Journal of Educational Measurement
Summer 1992, Vol. 29, No. 2, pp. 133–161

Estimating Population Characteristics From Sparse Matrix Samples of Item Responses

Robert J. Mislevy
Educational Testing Service
Albert E. Beaton
Boston College and Educational Testing Service
Bruce Kaplan and **Kathleen M. Sheehan**
Educational Testing Service

The multiple-matrix item sampling designs that provide information about population characteristics most efficiently administer too few responses to students to estimate their proficiencies individually. Marginal estimation procedures, which estimate population characteristics directly from item responses, must be employed to realize the benefits of such a sampling design. Numerical approximations of the appropriate marginal estimation procedures for a broad variety of analyses can be obtained by constructing, from the results of a comprehensive extensive marginal solution, files of plausible values of student proficiencies. This article develops the concepts behind plausible values in a simplified setting, sketches their use in the National Assessment of Educational Progress (NAEP), and illustrates the approach with data from the Scholastic Aptitude Test (SAT).

Sample surveys typically approximate distributions of survey variables from the values of a sample of units. In a simple random sample with replacement, for example, the sample mean and variance are unbiased estimates of the corresponding population parameters. Complications arise, however, if the variables of interest cannot be observed directly. In classical test theory, for instance, simple measurement error models precipitate corrections to estimate "true" variances and correlations from their "observed" counterparts. An analogous problem in the context of item response theory (IRT) arises in the National Assessment of Educational Progress (NAEP).

It is desirable to use IRT in NAEP to link observations from responses to disparate sets of test items, but each examinee is presented too few items to permit an accurate estimate of his or her proficiency. Population characteristics, such as subpopulation means, proportions of pupils above specified proficiency levels, and relationships between proficiency and various social and educational variables, are estimated on IRT scales directly from survey responses through marginal estimation procedures, circumventing the need for calculating scores for individual students. The computing approximation that affects this approach, Rubin's (1987) multiple imputations, or plausible values

The work on which this article is based was performed for the National Center for Education Statistics, Office of Educational Research and Improvement, by Educational Testing Service. The opinions expressed are those of the authors.

as they are called in NAEP, additionally provides synthetic data sets that secondary researchers can analyze with the standard techniques that would be appropriate if respondents' IRT proficiencies (inherently unobservable variables) could be ascertained without error.

Plausible values are constructed from the results of a comprehensive extensive marginal analysis, with the express property of reproducing the population characteristics implicit in the results of this analysis. This is exactly the opposite of the typical analysis, in which values first obtained independently for each respondent are aggregated to approximate population characteristics. If viewed from the perspective of the typical analysis, plausible values seem to possess some rather paradoxical properties; for example,

(a) Several plausible values are provided for a given student. They differ because one component in their construction is a random number for each. Obviously, it would be possible to obtain a better score for this student. Should not stripping the random elements from plausible values, or using their average, produce a better estimate for each student and lead therefore to better estimates of population characteristics?

(b) Background information, such as gender, is utilized in the construction of plausible values. If girls scored higher than boys, for example, the plausible values associated with a girl who made the same item responses as a boy would tend to be higher than the plausible values associated with him. Doesn't using background information in this way exaggerate the girl–boy difference?

This article develops the concepts behind plausible values from the perspective of marginal analysis, with the goal of resolving these and other paradoxes that arise when they are viewed from the perspective of individual measurement. Much of the presentation is devoted to constructing and using plausible values in a context where they are not needed (simple random samples from normal populations, with a normal measurement error model from classical test theory) in order to provide some intuition for more complex applications. The results here agree with familiar formulas, as found in Gulliksen (1950). We then describe the extensions that are required for the NAEP setting, with its IRT measurement model and complex student sampling and item sampling designs. Finally, we illustrate the procedures with data from the Scholastic Aptitude Test (SAT), in which matrix sampling was simulated from complete response vectors to an 85-item reading test. For additional information, the reader is referred to Rubin (1987) for the theoretical underpinnings of the approach; to Mislevy (1991) for its application in measurement error models; to Mislevy, Johnson, and Muraki (1992) for a general discussion of NAEP plausible values; and to NAEP technical reports (Beaton, 1987, 1988; Johnson & Zwick, 1990) for the details of their implementation in specific assessments.

Background

NAEP's purpose is to report on the competencies of students in the nation's school systems as a whole, rather than in detail about specific individuals. The difference is akin to that of the Census Bureau's interest in the distribution of

income in California and the Internal Revenue Service's interest in the income of each specific individual in that state. A key result of sampling theory is that less information is required for the former purpose than the latter. In NAEP, precise estimates of proficiency distributions can be obtained by surveying only a sample of students and administering only a relatively small sample of test exercises to each of them. A turning point in the development of large-scale educational assessment was the discovery that population characteristics can be estimated accurately without first obtaining accurate estimates for individual students (Lord, 1962; Sirotnik & Wellington, 1977). In fact, Lord (1962) has shown that, for a fixed number of item responses, the most efficient sampling designs for estimating the population average solicit only one response per sampled student!

NAEP typically administers more than one item per student—somewhere between 2 and 40 in a given reporting domain, depending on the nature of the tasks and the specificity of the domains. We shall not go into the details of NAEP item sampling designs here (again, see the NAEP technical reports), but the fact that different students receive different items, and often different numbers of items, in a reporting domain and that the collection of items in one assessment year overlaps only partially with the collection in other years is central to our purposes. The success of IRT in linking results from disparate samples of items in the context of individual measurement suggested its use in the context of assessment as well, as Bock, Mislevy, and Woodson (1982) and Messick, Beaton, and Lord (1983) argue.

The essential idea of IRT is that observed item responses are driven by an unobservable proficiency variable, often denoted θ. We shall limit our discussion to dichotomous items for convenience, although the same ideas apply with IRT models for rating-scale data. Let x_j represent the response to item j, 1 if correct and 0 if incorrect. Under the 3-parameter logistic IRT model used in several NAEP applications, the probability of a correct response to item j, given θ, is modeled as

$$P_j(\theta) = c_j + \frac{1 - c_j}{1 + \exp\left[-1.7a_j(\theta - b_j)\right]}, \tag{1}$$

where b_j reflects the difficulty of item j; a_j, its sensitivity to changes in θ; and c_j, the chances that examinees with very low proficiency will answer it correctly. Under the usual IRT assumption of local or conditional independence, the probability of a vector of responses to n items is the product of terms built up from (1):

$$p(\mathbf{x}|\theta,\boldsymbol{\beta}) = \prod_j P_j(\theta)^{x_j}(1 - P_j(\theta))^{1-x_j}, \tag{2}$$

where $\mathbf{x} = (x_1, \ldots, x_n)$ is a vector of responses to n items, and $\boldsymbol{\beta} = (a_1,b_1,c_1, \ldots, a_n, b_n, c_n)$ is the vector of parameters of all n items. A student's responses to any subset of items can be related to the values of θ that might have produced those responses via the likelihood function $L(\theta|\mathbf{x},\boldsymbol{\beta})$, which takes the same form as (2) but is viewed as a function of θ for fixed $\mathbf{x}$.

This formulation sets the stage for a common reporting scale despite matrix sampling and evolving item pools. In the setting of individual measurement, a student is presented many items (often 50 to 100), and the likelihood function is sharply peaked around the point where it reaches its maximum value—the maximum likelihood estimate (MLE), $\hat{\theta}$. Under these circumstances, the distribution of $\hat{\theta}$ obtained from a sample of examinees might provide a serviceable estimate of θ in the population. As we shall see, however, this is not the case when examinees are administered substantially fewer items, as in NAEP; the distribution of $\hat{\theta}$ does not converge to the distribution of θ, even as the examinee sample size increases without limit (see Lord, 1969). Some serious problems become evident even in the simpler setting of classical test theory, to which we now turn.

Classical Test Theory With a Single Normal Population

Suppose we wish to estimate the mean μ and variance σ^2 of the proficiencies θ in a large population known to follow a normal distribution from a simple random sample of size N. Rather than observing θ directly from sampled students, however, we observe a noisy version, $x = \theta + e$, where e is a random $N(0,1)$ variate. We shall see that, in this context, optimal point estimates of individual θs lead to correct estimates of μ but not σ^2.

The Complete Data Solution

We begin by reviewing the maximum likelihood solution for μ and σ^2 that would obtain if θ values could be observed without error—the complete data solution, in Dempster, Laird, and Rubin's (1977) terminology. A sample of N values of θ is observed, which we denote $\boldsymbol{\Theta} = (\theta_1, \ldots, \theta_N)$. The complete data likelihood function is

$$L(\mu,\sigma^2|\boldsymbol{\Theta}) = \prod_{i=1}^{N} p(\theta_i|\mu,\sigma^2) = (2\pi\sigma^2)^{-N/2} \prod_i \exp\left(-\frac{(\theta_i - \mu)^2}{2\sigma^2}\right), \qquad (3)$$

and the MLEs $\hat{\mu}$ and $\widehat{\sigma^2}$ are the values that maximize (3). It is equivalent, and more convenient, to maximize the logarithm of the likelihood function, or

$$l(\mu,\sigma^2|\boldsymbol{\Theta}) = \sum_{i=1}^{N} \ln p(\theta_i|\mu,\sigma^2). \qquad (4)$$

This is accomplished by taking its partial derivatives with respect to μ and σ^2 and finding the values that make them zero. This leads to the familiar expression

$$\mu = N^{-1} \sum_i \theta_i \qquad (5)$$

and

$$\sigma^2 = N^{-1} \sum_i \theta_i^2 - \mu^2. \qquad (6)$$

Evaluating the right sides of (5) and (6) with observed values of θ yields $\hat{\mu}$ and $\widehat{\sigma^2}$.

Estimating Individuals' Proficiencies

But, in this example, individuals' θs are not observed directly. The usual problem in educational measurement is to get a good estimate of each student's θ for making inferences about that specific individual. This section reviews two common methods of estimating individuals' θs and shows what happens when the distributions of those estimates are taken as an approximation of the underlying θ distribution in a population of students.

Maximum likelihood estimates. We have assumed a standard normal measurement error model, so that the probability density function of x for any given student—namely, $p(x|\theta)$—is $N(0,1)$, a unit normal distribution centered around the student's true proficiency. Observing a student's response x induces the likelihood function for his or her θ:

$$L(\theta|x) \propto \exp\left[\frac{-(x-\theta)^2}{2}\right], \tag{7}$$

a unit normal density centered at x. Figure 1 illustrates likelihood functions for a fictitious sample of 3 examinees from a $N(0, 1)$ population from whom the responses -1.51, $-.38$, and 1.89 are observed. (These values were selected to make sample estimates of population parameters equal to their true values, thereby providing clear illustrations of the paradoxes.) Because the likelihoods are normal and they are centered at the observed responses, a given examinee's x is in fact his or her MLE, $\hat{\theta}$. For each examinee, x is an unbiased estimate of θ. Its standard error of estimation is the variance of the likelihood function, (7), which takes the value 1 in this example.

The distribution of $\hat{\theta}$ in the population of examinees is simply the distribution of x. Because the observational errors e are assumed to be independent and have mean 0, the expected mean of x is the same as the mean of θ. With a large enough sample, then, the sample mean of the MLEs therefore converges to the

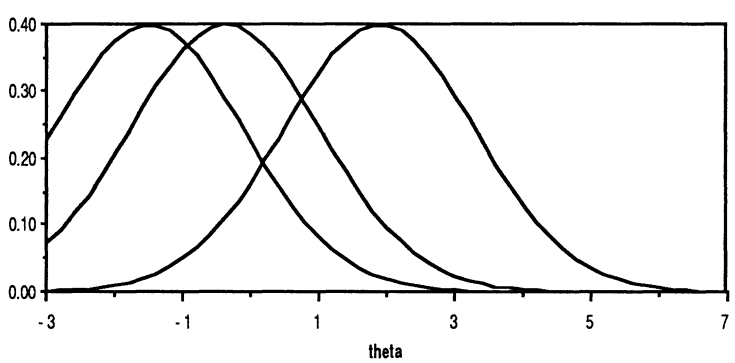

FIGURE 1. *Likelihood functions for three fictitious examinees*

correct population mean of θ, or μ. By construction, the mean of x, zero, is equal to μ in this example. The variance of $\hat{\theta}$ is not a consistent estimate of the variance of θ, however, because $\mathrm{Var}(\hat{\theta}) = \mathrm{Var}(x) = \mathrm{Var}(\theta) + \mathrm{Var}(e) = \sigma^2 + 1$. Therefore, approximating the variance of θ by the sample variance of $\hat{\theta}$ overestimates the correct value, no matter how many examinees the estimate is based on. Even though the $\hat{\theta}$s provide consistent estimates for each examinee individually, the sample variance of $\hat{\theta}$s is an inconsistent estimate of the population variance of θ.

How serious is the inconsistency? The ratio of the asymptotic estimate $\mathrm{Var}(\hat{\theta})$ to the true value $\mathrm{Var}(\theta)$ is the reciprocal of the classical test theory index of reliability, or the ratio of true-score variance to observed score variance:

$$\rho = \frac{\mathrm{Var}(\theta)}{\mathrm{Var}(\hat{\theta})} = \frac{\mathrm{Var}(\theta)}{\mathrm{Var}(\theta) + \mathrm{Var}(x|\theta)},$$

where the conditional variance of observed scores given true scores, or $\mathrm{Var}(x|\theta)$, is the variance of the measurement error e. Equal true-score and measurement error variance, as would occur in our example if $\sigma^2 = 1$, would yield a reliability coefficient of .50, and $\mathrm{Var}(\hat{\theta})$ would overestimate $\mathrm{Var}(\theta)$ by 100%. This is a rather low value for ρ in the context of educational testing—one which might be obtained from a 15-item quiz. Longer tests have higher reliabilities: perhaps .80 for a 30-item achievement test or .90 for an 85-item test like the SAT. $\mathrm{Var}(\hat{\theta})$ would overestimate $\mathrm{Var}(\theta)$ in these latter instances by 25% and 11%, respectively. It bears emphasis that this 11% overestimate holds for the SAT, even though it is 85 items long, regardless of how many examinees the estimate is based on.

This problem is less serious when the inferences to be made concern general comparisons of subpopulations, such as which has a higher mean or a larger standard deviation, as long as the same test is used for all groups being compared. If variance at two time points is always overestimated by about 11%, one can gauge whether the variance has increased or decreased. But, if samples are tested with a 30-item test at Time 1 and an 80-item test at Time 2, the variance of the maximum likelihood estimates of examinees will decrease by 11% when variance of the latent variables remains identical.

Bayesian estimates. While maximum likelihood estimation of an examinee's θ involves only that individual's data, Bayesian estimation additionally takes into account the distribution in the population to which the examinee belongs. In particular, the posterior distribution of a student with observed score x is proportional to the product of the likelihood induced by x through the response model (7) and the population density:

$$p(\theta|x) \propto L(\theta|x)p(\theta).$$

When both the likelihood function and the prior distribution are normal density functions, the posterior density is normal too. To express this result in general terms, denote the mean and variance of the likelihood function by μ_L

and σ_L^2, which take the values x and 1 in our example. The mean and variance of the posterior are

$$\mu_{\text{post}} = \frac{\sigma_L^{-2}\,\mu_L + \sigma^{-2}\mu}{\sigma_L^{-2} + \sigma^{-2}} \tag{8}$$

and

$$\sigma^2_{\text{post}} = [\sigma_L^{-2} + \sigma^{-2}]^{-1}. \tag{9}$$

The inverse of the variance of a distribution is sometimes referred to as *precision.* Thus, when both the prior and the likelihood are normal, the mean of the posterior is the precision-weighted average of the means of the prior and the likelihood. The precision of the posterior is the sum of the precisions of the prior and the likelihood. Using the algebraic result associated with test reliability

$$\frac{\sigma_L^{-2}}{\sigma_L^{-2} + \sigma^{-2}} = \frac{\sigma^2}{\sigma_L^2 + \sigma^2} \equiv \rho,$$

we obtain from (8) and (9) Kelley's formulas from classical test theory:

$$\bar{\theta} \equiv \mu_{\text{post}} = \rho x + (1 - \rho)\mu \tag{10}$$

and

$$\sigma^2_{\text{post}} = \rho\sigma^2, \tag{11}$$

so that $p(\theta|x)$ is $N[\rho x + (1 - \rho)\mu, \rho\sigma^2]$.

Figure 2 adds, to the likelihoods depicted in Figure 1, the $N(0,1)$ population distribution and the three resulting posterior distributions. The posterior means are, by (10), −.755, −.190, and .945, respectively—equally weighted averages of x and the population mean, 0. The posterior variances are, by (11), .5 in each case; the posterior standard deviations are thus approximately .71. Posterior means are sometimes preferred over MLEs as point estimates for individual students, because they minimize the expected mean squared error in the population. That is, the population average of the squared difference between students' true θs and estimates of those θs is smallest when the estimates are the students' posterior means.

How well are the population parameters estimated when Bayesian posterior means $\bar{\theta}$ are used as proxies for θs? First, for the mean,

$$E(\bar{\theta}) = E(\rho x + (1 - \rho)\mu) = \rho E(x) + (1 - \rho)\mu = \mu.$$

The mean of Bayesian estimates is the correct population mean, assuming, as we have, that the correct population mean was used to construct them in the first place. But, even under these most favorable circumstances, the variance of $\bar{\theta}$ does not fare as well as an approximation of the variance of θ:

$$\text{Var}(\bar{\theta}) = \text{Var}(\rho x + (1 - \rho)\mu) = \rho^2\text{Var}(x) = \rho\sigma^2.$$

The variance of the Bayesian estimates underestimates the variance of the

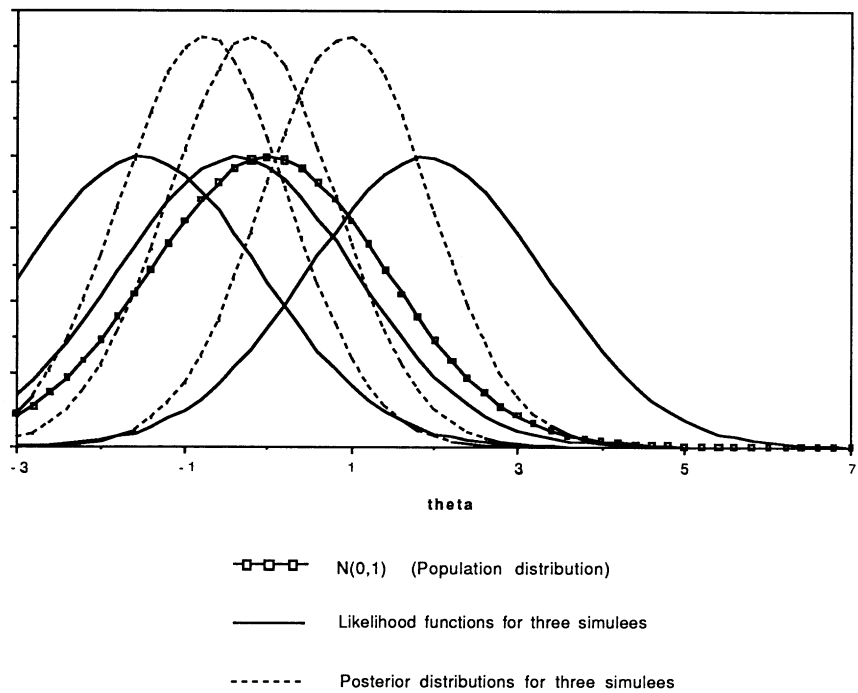

FIGURE 2. *A population distribution, likelihood functions, and posterior distributions*

underlying variable by $(1 - \rho)$ percent. Again, this bias remains as the examinee sample size increases.

Figure 3 summarizes these results by plotting the expected distributions of θ, $\bar{\theta}$, and $\hat{\theta}$ in large samples of subjects, with $\theta \sim N(0,1)$ and $e \sim N(0,1)$. Although the distributions of both $\bar{\theta}$ and $\hat{\theta}$ have the same center as that of θ, they give

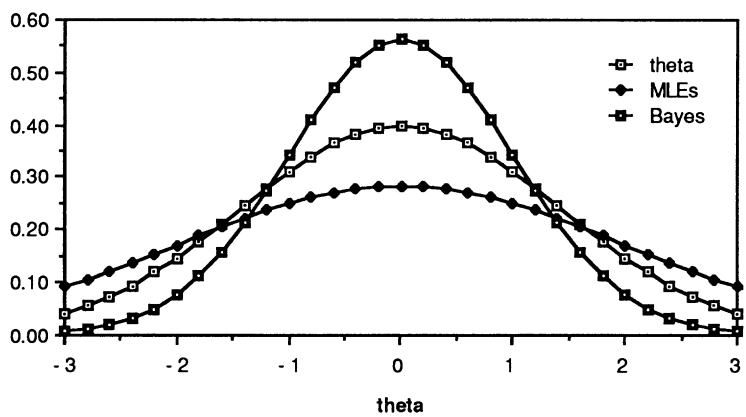

FIGURE 3. *Expected distributions of maximum likelihood and Bayesian estimates*

biased estimates of many other population characteristics—not only the variance, but quartiles, percentile points, proportions of examinees above specified θ values, and so on.

The Incomplete Data Solution

How, then, should μ and σ^2 be estimated? Maximum likelihood estimation proceeds from basic principles by finding those values of the parameters that maximize the likelihood of the observations, taking into account both the sampling of θ from the $N(\mu,\sigma^2)$ population and of x from the $N(\theta,1)$ distributions of potential responses from each sampled student. This requires the marginal, or in Dempster, Laird, and Rubin's (1977) terms the incomplete data, density in which the data are $\mathbf{X} = (x_1, \ldots, x_N)$ rather than $\boldsymbol{\Theta} = (\theta_1, \ldots, \theta_N)$:

$$p(\mathbf{X}|\mu,\sigma^2) = \prod_{i=1}^{N} \int p(x_i|\theta)p(\theta|\mu,\sigma^2)\, d\theta. \tag{12}$$

Equation 12 is the probability of observing values of x from N randomly selected students, obtained as the product of probabilities of observing x given θ, averaged over the θ distribution. MLEs for μ and σ^2 are again obtained by finding the zeros of the partial derivatives of the log of (12). Working through the algebra (see Mislevy, 1984), we obtain the following expressions:

$$\mu = N^{-1} \sum_i \int \theta p(\theta|x_i,\mu,\sigma^2)\, d\theta \tag{13}$$

and

$$\sigma^2 = N^{-1} \sum_i \int \theta^2 p(\theta|x_i,\mu,\sigma^2)\, d\theta - \mu^2. \tag{14}$$

The EM algorithm. The marginal maximum likelihood (MML) equations, (13) and (14), express $\hat{\mu}$ and $\widehat{\sigma^2}$ in terms of functions that involve μ and σ^2. The solutions, $\hat{\mu}$ and $\widehat{\sigma^2}$, have the property of self-consistency; evaluating (13) and (14) with the observed data $\mathbf{X}$ and $\hat{\mu}$ and $\widehat{\sigma^2}$ on the right sides of the expressions gives $\hat{\mu}$ and $\widehat{\sigma^2}$ back. Equations 13 and 14 can be used iteratively to obtain $\hat{\mu}$ and $\widehat{\sigma^2}$ from trial values $\mu^{(0)}$ and $\sigma^{2(0)}$. This particular example simplifies because closed-form solutions exist for the integrals. New values are obtained by evaluating the right sides of (13) and (14) with the trial values as follows:

$$\begin{aligned}
\mu^{(1)} &= N^{-1} \sum_i \int \theta p(\theta|x_i,\mu^{(0)},\sigma^{2(0)})\, d\theta \\
&= N^{-1} \sum_i E[\theta|x_i,\mu^{(0)},\sigma^{2(0)}] \\
&= N^{-1} \sum_i \rho^{(0)} x_i + (1 - \rho^{(0)})\mu^{(0)},
\end{aligned}$$

where

$$\rho^{(0)} = \frac{\sigma^{2(0)}}{\sigma_L^2 + \sigma^{2(0)}},$$

and

$$\begin{aligned}\sigma^{2(1)} &= N^{-1}\sum_i \int \theta^2 p(\theta|x_i,\mu^{(0)},\sigma^{2(0)})\,d\theta - (\mu^{(1)})^2 \\ &= N^{-1}\sum_i E[\theta^2|x_i,\mu^{(0)},\sigma^{2(0)}] - (\mu^{(1)})^2 \\ &= N^{-1}\sum_i \{E[\theta|x_i,\mu^{(0)},\sigma^{2(0)}]\}^2 + \mathrm{Var}[\theta|x_i,\mu^{(0)},\sigma^{2(0)}] - (\mu^{(1)})^2 \\ &= N^{-1}\sum_i \{\rho^{(0)}x_i + (1-\rho^{(0)})\mu^{(0)}\}^2 + \rho^{(0)}\sigma^{2(0)} - (\mu^{(1)})^2.\end{aligned}$$

Iterating until convergence in this example gives MLEs as the realized solutions of (13) and (14):

$$\hat{\mu} = N^{-1}\sum_i \int \theta p(\theta|x_i,\hat{\mu},\widehat{\sigma^2})\,d\theta \tag{15}$$

and

$$\widehat{\sigma^2} = N^{-1}\sum_i \int \theta^2 p(\theta|x_i,\hat{\mu},\widehat{\sigma^2})\,d\theta - \hat{\mu}^2. \tag{16}$$

Table 1 shows the results of applying these formulas repeatedly with the three fictitious values introduced earlier, starting from $\mu^{(0)} = 1.70$ and $\sigma^{2(0)} = 3.89$. This is a special case of Dempster, Laird, and Rubin's (1977) EM algorithm. The resulting values are the maximum likelihood estimates of μ and σ^2, consistent in N regardless of test length. They are not Bayesian estimates of μ and σ^2, even though intermediate steps in their calculation use values that could be used as Bayesian estimates for individuals in the context of individual measurement.

We thus arrive at the heart of the first paradox: Under practically any definition of *best,* the distribution of point estimates that are best individual by individual is generally not the best estimate of the distribution of the underlying latent variable. Obtaining a best estimate of the parameters of the underlying distribution can proceed directly from the imperfect response data, via an appropriate marginal procedure, without attempting to produce point estimates for individuals along the way. The results generalize beyond our illustrative normal solution, to any parametric distribution or to a nonparametric estimate of the θ distribution (e.g., Laird, 1978), and beyond marginal maximum likelihood (MML) to marginal least squares, marginal Bayesian estimation (Deely & Lindley, 1981), and so on.

Imputation. Having seen that traditional "good" estimates for individuals don't reproduce the "good" results of marginal analyses, we now consider the genesis of plausible values. The initially paradoxical role of their random components is clarified when plausible values are viewed as elements in a numerical approximation of the integrals in (15) and (16), rather than as scores for individuals.

The MML formulas for the population mean and variance take uncertainty about individuals' θs into account by averaging statistics involving θ over all the

Table 1
Trace of an EM Solution for μ and σ^2 in the Single-Population Example

Iteration	Population Mean	Population Variance	Mean for Subject 1	Mean for Subject 2	Mean for Subject 3	Posterior Variance
0	1.700	3.890	-0.854	0.045	1.851	0.796
1	0.348	2.060	-0.854	0.045	1.851	0.796
2	0.114	1.579	-0.903	-0.142	1.386	0.673
3	0.044	1.362	-0.881	-0.189	1.201	0.612
4	0.019	1.241	-0.852	-0.200	1.108	0.577
5	0.008	1.167	-0.828	-0.202	1.055	0.554
6	0.004	1.118	-0.809	-0.201	1.022	0.538
7	0.002	1.085	-0.795	-0.199	1.000	0.528
8	0.001	1.062	-0.785	-0.197	0.984	0.520
9	0.000	1.045	-0.777	-0.195	0.974	0.515
10	0.000	1.033	-0.771	-0.194	0.966	0.511
11	0.000	1.024	-0.767	-0.193	0.960	0.508
12	0.000	1.018	-0.764	-0.192	0.956	0.506
13	0.000	1.013	-0.762	-0.192	0.953	0.504
14	0.000	1.009	-0.760	-0.191	0.951	0.503
15	0.000	1.007	-0.759	-0.191	0.949	0.502
16	0.000	1.005	-0.758	-0.191	0.948	0.502
17	0.000	1.003	-0.757	-0.190	0.947	0.501
18	0.000	1.002	-0.756	-0.190	0.947	0.501
19	0.000	1.001	-0.756	-0.190	0.946	0.501
20	0.000	1.001	-0.756	-0.190	0.946	0.500
21	0.000	1.000	-0.755	-0.190	0.945	0.500
22	0.000	1.000	-0.755	-0.190	0.945	0.500
23	0.000	1.000	-0.755	-0.190	0.945	0.500
24	0.000	0.999	-0.755	-0.190	0.945	0.500
25	0.000	0.999	-0.755	-0.190	0.945	0.500

possible values θ might take, weighted in proportion to the posterior probability at each value. By the definitions of the integrals, unbiased numerical approximations of the left sides of (15) and (16) can be obtained after $\hat{\mu}$ and $\widehat{\sigma^2}$ have been calculated by drawing a value at random from each $p(\theta|x_i,\hat{\mu},\widehat{\sigma^2})$—call it, $\tilde{\theta}_i$—and evaluating (5) and (6) as if each $\tilde{\theta}_i$ were θ_i itself. One such set of draws for our artificial example, one from the posterior distribution of each simulee, is -1.593, .109, and $-.409$. The resulting mean and variance obtained through (5) and (6) are $-.631$ and .507—unbiased, but not very

accurate, approximations of the actual values 0 and 1 of (15) and (16).[1] Repeating the process to improve the accuracy of the approximation is discussed below. The key result at this point, though, is that the random draws, $\tilde{\theta}_i$, have the same expected distribution as θ while point estimates that are in some way optimal for estimating individuals' proficiencies do not. If the sample of simulees were large, we could construct a file of these imputations from which our marginal maximum likelihood estimates of μ and σ^2 could be recovered (as opposed to discovered).

We repeated the imputation procedure 25 times, each time calculating the mean via (5) and, with this estimate of μ, similarly calculating the variance via (6). The results are shown in Table 2. The averages for μ and σ^2 over the 25 replications are .131 and .984. The variation among the results for the 25 replications is due solely to not knowing the θs of the three simulees; everything is conditional on the responses of the three simulees in the realized sample. Rubin's (1987) multiple imputation procedures, discussed below, account for uncertainty about population characteristics from these two distinct sources, the sampling of θs from the population and the sampling of xs given the sampled but unobserved θs, to properly reflect the information about μ and σ^2 in $\mathbf{X}$. First, however, we investigate the second paradox.

Classical Test Theory With Two Normal Subpopulations

Let us extend the example to a population comprised of two subpopulations. Again, we assume normal θ distributions in both, and we assume the same $N(0,1)$ measurement error model for all students regardless of group membership. Let y_i indicate student i's group membership. The marginal likelihood function for $(\mu_1,\sigma_1^2,\mu_2,\sigma_2^2)$, given students' item responses $\mathbf{X}$ and group memberships $\mathbf{Y}$, takes the following form:

$$L(\mu_1,\sigma_1^2,\mu_2,\sigma_2^2|\mathbf{X},\mathbf{Y}) = \prod_{\text{Group 1}} \int p(x_i|\theta)p(\theta|\mu_1,\sigma_1^2)\,d\theta \times \prod_{\text{Group2}} \int p(x_i|\theta)p(\theta|\mu_2,\sigma_2^2)\,d\theta.$$

The likelihood equations are generalizations of (13) and (14); for $k = 1, 2$,

$$\mu_k = N^{-1} \sum_{i\in \text{ group k}} \int \theta p(\theta|x_i,\mu_k,\sigma_k^2)\,d\theta \tag{17}$$

and

$$\sigma_k^2 = N^{-1} \sum_{i\in \text{ group k}} \int \theta^2 p(\theta|x_i,\mu_k,\sigma_k^2)\,d\theta - \mu_k^2. \tag{18}$$

The EM algorithm leads, in the same way as in the preceding section, to consistent estimates of the subpopulation parameters $(\mu_1,\sigma_1^2,\mu_2,\sigma_2^2)$. Note that, in this example, only the data of students in group k are involved in estimating the parameters of group k.

Let our original three simulees represent a sample from Subpopulation 1, and add a sample of x values from Subpopulation 2: 1.89, 3.02, 5.29. These are

Table 2
Results from Twenty-five Sets of Imputations

Set of Draws	Subject 1	Subject 2	Subject 3	Estimated Mean	Estimated Variance
1	-1.593	0.109	-0.409	-0.631	0.888
2	-0.816	-0.424	0.852	-0.129	0.506
3	-1.780	-1.055	0.839	-0.665	1.645
4	-0.152	1.436	1.276	0.853	1.221
5	-1.589	-0.994	1.453	-0.376	1.858
6	-1.437	0.263	0.730	-0.147	0.872
7	-1.390	-0.572	1.947	-0.005	2.000
8	-0.924	0.137	1.217	0.143	0.768
9	-0.044	0.861	1.813	0.876	1.326
10	-0.939	-1.121	1.728	-0.110	1.691
11	-0.631	-0.107	0.155	-0.194	0.127
12	-1.519	-0.397	1.381	-0.178	1.441
13	-0.750	0.117	0.762	0.043	0.369
14	0.321	0.541	1.242	0.701	0.629
15	-0.557	-0.178	1.323	0.195	0.680
16	0.089	0.654	1.718	0.820	1.112
17	0.231	-0.130	0.824	0.308	0.232
18	-1.234	1.042	0.880	0.229	1.111
19	-0.923	-1.357	0.616	-0.554	1.007
20	0.129	-0.176	1.798	0.583	1.076
21	0.372	0.929	1.494	0.932	1.061
22	-1.428	-0.205	1.916	0.093	1.901
23	-0.910	0.589	0.465	0.048	0.447
24	-0.617	-0.035	0.969	0.105	0.423
25	-0.016	0.355	0.724	0.354	0.199
Average	-.724	0.011	1.108	0.131	0.984

just the values of the first sample, with 3.40 added to each. The grand mean and variance for the two subpopulations combined are 1.70 and 4.89. Table 1 shows the trace of an EM solution for μ_1 and σ_1^2, which leads to values of 0 and 1. In exactly the same way, using only the data from Subpopulation 2, similar calculations from 1.70 and 3.89 lead to values of 3.40 and 1.00 for μ_2 and σ_2^2. Constructing one of a number of files of plausible values from which the subpopulation parameters could be recovered requires a draw for each simulee i from the conditional distribution $p(\theta|x_i, y_i, \widehat{\mu_1}, \widehat{\sigma_1^2}, \widehat{\mu_2}, \widehat{\sigma_2^2})$. If simulee i belongs to

subpopulation k, we draw from $p(\theta|x_i,\widehat{\mu_k},\widehat{\sigma_k^2})$. Notice what this implies for the two simulees with observed values of 1.89, the highest score in Subpopulation 1 and the lowest in Subpopulation 2: Values in the plausible values file corresponding to these individuals are drawn from $N(.945,.50)$ and $N(2.645,.50)$, respectively!

The utilization of collateral information such as subpopulation membership is often eschewed in the measurement of individuals, particularly when a test is being used "as a contest." If, for example, admission to a college were based on Bayesian estimates conditioned on group membership, preference would be granted to members of higher scoring groups whenever students with the same observed score were compared. In this setting, one would probably prefer unbiased MLEs to Bayesian estimates (not to mention random draws from posterior distributions!), and it would be incumbent on the decision maker to obtain satisfactorily precise information about each individual. Of course, estimates of subpopulation distributions from these preferable point estimates would be inconsistent, but these are not the estimates that count in that setting.

What happens when collateral variables are ignored in the construction of plausible values but population characteristics involving those variables are subsequently calculated? In our example, the most favorable results under this restriction will occur when plausible values are constructed using the correct population distribution, a 50–50 mixture of the two normal distributions $N(0,1)$ and $N(3.4,1)$. Its overall mean and variance turn out to be 1.70 and 3.89. The posterior density for a simulee with observed score x is, regardless of subpopulation membership, obtained from Bayes theorem as the normalized product of the likelihood, $N(x,1)$, and the prior, the average of $N(0,1)$ and $N(3.4,1)$; that is,

$$p(\theta|x) \propto \exp\left(-\frac{(\theta-x)^2}{2}\right) \times \frac{1}{2}\left[\exp\left(-\frac{\theta^2}{2}\right) + \exp\left(-\frac{(\theta-3.4)^2}{2}\right)\right].$$

Figure 4 depicts this population distribution, the likelihood induced by the observation of $x = 1.89$, and the resulting posterior distribution. The mean of this posterior is 1.931; recall that the subject from Subpopulation 1 had a

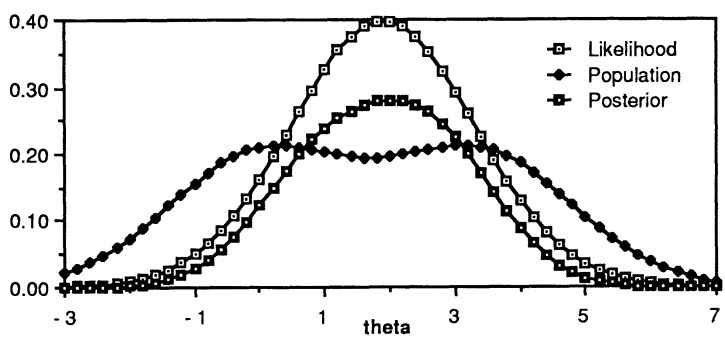

FIGURE 4. *An undifferentiated population distribution, a likelihood function, and a posterior distribution*

posterior mean of 1.851 in the single population example. The posterior means for all simulees are (−.748, −.142, 1.931) for Subpopulation 1 and (1.931, 3.047, 4.341) for Subpopulation 2. Altogether, the six values recover the global mean of 1.70. More importantly for our purposes, they underestimate the distance between the subpopulation means, giving .347 and 3.106 instead of 0 and 3.4. Subjects' posterior distributions are shifted away from their subpopulation distributions toward the composite population distribution. A file of plausible values constructed with equally large samples of simulees from both subpopulations would have the following characteristics:

1. The plausible values would echo back the correct mixed–normal shape for the population distribution, including not only the correct values for the global mean and variance but also its bimodality.
2. Simulees with the same observed scores would have the same distributions of plausible values, regardless of subpopulation membership.
3. The subpopulation means, distinguishing simulees according to a variable on which they differ but which was ignored when constructing plausible values, would be biased by a factor that depends on the magnitude of measurement error (see Mislevy, 1991, for closed-form results for special cases).

We thus arrive at the heart of the second paradox. We are not estimating population characteristics from plausible values but constructing plausible values to reflect the estimates of population characteristics obtained in a marginal analysis. Unless a characteristic were incorporated into the marginal analysis from which plausible values were constructed, it would generally not be recovered correctly from analyses of the completed data sets. (We shall discuss below, in the context of NAEP, designing marginal analyses that reduce secondary biases of this type.)

Why Use Plausible Values?

Having seen that plausible values are constructed from the results of marginal analyses, with the intent of echoing back the results of those analyses, why should one bother to create them in the first place? Wouldn't it be simpler just to carry out the marginal analyses and report their results? It certainly would in the examples we have examined above. The following arguments for producing plausible values, as opposed to reporting the results of marginal analyses, begin to carry force in more complex surveys.

Complex sample designs. The preceding examples were based on simple random sampling, where theory for marginal estimation has been worked out. The economics of large-scale surveys such as NAEP, however, necessitate complexities in the respondent sampling design such as unequal probabilities of selection for different students; stratification, to ensure prespecified rates of representation of targeted subpopulations; and clustering, which links the selection probabilities of students when their joint occurrence in the sample facilitates data collection (e.g., selecting schools, then students within schools). These features of the sampling designs must be taken into account both when

estimating population characteristics and when approximating their sampling variances. Appropriate designs and estimation techniques are readily available if survey variables are observed without error for all sampled units (e.g., Cochran, 1977), but, other than some work with the classical test theory model, no simple procedures are generally available. Rubin's (1987) multiple imputations approach provides the interface between the techniques of survey sampling, for handling uncertainty due to sampling respondents, and those of psychometrics, for handling uncertainty due to the latent nature of variables that may be of interest.

Convenience for secondary users. For the two-group normal mixture examined above, two means, two standard deviations, and the proportions of subpopulation membership convey all the information necessary to infer any population characteristic: moments, percentile points, proportions above specified proficiency levels, and so on. A list of parameter estimates and an asymptotic covariance matrix suffice for any question a secondary researcher might pose. This is not the case with NAEP, however, with its hundreds of test items and demographic, educational, and attitudinal variables. Droves of descriptive statistics, multiple regression analyses, and LISREL models might be entertained, and neither carrying out all of these analyses nor providing sufficient statistics for them is feasible. But a correct marginal analysis for a model involving latent variables often lies beyond the reach of secondary researchers whose expertise lies in substantive areas rather than in statistics. Providing files of plausible values constructed from a comprehensive primary marginal analysis allows these researchers to obtain reasonably good approximations for a wide variety of secondary analyses using only "complete data" statistical procedures, or those that would apply if θ values had been observed.

Improved estimation. We have stressed that plausible values are not raw data from which to discover population attributes but constructions from marginal analyses from which to recover those attributes. When the model for the original marginal analysis is specified correctly, plausible values convey nothing more or nothing less than its results. But, if the marginal analysis were specified incorrectly, the plausible values actually would provide better estimates. The two-group example above illustrates this point. Ignoring the group difference when constructing plausible values effectively sets the group difference at zero. The plausible values moved the estimate of the difference away from zero, more than three fourths of the way to the correct result. The improvement depends on the amount of information that observable variables carry about latent variables. At the extreme, as θ becomes better determined by x, analyses based on plausible values from virtually any primary analysis provide excellent approximations to the correct values. In general, the results obtained from plausible values constructed with a misspecified model move one EM step toward the correct values.

The General Form of the Procedures

This section summarizes the steps used to construct and analyze multiple imputations. It is a direct application of Rubin's (1987) procedures.

The Sampling Model

Consider a population of N identifiable units, indexed by i. Associated with each unit are three (possibly vector-valued) variables—θ_i, y_i, and z_i. The values of the design variables, z_i, are known for all units before observations are made, but the values of the survey variables, y_i and θ_i, are not. In the context of NAEP, θ represents proficiency in a subject area, y represents responses to demographic and educational questions, and z represents stratification and clustering variables such as school membership and region of the country. Let $(\boldsymbol{\Theta},\mathbf{Y},\mathbf{Z})$ denote the population matrix of values. Interest lies in the value of a function $S = S(\boldsymbol{\Theta},\mathbf{Y},\mathbf{Z})$. S would be calculable without error if all units' θ and y values were known. Under randomization-based survey-sampling inference, the locus of probability is the selection of the subset of units from which some or all of the values of the survey variables will be ascertained—the sample. A sample design assigns a probability to each of the 2^N subsets. These probabilities can depend on the intended size of the sample, or on z values, as in the case of stratification or clustering. One subset, the realized sample $\mathscr{S}$, is selected at random accordingly.

Denote by $(\boldsymbol{\theta}_s,\mathbf{y}_s)$ the values of the survey variables in the realized sample. If both θ and y were observed without error, inference about S would be based on the distribution of a statistic $s = s(\boldsymbol{\theta}_s,\mathbf{y}_s,\mathbf{Z})$ in repeated samples under the specified sample design. Popular designs are devised so as to support an (at least approximately) unbiased statistic, s, where bias is measured as the discrepancy between the fixed, but unknown, population value S and the average value of the sample statistic s over all potential samples, each weighted by the probability under the specified sampling design (e.g., Cochran, 1977). Widely used designs and estimators are accompanied by another sample statistic $U = U(\boldsymbol{\theta}_s,\mathbf{y}_s,\mathbf{Z})$, which approximates the variance of s over all potential samples. Design features, such as clustering and stratification, are taken into account in the calculation of both s and U. Inferences are typically based on the normal approximation

$$(s - S)/\sqrt{U} \sim N(0,1). \tag{19}$$

Because design variables are considered fixed and known, the possible dependence of S, s, and U on $\mathbf{Z}$ will be implicit in the sequel.

The Measurement Model

Measurement error exists when the values of the survey variables of the sampled units are not determined with certainty. Suppose that survey variable y will be ascertained with certainty but θ will not be. Under the classical form of measurement error in sample surveys (e.g., Hansen, Hurwitz, & Bershad, 1961), the surveyor observes for unit i, not θ_i itself, but $x_i = \theta_i + e_i$—the measurement error model discussed in the preceding sections. Standard analyses of survey error would address the impact of various characteristics of the values of e (e.g., its mean over units and whether it is correlated with θ) on the distribution of s, if s were to be evaluated with $\mathbf{x}_s$ in place of $\boldsymbol{\theta}_s$ (Cochran, 1977, secs. 13.8–13.13).

A model-based framework for measurement error introduces the distribution, $p(x|\theta;\boldsymbol{\beta})$, with possibly unknown parameters $\boldsymbol{\beta}$. An example pertinent to present purposes is that of latent variable modeling. Latent variables are posited to account for regularities in observable variables, such as examinees' tendencies to give correct responses to test items. In IRT, the probability that subject i, with unobservable proficiency parameter θ_i, will make a correct response to item j (i.e., $x_{ij} = 1$ as opposed to 0) depends on θ_i and a (possibly vector-valued) item parameter β_j, as $p(x_{ij} = 1|\theta_i,\beta_j)$. Under the assumption of conditional independence, the latent variable accounts for all associations among responses to various items in a specified subject area. Moreover, the latent variable is assumed to account for associations between response variables and collateral variables such as demographic or educational standing. Letting $\mathbf{x}_i = (x_{i1}, \ldots, x_{in})$ be a vector of responses to n items, conditional independence posits

$$p(\mathbf{x}_i|\theta_i,\boldsymbol{\beta},y_i,z_i) = p(\mathbf{x}_i|\theta_i,\boldsymbol{\beta}) = \prod_{j=1}^{n} p(x_{ij}|\theta_i,\beta_j), \tag{20}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)'$. Of course, these assumptions must be checked in practice; see, for example, Muthén and Lehman (1985) and Thissen, Steinberg, and Wainer (1988).

Under such a model, the responses to any subset of items induce a likelihood function for θ_i via (20). If the focus is on measuring individuals for placement or selection decisions—as it is in the context under which IRT evolved—one administers enough items to each examinee to make the likelihood function for his or her θ peak sharply. A precise point estimate of each θ, such as the MLE or the Bayes mean estimate, can be obtained under these circumstances. We have seen that the same point estimates can prove seriously misleading if the focus is on the distribution of θ in a population. That is, $s(\hat{\boldsymbol{\theta}}_s,\mathbf{y}_s)$ is not generally a defensible estimate of $s(\boldsymbol{\theta}_s,\mathbf{y}_s)$.

Although we cannot calculate $s(\boldsymbol{\theta}_s,\mathbf{y}_s)$ directly, Rubin (1987) points out that we can profitably shift our attention to its expected value given the data that are actually observed—namely, $(\mathbf{x}_s,\mathbf{y}_s)$, as follows:

$$s^*(\mathbf{x}_s,\mathbf{y}_s) \equiv E[s(\boldsymbol{\theta}_s,\mathbf{y}_s)|\mathbf{x}_s,\mathbf{y}_s,\mathbf{Z}] = \int s(\boldsymbol{\theta}_s,\mathbf{y}_s)p(\boldsymbol{\theta}_s|\mathbf{x}_s,\mathbf{y}_s,\mathbf{Z})\,d\boldsymbol{\theta}_s. \tag{21}$$

A numerical approximation of this integral is obtained by drawing a value at random from the predictive distribution of the vector of latent variables of all respondents, or $p(\boldsymbol{\theta}_s|\mathbf{x}_s,\mathbf{y}_s,\mathbf{Z})$. This is a file of plausible values. In addition to the measurement model, one additional component is required to construct this predictive distribution: the population-structure model.

The Population-Structure Model

As we have noted, the randomization approach to survey analysis considers the sample selection as the only random element in the problem. The contrasting model-based approach would view the existing finite population itself as a sample from a hypothetical superpopulation, in which variables are distributed

according to a specified model, possibly depending on unknown parameters (see Cassel, Särndal, & Wretman, 1977). The realized sample provides a basis for inferences about those parameters, about the structure of the superpopulation, and ultimately about the value of S. The key assumption of the model-based approach is the conditional exchangeability of units; that is, simple random sampling procedures are appropriate to estimate superpopulation parameters conditional on the values of design variables. In the present context, constructing $p(\boldsymbol{\theta}_s|\mathbf{x}_s,\mathbf{y}_s,\mathbf{Z})$ requires a model of the form $p(\theta|y,z;\boldsymbol{\alpha})$—the conditional distribution of proficiency, given collateral variables y and z. The $\boldsymbol{\alpha}$ denotes any parameters required in the population-structure model, such as (μ,σ^2) in the univariate normal case. Assuming experimental independence across sampled units, we obtain

$$p(\boldsymbol{\theta}_s|\mathbf{x}_s,\mathbf{y}_s,\mathbf{Z}) \propto \int p(\mathbf{x}_s|\boldsymbol{\theta}_s;\boldsymbol{\beta})p(\boldsymbol{\theta}_s|\mathbf{y}_s,\mathbf{Z};\boldsymbol{\alpha})\, d(\boldsymbol{\alpha},\boldsymbol{\beta}|\mathbf{x}_s,\mathbf{y}_s)$$

$$= \int \prod_{i\in\mathscr{S}} p(x_i|\theta;\boldsymbol{\beta})\, p(\theta_i|y_i,z_i;\boldsymbol{\alpha})\, d(\boldsymbol{\alpha},\boldsymbol{\beta}|\mathbf{x}_s,\mathbf{y}_s), \quad (22)$$

where the integration over the measurement-model parameters $\boldsymbol{\beta}$ and population-structure parameters $\boldsymbol{\alpha}$ accounts for the fact that they are inferred only imperfectly, from the information in $\mathscr{S}$. The effect is to incorporate additional noise into the imputations, although typically much less than is caused by uncertainty about individuals' θs. In applications where they are estimated precisely, it may be permissible to treat estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as known true values, as was done in our preceding examples; whence

$$p(\boldsymbol{\theta}_s|\mathbf{x}_s,\mathbf{y}_s,\mathbf{Z}) \propto \prod_{i\in\mathscr{S}} p(x_i|\theta;\hat{\boldsymbol{\beta}})p(\theta|,y_i,z_i;\hat{\boldsymbol{\alpha}}).$$

The more general form, (22), accounts for uncertainty about $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and will be described in the step-by-step procedures outlined below.

Estimating the parameters of the structural model, or $\boldsymbol{\alpha}$ in $p(\theta|y,z;\boldsymbol{\alpha})$, constitutes the marginal analysis phase of the approach. As we have seen, this can be accomplished via marginal maximum likelihood or marginal Bayesian procedures. Various models and procedures are discussed by Andersen and Madsen (1977), Dempster, Laird, and Rubin (1977), Laird (1978), Mislevy (1984, 1985), and Rigdon and Tsutakawa (1983). In the two-group example, the collateral variable y was group membership, and this phase of the analysis consisted of estimating the group difference and within-group variances. Ignoring the group difference when constructing plausible values replaced the correct term $p(\theta|y_i,\boldsymbol{\alpha})$ with $p(\theta|\boldsymbol{\alpha}^*)$, where $\boldsymbol{\alpha}^*$ describes the mixture of the two normal distributions without specifying individuals' group memberships.

Constructing and Working With Multiple Imputations

Inferences about S are drawn by carrying out the following steps:

1. Obtain the posterior distribution of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, or $p(\boldsymbol{\alpha},\boldsymbol{\beta}|\mathbf{x}_s,\mathbf{y}_s)$.
2. Produce M completed data sets $(\boldsymbol{\theta}_{s(m)},\mathbf{x}_s,\mathbf{y}_s)$. For the mth,
   a. draw a value $(\boldsymbol{\alpha},\boldsymbol{\beta})_{(m)}$ from $p(\boldsymbol{\alpha},\boldsymbol{\beta}|\mathbf{x}_s,\mathbf{y}_s)$;

b. for each sampled respondent, draw a value from the predictive distribution $p(\theta|x_i,y_i,z_i;(\boldsymbol{\alpha},\boldsymbol{\beta})_{(m)})$. (As noted above, it may suffice to draw from $p(\theta|x_i,y_i,z_i;(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}}))$ if the structural parameters are well determined.)

3. Using each completed data set in turn, calculate $s_{(m)} = s(\boldsymbol{\theta}_{s(m)},\mathbf{x}_s,\mathbf{y}_s)$ and $U_{(m)} = U(\boldsymbol{\theta}_{s(m)},\mathbf{x}_s,\mathbf{y}_s)$.
4. The final estimate of S is a numerical approximation of (21), the average of the M estimates from the completed data sets:

$$s_M^* = M^{-1}\sum_m s(m). \tag{23}$$

5. The estimated variance of s_M^* —namely, V_M—is the sum of two components:

$$V_M = U_M + (1 + M^{-1})B_M, \tag{24}$$

where

$$U_M = M^{-1}\sum_m U_{(m)}$$

quantifies uncertainty due to sampling respondents and

$$B_M = (M-1)^{-1}\sum_m (s_{(m)} - s_M^*)^2$$

quantifies uncertainty due to not observing θ from the sampled subjects. The factor $(1 + M^{-1})$ preceding B_M in (24) accounts for using a finite number of draws from the predictive distribution.

6. Inferences about S are based on

$$(s_M^* - S)/\sqrt{V_M} \sim t_\nu(0,1),$$

a t distribution with degrees of freedom given by

$$\nu = (M-1)(1 - r_M^{-1})^2,$$

where r_M is the proportion of total variance due to the latent nature of θ:

$$r_M = (1 + M^{-1})\frac{B_M}{U_M}.$$

For t-dimensional $\mathbf{s}$, such as the vector of coefficients in a multiple regression, $\mathbf{U}_M$ and each $\mathbf{U}_{(m)}$ will be covariance matrices, and $\mathbf{B}_M$ is an average of squares and cross products. The quantity

$$(\mathbf{S} - \mathbf{s}_M)'\mathbf{V}_{M^{-1}}(\mathbf{S} - \mathbf{s}_\mathbf{M})$$

is approximately F distributed with t and ν degrees of freedom, with ν defined as above but with a matrix generalization of r_M:

$$r_M = (1 + M^{-1})\ \text{Trace}\ (\mathbf{B}_M\mathbf{V}_{M^{-1}})/k.$$

Implementation in the National Assessment

Beginning in the 1983–1984 school year, NAEP has employed the plausible values approach to provide estimates of characteristics of the distributions of the nation's school children. In 1983–1984, for example, reading and writing were assessed with national probability samples of students at ages 9, 13, and 17 and in the modal grades associated with those ages—namely, 4, 8, and 11. Approximately 25,000 students were assessed from each grade/age cohort. At intervals of 2 years, assessments were similarly carried out in these and other subject areas, including science, mathematics, literature, and computer competence. Highlights of the multiple imputation procedures employed with IRT in these assessments follow.

The Sampling Model

Students are selected in NAEP in a multistage sampling design, with counties or groups of counties as the primary sampling units (PSUs). Schools are the second-stage sampling unit, and students within schools are the third. The design variables (z) for selecting students are PSU and school membership, region of the country, and size and type of community. Sampling weights were provided from which to estimate population characteristics on directly observed survey variables, and a multiweight jackknife procedure was developed to approximate sampling variances (U) of these estimates. Inferences about characteristics of the distributions of y variables could thus be carried out via (19).

Each student responds to roughly 50 demographic, educational, and attitudinal survey items (y), some of which are administered to all students and others of which only a subsample receive. In addition, each sampled student was administered a number of cognitive exercises (test items x) under a multiple matrix item sampling design. There were a total of 340 multiple-choice and free-response reading items in the 1983–1984 reading assessment, for example, of which a given student would receive between 5 and 50.

The IRT Model

IRT models were fit to data in several assessments. In the reading assessment in 1983–1984 and in subsequent reading assessments, a priori considerations and dimensionality analyses supported summarizing the responses to a majority of items by a single IRT scale (Zwick, 1987). In mathematics and sciences assessments, in contrast, IRT scaling was carried out separately in as many as six distinct subscales. The latent variables defined in these scalings play the role of θ in the preceding discussions. For multiple-choice items, the 3-parameter logistic IRT model presented in (1) was employed. For free-response items, c_j was fixed at 0. These item parameters, over all the items in a scale, play the role of β. Item parameters in a given scale were estimated using Mislevy and Bock's (1983) BILOG computer program from samples of approximately 10,000 students and treated as known values thereafter (more on this below). The likelihood function induced by this model takes the form of (20),

where the product runs over only the items administered to the student in question.

In several subject areas, the IRT models extend over grade/age cohorts. Checks for the fit of the model include analyses of residuals by gender, race/ethnicity, and grade/age cohort. About 5% of the items exhibit different trace lines for different grade/age cohorts, usually explicable in terms of the timing of the introduction of their topics in the curriculum. These items are modeled as distinct items in the different cohorts, and the linking of the scale between cohorts is thus defined by the set of items whose response curves can be matched between adjacent cohorts. This phenomenon is minimized by carrying out scaling in more narrowly defined subject areas, such as numerical operations, algebra, measurement, and geometry within mathematics.

When scaling is carried out within subscales in a given subject area, distinct IRT models are fit to each subscale. Under standard IRT assumptions, responses to items in different subscales are modeled as conditionally independent, given the multivariate θ parameter, even though the subscale proficiencies might be substantially correlated in a population of students. This correlation is accounted for with a multivariate population-structure model, or $p(\theta|y,z)$.

The Population-Structure Model

The form of $p(\theta|y,z;\boldsymbol{\alpha})$, employed in NAEP analyses, is analogous to that of a multiple regression model with homoscedastic residuals, except that the dependent variable θ is not directly observable. The parameter $\boldsymbol{\alpha}$ consists of two components, $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$. If θ is unidimensional, $\boldsymbol{\Gamma}$ is a vector of effects, and $\boldsymbol{\Sigma}$ is the residual variance; if θ is multidimensional, $\boldsymbol{\Gamma}$ is a matrix, and $\boldsymbol{\Sigma}$ is the residual covariance matrix. NAEP estimates them via marginal maximum likelihood (see Mislevy, 1985, for formulas). The collateral variables y and z are contrasts defined from students' design variables; responses to demographic, educational, and attitudinal items; and, in some assessments, responses from their teachers about classroom practices. Most of the effects were contrasts among categories of possible responses, such as boys–girls and among levels of parental education. In the most recent assessment—1990 mathematics—nearly 200 effects were included. They were mostly main effects, but they included interaction terms found important in previous studies. As with the item parameters, $\boldsymbol{\alpha}$ was estimated with large samples of students and treated as known thereafter. This is done separately in each grade/age, thus allowing for interactions of all other contrasts with grade/age cohort.

Five sets of plausible values are constructed and analyzed using the basic procedure outlined in the preceding section, with two minor differences. First, item parameters and population-structure parameters were treated as known for all five completed data sets, rather than drawn from their posterior distribution. This expedient is based on the large sample sizes with which they have been estimated; we are, however, currently investigating whether the improvement in the variance estimation to be gained by incorporating this uncertainty is worth the effort it will require. Second, because the calculation of

the multiweight jackknife estimate of sampling variance is computationally burdensome, U_M is approximated by the value of U from one completed data set, rather than the average of values from all five.

The Evolution of Plausible-Values Procedures in NAEP

Plausible values were first used in NAEP for the 1983–1984 reading assessment. A single IRT scale was employed, spanning the three grade/age cohorts. It was only possible at that time to fit a population-structure model with 16 effects in each grade/age cohort, a set based on the traditional NAEP reporting categories of region of the country, size and type of community, parents' education, gender, and race/ethnicity. Analyses of the marginal distributions of proficiency with respect to these variables were thus optimized. Analyses involving other collateral variables were subject to the secondary biases discussed above in the two-group classical test theory example. Simple regression analyses based on nonincluded variables, for example, exhibited an average of 15% shrinkage. Since that time, a continuing program of research has born a variety of techniques for minimizing biases for an increasingly broad array of secondary analyses. Of course, effects included in the marginal estimation phase continue to be optimized and to yield model-consistent estimates. Secondary biases for nonincluded variables have been reduced to approximately 5% by such techniques as those listed below (see the NAEP technical reports for 1986–1987 [Beaton, 1988] and 1988–1989 [Johnson & Zwick, 1990] for details):

Increased numbers of effects in the population-structure model. From the initial analysis with 16 effects, improved computational routines (e.g., Sheehan, 1985) and contrast-coding strategies have made it possible to model some 200 effects. One strategy, for example, has been to carry out estimation with the leading principal components of original contrast codings in order to enhance the stability of estimation. Another has been to introduce a contrast based on schools' average performance which improves the accuracy of secondary multilevel analyses.

Multivariate procedures. The multivariate population-structure models mentioned above were introduced in the NAEP Survey of Young Adult Literacy (Kirsch & Jungeblut, 1986). Exploiting population correlations of .6 among the four IRT scales sharpened respondents' predictive distributions and thereby reduced secondary biases by about as much as doubling test lengths would have.

Revised item-sampling designs. In the 1983–1984 and 1985–1986 assessments, each sampled student was presented items from several subject areas. This practice increases the accuracy of estimates of population characteristics that are included in the marginal analysis at the expense of greater secondary bias for characteristics that are not. It might therefore be the design of choice if analyses were limited to a prespecified set, but NAEP has as one objective providing data for secondary analysts. The item-sampling designs were therefore revised so as to present items from a single content area to any given student.

A Numerical Example With Data From the SAT

At the suggestion of Professor Robert Linn, Chair of the NAEP Technical Advisory Committee from 1984 to 1990, a demonstration of the plausible values approach vis-à-vis point estimation was undertaken with data from the verbal section of the SAT (SAT-V). The SAT-V comprises 85 items, making it possible to compare the results of analyses from the full test with those that would be obtained from artificially sampled-down matrix samples of item responses. Data from 10,000 high school seniors were obtained from the SAT public use data tape, Test Form 3GSA084. Of these, 9075 also provided background information through the Student Demographic Questionnaire (SDQ).

IRT parameters were estimated for a baseline set of 80 items using the LOGIST computer program (Wingersky, Barton, & Lord, 1982). To simulate the NAEP item-sampling design, the items were divided into 16 five-item parcels. A very sparse data set with 5 items per student was created by assigning a single parcel to each student in a spiraled pattern. The process was repeated with pairs of parcels to create a data set with 10 items per student. In the same way, data sets were created based on matrix samples with four sets of 20 items each and two subsets of 40 items each.

With each data set, the population characteristics of interest were subpopulation means for the following subpopulations: gender (male, female); race/ethnicity (Asian, Black, Hispanic, Other); parents' income ($<$\$15,000, \$15,000–\$26,999, \$27,000–\$45,000, $>$\$45,000, and missing); and region (New England, Middle, South, Midwest, and Southwest/West). These were estimated in three ways:

- Maximum likelihood estimates for each student.
- Plausible values, based on a marginal analysis for an undifferentiated population—that is, ignoring students' status on the demographic variables. We refer to these as *unconditional* plausible values.
- Plausible values, based on a marginal analysis with a population-structure model like that used in NAEP: a main-effects model θ given status on the demographic variables. We refer to these as *conditional* plausible values.

Results

The first problem that arises when computing MLEs for each student is that some response patterns have no finite MLE at all. This occurs for patterns with

Table 3

Frequencies and Proportions of Infinite Maximum Likelihood Estimates

	Items per Student				
	5	10	20	40	80
$\hat{\theta}=-\infty$	770 (8.5)	355 (3.9)	116 (1.3)	27 (0.3)	13 (0.1)
Valid $\hat{\theta}$	7552 (83.2)	8485 (93.5)	8903 (98.1)	9032 (99.5)	9056 (99.8)
$\hat{\theta}=+\infty$	753 (8.3)	235 (2.6)	56 (0.6)	16 (0.2)	6 (0.1)

all correct responses and for most of those with percent-correct scores below chance level (i.e., the sum of the items' c parameters), including patterns with all incorrect responses. Table 3 shows the frequency with which this occurred at each of the test lengths—from practically none with 80 items up to 17% of the respondents with 5 items. Estimating means is obviously problematic when the data contain infinite values. We somewhat arbitrarily replaced MLEs of $-\infty$ and $+\infty$ with -4 and $+4$, respectively, although more sophisticated choices could be made. From the data in Table 3 and the resulting mean estimates, the enterprising reader can explore the implications of alternatives. No such

Table 4
Estimated Race/Ethnicity Subpopulation Means

			Type of Estimation		
				Unconditional	Conditional
Group	N	# items	MLE	Plausible Values	Plausible Values
Black	615	5	-.76	-.15	-.61
		10	-.72	-.30	-.57
		20	-.69	-.40	-.62
		40	-.64	-.49	-.61
		80	-.65	-.57	-.63
Hispanic	236	5	-.60	-.12	-.49
		10	-.73	-.26	-.56
		20	-.61	-.37	-.54
		40	-.61	-.46	-.57
		80	-.57	-.50	-.56
Asian	388	5	-.14	.02	-.11
		10	-.31	-.01	-.10
		20	-.27	-.13	-.17
		40	-.22	-.11	-.17
		80	-.23	-.17	-.18
Other	7836	5	.20	.13	.18
		10	.15	.12	.15
		20	.16	.13	.16
		40	.17	.15	.16
		80	.16	.15	.16

problems arise with the plausible values, because predictive distributions are well defined for all response patterns.

Tables 4 and 5 present the means estimated by the three approaches for race/ethnicity and gender breakdowns at the various test lengths. Figure 5 depicts the race/ethnicity results graphically. The subpopulation samples comprise the following proportions of the total sample: Asian students, 2.6%; Black students, 6.8%; Hispanic students, 2.6%; and Other students, 86.3%. The following patterns can be discerned:

- MLEs overestimate group differences when few items are used. The consistency of MLEs for individuals' θs is an asymptotic property, realized as test length increases without limit; it promises nothing for short tests. The presence of extreme, including infinite, MLEs for shorter tests produces this somewhat unexpected result. The biases tend to decrease as test length increases.
- Unconditional plausible values underestimate group differences. This is an example of the phenomenon discussed above in the two-group classical test theory example. Estimated subpopulation means shrink toward the overall population mean, an effect more severe for the smaller and more extreme subpopulations. These biases also tend to decrease as test length increases.
- The conditional plausible values provide the most stable estimated means across different test lengths.

Table 5

Estimated Gender Subpopulation Means

			Type of Estimation		
				Unconditional	Conditional
Group	N	# items	MLE	Plausible Values	Plausible Values
Male	4244	5	.14	.11	.12
		10	.07	.09	.09
		20	.09	.09	.09
		40	.10	.10	.10
		80	.10	.10	.10
Female	4831	5	.07	.10	.08
		10	.03	.07	.06
		20	.05	.06	.05
		40	.05	.06	.05
		80	.05	.05	.05

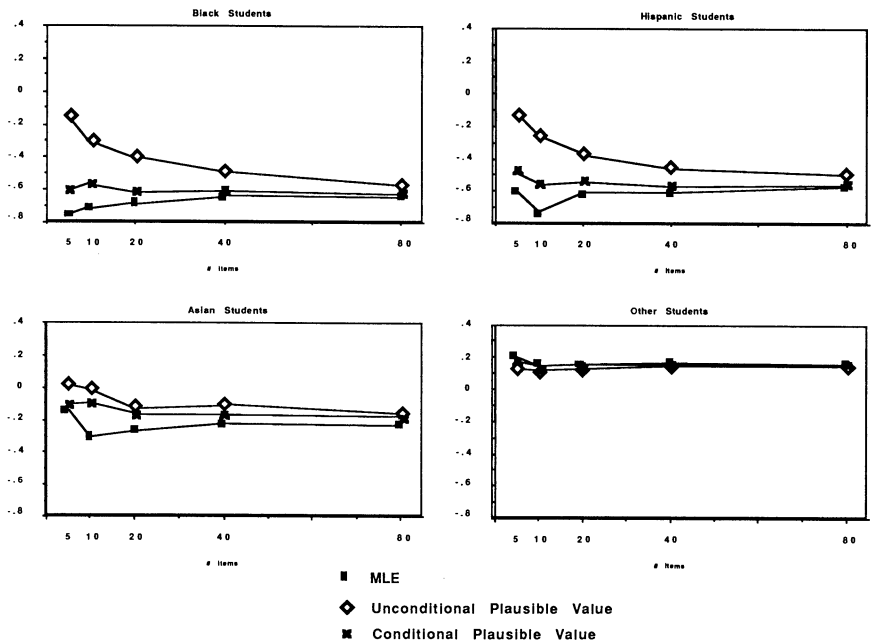

FIGURE 5. *Race/ethnicity means as estimated by alternative procedures*

- By the time test length reaches 80 items, all three methods agree fairly well.

Summary

Charles Stein shocked the statistical world by proving that the "best" estimate of the individual means in a multivariate normal distribution is an inadmissible estimate of the mean vector (Stein, 1956). Since that time, considerable research has supported the basic finding: In the presence of uncertainty at lower levels of hierarchical designs, different estimation procedures are better suited for different inferences; no single procedure gives the best possible answer to all possible questions. This problem arises in large-scale educational assessment, when interest lies in population distributions of latent variables. The distributions of point estimates that would be preferred for making inferences about individuals can depart substantially from the distribution of the underlying variable. The marginal procedures that possess superior properties for population-level analyses are less familiar in the educational measurement literature, however, and possess rather paradoxical characteristics from the point of view of individual measurement. The purposes of this presentation have been to explicate connections between the two types of procedures, emphasizing their suitabilities for distinct problems; to unravel some of the paradoxes; to describe the implementation of marginal analyses in the National Assessment for Educational Progress; and to illustrate their use in a simulation setting that facilitates the comparison of alternative approaches.

Note

[1]To simplify the example, we proceed as if the true population parameters were known. We address in a later section the procedures to be followed when parameters are not sufficiently well estimated to be treated as known.

References

Andersen, E. B., & Madsen, M. (1977). Estimating the parameters of a latent population distribution. *Psychometrika, 42,* 357–374.

Beaton, A. E. (1987). *Implementing the new design: The NAEP 1983-84 technical report* (No. 15-TR-20). Princeton, NJ: Educational Testing Service.

Beaton, A. E. (1988b). *Expanding the new design: The NAEP 1985-86 technical report* (No. 17-TR-20). Princeton, NJ: Educational Testing Service.

Bock, R. D., Mislevy, R. J., & Woodson, C. E. M. (1982). The next stage in educational assessment. *Educational Researcher, 11*(2), 4–11, 16.

Cassel, C-M, Särndal, C-E., & Wretman, J. H. (1977). *Foundations of inference in survey sampling.* New York: Wiley.

Cochran, W. G. (1977). *Sampling techniques.* New York: Wiley.

Deely, J. J., & Lindley, D. V. (1981). Bayes empirical Bayes. *Journal of the American Statistical Association, 76,* 833–841.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, 39,* 1–38.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Hansen, M. H., Hurwitz, W. N., & Bershad, M. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute, 38,* 359–374.

Johnson, E. G., & Zwick, R. (1990). *Focusing the new design: The NAEP 1988 technical report* (No. 19-TR-20). Princeton, NJ: Educational Testing Service.

Kirsch, I. S., & Jungeblut, A. (1986). *Literacy: Profiles of America's young adults.* Princeton, NJ: National Assessment of Educational Progress/Educational Testing Service.

Laird, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association, 73,* 805–811.

Lord, F. M. (1962). Estimating norms by item-sampling. *Educational and Psychological Measurement, 22,* 259–267.

Lord, F. M. (1969). Estimating true score distributions in psychological testing (An empirical Bayes problem). *Psychometrika, 34,* 259–299.

Messick, S., Beaton, A. E., & Lord, F. M. (1983). *National Assessment of Educational Progress reconsidered: A new design for a new era* (Report No. 83-1). Princeton, NJ: National Assessment for Educational Progress.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49,* 359–381.

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80,* 993–997.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56,* 177–196.

Mislevy, R. J., & Bock, R. D. (1983). *BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville, IN: Scientific Software.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17,* 131–154.

Muthén, B., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics, 10,* 133–142.

Rigdon, S. E., & Tsutakawa, R. K. (1983). Parameter estimation in latent trait models. *Psychometrika, 48,* 567–574.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Sheehan, K. M. (1985). *M-GROUP: Estimation of group effects in multivariate models* [Computer program]. Princeton, NJ: Educational Testing Service.

Sirotnik, K., & Wellington, R. (1977). Incidence sampling: An integrated theory for matrix sampling. *Journal of Educational Measurement, 14,* 343–399.

Stein, C. (1956). Inadmissibility of the usual estimator of the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium of Mathematical Statistics and Probability, Vol. 1* (pp. 197–206). Berkeley: University of California Press.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity.* Hillsdale, NJ: Erlbaum.

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

Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement, 24,* 293–308.

Authors

ROBERT J. MISLEVY is Principal Research Scientist, Educational Testing Service, Rosedale Rd., Princeton, NJ 08541. *Degrees:* BS, MS, Northern Illinois University; PhD, University of Chicago. *Specializations:* educational testing, large-scale assessment, and Bayesian inference.

ALBERT E. BEATON is Professor, Boston College, School of Education, Campion Hall 320, Chestnut Hill, MA 02167. *Degree:* EdD, Harvard University. *Specializations:* statistics, psychometrics, and data analysis.

BRUCE KAPLAN is Advanced Research Systems Specialist, Educational Testing Service, 18T, Rosedale Rd., Princeton, NJ 08541. *Degrees:* BS, SUNY at Stony Brook; MS, Cornell University. *Specializations:* statistical computing and sampling.

KATHLEEN M. SHEEHAN is Advanced Research Systems Specialist, Educational Testing Service, 03-T, Rosedale Rd., Princeton, NJ 08541. *Degree:* MS, West Virginia University. *Specialization:* item response theory.