater than two in order to approximate $w(\theta)$.

Lord (1984) states that leading Bayesian testing practitioners prefer to use priors more diffuse than the standard normal prior due to practical considerations. Thus, it can be argued that *the appeal of BME(θ) in IRT is not due to the properties of the posterior distribution, but rather is due to the coincidence that in IRT, $w(\theta)$ can sometimes be approximated by a diffuse normal curve.*

WLE(θ) also "regresses" estimates. However, the regressing function, $w(\theta)$, is a function of the test, and is in no way arbitrary. Its action is *always* to regress estimates in the direction opposite of the bias of MLE(θ), and into an interval for which the test has sufficient information to reduce bias, variance, and MSE.

Tailored tests. For both tailored tests, WLE(θ) was only slightly less biased than MLE(θ) (and at high θ only), while BME(θ) was very biased. For the tailored test with a simulated finite item bank, WLE(θ) and MLE(θ) were also about the same on the standard deviation and MSE criteria, although MLE(θ) may have some small advantage at low θ. However, for the tailored test with an infinite item bank ($a = 2$), MLE(θ) was considerably worse than WLE(θ) over a wide central region of θ (using the standard deviation and MSE criteria). This effect is apparently due to the high values of the a-parameters, and would seem to be a conditional (on θ) analogy to the "attenuation paradox" (Lord & Novick, 1967); that is, the conditional variance of MLE(θ) increases (for central θ) when the a-parameters are as high as two. If so, then this result suggests that the selection of items from item banks so as to maximize item information may not be optimal for MLE(θ).

Since WLE(θ), over a wide range of θ, used many fewer items than MLE(θ) in order to achieve the stopping criteria, tailored tests using WLE(θ) would, in general, be much shorter than if MLE(θ) were used to estimate θ. This advantage translates into savings in terms of testing time and the exposure of items to potential compromise.

Computation time for estimating θ (in the interval between items in tailored testing) is an important consideration for tailored tests. The time required for WLE(θ) is always more than for MLE(θ) at all values of θ—ranging from a slight increase in time to more than three times as long. This delay may or may not be significant. The times found in this study are only relative. The absolute times required for applications will depend upon several factors, such as the speed of the computer, the programming language, and the cleverness of the programmer. If the actual computation times can be decreased to one second or less, then no harm is caused by the extra calculations required for WLE(θ).

The nature of $w(\theta)$. Some readers may think of $w(\theta)$ as a Bayesian prior, but it is easy to show that interpretation is incorrect. Suppose the same test was given to two groups of examinees, whose distributions of θ are different and known. Bayesians will be obliged to use different priors for the two groups. For WLE(θ) the same $w(\theta)$ would be used for both groups, because $w(\theta)$ is a function of the test only. Similarly, if two different tests of the same ability are given to a single group, Bayesians would use the same prior for both tests, whereas for WLE(θ) a different $w(\theta)$ would be used for each test.

The shape of $w(\theta)$ may be of interest to some readers. When $c_i = 0$, all i, as in the

one- and two-parameter models of IRT, $w(\theta) = I^{1/2}$, the square root of test information. When $c_i \neq 0$, any i,

$$w(\theta) = I^{1/2} \cdot \exp\left(-(1/2) \cdot \int I^{-1} \mathscr{E}(l_1 l_2)\, \partial\theta\right),$$

which is similar to $I^{1/2}$, but with a fat tail (nonzero asymptote) at negative θ, and with its peak to the left of the peak of $I^{1/2}$. The $\mathscr{E}(l_1 l_2)$ in the argument of the integral is the expected value ($\mathscr{E}$) of the product of the 1st (l_1) and 2nd (l_2) derivatives of the log likelihood. It is closely related to the residual variation of the linear regression of the 2nd derivative on the 1st derivative, and corrects for the deviation of the model from the exponential family (Efron & Hinkley, 1978).

WLE extends naturally to the multiple parameter case. Jeffreys (1961) and Akaike (1978) have shown $w(\boldsymbol{\theta})$ to be equal to the square root of the determinant of the information matrix for certain models (which include the 1- and 2-parameter, but not the 3-parameter model of IRT). However, their motivations were Bayesian, which was shown above to be an inappropriate interpretation. Barndorff-Nielsen (1983) proposed using the square root of the determinant of observed information $(-l_2)^{1/2}$ as a weighting function, which is identical to $w(\theta)$ in the 1- and 2-parameter models of IRT (but, again, not the 3-parameter model).

Rational bounds for MLE(θ). Even though pure MLE has very attractive asymptotic properties, its performance with tests of finite length is miserable by most criteria. The unbounded nature of MLE(θ) is one source of its difficulties with finite tests. It is well known that if the response vector $\mathbf{u} = \mathbf{1}$, (all items answered correctly), then MLE(θ) $= +\infty$, and if $\mathbf{u} = \mathbf{0}$ (all items answered incorrectly), then MLE(θ) $= -\infty$. Thus, with finite tests MLE(θ) has infinite variance and indeterminate mean. In practical applications pure MLE(θ) is never used. Upper and lower bounds on MLE(θ) are always set. Even in theoretical work these bounds are necessary; for example, in Lord's (1983a) derivation of the bias of MLE(θ), θ is assumed to be bounded. The practical bounds are set arbitrarily at values that are not expected to have a significant impact.

It has not been recognized that there exist rational upper and lower bounds for MLE(θ) in IRT. The rational bounds occur at the points where the slope of BIAS(MLE(θ)) is equal to one. That is, MLE(θ) should be defined only in the interval(s) where

$$\theta^- \leq \text{MLE}(\theta) \leq \theta^+,$$

where θ^- is defined such that

$$\frac{\partial\, \text{BIAS}(\text{MLE}(\theta^-))}{\partial\theta} = 1,$$

and

$$\frac{\partial^2\, \text{BIAS}(\text{MLE}(\theta^-))}{\partial\theta^2} < 0;$$

and, furthermore, where θ^+ is defined such that

TABLE 1

Minimum, MLE(θ^-), and Maximum, MLE(θ^+), of "Sensible" MLE(θ), and of WLE(θ) and BME(θ), for Conventional Tests with n items, All a = 1 or 2, Normally Distributed b, and All c = 0.20 .

	a = 1			a = 2		
n	MLE(θ-)	Min WLE(θ)	Min BME(θ)	MLE(θ-)	Min WLE(θ)	Min BME(θ)
10	-1.907	-2.543	-1.769	-1.685	-2.002	-1.763
20	-2.284	-2.989	-2.175	-2.028	-2.337	-2.068
30	-2.486	-3.242	-2.410	-2.208	-2.519	-2.237
40	-2.623	-3.418	-2.574	-2.329	-2.642	-2.354
50	-2.725	-3.553	-2.701	-2.419	-2.735	-2.444
60	-2.807	-3.664	-2.803	-2.491	-2.809	-2.515
n	MLE(θ+)	Max WLE(θ)	Max BME(θ)	MLE(θ+)	Max WLE(θ)	Max BME(θ)
10	22.847	2.347	1.587	12.246	1.903	1.630
20	23.163	2.800	2.009	12.561	2.229	1.941
30	23.331	3.058	2.252	12.729	2.408	2.116
40	23.444	3.238	2.422	12.842	2.530	2.238
50	23.529	3.376	2.553	12.927	2.622	2.330
60	23.596	3.489	2.659	12.995	2.696	2.404

$$\frac{\partial\, \text{BIAS}(\text{MLE}(\theta^+))}{\partial\theta} = 1,$$

and

$$\frac{\partial^2\, \text{BIAS}(\text{MLE}(\theta^+))}{\partial\theta^2} > 0.$$

Outside these bounds the slope of BIAS(MLE(θ)) is always greater than one, that is, the magnitude of BIAS(MLE(θ)) increases faster than θ changes. Therefore, the expected value of MLE(θ) is always farther from true θ than is the closer of the two bounds, and the closer bound is a less biased estimate of θ than is MLE(θ). Moreover, the replacement of the closer bound for MLE(θ) as the estimate of θ reduces the variance of the estimates, and, consequently, also the MSE of the estimates. Thus, for all three of the criteria used in this study (bias, variance, and MSE), MLE(θ) is always degraded by permitting it to fall outside the closed interval [θ^-, θ^+]. These bounds are proposed as reasonable or "sensible" limits on MLE(θ). Therefore, MLE(θ) with these bounds may be termed "sensible MLE(θ)", or SMLE(θ).

Table 1 lists the minimum and maximum values of SMLE(θ), and the actual minimum and maximum values of WLE(θ) and BME(θ) for the 12 conventional tests (all items answered wrong and correctly, respectfully). Note that the minimum of WLE(θ) is always more extreme (more negative) than the minimum of the other two estimators. In contrast, the maximum of SMLE(θ) is by far the most extreme. This extreme upper

bound on SMLE(θ) would do little to reduce bias, but at least would prevent infinite estimates of θ. The extraordinary asymmetry of the bounds of SMLE(θ) is apparently due to the loss of test information at low θ, caused by the nonzero c-parameters.

Summary and Conclusions

A new method of estimation, weighted likelihood estimation (WLE), was derived, and proved to yield asymptotically normally distributed estimates, with finite variance, and with bias of only $o(n^{-1})$. This property of WLE is in contrast to maximum likelihood estimation (MLE) and Bayesian modal estimation (BME), both of which are biased to $O(n^{-1})$. The new estimator was applied to ability estimation in IRT, assuming the true values of the item parameters were known. Using Monte Carlo methods, WLE(θ) was compared to MLE(θ) and BME(θ) on 12 conventional tests with 10 to 60 items, and a-parameters of 1 or 2. The three estimators were also compared on two tailored tests. One tailored test had an infinite item bank and all $a = 2$. The other tailored test simulated a finite item bank with declining a-parameters.

In all tests WLE(θ) was less biased than both of the other estimators. In addition WLE(θ) had small variance over the entire range of the θ-scale, as well as small mean squared error even at noncentral θ. The relative unbiasedness of WLE(θ) makes this estimator particularly appropriate in applications of item response theory (IRT) for which the parameter invariance property is important.

Two new insights for MLE(θ) were discovered: (a) rational bounds, and (b) a conditional analogy to the attenuation paradox in tailored tests with high a-parameters.

The heart of WLE is a weighting function, $w(\theta)$, which is multiplied times the likelihood function, and the product maximized. This weighting function, which reduces the bias and uncontrolled variance of MLE(θ), is a function of θ and the item parameters, and is specific to each test. It was shown to be equal to the square root of test information for the one- and two-parameter models of IRT, and equal to a closely related function for the three-parameter model.

Appendix
Proof That the Weighted Likelihood Estimate is Unbiased to Order n^{-1}

The approach and techniques of this derivation were taken from, and parallel closely, the derivations in Lord (1983a, 1983b, & 1984) of the biases of the maximum likelihood and Bayesian modal estimates in item response theory, both of which are $O(n^{-1})$. The derivation is limited to a single parameter for a multinomially distributed variable and a regular, "smooth" mathematical model with rather restrictive assumptions. However, it would seem that the extension to other distributions, multiple parameters, and less regular models with less restrictive assumptions should be straight forward.

Preliminaries. For a set of n independent experiments, H_i ($i = 1, 2, \ldots, n$) with binary outcomes, u_i, (success or failure), let $P \equiv P_i(\theta)$ denote the probability of a success ($u_i = 1$), and let $Q \equiv Q_i(\theta) = 1 - P_i(\theta)$ denote the probability of failure ($u_i = 0$), where $P_i(\theta)$ is a strictly increasing function of the common parameter θ for all n experiments. $P_i(\theta)$ is not necessarily equal to $P_h(\theta)$, $h \neq i$. Let $\mathbf{u} = \{u_i\}$ denote the multinomially distributed, $n \times 1$ vector of outcomes of the n experiments.

Assumptions. (a) θ is a bounded variable on a continuous scale. (b) $P_i(\theta)$ is continuous and bounded away from 0 and 1 at all values of θ, $i = 1, 2, \ldots, n$. (c) At least

the first five derivatives with respect to θ of $P_i(\theta)$ exist at all values of θ, and are bounded. (d) For asymptotic considerations n is considered to be incremented with replications of all of the original n experiments.

Assumption (d) is the one usually made in mental test theory (Lord, 1983a). Bradley and Gart (1962), Hoadley (1971), and Basu and Ghosh (1980) present more general alternatives to assumption (d).

From these assumptions and theorems 1(i) and 1(iv) of Bradley and Gart (1962) it follows that the maximum likelihood estimate of θ, MLE(θ) $\equiv \hat{\theta}$, is a consistent estimator of θ, and that $n^{1/2}$. $(\hat{\theta} - \theta)$ is asymptotically normally distributed with zero mean and with variance given by

$$\lim_{n \to \infty} (nI)^{-1},$$

where I is Fisher's information. Assumption (d) guarantees the existence of this limit (Lord, 1983a).

Maximum Likelihood Estimation

The likelihood function, $L(\mathbf{u}|\theta)$, is given by (2). Let $l_s = \partial^s \ln L(\mathbf{u}|\theta)/\partial\theta^s$, where $\partial^s/\partial\theta^s$ indicates the s-th partial derivative with respect to θ, and ln indicates the natural logarithm.

The maximum likelihood estimate of θ is defined as the value of θ that maximizes (2). Usually $\hat{\theta}$ is found by setting l_1 equal to zero, and solving for θ, as in (A1).

$$\frac{\partial \ln L(\mathbf{u}|\theta)}{\partial\theta} \equiv l_1 = \Sigma \frac{(u - P)P'}{PQ} = 0, \tag{A1}$$

evaluated at $\hat{\theta}$. In (A1) and hereafter the argument (θ) and index i are usually dropped for convenience.

The asymptotic variance of MLE(θ) is the reciprocal of Fisher's information (Kendall & Stuart, 1973).

$$I = \mathscr{E}(l_1^2) = -\mathscr{E}(l_2) = \Sigma \frac{P'^2}{PQ},$$

where $\mathscr{E}$ is the expectation operator, and $P' = \partial P/\partial\theta$.

The bias of MLE(θ), BIAS(MLE(θ)), from Cox and Hinkley (1974) is given by (A2).

$$\text{BIAS}(\text{MLE}(\theta)) \equiv \mathscr{E}(\hat{\theta} - \theta) = \frac{-J}{2I^2}, \tag{A2}$$

where

$$J = -2\mathscr{E}(l_1 l_2) - \mathscr{E}(l_3) = \Sigma \frac{P'P''}{PQ}, \quad \text{and} \quad P'' = \frac{\partial^2 P}{\theta^2}.$$

Equation (A2) is equivalent to Lord's (1983a) Equation (28) for BIAS(MLE(θ)). Note that I and J are $O(n)$, and that since neither are a function of $\mathbf{u}$, $J/2I^2$ is $O(n^{-1})$.

Weighted Likelihood Estimation

The weighted likelihood estimate of θ, $\underline{\text{WLE}}(\theta) \equiv \theta^*$, is defined as the value of θ, such that the weighted likelihood function, given in (A3) is maximized,

$$w(\theta) \cdot L(\mathbf{u}|\theta), \tag{A3}$$

where $\partial \ln w(\theta)/\partial\theta = J/2I$. WLE($\theta$) is found by solving the weighted likelihood equations as in (A4),

$$l_1 + \frac{J}{2I} = 0, \tag{A4}$$

evaluated at θ^*, or, letting $d_s = \partial^s \ln w(\theta)/\partial\theta^s$,

$$l_1 + d_1 = 0.$$

Note that $d_1 = J/2I = -I \cdot \text{BIAS}(\text{MLE}(\theta))$.

Rather than finding WLE(θ) by maximizing (A3), it will be useful to maximize (A3) raised to the n^{-1} power, which will always yield the same estimate for any given set of data, since n is always positive. The reason for doing so is to help keep track of the order of the terms. Letting T_S = the s-th derivative of the log of n-th root of (A3), and $T_1^* = T_1$, evaluated at θ^*,

$$T_s = \frac{\partial^s \ln [w(\theta) \cdot L(\mathbf{u}|\theta)]^{1/n}}{\partial\theta^s}$$

$$= \frac{l_s}{n} + \frac{d_s}{n}.$$

Theorem. WLE(θ) is unbiased to order n^{-1}, that is,

$$\text{BIAS}(\text{WLE}(\theta)) = 0 + o(n^{-1}),$$

where $o(n^{-r})$ represents terms such that

$$\lim_{n\to\infty} n^r \cdot o(n^{-r}) = 0.$$

Proof. It is sufficient to maximize the log of the n-th root of (A3):

$$T_1^* = \frac{l_1}{n} + \frac{d_1}{n} = 0, \tag{A5}$$

evaluated at θ^*. Letting $x = (\theta^* - \theta)$, expand (A5) in terms of x and θ.

$$T_1^* = 0$$

$$= T_1 + xT_2 + \frac{x^2T_3}{2} + \frac{x^3T_4}{6} + \frac{\tau x^4 V_5}{24}, \tag{A6}$$

where $V_5 \equiv \text{Max}\ (T_5)$ over θ, and $|\tau| < 1$. This closed form of the expansion is always valid, making the proof of the convergence of the Taylor series unnecessary. Letting $g_s = \mathscr{E}l_s/n$, and $e_s = l_s/n - g_s$, then $l_s/n = g_s + e_s$, and

$$T_s = g_s + e_s + \frac{d_s}{n}. \tag{A7}$$

The purpose of (A7) is to separate T_s into a sum g_s of terms not containing $(u - P)$, a sum e_s of terms containing $(u - P)$, and a term d_s/n that is a ratio of sums not containing $(u - P)$.

Substituting (A7) into (A6), $s = 1, 2, 3, 4$, expresses the expansion of the weighted likelihood equation in terms of g_s, e_s, d_s/n, V_5, and powers of x in (A8).

$$\begin{aligned} -\left(e_1 + \frac{d_1}{n}\right) = g_1 &+ x\left(g_2 + e_2 + \frac{d_2}{n}\right) \\ &+ \frac{x^2\left(g_3 + e_3 + \frac{d_3}{n}\right)}{2} \\ &+ \frac{x^3\left(g_4 + e_4 + \frac{d_4}{n}\right)}{6} \\ &+ \frac{\tau x^4 V_5}{24}. \end{aligned} \tag{A8}$$

We now need to evaluate some of the terms in (A8), and their expected values.

$$g_1 = 0.$$

$$g_2 = -n^{-1}\Sigma \frac{P'^2}{PQ} = \frac{-I}{n}. \tag{A9}$$

$$g_3 = \frac{-3J + 2K}{n}, \tag{A10}$$

where $K = \Sigma(1 - 2P)P'^3/(PQ)^2 = -3\mathscr{E}(l_1 l_2) - \mathscr{E}(l_3)$.

$$d_1 = \frac{J}{2I}. \tag{A11}$$

$$d_2 = \frac{IJ' - I'J}{2I^2}.$$

$$d_3 = \frac{I^2J'' - II''J - 2II'J' + 2I'^2J}{2I^3}.$$

where $(')$ and $('')$ indicate first and second derivatives with respect to θ, respectively.

Since g_s and d_s do not contain $(u - P)$,

$$\mathscr{E}g_s = g_s, \quad \text{and} \quad \mathscr{E}d_s = d_s.$$

$$e_1 = n^{-1} \cdot \Sigma(u - P)P'/PQ. \tag{A12}$$

$$e_2 = n^{-1} \cdot \Sigma\left\{(u - P) \cdot \left[\frac{P''}{PQ} - \frac{(1 - 2P)P'^2}{(PQ)^2}\right]\right\}. \tag{A13}$$

$$\mathscr{E}e_s = 0.$$

Since e_s is n^{-1} times the sum of the terms of l_s which contain $(u - P)$, e_s may be expressed as

$$e_s = n^{-1} \cdot \Sigma(u - P) \cdot R_{si}, \; s = 1, 2, 3, 4, 5,$$

where R_{si} is the $(s - 1)$-th derivative of P'/PQ, and does not depend on n nor on u_i. Since by assumption (b) P and Q are bounded, and by assumption (c) the required derivatives of P are bounded, the R_{si} and, thus e_s, are bounded. By assumption (d) the bound does not depend on n. The same conclusion is true of g_s.

Since by assumption (d) l_1/n in (A5) is $O(n^0)$, and d_1/n is $O(n^{-1})$, (A1) and (A4) are asymptotically equivalent, and, asymptotically, $n^{1/2} \cdot (\theta^* - \theta) = n^{1/2} \cdot (\hat{\theta} - \theta)$. Because $n^{1/2} \cdot (\hat{\theta} - \theta)$ is asymptotically normally distributed with zero mean and finite variance, so is $n^{1/2} \cdot (\theta^* - \theta)$. Therefore, $\mathscr{E}x^r$ $(r = 1, 2, \ldots)$ is $O(n^{-r/2})$. By similar logic e_s^r is of the same order. By the Cauchy-Schwartz inequality $\mathscr{E}x^r e_s^t <= (\mathscr{E}x^{2r} e_s^{2t})^{1/2}$, and therefore $\mathscr{E}x^r e_s^t$ is $O(n^{-(r+t)/2})$ $(r, t = 1, 2, \ldots)$.

The Variance of WLE(θ)

To get the variance of θ^*, VAR (θ^*), square (A8) and take expectations.

$$\begin{aligned}\mathscr{E}e_1^2 + \frac{2d_1\mathscr{E}e_1}{n} + \frac{d_1^2}{n^2} = {} & g_2^2\mathscr{E}x^2 \\ & + g_2\mathscr{E}x^2e_2 + g_2g_3\mathscr{E}x^3/2 \\ & + \mathscr{E}x^2e_2^2 + \frac{g_2\mathscr{E}x^3e_3}{2} + \frac{g_2g_3\mathscr{E}x^3}{2} + \frac{g_3\mathscr{E}x^4}{4} + \cdots\end{aligned} \tag{A14}$$

The terms in the first line of (A14) are $O(n^{-1})$, in the second line $O(n^{-3/2})$, and in the third line $O(n^{-2})$ with the remaining terms $o(n^{-2})$. Dropping all terms of $o(n^{-1})$, we can rewrite (A14) as

$$\mathscr{E}e_1^2 + \frac{2d_1\mathscr{E}e_1}{n} + \frac{d_1^2}{n^2} = g_2^2\mathscr{E}x^2 + o(n^{-1}).$$

Since $\mathscr{E}e_1^2 = I/n^2$, $\mathscr{E}e_1 = 0$, $g_2^2 = (-I)^2/n^2$, and $\mathscr{E}x^2 =$ VAR (θ^*), (A14) evaluates as (A15).

$$\frac{I + d_1^2}{n^2} = \left(\frac{I^2}{n^2}\right) \text{VAR } (\theta^*) + o(n^{-1}). \tag{A15}$$

Solving for VAR (θ^*) gives (A16),

$$\text{VAR}\,(\theta^*) = \frac{I + d_1^2}{I^2} + o(n^{-1})$$

$$= I^{-1} + o(n^{-1}), \tag{A16}$$

which proves that the asymptotic variance of WLE(θ) is equal to the asymptotic variance of MLE(θ).

The Bias of WLE(θ)

Let $\mathscr{E}_1$ indicate the expectation operator in which only $O(n^{-1})$ terms are retained. To get the first order statistical bias of θ^* take the first order expectation of (A8) to obtain (A17).

$$\frac{-d_1}{n} = g_2\mathscr{E}_1 x + \mathscr{E}_1 x e_2 + \frac{g_3\mathscr{E}_1 x^2}{2}. \tag{A17}$$

To evaluate $\mathscr{E}_1 x e_2$ multiply (A8) by e_2, and take first order expectations.

$$-\mathscr{E}_1 e_1 e_2 = g_2\mathscr{E}_1 x e_2. \tag{A18}$$

To evaluate the LHS of (A18) substitute (A12) and (A13), and take the expectation. Both (A12) and (A13) are sums of n terms indexed with i, each containing the factor $(u_i - P_i)$. The product is a double sum of n^2 terms, each, the product of a term in (A12), indexed with i, and a term in (A13), indexed with i', say. Because the n experiments, H_i, are independent, the expected value of all terms are equal to zero, except the n terms where $i = i'$. Noting from (A10) that $K = \Sigma(1 - 2P)P'^3/(PQ)^2$, then

$$-\mathscr{E}_1 e_1 e_2 = \frac{-J + K}{n^2}. \tag{A19}$$

Substituting (A9) and (A19) into (A18), and solving for $\mathscr{E}_1 x e_2$ gives

$$\mathscr{E}_1 x e_2 = \frac{J - K}{nI}. \tag{A20}$$

Substituting (A9), (A10), (A11), (A16), and (A20) into (A17) obtains

$$\frac{-J}{2nI} = \left(\frac{-I}{n}\right)\mathscr{E}x + \frac{J - K}{nI} + \frac{-3J + 2K}{2nI}. \tag{A21}$$

Finally, solving (A21) for $\mathscr{E}x$,

$$\mathscr{E}x = \mathscr{E}(\theta^* - \theta) = 0 + o(n^{-1}),$$

which completes the proof. □

It is interesting to note that, if the mathematical model is such that $P'' = P' \cdot \partial(\ln PQ)/\partial\theta$ as in item response theory when $c_i = 0$, all i, then

$$l_2 = -I, \quad \text{exactly},$$

$$\mathscr{E}(l_1 l_2) = 0,$$

$$I' = J = K = -\mathscr{E}(l_3), \quad \text{and}$$

$$w(\theta) = I^{1/2}$$

Otherwise, $w(\theta) = I^{1/2} \cdot \exp((-1/2)\int(J - K)/I\,\partial\theta)$ for which there is no known closed form solution for the indefinite integral.

References

Akaike, H. (1978). A New Look At The Bayes Procedure. *Biometrika*, *65*(1), 53–59.

Anderson, J. A., & Richardson, S. C. (1979). Logistic discrimination and bias correction in maximum likelihood estimation. *Technometrics*, *21*(1), 71–78.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., & Tukey, J. W. (1972). *Robust estimates of location*. Princeton, NJ: Princeton University Press.

Baker, F. B. (1984). Ability metric transformations involved in vertical equating under item response theory. *Applied Psychological Measurement*, *8*(3), 261–271.

Barndorff-Nielsen, O. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, *70*(2), 343–365.

Basu, A. P., & Ghosh, J. K. (1980). Asymptotic properties of a solution to the likelihood equation with life-testing applications, *Journal of the American Statistical Association*, *75*(370), 410–414.

Bock, R. D. (1983). The discrete Bayesian. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement. A festschrift for Frederick M. Lord* (pp. 103–115). NJ: Lawrence Erlbaum.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443–459.

Bock, R. D., & Mislevy, R. J. (1981). *Biweight estimates of latent ability*. Unpublished manuscript.

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

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. New York: Chapman & Hall. (Distributed by Halsted Press, New York)

Efron, B., & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, *65*(3), 457–487.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology*, *33*, 234–246.

Hoadley, B. (1971). Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *Annals of Mathematical Statistics*, *42*(6), 1977–1991.

Jeffreys, H. (1961). Theory of Probability (3rd ed.). Oxford: Clarendon.

Jones, D. H. (1982). *Redescending M-type estimators of latent ability* (Program Statistics Tech. Rep. No. 82-30). Princeton, NJ: Educational Testing Service.

Kendall, M. G., & Stuart, A. (1973). *The advanced theory of statistics* (Vol. 2). New York: Hafner.

Kendall, M. G., & Stuart, A. (1977). *The advanced theory of statistics* (Vol. 1). London: Charles Griffin.

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

Lord, F. M. (1983a). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, *48*, 233–245.

Lord, F. M. (1983b). *Memorandum for: Ms. Stocking, Ms. M. Wang, Ms. Wingersky. Subject: Sampling variance and bias for MLE and Bayesian estimation of* θ. August 26, 1983 (Internal Memorandum). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1984). *Maximum likelihood and Bayesian parameter estimation in item response theory* (Research Rep. No. RR-84-30-ONR). Princeton, NJ: Educational Testing Service.

Lord, F. M., & Novick, M. (1967). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, *70*, 351–356.

Quenouille, M. H. (1956). Notes on Bias in Estimation. *Biometrika*, *43*, 353–360.

Samejima, F. (1980). *Is Bayesian estimation proper for estimating the individual's ability* (Research Rep. 80-3). Knoxville, TN: University of Tennessee, Department of Psychology.

Schaefer, R. L. (1983). Bias correction in maximum likelihood logistic regression. *Statistics in Medicine*, *2*, 71–78.

Wainer, H., & Wright, B. D. (1980). Robust estimation of ability in the Rasch model. *Psychometrika*, 45, 373–391.

Weiss, D. J., & McBride, J. R. (1984). Bias and information of Bayesian adaptive testing. *Applied Psychological Measurement*, *8*(3), 273–285.

Manuscript received 4/17/85
Final version received 6/24/88