PSYCHOMETRIKA—VOL. 56, NO. 2, 177–196
JUNE 1991

RANDOMIZATION-BASED INFERENCE ABOUT LATENT VARIABLES FROM COMPLEX SAMPLES

ROBERT J. MISLEVY

EDUCATIONAL TESTING SERVICE

Standard procedures for drawing inferences from complex samples do not apply when the variable of interest θ cannot be observed directly, but must be inferred from the values of secondary random variables that depend on θ stochastically. Examples are proficiency variables in item response models and class memberships in latent class models. Rubin's "multiple imputation" techniques yield approximations of sample statistics that would have been obtained, had θ been observable, and associated variance estimates that account for uncertainty due to both the sampling of respondents and the latent nature of θ. The approach is illustrated with data from the National Assessment for Educational Progress.

Key words: complex samples, item response theory, latent structure, missing data, multiple imputation, National Assessment of Educational Progress, sample surveys.

Introduction

Latent variable models can provide parsimonious explanations of associations among observed variables in terms of theoretical constructs. Using latent variable models, one might compare examinees who have taken different tests, or track consumer satisfaction over time with different survey questions. This paper concerns the estimation of distributions of latent variables in finite populations, when the data are obtained in complex sampling designs. The solution proposed here is based on Rubin's (1987) "multiple imputation" procedures for missing data. This approach provides consistent estimates of population characteristics, statements of precision that account for both the sampling of subjects and the latent nature of variables, and filled-in datasets that are easy for secondary researchers to analyze correctly.

The following section reviews randomization-based inference from complex samples, including the handling of missing responses. These ideas are then applied to latent variable models and illustrated in the context of classical test theory. Computing approximations are discussed. The paper concludes by sketching the implementation, the results, and some lessons learned in applying the procedures to the 1984 reading survey of the National Assessment of Educational Progress (NAEP).

This research was supported by Grant No. NIE-G-83-0011 of the Office for Educational Research and Improvement, Center for Education Statistics, and Contract No. N00014-88-K-0304, R&T 4421552 from the Cognitive Sciences Program, Cognitive and Neural Sciences Division, Office of Naval Research. It does not necessarily reflect the views of either agency. I am grateful to R. Darrell Bock for calling my attention to the applicability of multiple imputation to the assessment setting; to Albert Beaton and Eugene Johnson for enlightening discussions on the topic; and to Henry Braun, Ben King, Debra Kline, Gary Phillips, Paul Rosenbaum, Don Rubin, John Tukey, Ming-Mei Wang, Kentaro Yamamoto, Rebecca Zwick, and two anonymous reviewers for comments on earlier drafts. Example 4 is based on the analysis of the 1984 National Assessment for Educational Progress reading survey, carried out at Educational Testing Service through the tireless efforts of too many people to mention by name, under the direction of Albert Beaton, Director of NAEP Data Analyses. David Freund, Bruce Kaplan, and Jennifer Nelson conducted additional analyses of the 1984 and 1988 data for the example.

Requests for reprints should be sent to Robert J. Mislevy, Educational Testing Service, Princeton, NJ 08541.

0033-3123/91/0600-88030$00.75/0

Inference from Complex Samples

Standard methods for sample survey data (e.g., Cochran, 1977) provide estimates of population characteristics, and associated sampling variances, when the values of survey variables are observed from each respondent in the realized sample. Consider a population of N identifiable units, indexed by i. Associated with each unit are two possibly vector-valued variables $\mathbf{y}$ and $\mathbf{z}$. The values of the design variables, $\mathbf{z}$, are known for all units before observations are made, but the values of the survey variables, y, are not. Let $(\mathcal{Y}, \mathcal{Z})$ denote the population matrix of values. Interest lies in the value of a function $S \equiv S(\mathcal{Y}, \mathcal{Z})$ of the population values. A sample design assigns a probability to each of the 2^N possible subsets of units. One subset is selected accordingly, and the y-values of these units, say $\mathbf{Y}$, are ascertained. The result is a "complex sample" if the sample design has at least one of the following attributes: unequal probabilities of selection for different units; stratification, which ensures prespecified rates of representations for particular values of $\mathbf{z}$; or clustering, which uses values of $\mathbf{z}$ to link selection probabilities of units when their joint occurrence facilitates data collection.

Randomization-based inference about S is based on the distribution of a statistic $s \equiv s(\mathbf{Y}, \mathcal{Z})$ in repeated samples of $\mathbf{Y}$ under a specified sample design. Sample designs are usually constructed so as to support an (at least approximately) unbiased statistic s and an estimate of its variance in the form of another statistic $U \equiv U(\mathbf{Y}, \mathcal{Z})$. Complex design features such as clustering and stratification are reflected in the calculation of both s and U. Inferences are typically based on the normal approximation

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

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

Under the randomization approach, population values $\mathcal{Y}$ are considered fixed unknown quantities, and the notion of randomness enters only in the selection of a sample in accordance with the sample design. Under the contrasting model-based approach, $\mathcal{Y}$ itself is viewed as a sample from a hypothetical "superpopulation" in which variables are distributed according to some model $p(\mathbf{y}|\mathbf{z})$ (Cassel, Särndal, & Wretman, 1977). The randomization approach dominates current practice, and will be used here to deal with uncertainty due to sampling subjects. It will be seen that superpopulation concepts are required nonetheless to handle missingness and latency.

Often in practice, values of one or more survey variables are not observed from some subjects in the realized sample, for reasons that may or may not be related to their values. Let the partitioning $\mathbf{y}' = (\mathbf{y}'_{\text{mis}}, \mathbf{y}'_{\text{obs}})$ indicate the elements of a respondent's survey variables that were missing and observed, and let $(\mathbf{Y}_{\text{mis}}, \mathbf{Y}_{\text{obs}})$ denote the matrices of missing and observed survey variables in a realized sample. It is no longer possible to calculate s directly, but Rubin (1977) points out that it may be possible to calculate its conditional expectation:

$$E[s(\mathbf{Y})|\mathbf{Y}_{\text{obs}}] = \int s(\mathbf{Y}_{\text{mis}}, \mathbf{Y}_{\text{obs}}) p(\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}, \mathcal{Z}) d\mathbf{Y}_{\text{mis}}. \tag{1}$$

The predictive density $p(\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}, \mathcal{Z})$ expresses what is known about what the missing responses might have been, given the observed responses and the survey variables. This distribution lies outside the framework of randomization-based survey inference.

It can be introduced by viewing the finite population as a sample from an infinite superpopulation with a particular structure, or erecting a Bayesian exchangeability structure to relate sampled units' observed and missing responses.

Missing responses are *missing at random* (MAR) when the probability that the elements in $\mathbf{y}_{mis}$ are missing does not depend on the value of $\mathbf{y}_{mis}$, given the values of $\mathbf{y}_{obs}$ and $\mathbf{z}$ (Little & Rubin, 1987). If MAR holds, the predictive distribution of a missing element for Subject i—say $\mathbf{y}_{i,mis}$—is the same as the distribution of responses to that element from subjects who did respond to it and have the same values as Subject i for $\mathbf{z}$ and the elements in $\mathbf{y}_{i,obs}$. In large surveys with few missing values, one can use empirical distributions directly. The Census Bureau's "hot deck" procedure (Ford, 1983) calculates s using each person's observed responses and, for those he/she is missing, random draws from the actual responses of suitably similar respondents. Alternatively, one can posit a functional form for $p(\mathbf{y}_{mis}|\mathbf{y}_{obs}, \mathbf{z})$ and estimate its parameters from the observed data. An approximation of (1)—an unbiased estimate of the expectation of s—can be obtained in either case by filling in each missing response with a random draw from its predictive density, then calculating s as if the imputations were observed responses.

An analogous approximation of U from these single imputations underestimates the variability of the resulting estimate of S, however. It accounts for uncertainty due to sampling subjects, but not uncertainty due to imputing values for missing responses. To remedy this deficiency, Rubin (1987) suggests *multiple* imputations. As will be described below in the context of latent variables, the variance among repeated evaluations of s obtained with different imputations reflects uncertainty from this latter source.

Latent Variables in Sample Surveys

Latent variables are posited to account for regularities in observable variables, such as examinees' tendencies to give correct responses to test items. The probability that a subject with latent parameter θ will make the response x_j to item j depends on θ and a (possibly vector-valued) item parameter β_j, as $p(x_j|\theta, \beta_j)$. Under the assumption of conditional independence, the latent variable accounts for all associations among responses to various items in a specified domain. Moreover, the latent variable is typically assumed to account for associations between response variables and collateral variables such as demographic or educational standing. Denoting collateral variables by $\mathbf{y}$ and $\mathbf{z}$ to anticipate developments below, and letting $\mathbf{x}' = (x_1, \ldots, x_n)$ be a vector of responses to n items, conditional independence states

$$p(\mathbf{x}|\theta, \boldsymbol{\beta}, \mathbf{y}, \mathbf{z}) = p(\mathbf{x}|\theta, \boldsymbol{\beta}) = \prod_{j=1}^{n} p(x_j|\theta, \beta_j), \tag{2}$$

where $\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_n)$.

Under such a model, the responses to any subset of items induce a likelihood function for θ via (2). If the focus is on measuring individuals for placement or selection decisions, one administers enough items to each respondent to make the likelihood function for his or her θ peak sharply. A precise point estimate of each θ, such as the MLE $\hat{\theta}$ or the Bayes mean estimate $\bar{\theta} = E(\theta|\mathbf{x})$, can be obtained under these circumstances. If the focus is on the parameters $\boldsymbol{\alpha}$ of a population distribution $p(\theta|\boldsymbol{\alpha})$ of θ, however, point estimates that are optimal for decisions about individuals can be seri-

ously misleading. The distributions of $\hat{\theta}$ and $\bar{\theta}$ from a fixed test do not converge to the distribution of θ as the size of the examinee sample increases (Lord, 1965, 1969).

If the data consisted of response vectors $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ of a simple random sample (SRS) of respondents from a very large population, the starting point for estimating $\boldsymbol{\alpha}$ directly—that is, not from point estimates of θ—would be the marginal probability

$$p(\mathbf{X}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{N} \int p(\mathbf{x}_i|\theta, \boldsymbol{\beta}) p(\theta|\boldsymbol{\alpha})\, d\theta. \tag{3}$$

Maximum likelihood and Bayesian estimation of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, or both via (3) have been discussed by Andersen and Madsen (1977), Bock and Aitkin (1981), Dempster, Laird, and Rubin (1977), Laird (1978), Mislevy (1984), Sanathanan and Blumenthal (1978), and Rigdon and Tsutakawa (1983). As presented, the methods are not well suited to general analyses of survey data. As well as being limited to SRS data, they require statistical concepts and computing methodologies that are beyond the typical secondary analyst.

A key insight for dealing with latent variables in sample surveys is to view them as survey variables whose responses are missing for every respondent. Because they are missing regardless of their values, they satisfy MAR and are thus amenable to the MAR missing data procedures described above. Knowledge about subjects' latent variables θ can be conveyed by predictive distribution given the observed data, namely the design variables $\mathbf{z}$, background survey variables $\mathbf{y}$, and item responses $\mathbf{x}$.

Let the object of inference be the scalar $S \equiv S(\boldsymbol{\Theta}, \mathcal{X}, \mathcal{Y}, \mathcal{Z})$, some function of the population values of latent variables, item responses, background survey variables, and design variables. Consider $\mathcal{Z}$ fixed and known, and suppose that a design has been specified to sample subjects whose $\mathbf{x}$ and $\mathbf{y}$ values will be ascertained. Inference about S can be based on three building blocks. The first is that one could carry out randomization-based inference about S if θ values could be observed:

The sampling model. Assume that if values of θ were observed from sampled subjects along with $\mathbf{x}$ and $\mathbf{y}$, randomization-based inference about S could be based on a sample statistic $s \equiv s(\boldsymbol{\theta}, \mathbf{X}, \mathbf{Y}, \mathcal{Z})$ and a variance estimator $U \equiv U(\boldsymbol{\theta}, \mathbf{X}, \mathbf{Y}, \mathcal{Z})$ via $(s - S)/\sqrt{U} \sim N(0, 1)$. (The dependence of s and U on $\mathcal{Z}$ will again be implicit below.)

One cannot calculate s directly because all values of θ are missing. One can, however, base inferences on the conditional distribution of s given what is known, namely, sample values $\mathbf{X}$ and $\mathbf{Y}$ and population values $\mathcal{Z}$. Its conditional expectation, for example, is obtained by adapting (1) as follows:

$$E(s|\mathbf{X}, \mathbf{Y}, \mathcal{Z}) = \int s(\boldsymbol{\theta}, \mathbf{X}, \mathbf{Y}) p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}, \mathcal{Z})\, d\boldsymbol{\theta}. \tag{4}$$

The remaining two building blocks are needed to define the predictive distribution $p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}, \mathcal{Z})$. They are a population model for θ given $\mathbf{y}$ and $\mathbf{z}$, and a latent variable model for $\mathbf{x}$ given θ. The fixed but unknown θ values in the finite population are treated as draws from a θ superpopulation, justifying the estimation of parameters of their distribution. The observed $\mathbf{x}$ values are treated as draws from superpopulation of responses from respondents with their fixed but unknown θ values, justifying the estimation of parameters of the latent variable model.

The population model. Assume that the distribution of latent variables θ given collateral survey variables $\mathbf{y}$ and design variables $\mathbf{z}$ follows a known functional form,

$p(\theta|\mathbf{y}, \mathbf{z}, \boldsymbol{\alpha})$, characterized by possibly unknown parameters $\boldsymbol{\alpha}$. Additionally assuming independence over subjects,

$$p(\boldsymbol{\theta}|\mathbf{Y}, \mathfrak{X}, \boldsymbol{\alpha}) = \prod_{i}^{N} p(\theta_i|\mathbf{y}_i, \mathbf{z}_i, \boldsymbol{\alpha}). \tag{5}$$

The latent variable model. Assume that item responses $\mathbf{x}$ are governed by the latent variable θ through a model of known functional form, $p(\mathbf{x}|\theta, \boldsymbol{\beta})$, characterized by parameters $\boldsymbol{\beta}$ and satisfying local independence as in (2). Additionally assuming independence over subjects,

$$p(\mathbf{X}|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{Y}, \mathfrak{X}) = p(\mathbf{X}|\boldsymbol{\theta}, \boldsymbol{\beta}) = \prod_{i}^{N} p(\mathbf{x}_i|\theta_i, \boldsymbol{\beta}). \tag{6}$$

To see how the population and latent variable models lead to $p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}, \mathfrak{X})$, first consider relationships conditional on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Using Bayes theorem, then (2), (5), and (6), gives

$$p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}, \mathfrak{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = K_{\alpha\beta} p(\mathbf{X}|\boldsymbol{\theta}, \mathbf{Y}, \mathfrak{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}) p(\boldsymbol{\theta}|\mathbf{Y}, \mathfrak{X}, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

$$= \prod_{i}^{N} K_{i\alpha\beta} p(\mathbf{x}_i|\theta_i, \boldsymbol{\beta}) p(\theta_i|\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\alpha}),$$

where the normalizing constants $K_{i\alpha\beta} = 1/p(\mathbf{x}_i|\mathbf{y}_i, \mathbf{z}_i, \boldsymbol{\alpha}, \boldsymbol{\beta})$ depend on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ but not on θ_i. Given $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, then, the predictive distribution for the latent variable θ_i of subject i, or $p(\theta_i|\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i, \boldsymbol{\alpha}, \boldsymbol{\beta})$, is obtained by normalizing the product of (i) the likelihood function for θ induced by $\mathbf{x}_i$ via the latent variable model (6), and (ii) the conditional distribution for θ implied by his or her background and design variables $\mathbf{y}_i$ and $\mathbf{z}_i$.

The preceding paragraphs set the stage for randomization-based inference with latent variables in sample surveys. To carry out an analysis with multiple imputations, one implements Rubin's (1987) procedures as follows:

1. Obtain the posterior distribution of the parameters $\boldsymbol{\beta}$ of the latent variable model and $\boldsymbol{\alpha}$ of the conditional distributions of θ, or $p(\boldsymbol{\alpha}, \boldsymbol{\beta}|\mathbf{X}, \mathbf{Y}, \mathfrak{X})$, by the methods discussed in connection with (3)—for example, a large sample normal approximation based on the MLE $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ and asymptotic covariance matrix $\boldsymbol{\Sigma}_{\alpha\beta}$.

2. Produce M "completed" datasets $(\boldsymbol{\theta}_{(m)}, \mathbf{X}, \mathbf{Y})$. For the m-th,

   a. Draw a value $(\boldsymbol{\alpha}, \boldsymbol{\beta})_{(m)}$ from $p(\boldsymbol{\alpha}, \boldsymbol{\beta}|\mathbf{X}, \mathbf{Y}, \mathfrak{X})$. (If $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ have been estimated very precisely, it may be acceptable to use $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ for each completed dataset—an "empirical Bayes" approximation that introduces a tendency to underestimate the uncertainty of final estimates of S.)

   b. For each respondent, draw a value from the predictive distribution $p(\theta|\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i, (\boldsymbol{\alpha}, \boldsymbol{\beta}) = (\boldsymbol{\alpha}, \boldsymbol{\beta})_{(m)})$. The resulting imputation for θ_i and the observed values $\mathbf{x}_i$, $\mathbf{y}_i$, and $\mathbf{z}_i$ are a "completed" response for subject i.

3. Using each completed dataset, calculate $s_{(m)} \equiv s(\boldsymbol{\theta}_{(m)}, \mathbf{X}, \mathbf{Y})$ and $U_{(m)} \equiv U(\boldsymbol{\theta}_{(m)}, \mathbf{X}, \mathbf{Y})$.

4. The final estimate of S is a numerical approximation of (4), the average of the M estimates from the completed datasets:

$$s_M = \frac{\sum s_{(m)}}{M}. \tag{7}$$

5. The variance of s_M, namely V_M, is the sum of two components:

$$V_M = U_M + (1 + M^{-1})B_M, \tag{8}$$

where

$$U_M = \frac{\sum U_{(m)}}{M}$$

quantifies uncertainty due to sampling subjects, and

$$B_M = \frac{\sum (s_{(m)} - s_M)^2}{(M - 1),}$$

the variance among the estimates of s from the M completed datasets, quantifies uncertainty due to not observing θ from the sampled subjects. From the randomization perspective, V_M approximates the variance of s_M around S in repeated samples of (**X**, **Y**). From the Bayesian perspective, it approximates the posterior variance of s_M conditional on the realized sample.

6. Inferences about S are based on

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

a t-distribution with degrees of freedom given by

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

here r_M is the proportional increase in variance due to the latent nature of θ:

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

When B_M is small relative to U_M, ν is large and the normal approximation to the t-distribution suffices.

For k-dimensional **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

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

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

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

As with the normal approximation for scalar s, a chi-square distribution on k degrees of freedom suffices when r_M is small.

Steps 1 and 2 above produce M completed datasets that can be used to draw inferences about any number of statistics by applying Steps 3–6 repeatedly. The so-

phisticated methodologies and heavy computation are isolated in Steps 1 and 2, which can be carried out just once—probably by the institution held responsible for primary data analysis, where the expertise and resources are more likely to be available. The completed datasets are then provided to secondary researchers, who need only apply standard routines for complete data M times and combine the results in simple ways.

Example 1: Classical Test Theory, Fixed Effects

This example lays out imputation procedures for a true-score test model, in the presence of two collateral survey variables. The same analysis applies to stratification design variables. To focus on the construction and the nature of imputations, we assume that a large SRS will be drawn from a large multivariate normal population with known parameters. There are four variables for each subject: θ, the latent variable, is "true score"; x is the observed score; and $\mathbf{y}' = (y_1, y_2)$ are collateral variables.

Suppose that $x = \theta + e$, where the residual or error term e is distributed $N(0, \sigma_e^2)$ independently of θ and $\mathbf{y}$, so that

$$x|\theta \sim N(\theta, \sigma_e^2). \tag{9}$$

Suppose also that $(\theta, \mathbf{y}')$ follows a standard model distribution in the population, so that jointly, $(x, \theta, y_1, y_2) \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 + \sigma_e^2 & & & \\ & & \text{(symmetric)} & \\ 1 & 1 & & \\ r_{\theta 1} & r_{\theta 1} & 1 & \\ r_{\theta 2} & r_{\theta 2} & r_{12} & 1 \end{bmatrix}.$$

The conditional distribution of θ given $\mathbf{y}$ is

$$\theta|\mathbf{y} \sim N(\boldsymbol{\beta}'\mathbf{y}, \sigma_{\theta|\mathbf{y}}^2), \tag{10}$$

where

$$\boldsymbol{\beta}'\mathbf{y} = E(\theta|\mathbf{y}) = \beta_{1|2} y_1 + \beta_{2|1} y_2,$$

with $\beta_{1|2} \equiv (r_{\theta 1} - r_{\theta 2} r_{12})/(1 - r_{12}^2)$ and $\beta_{2|1}$ defined similarly, and

$$\sigma_{\theta|\mathbf{y}}^2 \equiv \text{Var}(\theta|\mathbf{y}) = 1 - R^2,$$

with R^2 the proportion of variance of θ accounted for by $\mathbf{y}$. Define ρ_c, the "conditional reliability" of x given $\mathbf{y}$, as

$$\rho_c \equiv \frac{\sigma_{\theta|\mathbf{y}}^2}{\sigma_{\theta|\mathbf{y}}^2 + \sigma_e^2}.$$

Values of x and $\mathbf{y}$ are observed from a sampled subject. An imputation is to be drawn from the predictive distribution for θ, or $p(\theta|x, \mathbf{y}) \propto p(x|\theta)\, p(\theta|\mathbf{y})$. The first factor is the likelihood, $n(\theta; x, \sigma_e^2)$; the second is the conditional density $n(\theta; \boldsymbol{\beta}'\mathbf{y}, \sigma_{\theta|\mathbf{y}}^2)$. Thus,

$$\theta|x, \mathbf{y} \sim N(\bar{\theta}, \sigma_{\theta|x\mathbf{y}}^2),$$

TABLE 1

Expectations of Secondary Analyses

Population Attribute	Dependent Variable			
	θ and $\tilde{\theta}$	$\hat{\theta}$	$\bar{\theta}_{xy}$	$\bar{\theta}_{x}$
Mean	0	0	0	0
Variance	1	$1+\sigma_e^2$	$1-\sigma_{\theta\|xy}^2$	$1-\sigma_{\theta\|x}^2 = \rho$
Simple Regression				
Coefficient	$r_{\theta 1}$	$r_{\theta 1}$	$r_{\theta 1}$	$\rho\, r_{\theta 1}$
Residual Variance	$\sigma_{\theta\|y_1}^2$	$\sigma_{\theta\|y_1}^2+\sigma_e^2$	$\sigma_{\theta\|y_1}^2-\sigma_{\theta\|xy}^2$	$\rho^2(\sigma_{\theta\|y_1}^2+\sigma_e^2)$
% Variance Accounted for	$1-\sigma_{\theta\|y_1}^2$	$\dfrac{1-\sigma_{\theta\|y_1}^2}{1+\sigma_e^2}$	$\dfrac{1-\sigma_{\theta\|y_1}^2}{1-\sigma_{\theta\|xy}^2}$	$\rho^2(\sigma_{\theta\|y_1}^2+\sigma_e^2)$

where, as in Kelley (1923, pp. 212–216),

$$\bar{\theta} \equiv E(\theta|x, \mathbf{y}) = \rho_c x + (1 - \rho_c)\boldsymbol{\beta}'\mathbf{y},$$

and

$$\sigma_{\theta|xy}^2 \equiv \text{Var}\,(\theta|x, \mathbf{y}) = (1 - \rho_c)\sigma_{\theta|\mathbf{y}}^2 = (1 - \rho_c)(1 - R^2).$$

An imputation $\bar{\theta} \equiv \bar{\theta}(x, \mathbf{y})$ can thus be constructed as

$$\bar{\theta} = \bar{\theta} + f,$$

where f is *drawn at random* from $N(0, \sigma_{\theta|xy}^2)$.

For a given individual, an imputation is neither unbiased nor efficient, as is the MLE $\hat{\theta} = x$; nor does it minimize mean square error over the population, as does $\bar{\theta}$. But it can be shown that the distribution of $(\bar{\theta}, \mathbf{y})$ is multivariate normal with the same mean vector and covariance matrix as that of $(\theta, \mathbf{y})$. Treating $\bar{\theta}$ as θ thus gives the correct expectations for estimates of all population attributes such as means, variances, covariances, and regression coefficients. As shown in Table 1, however, treating either $\hat{\theta}$ or $\bar{\theta}$ as θ leads to estimates that have the correct expectations for some attributes, but not for others. The Bayes estimate referred to above as $\bar{\theta}$ is denoted $\bar{\theta}_{xy}$ in the table to emphasize that it is conditional on both x and $\mathbf{y}$. Another Bayes estimate often used in practice, $E(\theta|x)$, ignores $\mathbf{y}$; it is denoted as $\bar{\theta}_x$. Tables 2 and 3 give numerical values for the expressions in Table 1, as obtained from a test with $\sigma_e^2 = 1$ and a test with

TABLE 2

Numerical Values for a Short Assessment Instrument

Population Attribute	Dependent Variable			
	θ and $\bar{\theta}$	$\hat{\theta}$	$\bar{\theta}_{xy}$	$\bar{\theta}_{x}$
Mean	.000	.000	.000	.000
Variance	1.000	2.000	.600	.500
Simple Regression				
Coefficient	.500	.500	.500	.250
Residual Variance	.750	1.750	.350	.438
% Variance Accounted for	.250	.125	.417	.125

$\sigma_e^2 = .1$. The first test has a reliability of .5, typical of ten items on a topic that might appear on an educational assessment. The second has a reliability of .91, more like that of a 60-item achievement test. In both cases, $r_{\theta 1} = r_{\theta 2} = r_{12} = .50$.

Example 2: Classical Test Theory, Random Effects

Suppose that $x|\theta \sim N(\theta, \sigma_e^2)$ as in Example 1, but now the population comprises a large number of clusters indexed by k, and z indicates the cluster to which a subject belongs. Assume normality for cluster means ν and for subjects within clusters:

$$\nu \sim N(\mu, \sigma_b^2) \text{ and } \theta|(z = k) \sim N(\nu_k, \sigma_w^2).$$

Suppose that a sample of clusters is drawn, a sample of m subjects is drawn from each selected cluster, and a value of x is observed from each sampled subject. Assuming for the moment that the structural parameters μ, σ_e^2, σ_b^2, and σ_w^2 are known, the posterior distribution for subject i of cluster k *conditional on* ν_k is

$$\theta_{ik}|(x_{ik}, \nu_k) \sim N[\rho x_{ik} + (1 - \rho)\nu_k, (1 - \rho)\sigma_w^2],$$

where $\rho \equiv \sigma_w^2/(\sigma_w^2 + \sigma_e^2)$. The posterior distribution for ν_k is

$$\nu_k|(x_{1k}, \ldots, x_{mk}) \sim N[\lambda \bar{x}_k + (1 - \lambda)\mu, (1 - \lambda)\sigma_b^2],$$

TABLE 3

Numerical Values for a Long Test

Population Attribute	Dependent Variable			
	θ and $\tilde{\theta}$	$\hat{\theta}$	$\bar{\theta}_{xy}$	$\bar{\theta}_{x}$
Mean	.000	.000	.000	.000
Variance	1.000	1.100	.940	.910
Simple Regression				
Coefficient	.500	.500	.500	.455
Residual Variance	.750	.850	.690	.703
% Variance Accounted for	.250	.227	.266	.227

where $\bar{x}_k$ is the mean of the x values observed from cluster k and $\lambda \equiv \sigma_b^2/[\sigma_b^2 + (\sigma_w^2 + \sigma_e^2)/m]$. An imputation for subject i from cluster k is then constructed as

$$\bar{\theta}_{ik} = \rho x_{ik} + (1 - \rho)[\lambda \bar{x}_k + (1 - \lambda)\mu + g_k] + f_{ik},$$

where f_{ik} is a draw from $N[0, (1 - \rho)\sigma_w^2]$ used uniquely with subject i, and g_k is a draw from $N[0, (1 - \lambda)\sigma_b^2]$ used with all subjects from cluster k. It can be verified that the expected mean of clusters and the within and between variance components of imputations constructed in this manner are μ, σ_w^2, and σ_b^2.

In practice, the values of μ, σ_e^2, σ_b^2, and σ_w^2 are not known, but must be estimated from the observed data. The steps in constructing imputations outlined above must be preceded by drawing values from their joint posterior distribution. Working with variants of the classical test theory latent variable model, Goldstein and McDonald (1988) and Muthén and Satorra (1989) give large-sample estimates of structural parameters in designs with both random-effect and fixed-effect collateral variables. Although Monte Carlo approaches such as that of Tanner and Wong (1987) hold promise of true Bayesian solutions with general latent variable models, procedures are currently less well-developed for random effects than for fixed effects. Treating clusters as fixed effects in the present example is a good approximation to the proper random-effects treatment if m is large, because λ approaches unity and the variance of g_k approaches zero.

Computing Approximations and Secondary Biases

The imputation-based estimate $s_M(\mathbf{X}, \mathbf{Y})$ approximates the expectation of $s(\boldsymbol{\theta}, \mathbf{X}, \mathbf{Y})$ defined in (4) by evaluating s with draws $\tilde{\boldsymbol{\theta}}$ from the predictive distribution

$$p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}, \mathfrak{X}) \propto \int\int p(\mathbf{X}|\boldsymbol{\theta}, \boldsymbol{\beta}) p(\boldsymbol{\theta}|\mathbf{Y}, \mathfrak{X}, \boldsymbol{\alpha}) p(\boldsymbol{\alpha}, \boldsymbol{\beta}|\mathbf{X}, \mathbf{Y}, \mathfrak{X})\, d\boldsymbol{\alpha}\, d\boldsymbol{\beta}$$

$$= \int\int \prod_i p(\mathbf{x}_i|\theta_i, \boldsymbol{\beta}) p(\theta_i|\mathbf{y}_i, \mathbf{z}_i, \boldsymbol{\alpha}) p(\boldsymbol{\alpha}, \boldsymbol{\beta}|\mathbf{X}, \mathbf{Y}, \mathfrak{X})\, d\boldsymbol{\alpha}\, d\boldsymbol{\beta}.$$

Obtaining consistent estimates of population attributes requires drawing imputations from consistent estimates of $p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}, \mathfrak{X})$. As noted above, however, procedures for characterizing $p(\theta|\mathbf{y}, \mathbf{z}, \boldsymbol{\alpha})$ are neither numerous nor easily applied. If the dimensionalities of $\mathbf{y}$ or $\mathbf{z}$ are high, or if the sampling design involves multiple levels of clustering, then simplifications and computing approximations cannot be avoided. The correct conditional distribution is replaced by the computing approximation $p^*(\theta|\mathbf{y}, \mathbf{z}, \boldsymbol{\alpha}^*)$, which, when combined with the latent variable model $p(\mathbf{x}|\theta, \boldsymbol{\beta})$, yields the computing approximation from which imputations $\tilde{\boldsymbol{\theta}}^*$ are drawn:

$$p^*(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}, \mathfrak{X}) \propto \int\int p(\mathbf{X}|\boldsymbol{\theta}, \boldsymbol{\beta}) p^*(\boldsymbol{\theta}|\mathbf{Y}, \mathfrak{X}, \boldsymbol{\alpha}^*) p(\boldsymbol{\alpha}^*, \boldsymbol{\beta}|\mathbf{X}, \mathbf{Y}, \mathfrak{X})\, d\boldsymbol{\alpha}^*\, d\boldsymbol{\beta}$$

$$= \int\int \prod_i p(\mathbf{x}_i|\theta_i, \boldsymbol{\beta}) p^*(\theta_i|\mathbf{y}_i, \mathbf{z}_i, \boldsymbol{\alpha}^*) p(\boldsymbol{\beta}^*, \boldsymbol{\beta}|\mathbf{X}, \mathbf{Y}, \mathfrak{X})\, d\boldsymbol{\alpha}^*\, d\boldsymbol{\beta}.$$

While the expectation of s_M based on imputations $\tilde{\boldsymbol{\theta}}$ is the correct value s, that of the expectation of s^*_M based on imputations $\tilde{\boldsymbol{\theta}}^*$ may not be. Its expectation is

$$E(s^*_M) = \int s(\boldsymbol{\theta}, \mathbf{X}, \mathbf{Y}) p^*(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}, \mathfrak{X})\, d\boldsymbol{\theta}.$$

For a fixed observed sample $(\mathbf{X}, \mathbf{Y})$, the bias in estimating s caused by using $p^*(\theta|\mathbf{y}, \mathbf{z}, \boldsymbol{\alpha}^*)$ rather than $p(\theta|\mathbf{y}, \mathbf{z}, \boldsymbol{\alpha})$ is thus

$$\text{Bias} = E(s^*_M|\mathbf{X}, \mathbf{Y}) - E(s|\mathbf{X}, \mathbf{Y})$$

$$= \int s(\boldsymbol{\theta}, \mathbf{X}, \mathbf{Y})[p^*(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}, \mathfrak{X}) - p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}, \mathfrak{X})]\, d\boldsymbol{\theta}.$$

The adjective "secondary" distinguishes these biases from the usual ones that could occur in analyses of true θ values.

Example 3: Secondary Biases in Regression Coefficients

Consider again the setup from Example 1, with the latent variable model $x|\theta \sim N(\theta, \sigma_e^2)$ and population model $\theta|\mathbf{y} \sim N(\boldsymbol{\beta}'\mathbf{y}, \sigma^2_{\theta|\mathbf{y}})$. The correct distribution for imputation is

$$p(\theta|x, \mathbf{y}) = n[\theta; \rho_c x + (1 - \rho_c)\boldsymbol{\beta}'\mathbf{y}, \sigma^2_{\theta|x\mathbf{y}}],$$

but suppose the imputer instead draws imputations $\tilde{\theta}^*$ from a predictive distribution that is conditional on y_1 but not y_2:

$$p^*(\theta|x,\mathbf{y}) = n[\theta;\rho_c^* x + (1-\rho_c^*)\beta_1 y_1,\ \sigma^2_{\theta|xy_1}],$$

where ρ_c^* is the conditional reliability of x given y_1 only, or

$$\rho_c^* = \frac{\sigma^2_{\theta|y_1}}{\sigma^2_{\theta|y_1} + \sigma_e^2} = \frac{1 - r^2_{\theta 1}}{1 - r^2_{\theta 1} + \sigma_e^2},$$

and

$$\sigma^2_{\theta|xy_1} = \text{Var}\ (\theta|x,\ y_1) = (1-\rho_c^*)\sigma^2_{\theta|y_1} = (1-\rho_c^*)(1-r^2_{\theta 1}).$$

It can be shown that $(\bar{\theta}^*,\ \mathbf{y})$ is normal with mean vector $\mathbf{0}$, as is $(\theta,\ \mathbf{y})$, and its covariance matrix agrees with that of $(\theta,\ \mathbf{y})$ for all elements except the one for $\bar{\theta}^*$ with y_2. Rather than $r_{\theta 2}$,

$$\text{Cov}\ (\bar{\theta}^*,\ y_2) \equiv r^*_{\theta 2} = \rho_c^* r_{\theta 2} + (1-\rho_c^*) r_{\theta 1} r_{12}.$$

The characteristics of the joint distribution of $\bar{\theta}^*$ and y_1 are identical to those of θ and y_1, including the regression coefficient and residual variance in the distribution of θ given y_1. Analyses involving y_2 do not fare as well. Whereas

$$E(\theta|\mathbf{y}) = E(\bar{\theta}|\mathbf{y}) = \beta_{1|2} y_1 + \beta_{2|1} y_2,$$

it is found that

$$E(\bar{\theta}^*|\mathbf{y}) = \beta_{1|2} y_1 + \beta_{2|1} y_2 - (1-\rho_c^*)\beta_{2|1}(y_2 - r_{12} y_1). \tag{11}$$

A bias is thus introduced into the imputations, which, as seen in (11), can be driven to zero in three ways:

1. As $\rho_c^* \to 1$; x is a perfectly reliable measure of θ;
2. As $\beta_{2|1} \to 0$; there is no contribution from y_2 anyway;
3. As $|r_{12}| \to 1$; y_2 is perfectly predictable from y_1.

Table 4 shows the impact on regression coefficients in secondary analyses. As noted above, secondary biases are observed for analyses involving θ and y_2, but not for analyses involving only θ and y_1. Biases in simple regression are smaller than those in multiple regression. If the bias of the *multiple* regression coefficient for a variable left out of the construction of $p^*(\theta|\mathbf{y},\ \mathbf{z})$ is, say, 30%, the bias of its *simple* regression coefficient is the 30% times $1 - r^2_{12}$, the proportion of its variance not accounted for by the survey variables that were included. Higher values of ρ^* reduce all biases that do exist. Higher values of r^2_{12} *mitigate* bias for the simple regression of θ on the excluded variable and for its coefficient in multiple regression, but they *exacerbate* the bias for the coefficient of the included variable in the multiple regression. Illustrative numerical values appear in Table 5. Evidence of such effects, and suggestions for avoiding them, are discussed in the following example.

Example 4: The 1984 NAEP Reading Assessment

During the 1983–84 school year, the National Assessment of Educational Progress (NAEP) surveyed the reading and writing skills of national probability samples of students at ages 9, 13, and 17, and in the modal grades associated with those ages 4, 8, and 11. Beaton (1987) details in assessment procedures and analyses. Highlights of the multiple-imputations procedures used in the analysis of the reading data follow.

Students were selected in a multistage sampling design, with counties or groups of

TABLE 4

Secondary Biases in Regression Coefficients

Coefficient for...	True Value	Expectation Under Incomplete Conditioning	Bias
Simple Regression			
y_1	$r_{\theta 1}$	$r_{\theta 1}$	0
y_2	$r_{\theta 2}$	$\rho_c^* r_{\theta 2} + (1-\rho_c^*) r_{\theta 1} r_{12}$	$(1-\rho_c^*)(1-r_{12}^2) \beta_{2\|1}$
Multiple Regression			
y_1	$\beta_{1\|2} = \frac{r_{\theta 1} - r_{\theta 2} r_{12}}{1-r_{12}^2}$	$\frac{r_{\theta 1} - r_{\theta 2}^* r_{12}}{1-r_{12}^2}$	$(1-\rho_c^*) \beta_{2\|1} r_{12}$
y_2	$\beta_{2\|1}$	$\rho_c^* \beta_{2\|1}$	$-(1-\rho_c^*) \beta_{2\|1}$

Notes: Imputations constructed by conditioning on y_1 but not y_2.

$r_{\theta 2}^* = \rho_c^* r_{\theta 2} + (1-\rho_c^*) r_{\theta 1} r_{12}$.

counties as the primary sampling units (PSUs). Schools were second-stage sampling units, and students within schools were the third. There were 64 PSUs in the sample, 1,465 schools, and, at each grade/age, about 20,000 students. Neglecting minor adjustments for nonresponse and poststratification, the design variables **z** were region of the country, size-and-type-of-community (STOC), PSU, and school membership. Population means and totals of survey variables were estimated by weighted sample means and totals. A multiweight jackknife procedure was used to calculate U, the estimated variance of a statistic due to sampling students.

Each student responded to survey items **y** on demographic status, educational experiences, and attitudes about reading and writing. About 50 were common to all assessment forms. Another 300, of which a student received from 10 to 30 under a balanced incomplete block (BIB) item-sampling design, addressed reading activities at home and school.

A total of 340 multiple-choice and free-response reading exercises were used in the assessment. A student who received any reading exercises received from 5 to 50 of them under the BIB design. A priori considerations and dimensionality analyses supported summarizing responses to 228 of the 340 items by the 3 parameter logistic (3PL) IRT model, with its single latent proficiency variable θ (Zwick, 1987). Responses to these items will be denoted by **x**. The 3PL gives the probability of a correct response to Item j from a student with proficiency θ as

TABLE 5

Expected Regression Coefficients for a Short and a Long Test, with Complete and Incomplete Conditioning for Imputation

Population Attribute	Dependent Variable		
	θ and $\tilde{\theta}$	$\tilde{\theta}^*$ (ρ=.50)	$\tilde{\theta}^*$ (ρ=.91)
Simple Regression			
$r_{\theta 1}$	.500	.500	.500
$r_{\theta 2}$	.500	.357	.471
Multiple Regression			
$\beta_{1\|2}$	.333	.429	.353
$\beta_{2\|1}$	.333	.143	.294

Note: Imputations constructed by conditioning on y_1 but not y_2.

$$P(x_j = 1|\theta, a_j, b_j, c_j) = c_j + \frac{1 - c_j}{1 + \exp[-1.7a_j(\theta - b_j)]},$$

where $x = 1$ indicates a correct response and $x = 0$ an incorrect one, and the item parameters (a_j, b_j, c_j) give the regression of x_j on θ. This expression for a single response and conditional independence (2) give the likelihood function $p(\mathbf{x}|\theta, \boldsymbol{\beta})$ $[\equiv p(\mathbf{x}|\theta, \boldsymbol{\beta}, \mathbf{y}, \mathbf{z})]$ for responses to any subset of the 228 items in the scale. Here $\boldsymbol{\beta}$ denotes the vector of 228 × 3 item parameters.

Students receiving items in the reading scale received from 5 to 40 of them—about 17 on average. Item parameters were estimated from the responses of a sample of 10,000 students with Mislevy and Bock's (1983) BILOG computer program. These estimates were treated as known true values thereafter, with the origin and unit-size of the θ scale set so as to make the weighted mean and standard deviation of the combined samples 250.5 and 50.

The computer program available for estimating $p(\theta|\mathbf{y}, \mathbf{z})$ at that time was a prototype developed for Mislevy's (1985) 4-group example, with ANOVA-type structures on means and homoscedastic normal residuals. It was possible within NAEP's timelines to extend the program to a main effects model based on traditional NAEP reporting categories: sex, ethnicity, STOC, region, parental education, and indicators for at, above, or below modal grade and age. These variables constitute $(\mathbf{y}', \mathbf{z}')^c$. The model was

$$\theta|(\mathbf{y}, \mathbf{z}) \sim N[(\mathbf{y}', \mathbf{z})^c\boldsymbol{\Gamma}, \sigma^2],$$

where Γ is a vector of main effect parameters and σ^2 is the residual variance. Γ and σ^2 constitute the population parameter α. The MLEs of α were obtained separately within ages using all available reading data, implicitly allowing for all two-way interactions between age cohorts and the other effects. The resulting estimates were also treated as true values thereafter.

Note that among the design variables $\mathbf{z}$, STOC and region are explicitly included in the conditioning distributions, but PSU and school membership are not. Secondary biases in estimates of statistics involving these effects can thus occur, although they are mitigated to the degree that between-PSU and between-school variation are explained by the conditioned-upon variables such as region, STOC, and parental education. One such statistic of particular interest is a jackknife estimate of sampling variance, which will be discussed in due course.

Each student's posterior was approximated by a histogram over 40 points, denoted $\Theta_1, \ldots, \Theta_{40}$, from -5 to $+5$ standard deviations around the grand mean. The height of the q-th bar for the i-th student was obtained as

$$P(\Theta_q|\mathbf{x}_i, \mathbf{y}_i^c, \mathbf{z}_i^c) = \frac{P(\mathbf{x}_i|\theta = \Theta_q, \boldsymbol{\beta} = \hat{\boldsymbol{\beta}})P(\theta = \Theta_q|\mathbf{y}_i^c, \mathbf{z}_i^c, \boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}})}{\sum_s P(\mathbf{x}_i|\theta = \Theta_s, \boldsymbol{\beta} = \hat{\boldsymbol{\beta}})P(\theta = \Theta_s|\mathbf{y}_i^c, \mathbf{z}_i^c, \boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}})} \tag{12}$$

where $P(\theta = \Theta_s|\mathbf{y}_i^c, \mathbf{z}_i^c, \boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}})$ is the density at Θ_s of the normal pdf with mean $(\mathbf{y}', \mathbf{z}')_i^c\hat{\Gamma}$ and the residual variance for that student's cohort. Each of five imputations was drawn in two steps. First, a bar was selected at random in accordance with (12). Second, a point was selected at random from the range spanned by that bar.

The main results in this first analysis were the mean proficiencies in the traditional reporting categories. A given mean was calculated five times, once from each completed data set, and the average of the five was reported. A jackknife variance was also calculated from each completed dataset, with the imputations in the set treated as true values of θ; these values were averaged to yield U_M. The reported variance was the sum of U_M and $1 + M^{-1} = 1.2$ times the variance of the five aforementioned means. Table 6 illustrates some results for Age 17. The full report is available as *The Reading Report Card* (NAEP, 1985).

The proportional increase in variance due to the latency of θ, or r_M, varies from 2% up to nearly 30%. The largest increase is for a low scoring group, for whom the test items were rather difficult. Proficiencies of members of this group were determined less precisely by their item responses, so their likelihoods for θ, and the consequent posterior distributions from which their imputations were drawn, were more dispersed.

The original completed datasets focused their accuracy on the subpopulation means featured in *The Reading Report Card*. But as Example 3 shows, analyses involving survey variables that were not built into the conditional distribution $p^*(\theta|\mathbf{y}, \mathbf{z})$ can be biased. An opportunity to examine such biases arose in a subsequent study of reading levels of pupils whose primary language was not English; it required multiple regressions involving some variables upon which θ had been conditioned originally, and others that had not. New completed datasets were constructed in which all the variables required in the analyses—some fifty effects per age—were included in $p^*(\theta|\mathbf{y}, \mathbf{z})$, using a new program (Sheehan, 1985). Table 7 compares multiple regression coefficients from the old and the new imputations.

The *multiple* regression coefficients for the two effects that were included in $p^*(\theta|\mathbf{y}, \mathbf{z})$ in the original analysis differed by only 4% and 9% of their new values. Coefficients for significant effects not originally included in conditioning, however, were underestimated by 10% to 40%. Since the classical conditional reliabilities of test

TABLE 6

Estimating Age 17 Means from Completed Datasets

	Total	Males	Black	West	Rural
(1) Imputation 1	288.005	282.644	266.195	287.177	282.990
(2) Imputation 2	288.258	283.201	265.104	288.338	283.499
(3) Imputation 3	288.208	282.869	265.259	288.018	283.285
(4) Imputation 4	288.135	282.554	264.832	287.745	282.854
(5) Imputation 5	287.819	282.314	264.241	287.196	282.360
(6) Average (1)-(5)	288.085	282.718	265.126	287.695	282.997
(7) Variance (1)-(5)	.025	.092	.406	.208	.152
(8) Average Jackknife Variance	1.248	1.225	1.742	4.333	9.218
(9) Total Variance 1.2 x (7) + (8)	1.278	1.335	2.229	4.583	9.400
(10) Proportional Increase [(9)-(8)]/(8)	.024	.090	.280	.058	.020

booklets averaged about .7 (Mislevy & Sheehan, 1987), this result is consistent with expectations for multiple regression from the simplified setting of Example 3 (see the last row of Table 4). *Simple* coefficients for the same variables suffered less bias—an average of 15%—due to the mitigating factor of the proportion to which the included effects account for each omitted effect (see the second row of Table 4).

The effect on jackknife variance estimates of the omission of school and PSU effects from $p^*(\theta|\mathbf{y}, \mathbf{z})$ can now be addressed. The jackknife estimate of the sampling variability of an effect depends on the variation of that effect within pairs of PSUs matched on geographic proximity, degree of urbanicity, and proportion of minority students (Johnson, 1987). With more than 300 pupils per PSU, the fixed-effect approximation to the random-effects estimation of within-pair differences is reasonable, so the question is the degree to which these differences tend to be underestimated. A rough answer is 30% times the complement of multiple R^2 for predicting pair membership from the variables included in conditioning. These R^2 values ranged from .22 to .98 in the 32 PSU pairs, with a median of .71. Thus, the population-sampling variance components of effects involving θ have probably been underestimated by about .30 $\times$ (1 − .71), or 9%. This problem can be rectified in future analyses by including PSUs as conditioning variables or boosting sampling variances by a factor calculated in the manner described above.

Neither biases of the magnitude seen in Table 7 nor redrawing imputations for each new secondary analysis are satisfactory. Clearly, care is required to impute values for latent variables that lead to acceptably accurate results across in a broad range of secondary analyses. If both y and z can take only a few distinct values and there are

TABLE 7

Multiple Regression Estimates Based on Imputations Constructed with Partial and Full Conditioning

	Partial	Full Conditioning			%-attenuation	
Effect	β	β	SE(β)	t	All	Significant
White; Language Minority	6.08	4.23	3.96	1.07	-43.74	
White; Language Non-minority	12.22	13.72	2.94	4.67	10.93	10.93
Hispanic; Lang. Non-minority	-.76	1.22	3.25	.38	162.30	
Asian; Language Minority	-2.90	-6.25	4.39	-1.42	53.60	
Asian; Language Non-minority	9.54	17.34	4.66	3.72	44.98	44.98
Black; Language Non-minority	-8.64	-10.82	2.95	-3.67	20.15	20.15
Sex: Male*	-8.55	-9.35	.80	-11.69	8.56	8.56
Parent Education*	6.03	5.80	.38	15.26	-3.97	-3.97
Home Language Minority	-9.41	-13.78	2.78	-4.96	31.71	31.71
Study Aids	2.63	3.89	.43	9.05	32.39	32.39
Homework	2.78	3.82	.30	12.73	27.23	27.23
Hours of TV	-1.22	-2.04	.24	-8.50	40.20	40.20
Pages Read	6.36	10.59	1.01	10.49	39.94	39.94
Years Academic Courses	.91	1.25	.14	8.93	27.20	27.20

*Effect included in partial conditioning set.

TABLE 8

Estimated Effects Based on Full, No, and Partial Conditioning

Effect	Conditioning				
	Full*	None	Bias	32 Components	Bias
Male-Female	-15.7	-14.4	-8%	-15.9	1%
White-Black	26.6	23.8	-11%	26.3	-1%
High Metropolitan - Low Metropolitan	32.6	30.5	-6%	32.8	1%
North-South	10.4	9.4	-10%	10.1	-3%
At Modal Grade - Below Modal Grade, Within Modal Age	32.7	29.2	-11%	32.7	0

* Imputations constructed with conditional distributions that included 64 contrasts, including those shown here.

many respondents at each (**y**, **z**) combination, a nonparametric estimate $p(\theta|\mathbf{y}, \mathbf{z})$ can be obtained for each combination by the methods of Laird (1978) or Mislevy (1984). This eliminates specification error in the population model as a potential source of secondary biases. Currently, this approach is infeasible for surveys such as NAEP with its hundreds of collateral variables, and a computing approximation $p^*(\theta|\mathbf{y}, \mathbf{z})$ must be employed. To optimize the accuracy of potential secondary analyses, the following suggestions can be offered:

Determine θ as well as is practical. Measuring each individual's θ more accurately increases testing time, but decreases biases in secondary analyses when the imputation model is incorrect. Tradeoffs thus arise among item-sampling designs. Compared to a design that administers five items to each subject, a design that administers ten items yields *less efficient* estimates of statistics involving only **y** and **z** variables that *have* been used to construct $p^*(\theta|\mathbf{y}, \mathbf{z})$, but *less biased* estimates of statistics involving variables that *have not*.

This point can be illustrated with data from the 1988 NAEP Age 13 reading assessment. To reduce biases in analyses of excluded background variables from the 30% levels of 1984, the average number of items per student was raised from 17 to about 40. Table 8 shows estimates of selected effects when imputations were constructed conditioning on all pertinent background variables, and imputations constructed without conditioning on any. The resulting shrinkages average around 10%, a substantial improvement in "worst case" performance.

Borrow information from related scales. Imputation methods can be applied with vector-valued latent variables. One combines multivariate conditional distributions $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z})$ with multi-dimensional likelihoods $p(\mathbf{x}|\boldsymbol{\theta})$, and draws vector-valued imputations from predictive distributions $p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}, \mathbf{z})$. Exploiting correlations of .6 among the

four IRT scales in NAEP's Survey of Young Adult Literacy (Kirsch & Jungeblut, 1986) sharpened respondents' predictive distributions within each scale as much as doubling test lengths.

Condition explicitly on particularly important contrasts, such as treatment group in a survey undertaken to compare treatment effects in a program evaluation. This ensures that the subpopulation means or regression coefficients involving key variables are estimated accurately.

Condition on well-chosen combination of variables. One can reduce biases for a large number of contrasts by conditioning on linear combinations of them—for example, the first h component scores from a principal components decomposition of the covariance matrix among effects. Example 3 implies that an upper bound for the bias for the regression coefficient of a contrast treated this way is the product of (a) the proportion of its variance not accounted for by the conditioned-upon components and (b) the complement of test reliability. Table 8 shows estimates that result from imputations constructed with the first 32 principal components of the 64 effects that were used in the baseline analysis. Biases are negligible, and average around zero.

Conclusion

In the early 1980s, Bock, Mislevy, and Woodson (1982) hailed item response theory as a cornerstone of progress for educational assessment. IRT does indeed make it *possible* to solve many practical problems, such as allowing item pools to evolve over time, providing results on a consistent scale in complex item-sampling designs, and reducing the numbers of items per pupil.

But *possible* doesn't necessarily mean *easy*. Familiar IRT procedures based on point estimates for individual examinees break down in efficient designs that solicit relatively few responses from each. Higher levels of theoretical and computational complexity are required to realize the benefits IRT offers.

This paper argues that multiple imputation provides a theoretical framework for analyses involving latent variables from sample survey data, and shows how to apply it. The approach places the burden of the complexities of the problem on the primary analyst, who must create multiple completed datasets. Secondary analysts can then draw correct inferences from them using standard routines for complete data.

References

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

Beaton, A. E. (1987). *The NAEP 1983/84 technical report* (NAEP Report 15-TR-20). Princeton: Educational Testing Service.

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

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

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.

Ford, B. L. (1983). An overview of hot-deck procedures. In W. G. Madow, I. Olkin, & D. B. Rubin (Eds.), *Incomplete data in sample surveys, Volume 2, Theory and bibliographies* (pp. 185–207). New York: Academic Press.

Goldstein, H., & McDonald, R. P. (1988). A general model for the analysis of multilevel data. *Psychometrika, 53*, 455–467.

Johnson, E. G. (1987). Estimation of uncertainty due to sampling variation. In A. E. Beaton, *The NAEP 1983/84 technical report* (NAEP Report 15-TR-20, pp. 505–512). Princeton: Educational Testing Service.

Kelley, T. L. (1923). *Statistical method*. New York: Macmillan.

Kirsch, I. S., & Jungeblut, A. (1986). *Literacy: Profile of America's young adults* (Final Report, No. 16-PL-01). Princeton, NJ: National Assessment of Education Progress.

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

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Lord, F. M. (1965). A strong true-score theory, with applications. *Psychometrika, 30*, 239–270.

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

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

Mislevy, R. J., & Sheehan, K. M. (1987). Marginal estimation procedures. In A. E. Beaton, *The NAEP 1983/84 technical report* (NAEP Report 15-TR-20, pp. 293–360). Princeton: Educational Testing Service.

Muthén, B., & Satorra, A. (1989). Multilevel aspects of varying parameters in structural models. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 87–99). San Diego: Academic Press.

National Assessment of Educational Progress (1985). *The reading report card: Progress toward excellence in our schools*. Princeton: Educational Testing Service.

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

Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association, 72*, 538–543.

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

Sanathanan, L., & Blumenthal, N. (1978). The logistic model and estimation of latent structure. *Journal of the American Statistical Association, 73*, 794–798.

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

Tanner, M., & Wong, W. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association, 82*, 528–550.

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

Manuscript received 10/16/85
Final version received 4/10/90