PSYCHOMETRIKA—VOL. 58, NO. 1, 79–85
MARCH 1993
NOTES AND COMMENTS

SHOULD "MULTIPLE IMPUTATIONS" BE TREATED AS "MULTIPLE INDICATORS"?

ROBERT J. MISLEVY

EDUCATIONAL TESTING SERVICE

Rubin's "multiple imputation" approach to missing data creates synthetic data sets, in which each missing variable is replaced by a draw from its predictive distribution, conditional on the observed data. By construction, analyses of such filled-in data sets as if the imputations were true values have the correct expectations for population parameters. In a recent paper, Mislevy showed how this approach can be applied to estimate the distributions of latent variables from complex samples. Multiple imputations for a latent variable bear a surface similarity to classical "multiple indicators" of a latent variable, as might be addressed in structural equation modelling or hierarchical modelling of successive stages of random sampling. This note demonstrates with a simple example why analyzing "multiple imputations" as if they were "multiple indicators" does not generally yield correct results; they must instead be analyzed by means concordant with their construction.

Key words: multiple imputation, multiple indicators, National Assessment of Educational Progress, plausible values.

Introduction

In the National Assessment of Educational Progress (NAEP), multiple imputations ("plausible values" as they are called in NAEP) are created to facilitate data analyses that involve latent variables, to deal with the uncertainty associated with each individual subject. Although latent variables cannot, by their nature, be determined for each sampled subject, a file of plausible values is constructed containing, for each subject, a numerical value corresponding to his or her latent variable. Analyzing these values as if they were the true values of the latent variable yields estimates of the population distributions of the latent variables that are consistent under the model used to construct them. Several such "pseudo-data files" are in fact produced; the variance of a statistic across these files quantifies the uncertainty associated with the latent nature of the variable of interest. This approach is founded on Rubin's (1987) multiple imputation technique for missing data, adapted for latent variables as described in Mislevy (1991). A didactic presentation of the approach appears in Mislevy, Beaton, Sheehan, and Kaplan (1992).

Each sampled subject will have associated with him or her several plausible values, one per pseudo file. On the surface, this situation resembles another one more familiar in educational and psychological measurement—that of multiple unbiased and conditionally independent indicators of a latent variable. Standard procedures are available for estimating population characteristics in this latter situation, such as structural analyses with LISREL (Jöreskog & Sörbom, 1989) or LISCOMP (Muthén, 1988), or in hierarchical analyses as another stage of sampling (e.g., Goldstein, 1987). What happens when NAEP plausible values are employed as input to LISREL, for example,

I am grateful to Frank Jenkins, John Mazzeo, Kentaro Yamamoto, and Rebecca Zwick for comments on earlier versions of this paper.

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

0033-3123/93/0300-91396$00.75/0

under standard LISREL assumptions for multiple indicators of a latent variable? The short answer is that some population characteristics are estimated correctly, but most are estimated incorrectly. Plausible values should be analyzed as multiple imputations, not as multiple indicators. To provide some insight into the reasons, the details of a simple special case that can be worked out directly are presented: simple linear regression with one observed variable and one latent variable.

Case 1: True Population Parameters

Let θ represent proficiency, and y a background variable. Assume these variables are distributed bivariate normally, with mean vector zero and the following covariance matrix, Σ:

$$\begin{array}{c|cc} & \theta & y \\ \hline \theta & \sigma_\theta^2 & \\ y & \sigma_{\theta y} & \sigma_y^2 \end{array}\ .$$

Let R denote the correlation between θ and y, or $\sigma_{\theta y}/(\sigma_\theta \sigma_y)$. The parameters of the linear regression models for θ on y, and for y on θ, are given in the first column of entries in Table 1. They would be the expected values of corresponding estimated quantities in large samples of observations of θ and y.

Case 2: Multiple, Conditionally-Independent, Unbiased Indicators

Suppose that rather than observing θ directly, one observes instead two indicators x_1 and x_2, with

$$x_1 = \theta + e_1, \tag{1a}$$

$$x_2 = \theta + e_2, \tag{1b}$$

and where $e_1 \sim N(0, \sigma_e^2)$, $e_2 \sim N(0, \sigma_e^2)$, and e_1 and e_2 are independent of each other, θ, and y. The covariance matrix among the observable variables is as follows:

$$\begin{array}{c|ccc} & x_1 & x_2 & y \\ \hline x_1 & \sigma_\theta^2 + \sigma_e^2 & & \\ x_2 & \boxed{\sigma_\theta^2} & \sigma_\theta^2 + \sigma_e^2 & \\ y & \boxed{\sigma_{\theta y}} & \sigma_{\theta y} & \boxed{\sigma_y^2} \end{array}\ .$$

A LISREL analysis for the regression of θ on y, or vice versa, correctly exploits the measurement model described above to estimate the relationship between θ and y. In this example, the keys to the LISREL solution are that (i) the observed covariance between x_1 and x_2 has as its expectation the true variance of θ, and (ii) the observed covariance between x_k and y has the same expectation as that between θ and y; that is, estimates of the values in the boxes approximate Σ. These values could be approximated as accurately as desired by increasing the examinee sample, and, subsequently, regression models for θ and y could be estimated as in Case 1 using these unbiased estimates of the correct population parameters. The expected results are again as given in the first column of entries in Table 1.

TABLE 1

Expected Regression Parameter Estimates

Statistic	Correct Values (Cases 1, 2, 3)	LISREL analysis of plausible values (Case 4)
Regression of θ on y		
Regression coefficient	$\frac{\sigma_{\theta y}}{\sigma_y^2}$	$\frac{\sigma_{\theta y}}{\sigma_y^2}$
Residual variance	$\sigma_\theta^2 - \frac{\sigma_{\theta y}^2}{\sigma_y^2}$	$\sigma_\theta^2\left(1 - (1-\rho_c)\left(1-\frac{\sigma_{\theta y}^2}{\sigma_\theta^2\sigma_y^2}\right)\right) - \frac{\sigma_{\theta y}^2}{\sigma_y^2}$
R^2	$\frac{\sigma_{\theta y}^2}{\sigma_\theta^2\sigma_y^2}$	$\frac{\sigma_{\theta y}^2}{\sigma_y^2\sigma_\theta^2\left(1 - (1-\rho_c)\left(1-\frac{\sigma_{\theta y}^2}{\sigma_\theta^2\sigma_y^2}\right)\right)}$
Regression of y on θ		
Regression coefficient	$\frac{\sigma_{\theta y}}{\sigma_\theta^2}$	$\frac{\sigma_{\theta y}}{\sigma_\theta^2\left(1 - (1-\rho_c)\left(1-\frac{\sigma_{\theta y}^2}{\sigma_\theta^2\sigma_y^2}\right)\right)}$
Residual variance	$\sigma_y^2 - \frac{\sigma_{\theta y}^2}{\sigma_\theta^2}$	$\sigma_y^2 - \frac{\sigma_{\theta y}^2}{\sigma_\theta^2\left(1 - (1-\rho_c)\left(1-\frac{\sigma_{\theta y}^2}{\sigma_\theta^2\sigma_y^2}\right)\right)}$
R^2	$\frac{\sigma_{\theta y}^2}{\sigma_\theta^2\sigma_y^2}$	$\frac{\sigma_{\theta y}^2}{\sigma_y^2\sigma_\theta^2\left(1 - (1-\rho_c)\left(1-\frac{\sigma_{\theta y}^2}{\sigma_\theta^2\sigma_y^2}\right)\right)}$

Case 3: Multiple Imputations

A key idea underlying multiple imputations appears in Rubin (1977). Suppose that if values of both θ and y were observed for each of N sampled subjects, one would be able to calculate a sample statistic $s(\boldsymbol{\theta}, \mathbf{y})$ for a parameter of interest, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$ and $\mathbf{y} = (y_1, \ldots, y_N)$. Observing values of x rather than θ renders it impossible to compute s directly, but one may be able to compute its expectation conditional on the data that were observed, **x** and **y**:

$$E[s(\boldsymbol{\theta},\mathbf{y})|\mathbf{x}, \mathbf{y}] = \int s(\boldsymbol{\theta},\mathbf{y})\, p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y})\, d\boldsymbol{\theta}. \tag{2}$$

The first stage of the analysis is to construct a representation of $p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y})$. In latent variable problems, Bayes theorem is employed to combine the measurement model $p(\theta|x)$ with a population-structure model, $p(\theta|y)$. Typically the distribution of unknown parameters of one or both of these components must be estimated. The second stage is to approximate the integral numerically by averaging the results obtained by evaluating $s(\boldsymbol{\theta}, \mathbf{y})$ with repeated draws from $p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y})$. One such draw consists of a "plausible value" of θ for each subject in the sample, the full vector of which we denote by $\tilde{\boldsymbol{\theta}}$. (If imperfectly known parameters for $p(\theta|x)$ and/or $p(\theta|y)$ are involved, a value from their posterior distribution is first drawn, and individual subjects' imputations in a given $\tilde{\boldsymbol{\theta}}$ are drawn from their predictive distributions conditional on these values.) For a properly constructed set of imputations,

$$E[s(\tilde{\boldsymbol{\theta}}, \mathbf{y})] = E[s(\boldsymbol{\theta}, \mathbf{y})|\mathbf{x}, \mathbf{y}).$$

The "multiple" aspect of multiple imputations is that the variance among evaluations of s with different imputation sets is a variance component due to the latent nature of θ, which must be added to the more familiar component due to sampling of subjects to capture both sources of uncertainty about s.

It should be pointed out that there would be little point in doing all this in the running example. The population structure model, $p(\theta|y)$, is $N[E(\theta|y), \text{Var}\ (\theta|y)]$; to construct it requires characterizing what the data convey about $\boldsymbol{\Sigma}$ in the first place. The LISREL analysis described in Case 2 would be appropriate, and, having carried it out, one knows essentially all there is to know. The approach becomes attractive (especially for secondary analysis) in a setting like that of NAEP, with its complex student sampling design, item response theoretic measurement models, and hundreds of background variables. A comprehensive model is then fit for $p(\theta|y)$ and used to construct imputations useful for a broad range of potential secondary analyses. (The interested reader is referred to NAEP Technical Reports for details of NAEP procedures; see Beaton, 1987, 1988; and Johnson & Zwick, 1990.) Files of multiple imputations thus *re*cover structure implicit in the initial comprehensive analysis. The point of the present example is to show, in a simplified setting, that appropriate analyses of properly-constructed imputations have as expected values the population parameters, whereas analyses of them as if they were "multiple indicators" do not.

Returning to the example, it is again supposed that rather than observing θ directly, one observes a fallible indicator $x \sim N(\theta, \sigma_e^2)$, as in (1), along with y. Now construct two files of multiple imputations for θ by the methods described in Example 1 of Mislevy (1991). [Because the focus here lies in large-sample expectations, the current presentation can be simplified by assuming sufficiently large examinee samples to treat population parameters as essentially known. The same results hold in the more complex but realistic analyses in which uncertainty about population parameters is taken into account (see Rubin, 1987).] With population parameters assumed known, a plausible value for a respondent with given values of x and y is simply a draw from $p(\theta|x, y) \propto p(x|\theta)\, p(\theta|y)$, which in this example is another normal distribution. The k-th plausible value, $\tilde{\theta}_k$, is a draw from this distribution:

$$\tilde{\theta}_k = \bar{\theta} + d_k, \tag{3}$$

where

$\bar{\theta}$ is the posterior mean of θ given the individual's x and y values, namely,

$$\bar{\theta} = (1 - \rho_c)\,\frac{\sigma_{\theta y}}{\sigma_y^2}\,y + \rho_c x,$$

with ρ_c the conditional reliability of x given y;

$$\rho_c = \frac{\text{Var}\,(\theta|y)}{\text{Var}\,(\theta|y) + \text{Var}\,(x|\theta)} = \frac{\sigma_\theta^2(1 - R^2)}{\sigma_\theta^2(1 - R^2) + \sigma_e^2}; \text{ and}$$

d_k is drawn at random from

$$N\left(0,\ \sigma_\theta^2(1 - \rho_c)\left(1 - \frac{\sigma_{\theta y}^2}{\sigma_\theta^2\sigma_y^2}\right)\right), \text{ or}$$

$$N(0,\ \sigma_\theta^2(1 - \rho_c)(1 - R^2)).$$

Mislevy (1991) shows that the covariance matrix of $\tilde{\theta}_k$ with y takes the following form:

$$\begin{array}{c|cc} & \tilde{\theta}_k & y \\ \hline \tilde{\theta}_k & \sigma_\theta^2 & \\ y & \sigma_{\theta y} & \sigma_y^2 \end{array}.$$

Thus, carrying out regression analyses with $\tilde{\theta}_k$ as if it were θ itself has as its expectation the same (correct) values as in Cases 1 and 2.

Note two vital distinctions between the e_k disturbance terms in (1) and the d_k terms in (3) used to construct plausible values:

1. The $e_k s$ are unbiased additions to θ *supplied by nature*. The more conditionally-independent indicators x one can arrange to observe, the more accurately one can triangulate in on the θ values of each of the sampled subjects. The d_k's are unbiased additions to $\bar{\theta}$ *supplied by the analyst*. The x's one observes set an upper limit on what one can known about a statistic s involving θs—specifically, that achievable through an exact evaluation of (2). We can generate as many sets of multiple imputations as we like, merely to the end of a better numerical approximation of (2).

2. The e_k's are unbiased additions to θ, so x's have the correct expectations for each examinee, but a larger variance than the θs. This is the situation addressed by the "multiple indicators" measurement model of Case 2. The d_k's are unbiased additions to $\bar{\theta}$, but they *cannot* be unbiased additions to θ if the $\tilde{\theta}_k$s are to have the same variance as θ's. This property is a consequence of the goal of constructing synthetic variates with the same population characteristics as θ, but it violates an assumption of the multiple indicators model.

Case 4: Multiple Imputations (Incorrectly) Treated as Multiple Indicators

Suppose that two sets of plausible values were constructed, as in Case 3, but were input into a LISREL analysis as if they were multiple, conditionally-independent, unbiased measures. The covariance matrix among $\tilde{\theta}_1$, $\tilde{\theta}_2$, and y is as follows:

TABLE 2

Numerical Illustration of Expected Regression Parameter Estimates

Statistic	Correct Values (Cases 1, 2, 3)	LISREL analysis of plausible values (Case 4)
Regression of θ on y		
Regression coefficient	.7071	.7071
Residual variance	.5	.25
R^2	.5	.6667
Regression of y on θ		
Regression coefficient	.7071	.9428
Residual variance	.5	.3333
R^2	.5	.6667

Note: Evaluated with $\sigma_\theta^2 = \sigma_y^2 = 1$, $\sigma_{\theta y} = .7071$, and ρ_c=.5.

$$\begin{array}{c|c|c|c}
 & \tilde{\theta}_1 & \tilde{\theta}_2 & y \\
\hline
\tilde{\theta}_1 & \boxed{\sigma_\theta^2} & & \\
\tilde{\theta}_2 & \boxed{\sigma_\theta^2\left(1-(1-\rho_c)\left(1-\frac{\sigma_{\theta y}^2}{\sigma_\theta^2\sigma_y^2}\right)\right)} & \sigma_\theta^2 & \\
y & \boxed{\sigma_{\theta y}} & \sigma_{\theta y} & \boxed{\sigma_y^2}
\end{array}$$

As in Case 2, the entries in the boxes are (now mistakenly) taken to approximate the covariance matrix for θ and y. The variance of θ is underestimated by a factor that depends on the amount of information about θ conveyed by x and y. The rightmost column of Table 1 gives the resulting regression estimates that LISREL would construct under this misspecification. Table 2 gives illustrative numerical values, with $\sigma_\theta^2 = \sigma_y^2 = 1$; $\sigma_{\theta y} = .7071$, so that $R^2 = .5$; and $\sigma_e^2 = 1$, so that $\rho_c = .5$.

Conclusion

Multiple imputations for latent variables are constructed so that analyses treating them as if they were true values have the correct expectations for population characteristics. The properties entailed by this requirement contradict the properties of multiple, unbiased, conditionally-independent indicators as might be addressed in LISREL, LISCOMP, or hierarchical analyses. Analyzing "multiple imputations" in accordance with their construction yields correct estimates of population characteristics, given the model used to construct them. Analyzing them as if they were "multiple indicators" generally does not.

References

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

Beaton, A. E. (1988). *The NAEP 1985/86 technical report* (NAEP Report 17-TR-20). Princeton: Educational Testing Service.

Goldstein, H. (1987). *Multilevel models in educational and social research.* London: Griffin, NY: Oxford University Press.

Johnson, E. G., & Zwick, R. (1990). *The NAEP 1987/88 technical report* (NAEP Report 19-TR-20). Princeton: Educational Testing Service.

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7: User's reference guide.* Mooresville, IN: Scientific Software.

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

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29,* 133–161.

Muthén, B. (1988). *LISCOMP* [computer program]. Mooresville, IN: Scientific Software.

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.

Manuscript received 11/13/91
Final version received 3/30/92