PSYCHOMETRIKA—VOL. 56, NO. 1, 55–67
MARCH 1991

ASSESSING STATISTICAL ACCURACY IN ABILITY ESTIMATION: A BOOTSTRAP APPROACH

MICHELLE LIOU

ACADEMIA SINICA
UNIVERSITY OF CALIFORNIA, BERKELEY

LIEN-CHI YU

NORTH CAROLINA STATE UNIVERSITY, RALEIGH

Given known item parameters, the bootstrap method can be used to determine the statistical accuracy of ability estimates in item response theory. Through a Monte Carlo study, the method is evaluated as a way of approximating the standard error and confidence limits for the maximum likelihood estimate of the ability parameter, and compared to the use of the theoretical standard error and confidence limits developed by Lord. At least for short tests, the bootstrap method yielded better estimates than the corresponding theoretical values.

Key words: ability estimates, item response theory, maximum likelihood, statistical accuracy, the bootstrap method.

Introduction

This paper addresses the use of the bootstrap method for assessing the accuracy of ability estimates in item response models, with an emphasis on estimating standard error and confidence limits. In earlier work, Lord (1983) derived theoretical formulas for estimating the standard error of ability estimates assuming items were fixed and their parameters known. Because the bootstrap method requires resampling of item responses from some estimated sampling populations, a distinction should be made between sampling responses from n fixed items and from a random sample of items. In this paper, the distinction between item selection procedures and its effect on the resampling techniques will be developed in detail, followed by examples of applying the resampling techniques to estimating standard error and confidence limits for maximum likelihood (ML) estimates of ability parameters assuming all item parameters are known. Through a Monte Carlo study, the results obtained from the bootstrap approach are compared to the use of the approximate formulas developed by Lord.

The Bootstrap Method

We begin by considering the bootstrap method within a framework of simple random samples, and discuss a generalization to nonidentically distributed samples in the next section. Explicitly, let $\mathbf{x}$ be a random vector containing n observations from an unknown probability distribution F,

$$\mathbf{x}' \equiv (x_1, x_2, \ldots, x_n) \sim F,$$

This paper was originally presented at the Annual Meeting of American Educational Research Association, Washington, D.C., April 1987. The authors are indebted to Patricia Busk, the editor, and two anonymous referees for helpful comments on the earlier drafts of the manuscript.

Requests for reprints should be sent to Michelle Liou, Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, R.O.C.

0033-3123/91/88036 $00.75/0

and let $\hat{\sigma}$ be a statistic based on **x** that estimates the true value of σ in the population. Our main concern is with the sampling error of $\hat{\sigma}$ as an estimate of σ, and the construction of approximate confidence limits around the unknown σ. As notation, we let $\hat{F}$ denote an estimate of F, $\mathbf{x}^*$ a bootstrap random sample drawn with replacement from $\hat{F}$, and $\hat{\sigma}^*$ a bootstrap random variable calculated on the basis of $\mathbf{x}^*$. If $\hat{F}$ is a close approximation to F, the sampling distribution of $\hat{\sigma}$ can be estimated through the repeated random sampling of $\mathbf{x}^*$ and the recalculation of $\hat{\sigma}^*$. Thus, the Monte Carlo algorithm leading to bootstrap estimates proceeds in three steps: (i) estimation of the probability model F from the observed data **x**, producing $\hat{F}$; (ii) construction of bootstrap random samples $\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_B^*$ from $\hat{F}$ according to the sampling rule by which the empirical **x** is drawn from F; and (iii) assessment of the accuracy of $\hat{\sigma}$ based on the observations $\hat{\sigma}^*$.

The crucial step for the bootstrap method obviously lies in an estimation of the unknown F (Chao, 1984; Efron & Tibshiran, 1986). The mechanism originally proposed by Efron (1979) is very general and merely uses the empirical distribution of the data as an estimate of F, that is,

$$\hat{F}\text{: probability mass } \frac{1}{n} \text{ on } x_1, x_2, \ldots, x_n. \tag{1}$$

It can be shown that $\hat{F}$ is a nonparametric ML estimate of F, in the sense that if a random sample of size n is drawn from $\hat{F}$ with replacement, the likelihood of observing a sample vector that will exactly match the empirical **x** is at a maximum. If a more specific model for F is assumed, other estimation strategies may be more appropriate. For instance, if F is assumed normal, it can be estimated with a normal distribution having the same mean and variance as the sample data (Efron & Tibshiran).

After the estimate $\hat{F}$ is available, the B bootstrap random samples can be drawn, and the bootstrap random variables $\hat{\sigma}^*$ constructed. A bootstrap estimate of standard error is simply the sample standard deviation of the $\hat{\sigma}^*$,

$$\text{SE}\,(\hat{\sigma}^*) = \left\{ \frac{\sum_{j=1}^{B} (\hat{\sigma}_j^* - \hat{\sigma}_{(\cdot)}^*)^2}{B-1} \right\}^{1/2}, \tag{2}$$

where

$$\hat{\sigma}_{(\cdot)}^* = \frac{\sum \hat{\sigma}_j^*}{B}.$$

A confidence interval for σ can be obtained through the percentile method (Efron, 1981a, 1982). Explicitly, if

$$\text{Cdf}\,(\sigma_0) \equiv \text{Prob}\,(\hat{\sigma}^* \leq \sigma_0)$$

is the cumulative distribution function of the bootstrap distribution of $\hat{\sigma}^*$, the $1 - \alpha$ confidence limits (σ_1, σ_2) are

$$\sigma_1 = \text{Cdf}^{-1}\,(1/2\alpha) \quad \text{and} \quad \sigma_2 = \text{Cdf}^{-1}\,(1 - 1/2\alpha), \tag{3}$$

or equivalently, the $100 \cdot 1/2\alpha$ and $100 \cdot (1 - 1/2\alpha)$ percentile points in the bootstrap distribution. The bootstrap sample size B is arbitrary, but we do note that values as small as 250 have been shown to generate reasonably accurate confidence intervals (assuming $\hat{\sigma}$ is approximately normally distributed; see Efron, 1984), and even smaller values may be sufficient for obtaining good estimates of the standard error. The reader is referred to Efron (1984, section 8) for a detailed discussion of the choice of bootstrap sample size.

The Statistical Accuracy of Ability Estimates

The 3-parameter (3P) logistic model posits that the distribution of the random observation u_i, representing a zero or one response to item i, has the form:

$$P_i(u_i = 1|\theta) = c_i + \frac{1 - c_i}{1 + \exp\{-1.7a_i(\theta - b_i)\}} \tag{4}$$

(Lord, 1980) where θ denotes the true ability value, and a_i, b_i, and c_i are known item parameters. Given n items, the log likelihood (L_0) for the response vector can be written as

$$L_0 \equiv \log \prod_{i=1}^{n} P_i^{u_i} Q_i^{1-u_i},$$

where $Q_i \equiv 1 - P_i$. The ML estimate, $\hat{\theta}$, is obtained by maximizing L_0 and typically by solving for θ through some iterative method on the basis of the likelihood equation:

$$L_1 \equiv \frac{\partial L_0}{\partial \theta} = \sum_{i=1}^{n} \frac{(u_i - P_i)P_i'}{P_i Q_i} = 0 \tag{5}$$

(Lord, 1983) where P_i' is the first derivative of P_i with respect to θ.

Theoretical Analysis of $\hat{\theta}$

Under regularity conditions and in a neighborhood of θ, the likelihood equation can be expanded into a Taylor series:

$$0 = L_1 + (\hat{\theta} - \theta)L_2 + \frac{(\hat{\theta} - \theta)^2}{2} L_3 + \frac{(\hat{\theta} - \theta)^3}{6} L_4 + \cdots,$$

where by definition

$$L_k = \sum_{i=1}^{n} \frac{\partial^k}{\partial \theta^k} \log P_i^{u_i} Q_i^{1-u_i}|_\theta (k = 1, 2, \ldots),$$

and θ in the equation is evaluated at its true value. For simplification, the symbol y is used to denote $(\hat{\theta} - \theta)$, and the Taylor series rewritten as

$$\begin{aligned} -L_1 &= yL_2 + \frac{y^2}{2} L_3 + \frac{y^3}{6} L_4 + \cdots \\ &= y\left\{L_2 + \frac{y}{2} L_3 + \frac{y^2}{6} L_4 + \cdots\right\} \\ &\equiv y\{\varphi(y)\}, \end{aligned}$$

where $\varphi(y)$ denotes a series in powers of y. By Lagrange's formula (Miller, 1960), the series can be inverted as follows:

$$\begin{aligned} y &= -L_1\left\{\frac{1}{\varphi(y)}\right\} \\ &= -L_1\left\{\frac{1}{\varphi(0)}\right\} + \sum_{k=2}^{n} (-L_1)^k \frac{\partial^{(k-1)}}{\partial y^{(k-1)}}\left\{\frac{1}{\varphi(y)}\right\}^k\Bigg|_{y=0} + \cdots \\ &= \frac{-L_1}{L_2} - \frac{L_1^2 L_3}{2L_2^3} - \frac{L_1^3(3L_3^2 - L_2 L_4)}{6L_2^5} + \cdots \end{aligned} \tag{6}$$

Following Lord (1983), we define

$$s_k \equiv \frac{1}{n} E(L_k) \quad \text{and} \quad \gamma_k \equiv \frac{1}{n}[L_k - E(L_k)],$$

and note that

$$\left(\frac{1}{n} L_2\right)^{-k} = (s_2 + \gamma_2)^{-k} = \left\{s_2\left(1 + \frac{\gamma_2}{s_2}\right)\right\}^{-k}.$$

According to the 3P logistic model, $E(u_i) = P_i$, and therefore, $s_1 = 0$. If y is replaced by $(\hat{\theta} - \theta)$, the inverted function in (6) becomes

$$\begin{aligned} (\hat{\theta} - \theta) = -\gamma_1\left(\frac{1}{s_2} - \frac{\gamma_2}{s_2^2} + \frac{\gamma_2^2}{s_2^3} - \cdots\right) - \frac{\gamma_1^2(s_3 + \gamma_3)}{2}\left(\frac{1}{s_2} - \frac{\gamma_2}{s_2^2} + \cdots\right)^3 \\ - \frac{\gamma_1^3\{3(s_3 + \gamma_3)^2 - (s_2 + \gamma_2)(s_4 + \gamma_4)\}}{6}\left(\frac{1}{s_2} - \cdots\right)^5 + \cdots \end{aligned} \tag{7}$$

If item responses are independent of each other and the P_i are all bounded away from 0 and 1, γ_k is simply the mean of n independent and finite random variables, and thus, by the general version of the Central Limit Theorem, $(n)^{1/2}\gamma_k$ is asymptotically normally distributed with zero mean and finite variance. Therefore, $E(\gamma_k)$ is of order $n^{-1/2}$.

The expansion in (7) leads to the following asymptotic expectations:

$$E(\hat{\theta} - \theta) = \frac{E(\gamma_1\gamma_2)}{s_2^2} - \frac{E(\gamma_1^2)s_3}{2s_2^3} + o\left(\frac{1}{n}\right),$$

and

$$E(\hat{\theta} - \theta)^2 = \frac{E(\gamma_1^2)}{s_2^2} + o\left(\frac{1}{n}\right).$$

Lord (1983) derived the same asymptotic expectations with the procedure called "adjusted order of magnitude" (Shenton & Bowman, 1977) which relies on a normal approximation of $\hat{\theta}$. With the local independence assumption, the first-order error variance can be reduced further to

$$E(\hat{\theta} - \theta)^2 \equiv \frac{1}{\tau^2(\theta)} + o\left(\frac{1}{n}\right) = \left(\sum_i \frac{P_i'^2}{P_iQ_i}\right)^{-1} + o\left(\frac{1}{n}\right). \tag{8}$$

In practice, $\hat{\theta}$ may be substituted for the unknown θ in (8) to estimate the standard error. We do note, however, that these asymptotic expectations are valid only if $\hat{\theta}$ is in a neighborhood of θ (e.g., when the bias in $\hat{\theta}$ is small).

The likelihood equation in (5) is the sum of the independent random variables, $\Gamma_i = (u_i - P_i)P_i'/P_iQ_i$ $(i = 1, \ldots, n)$. By a more general version of the Central Limit Theorem, L_1 is approximately normal with mean zero and finite variance, $\tau^2(\theta) = E(\Sigma_i \Gamma_i^2) = \Sigma_i (P_i'^2/P_iQ_i)$. It follows from the properties of the normal distribution that an asymptotic confidence interval with confidence coefficient $1 - \alpha$ can be obtained for L_1:

$$\text{Prob}\,\{z^{(1/2\alpha)}\tau(\theta) \leq L_1 \leq -z^{(1/2\alpha)}\tau(\theta)\} = 1 - \alpha,$$

where $z^{(1/2\alpha)}$ is the $100 \cdot 1/2\alpha$ percentile point of a standard normal variate. Assuming that the function L_1 is one-to-one (which is always true for the Rasch and the two-parameter logistic models) and that the sample size n is sufficiently large, the statement that L_1 lies in the interval $\pm z^{(1/2\alpha)}\tau(\theta)$ is equivalent to θ itself lying in the interval (θ_1, θ_2), where θ_1 and θ_2 are the roots of the equations $L_1 = \pm z^{(1/2\alpha)}\tau(\theta)$ (Lloyd, 1984). Given a response vector $\mathbf{u} = (u_1, \ldots, u_n)$, for instance, the 95% confidence limits satisfy

$$\sum_i^n \frac{(u_i - P_i)P_i'}{P_iQ_i} = \pm 1.96\left(\sum_i^n \frac{P_i'^2}{P_iQ_i}\right)^{1/2}. \tag{9}$$

Alternatively, if we are willing to make an assumption that $\hat{\theta}$ is asymptotically normally distributed (Bradley & Gart, 1962; Lord, 1983) with mean zero and variance $1/\tau^2(\theta)$, the approximate 95% confidence limits (Lloyd, 1984, p. 303) for θ are simply

$$\hat{\theta} \pm 2\{\tau(\hat{\theta})\}^{-1}. \tag{10}$$

The computation of standard error and confidence limits for $\hat{\theta}$ can be illustrated by considering a sample test with twenty items whose a, b, and c parameters have average values of 0.976, −0.197, and 0.137, respectively. With the response vector (11111011101011101101), for instance, the three functions, L_1, $+1.96\tau(\theta)$, and $-1.96\tau(\theta)$ in (9) are plotted in Figure 1 for different θ values. In this Figure, the function L_1 is plotted with the solid line and has a zero value at $\theta = 1.62$, which is also the ML estimate of θ. The function L_1 intersects the dashed curves for the two functions $\pm 1.96\tau(\theta)$ at the points (0.36, 2.61), which are approximately the 95% confidence limits for θ. The first-order standard error has an estimate 0.47, computed by substituting $\hat{\theta}$ into (8). If we rely on the normal approximation of $\hat{\theta}$, the central 95% confidence derived from (10) yields an interval (0.68, 2.57).

Bootstrap Analysis of $\hat{\theta}$

If the true values of P_i are all known, we can resample each bootstrap response from the corresponding P_i for $i = 1, \ldots, n$. For instance, an item score u_i^* is simulated by first generating a random number from a uniform distribution, and then assigning a score 1 to u_i^* if the random number is less than or equal to P_i, and a score 0 otherwise. After observing all the u_i^*, we then evaluate $\hat{\theta}^*$. This resampling procedure is designed primarily to simulate an examinee's random responses to the fixed n items. The resampling process is repeated until B bootstrap random variables, $\hat{\theta}_j^*$ $(j = 1, \ldots, B)$, are

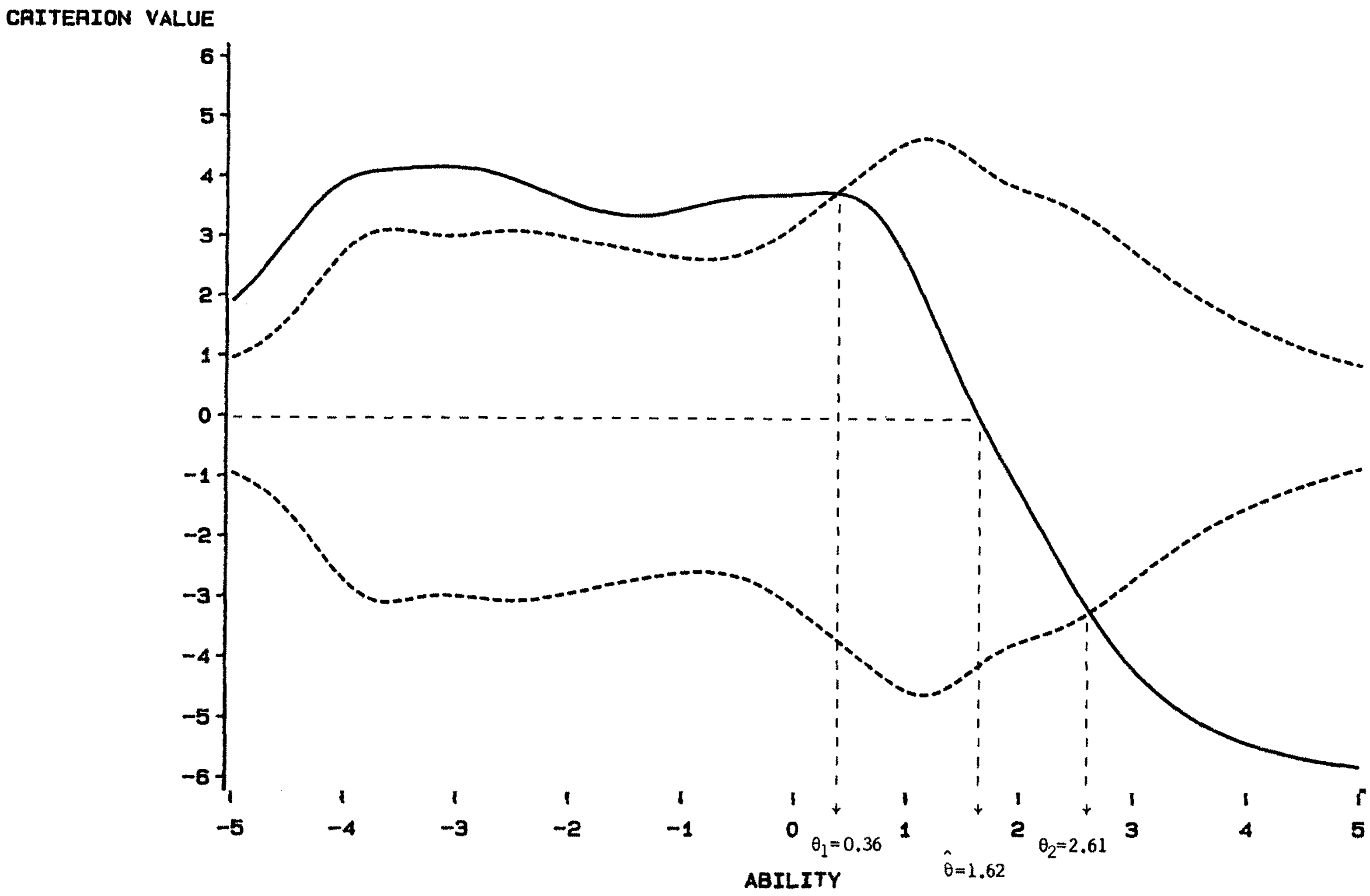

FIGURE 1

Plots of the three functions L_1 and $\pm 1.96\tau(\theta)$ for a sample 20-item test. The function L_1 is plotted with solid line, and the two functions $\pm 1.96\tau(\theta)$ plotted with the dashed lines.

observed, and the bootstrap estimates of standard error and confidence limits can then be calculated based on (2) and (3), respectively. The P_i, however, can never be known exactly and must be estimated from the empirical data. In this study, we suggest substituting the ML estimate, $\hat{\theta}$, for θ in (4) to estimate P_i. The obtained estimates, $\hat{P}_i$, have the desirable property that, if we repeat the random sampling of $\mathbf{u}^*$ from the $\hat{P}_i$ ($i = 1, \ldots, n$), the likelihood of observing a bootstrap response vector identical to the empirical $\mathbf{u}$ is at a maximum.

An obvious question arises regarding the alternative mechanism specified in (1) for obtaining the bootstrap random samples. In other words, we can place the probability mass $1/n$ on $u_1, u_2, \ldots, u_n$ and resample n item scores along with the corresponding item parameters from the empirical distribution to evaluate the bootstrap random variable $\hat{\theta}^*$. In fact, the sources of sampling error involved in $\hat{\theta}$ can be viewed from two independent perspectives—an examinee's random responses on the fixed n items and random selection of the n items from a population of items. In the literature, the standard error of $\hat{\theta}$ is typically considered for n fixed items (Lord, 1983), and we have implicitly assumed so. If we apply the random mechanism specified in (1) to u_i, the random error involved in $\hat{\theta}^*$ will primarily reflect the sampling of items. If the researcher wishes to incorporate variability in $\hat{\theta}$ due to the random choice of n items, we propose the use of two-stage sampling to obtain an estimate of the standard error of $\hat{\theta}$. Given that the n items are randomly selected from the population of items, we take advantage of the random mechanism proposed by Efron (1979) to approximate the unknown sampling population $\mathcal{F}$ with the nonparametric ML estimate, $\hat{\mathcal{F}}$, that is,

$$\hat{\mathcal{F}}\text{: probability mass } \frac{1}{n} \text{ on } \hat{P}_1, \hat{P}_2, \ldots, \hat{P}_n,$$

where the $\hat{P}_i$ are estimates of the P_i. First of all, we randomly select a bootstrap sample, $\hat{P}_i^*$, of size n from $\hat{\mathcal{F}}$ with replacement (the i-th item is selected, for instance, if the integer of $Rn + 1$ is equal to i, where R is a random number from a uniform distribution). This procedure is designed to simulate the sampling of n items from a population of items. Second, we resample u_i^* from the corresponding $\hat{P}_i^*$, for $i = 1, \ldots, n$, to simulate an examinee's random responses on the n random items. Based on $\mathbf{u}^*$, $\hat{\theta}^*$ is evaluated, and the two-stage sampling is repeated until B bootstrap random variables are observed. The bootstrap standard error is then computed with (2).

We consider the sample items used previously as an example, and substitute $\hat{\theta} = 1.62$ along with the item parameters into (4). A vector of $\hat{P}_i$ can be obtained as (.998, .997, .827, .852, .995, .388, .999, .711, .999, .779, .999, .419, .830, .985, .935, .432, .525, .282, .807, .960). To simulate the bootstrap item responses, uniformly distributed random numbers are generated and compared with the $\hat{P}_i$ to assign the u_i^* (the random numbers were generated by IMSL subroutine RNUN; International Mathematical and Statistical Libraries, 1987). A histogram of 1,000 bootstrap $\hat{\theta}^*$ values is shown in Figure 2 by a solid line. The 1,000 bootstrap random variables have a sample standard deviation of 0.50, which is an estimate of the standard error of $\hat{\theta}$. The 2.5th and 97.5th percentile points in the distribution are, respectively, 0.84 and 2.78, which are also the 95% confidence limits for θ. In Figure 2, the dashed histogram is the distribution of 1,000 bootstrap $\hat{\theta}^*$ values based on the method of two-stage sampling. The $\hat{\theta}^*$ values from the two-stage sampling yield a sample standard deviation 0.52 that is virtually identical to 0.50. Therefore, random selection of 20 items from the population of items seems to have little effect on the estimation of the underlying true ability.

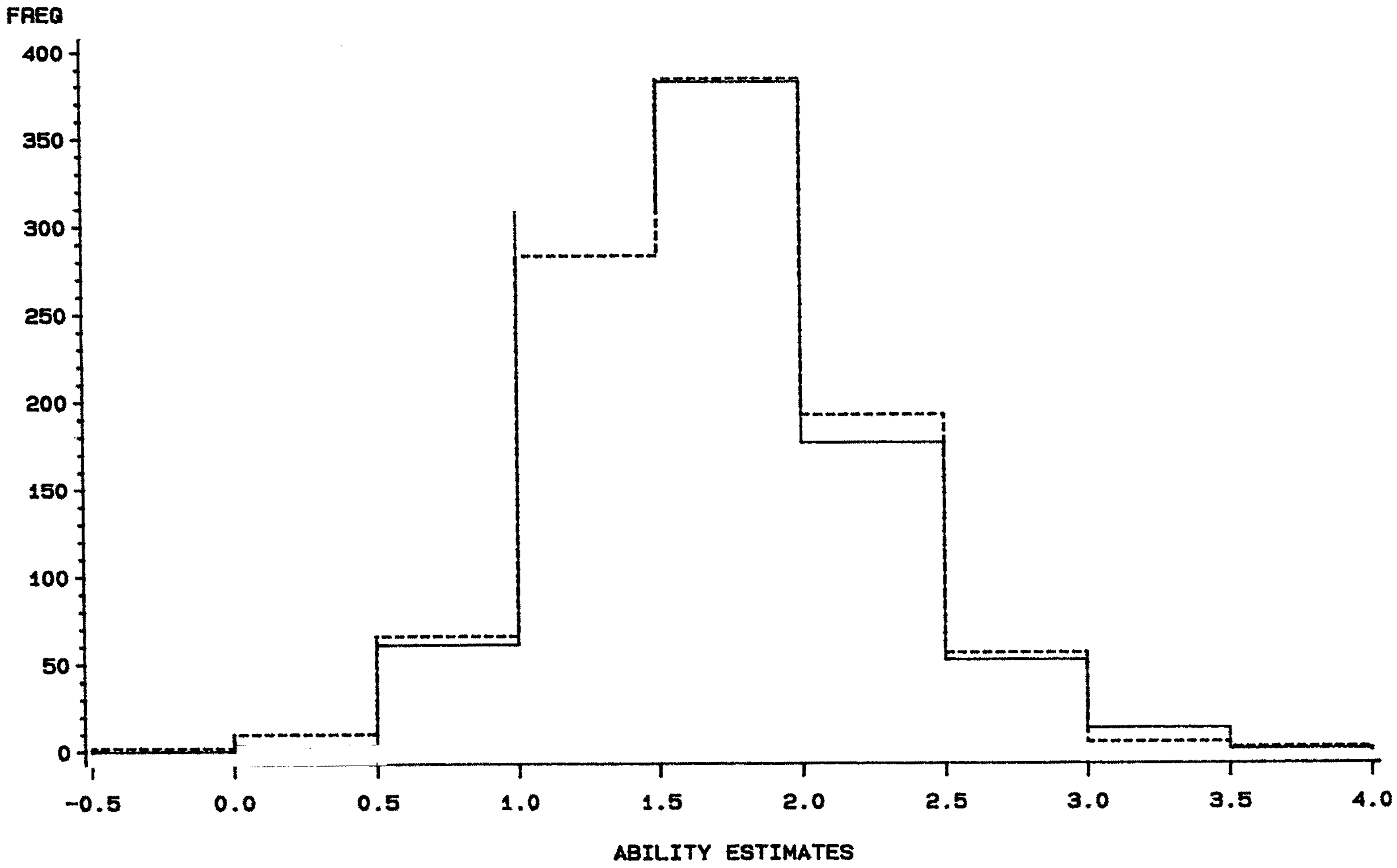

FIGURE 2

Histograms of the bootstrap $\hat{\theta}_1^*, \ldots, \hat{\theta}_{1000}^*$ for the sample 20-item test. The solid histogram is the distribution of 1,000 bootstrap $\hat{\theta}^*$ values, and the dashed histogram is the distribution of another set of 1,000 bootstrap $\hat{\theta}^*$ values computed on the basis of two-stage sampling.

TABLE 1

The Averages and Standard Errors of ML Ability Estimates (Trials = 500)

	20–Item		40–Item		60–Item	
θ	Ave($\hat{\theta}$)	SE($\hat{\theta}$)	Ave($\hat{\theta}$)	SE($\hat{\theta}$)	Ave($\hat{\theta}$)	SE($\hat{\theta}$)
2.50	2.464	0.448	2.626	0.476	2.583	0.433
2.00	2.039	0.467	2.046	0.367	2.026	0.286
1.50	1.492	0.426	1.515	0.308	1.518	0.255
1.00	1.016	0.447	1.018	0.324	0.995	0.256
0.50	0.525	0.534	0.506	0.320	0.511	0.267
0.00	-0.056	0.622	-0.005	0.338	-0.020	0.266
-0.50	-0.519	0.611	-0.562	0.426	-0.534	0.344
-1.00	-0.988	0.740	-1.026	0.508	-1.052	0.414
-1.50	-1.523	1.280	-1.520	0.622	-1.503	0.460
-2.00	-2.282	2.475	-2.011	0.725	-2.022	0.506
-2.50	-3.044	3.209	-2.629	1.783	-2.530	0.610

Simulation Study

To evaluate the bootstrap estimates of standard error and confidence limits, a simulation study was conducted by assuming the 3P logistic model. Three hypothetical tests containing 20, 40, and 60 items were designed to approximate the item parameters in the College Entrance Examination Board's Scholastic Aptitude Test, Verbal Section (see Lord, 1983, p. 238). The 20- and 40-item tests comprise items 1–20 and 1–40, respectively, of the 60-item test. For a given θ, each item response in **u** was generated from the corresponding P_i ($i = 1, \ldots, n$). An item score u_i was simulated by first generating a random number from a uniform distribution with the IMSL subroutine RNUN, and then assigning a score 1 to u_i if the random number was less than or equal to P_i, and a score 0 otherwise. The ML estimate $\hat{\theta}$ was found by solving (5) with Newton's method, with a convergence criterion of 0.1 percent change in $\hat{\theta}$. A simulated perfect score or zero score was eliminated from the evaluation, resulting in a truncation of the sampling distribution of $\hat{\theta}$. The averages (Ave) and standard errors (SE) of $\hat{\theta}$ over 500 simulation trials for selected ability levels are contained in Table 1. Because the simulated items provide little information at the lower end of the ability scale, the standard errors of ML ability estimates tend to be large for the low ability levels especially for the 20-item test.

By assuming the same item parameters, item responses were independently simulated, and θ estimated by the ML procedure. The observed first-order error variance was calculated by substituting $\hat{\theta}$ for θ in (8). One computational problem with the first-order error variance involves $\tau^2(\hat{\theta})$ in the denominator of (8), which cannot be close to zero, because otherwise, the standard error estimate would become extremely large. The simulation restricted the ability estimates within the range of (−5.0, 5.0), resulting in a further truncation of the sampling distribution of $\hat{\theta}$ especially for the

TABLE 2

The Theoretical Standard Errors and Empirical Sizes of the Theoretical Confidence Limits for ML Ability Estimates (Trials = 500)

	20–Item		40–Item		60–Item	
θ	$SE_1(\hat{\theta})$	%	$SE_1(\hat{\theta})$	%	$SE_1(\hat{\theta})$	%
2.50	0.602	0.94(0.98)	0.497	0.97(0.99)	0.404	0.96(0.97)
2.00	0.481	0.96(0.98)	0.359	0.94(0.97)	0.287	0.95(0.97)
1.50	0.440	0.95(0.96)	0.302	0.94(0.96)	0.257	0.95(0.95)
1.00	0.452	0.95(0.96)	0.304	0.95(0.97)	0.254	0.95(0.95)
0.50	0.498	0.94(0.96)	0.315	0.95(0.95)	0.258	0.93(0.94)
0.00	0.586	0.94(0.94)	0.340	0.95(0.97)	0.278	0.95(0.95)
-0.50	0.654	0.96(0.95)	0.420	0.95(0.95)	0.339	0.94(0.94)
-1.00	0.721	0.96(0.92)	0.513	0.95(0.93)	0.412	0.95(0.94)
-1.50	0.815	0.95(0.95)	0.570	0.95(0.92)	0.467	0.95(0.94)
-2.00	0.914	0.95(0.94)	0.662	0.95(0.95)	0.510	0.95(0.96)
-2.50	1.070	0.96(0.95)	0.830	0.95(0.97)	0.568	0.96(0.97)

Note. The values in parentheses are the empirical sizes of the confidence intervals computed from (10).

simulated ability at the lower end. The average of the first-order error variances over 500 simulation trials was computed. The symbol $SE_1(\hat{\theta})$ in Table 2 denotes the square root of the average value. In each trial, the 95% confidence limits were found by both (9) and (10), and the solutions for θ_1 and θ_2 in (9) were also found by Newton's method. Because the two solutions might not be unique, the numerical program selected the two solutions that were right above and below $\hat{\theta}$ as the confidence limits. Given a true value of θ, the proportion of trials in which the range of the confidence limits contains θ is reported in Table 2. For convenience, the value of the proportion is referred to as the empirical size of the confidence interval, and 95% as its nominal size.

In general, the first-order standard error is quite close to the actual standard error in Table 1 for the 60-item test. For shorter tests of 20 or 40 items, the theoretical estimator tends to underestimate the true sampling error for the lower ability levels. The empirical size of the confidence interval computed from (9) is quite close to its nominal size except for a few ability levels at the upper tail. The empirical size of the confidence interval calculated on the basis of normal approximation of $\hat{\theta}$, on the other hand, tends to be larger than its nominal size.

From the known P_i $(i = 1, \ldots, n)$, an empirical response vector was independently and randomly generated for evaluating the bootstrap estimators. The $\hat{\theta}$ value was also restricted within the range of $(-5.0, 5.0)$. The $\hat{P}_i$ were computed by substituting $\hat{\theta}$ for θ in (2), and the bootstrap sample $\mathbf{u}^*$ generated from the $\hat{P}_i$ for calculating the $\hat{\theta}_j^*$. For a given $\hat{\theta}$, the bootstrap sampling was repeated 1,000 times ($B = 1{,}000$). Based on the 1,000 bootstrap samples, the bootstrap error variance and 95% confidence limits were

TABLE 3

The Bootstrap Standard Errors and Empirical Sizes of the Percentile Confidence Limits for ML Ability Estimates (Trials = 500)

	20–Item		40–Item		60–Item	
θ	$SE(\hat{\theta}^*)$	%	$SE(\hat{\theta}^*)$	%	$SE(\hat{\theta}^*)$	%
2.50	0.436(0.461)	0.98	0.443(0.455)	0.92	0.421(0.425)	0.92
2.00	0.455(0.476)	0.95	0.383(0.395)	0.94	0.318(0.331)	0.95
1.50	0.458(0.485)	0.95	0.323(0.339)	0.94	0.267(0.274)	0.95
1.00	0.486(0.512)	0.94	0.310(0.324)	0.95	0.257(0.263)	0.94
0.50	0.535(0.700)	0.94	0.326(0.335)	0.95	0.263(0.268)	0.94
0.00	0.626(0.954)	0.95	0.366(0.390)	0.95	0.292(0.296)	0.94
-0.50	0.902(1.555)	0.95	0.466(0.548)	0.94	0.352(0.358)	0.95
-1.00	1.365(2.243)	0.94	0.577(0.917)	0.96	0.418(0.443)	0.96
-1.50	1.932(3.335)	0.94	0.935(1.480)	0.96	0.526(0.660)	0.95
-2.00	3.031(4.377)	0.93	1.630(2.444)	0.95	0.749(1.059)	0.94
-2.50	4.359(5.478)	0.95	2.826(3.719)	0.94	1.399(1.745)	0.95

Note. The values in parentheses are standard errors based on two-stage sampling.

computed with (2) and (3), respectively. The bootstrap scheme—model estimation, random sampling—was repeated with 500 simulation trials, and the average bootstrap error variance and the empirical size of the percentile confidence interval were computed for the 500 simulation trials. The square root of the average variance and the empirical size of the confidence interval are contained in Table 3.

The bootstrap standard error seems to overestimate the true sampling error throughout the simulation conditions. The bootstrap estimator, however, performs collectively better than the theoretical estimator especially for the 20-item test. This finding is consistent with earlier findings of Wainer and Wright (1980) who used another resampling scheme—a modified jackknife to correct guessing distortion in Rasch ability estimates and found that the resampling technique worked well for short tests of 10 or 20 items. The empirical size of the percentile confidence interval is also close to its nominal size except for a few ability levels at the upper tail. The standard errors based on two-stage sampling are listed in Table 3 and suggest that the ability estimates at the lower end are more likely affected by random selection of items when the test contains only 20 items.

Discussion

The validity of the first-order standard error in (8) is guaranteed by the Central Limit Theorem. Both the bootstrap and first-order standard errors rely on rationales independent of the usual assumption that $\hat{\theta}$ is normally distributed (Lord, 1983). If the true θ value is known, the bootstrap method with a sufficiently large number of bootstrap samples will yield a standard error that is exact. Under the same condition, the first-order standard error will still approximate the true standard error due to the truncation of higher degree terms in the Taylor series. In application, if bias in $\hat{\theta}$ is negligible, the $\hat{P}_i$ will yield the same information as its underlying true model, P_i. Therefore, both the bootstrap method and the Taylor series method need the same assumption that $\hat{\theta}$ is in a neighborhood of θ (Efron, 1981b, 1982; Parr, 1983). When this is violated, the simulation study suggests that the bootstrap method yields more accurate results than the theoretical estimator, which is a promising finding for circumstances where test length cannot be increased sensibly. The computation of the first-order standard error is much simpler than the resampling method, and, therefore, its use is still recommended for tests of reasonable length (e.g., 40 items). The simulation results also suggest that the percentile confidence interval computed from 1,000 bootstrap samples is equally useful as the theoretical confidence interval based on (9). Although validity of the theoretical confidence interval is guaranteed by the Central Limit Theorem, the assumption that the function L_1 is one-to-one should be noted, because it can be violated when a test has a short length or item responses are seriously distorted by guessing. It would be useful to plot the three functions L_1 and $\pm 1.96\tau(\theta)$ against θ before deciding on the confidence limits, θ_1 and θ_2.

In the framework so far outlined, individual ability is estimated with predetermined item parameters as in item banking and tailored testing. By knowing item parameters, the theoretical formulas and the bootstrap method suggest two competitive solutions for assessing the statistical accuracy of $\hat{\theta}$. Perhaps a potential use of the bootstrap method lies in the circumstance where the theoretical formulas are rather complicated to derive. With the bootstrap method, for instance, error resulting from random selection of n items for measuring ability can possibly be evaluated by including the item population, $\mathcal{F}$. Another useful example would be in estimating the standard error of $\hat{\theta}$ when item parameters are jointly estimated by the ML procedure. Furthermore, a practitioner might be more interested in the consequence of violation of the logistic model to estimation of ability. In that case, the response process proposed by Tucker, Humphreys, and Roznowski (1986) suggests a more realistic sampling procedure for simulating the bootstrap random sample $\mathbf{u}^*$. With more complicated problems, however, the validity of using the bootstrap method needs to be tested further.

References

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

Chao, M. T. (1984). *Generalized bootstrap methods* (TRB-84-001). Taipei: Institute of Statistical Science, Academia Sinica, R.O.C.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7*, 1–26.

Efron, B. (1981a). Nonparametric standard errors and confidence intervals. *The Canadian Journal of Statistics, 9*, 139–172.

Efron, B. (1981b). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods, *Biometrika, 68*, 589–599.

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. *Society of Industrial and Applied Mathematics CBMS-NSF Monographs, 38*.

Efron, B. (1984). *Better bootstrap confidence intervals* (Tech. Rep. No. 226). Palo Alto, CA: Stanford University, Department of Statistics.

Efron, B., & Tibshiran, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science, 17*, 1–35.

IMSL (1987). *International Mathematical and Statistical Libraries (Stat/Library Vol. I), User's Manual.* Houston, TX: Author.

Lloyd, E. (1984). Maximum likelihood estimates. In E. Lloyd (Ed.), *Handbook of applicable mathematics, Volume VI: Statistics (Part A)* (pp. 283–354). New York: Wiley.

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

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

Miller, K. S. (1960). *An introduction to the calculus of finite differences & difference equations*. New York: Holt.

Parr, W. C. (1983). A note on the jackknife, the bootstrap and the delta method estimators of bias and variance. *Biometrika, 70*, 719–722.

Shenton, L. R., & Bowman, K. O. (1977). *Maximum likelihood estimation in small samples*. New York: Macmillan.

Tucker, L. R., Humphreys, L. G., & Roznowski, M. A. (1986). *Comparative accuracy of five indices of dimensionality of binary items* (Technical Report No. AD-A172110). Springfield, VA: National Technology Information Service.

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

Manuscript received 7/18/88
Final version received 1/22/90