Multiple Imputation After 18+ Years

Donald B. Rubin

Journal of the American Statistical Association, Vol. 91, No. 434 (Jun., 1996), 473-489.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199606%2991%3A434%3C473%3AMIA1Y%3E2.0.CO%3B2-V

Journal of the American Statistical Association is currently published by American Statistical Association.

Multiple Imputation After 18+ Years

Donald B. RUBIN

Multiple imputation was designed to handle the problem of missing data in public-use data bases where the data-base constructor and the ultimate user are distinct entities. The objective is valid frequency inference for ultimate users who in general have access only to complete-data software and possess limited knowledge of specific reasons and models for nonresponse. For this situation and objective, I believe that multiple imputation by the data-base constructor is the method of choice. This article first provides a description of the assumed context and objectives, and second, reviews the multiple imputation framework and its standard results. These preliminary discussions are especially important because some recent commentaries on multiple imputation have reflected either misunderstandings of the practical objectives of multiple imputation or misunderstandings of fundamental theoretical results. Then, criticisms of multiple imputation are considered, and, finally, comparisons are made to alternative strategies.

KEY WORDS: Confidence validity; Missing data; Nonresponse in surveys; Public-use files; Sample surveys; Superefficient procedures.

1. THE PROBLEM MULTIPLE IMPUTATION WAS DESIGNED TO ADDRESS

Missing values are a problem in many data sets and seem especially common in the medical and social sciences. For nearly two decades I have been advocating and developing the use of multiple imputation to address aspects of this problem; early documents include Rubin (1977a, 1977b, 1978, 1980, 1983), Herzog and Rubin (1983), Rubin and Schenker (1986), and the basic reference Rubin (1987). There are situations where multiple imputation is appropriate, and, as with any statistical tool, there are others where its application is more questionable. Originally it was viewed as being most appropriate in complex surveys that are used to create public-use data sets to be shared by many ultimate users, although over the years, it has proven valuable in other settings as well.

For the context for which it was envisioned, with data-base constructors and ultimate users as distinct entities, I firmly believe that multiple imputation is the method of choice for addressing problems due to missing values: alternative methods either require special knowledge and techniques not available to typical users or produce answers that are generally not statistically valid for scientific estimands. This is a strong statement, and it is clear that its accuracy must depend on the class of problems to which it is applied. Consequently this article begins with a description of the assumed statistical computing environment for the ultimate users of shared data-bases and of our objectives for handling missing data in this environment. It is especially important to provide this background to emphasize that the goal of multiple imputation is to provide statistically valid inference (in the traditional complex survey sense of Neyman, Cochran, and Hansen) in the difficult real-world situation where (1) ultimate users and data-base constructors are distinct entities with different analyses, models, and capabilities, and (2) there typically is no one accepted reason for the missing data.

In Section 2 multiple imputation is reviewed, with particular emphasis given to how it was designed to satisfy the stated objectives in the assumed environment for ultimate users. This review of critical points of the theory and intended practice of multiple imputation minimizes technical details so that essential statistical points will be more transparent than in the theoretical material in Rubin (1987), which requires substantial familiarity with, and acceptance of the relevance of, both randomization-based and Bayesian inference. Then, in Section 3, current concerns about multiple imputation are discussed with the benefit of the simplified theory. Finally, competing techniques are evaluated for their utility in the assumed context and are found to be less effective than multiple imputation.

1.1 Assumed Environment for Ultimate Users

Public-use (shared) data bases are analyzed by many ultimate users with varying degrees of statistical expertise and computing power, and with different scientific questions and objectives. Typically such users have available to them a number of standard complete-data techniques. These include various stand-alone routines such as ones for ordinary least-squares regression, logistic regression, factor analysis, variance components estimation, proportional hazards models, etc., and various packages of programs such as SAS, BMDP, SPSS, etc. Also, there may be available routines for inference in the presence of missing data under particular models (e.g., Schafer 1995), complete-data management routines for merging files, subsetting data, deleting cases and variables, applying transformations, and creating new variables, or various resampling programs to create simulated replicate data, principally jackknife and bootstrap routines.

Donald B. Rubin is Professor, Department of Statistics, Harvard University, Cambridge, MA 02138. This work was partially supported by National Science Foundation Grant SES-92-07456 and partially by the U.S. Census Bureau through a subcontract to Datametrics Research, Inc. from NORC. Very helpful comments on earlier drafts were made by J. Brand, R. J. A. Little, X. L. Meng, F. Scheuren, and editorial reviewers. Also, thanks are due to R. E. Fay for his continuing interest in multiple imputation and for his special examples, which helped stimulate the formulation here of superefficient multiple imputation and the associated new results. Finally, David Binder's comments on presentations of this material are gratefully acknowledged.

Journal of the American Statistical Association
June 1996, Vol. 91, No. 434, Applications and Case Studies

Essentially all public-use data sets have missing values, typically not of any nice neat type. In general, ultimate users have neither the knowledge nor the tools to address missing data problems satisfactorily. Even if some ultimate users do have adequate resources for modeling and computation, data-base constructors typically know more about reasons for nonresponse and have better access to confidential and detailed information not released for public use (e.g., exact addresses and neighborhood relationships, hourly blood pressure readings and doctor indicators), information that can be useful for modeling missing data. Moreover, ultimate users should be focused on their substantive scientific analyses and for these, missing data are generally simply a nuisance. My conclusion is that "correctly" modeling the missing data must be, in general, the data constructor's responsibility.

We, that is, data-base constructors and statistical software designers, have no direct control over what ultimate users will do with their arsenal of tools. We cannot stop users from doing bad science, but if possible we should facilitate their ability to do good science with their available tools, even when data sets suffer from missing values.

1.2 Achievable Basic Objective

One achievable basic objective in such a setting is the following: Each tool in the ultimate users' existing arsenals can be applied to any data set with missing values using the same command structure and output standards as if there were no missing data. The only additional software that is allowed to be required comprises entirely general macros that can be applied to any complete-data analysis and incomplete data set. Certain ad hoc methods of handling missing data, such as "complete-case analysis," "available-case analysis," and "fill-in with means" (e.g., see Little and Rubin 1987, part I), satisfy this basic objective and so have a certain appeal. The problem with such methods is that they typically yield statistically invalid answers for scientific estimands; "scientific estimands" and "statistically valid" require definition.

1.3 Scientific Estimands

By a scientific estimand I mean a quantity of scientific interest that can be calculated in the population and does not change its value depending on the data collection design used to measure it (i.e., it does not vary with sample size and survey design, or the number of nonrespondents, or follow-up efforts). Letting X be the array of all background (e.g., stratification) information fully observed in a population and Y be the array of outcome information in the population that is to be sampled in the survey, a scientific estimand is a function of X and Y, say $Q = Q(X, Y)$. Scientific estimands include population means, variances, correlations, factor loadings, regression coefficients, and these quantities within strata or domains, but exclude the sampling variance of a sample mean under a particular sampling plan and the expectation of the complete-data sample mean when missing values are filled in with zero or the observed sample mean. These latter quantities can be important for inference and design, but they are not scientific estimands in my definition because they are functions of sample size, sample design, response rates for a particular survey, methods for handling missing values, *and* scientific estimands such as population means and variances.

The distinction between estimates of scientific estimands and measures of their uncertainty is an old one in statistics; see, for example, Fisher (1925, p. 724) where a measure of uncertainty associated with an estimate is called an "ancillary" statistic, that is, a subordinate or supplemental statistic. In Fisher's context, the estimate was the maximum likelihood estimate and the ancillary statistic was the second derivative of the log-likelihood, but the distinction is relevant to more general estimates and associated measures of uncertainty, as we see in the next section.

1.4 What is Meant by Statistically Valid?

In the context of shared data bases supporting analyses by many users, I believe that statistically valid must be a frequency concept, averaging over randomization distributions generated by known sampling mechanisms (used to collect data) and posited distributions for the response mechanisms (the processes underlying nonresponse). In standard scientific surveys, the sampling mechanism is known but the nonresponse mechanisms is rarely fully known and so typically must be posited, either implicitly or explicitly.

Bayesian validity is also important, but is far more difficult to achieve in this context because it requires far more compatibility between the data-base constructor and the analyst. In fact, in general I do not believe it can be achieved in any real sense in the context of the basic objective to use existing complete-data tools with shared data bases. In any case, no Bayesian should object to achieving frequentist validity; effectively, Bayesians want and promise much more: calibration conditional on the data in addition to unconditional calibration (e.g., in Rubin 1984, I call such frequency calculations "Bayesianly relevant and justifiable").

First and foremost, for statistical validity for scientific estimands, point estimation must be approximately unbiased for the scientific estimands, averaging over the sampling and the posited nonresponse mechanisms (e.g., filling in zeros or means is not generally acceptable). Second, interval estimation and hypothesis testing must be valid in the sense that nominal levels describe operating characteristics over sampling and posited response mechanisms. Two versions of such frequentist validity for nominal levels are especially important to distinguish when assessing multiple imputation.

Using terminology from Rubin (1987, pp. 117–118), "randomization validity" means that, for interval estimates, "actual interval coverage = nominal interval coverage," and for tests of hypotheses, "actual rejection rate = nominal rejection rate." Randomization validity is the natural objective in most survey contexts. In standard asymptotic situations, a complete-data estimate $\hat{Q}$ of an estimand Q has a normal sampling distribution centered at Q with sampling variance (or more generally, variance–covariance) consistently esti-

mated by the statistic U, where the randomization distribution is that generated by the sampling indicator I given fixed (X, Y)—the sampling mechanism. In this case we have

$$E(\hat{Q}|X, Y) \doteq Q \tag{1.1}$$

and

$$E(U|X, Y) \doteq \text{var}(\hat{Q}|X, Y), \tag{1.2}$$

and then randomization validity is not only desirable but theoretically achievable. The precision of $\hat{Q}$ is measured by U^{-1}, which plays the role of the ancillary statistic and can be used as a "true weight" (Fisher 1925, p. 724) for combining estimates.

A more generally achievable objective, however, is "confidence validity," meaning that for interval estimates, "actual interval coverage $\geq$ nominal interval coverage," and for tests of hypotheses, "actual rejection rate $\leq$ nominal rejection rate." For confidence validity with complete data, we replace (1.2) with

$$E(U|X, Y) \geq \text{var}(\hat{Q}|X, Y). \tag{1.3}$$

If (1.3) is satisfied but (1.2) is not, then U^{-1} is only an "approximate weight for the value of the estimate" (Fisher 1925, p. 724).

The distinction between randomization validity and confidence validity can be quite important when dealing with approximate procedures, which necessarily arise with nonresponse in public-use surveys, and this distinction appears in Neyman (1934), which is the foundation for statisticians' current view of frequentist validity in surveys. Here Neyman (1934, pp. 562–563) defined confidence intervals, confidence coefficients, and confidence limits, and these definitions remain the accepted mathematical definitions of these terms (e.g., Lehmann 1959). In particular, confidence limits are statistics defining an interval such that, in repeated experience, the estimand lies in the confidence interval with probability *greater than or equal to* the confidence coefficient; the shorter the interval satisfying this constraint, the better.

A simple example illustrates the wisdom implicit in Neyman's definition. Consider a particular situation with two different confidence-valid procedures for creating confidence intervals with confidence coefficient 95%. Procedure 1 produces intervals that are always *shorter* than the intervals produced by Procedure 2, and moreover, Procedure 1 has actually probability $> 95\%$ of covering the estimand, whereas Procedure 2 has only the nominal 95% probability of covering the estimand. Clearly, Procedure 1 is scientifically and statistically superior to Procedure 2 because it provides tighter inferences with greater confidence, and Neyman's definition and desiderata agree with this fact. Requiring exact agreement between nominal and actual levels as a desideratum for validity would lead one to reject Procedure 1 as invalid and choose Procedure 2, clearly a mistake. It is for this reason that confidence validity is more fundamental than randomization validity for interval estimation.

Of course, if we have a procedure that is confidence valid but not randomization valid, there is the hope that a better confidence-valid procedure exists (i.e., one with shorter intervals), which is also randomization valid, but in general this is not achievable. An attendant advantage, when the best confidence interval is randomization valid, is that the associated measure of precision can be thought of as a true rather than approximate weight (again, in the sense of Fisher 1925, p. 724—also see Fisher 1934, criticizing Neyman 1934, on this point).

1.5 Supplemental Objective Concerning Statistical Validity

We are now prepared to supplement the Achievable Basic Objective when faced with missing values, regarding the ability to apply standard complete-data statistical tools, with an objective concerning statistically valid inference for scientific estimands. It is easy to ask for more than is possible and then do something misguided when attempting the impossible. We first consider a hopeless objective, which is commonly sought, and then state an achievable one.

Hopeless Supplemental Objective. Each complete-data statistical tool can be applied to each incomplete data set to obtain the same inference as if the data set had no missing values.

This objective is clearly impossible because of the lost information, but nevertheless, it guides some thinking about how to handle missing data. It is analogous to saying that the objective of a survey is to obtain the same answer as a complete census, and it can lead to an "operations research" objective of creating imputations for missing values that are as close as possible to the truth (i.e., fill in missing values to minimize some objective function). Our actual objective is valid statistical inference not optimal point prediction under some loss function, and replacing the former with the latter can lead one badly astray. For example, suppose we have a coin that, in truth, is biased .6 heads and .4 tails. This known truth is model A, whereas model B asserts that the coin has two heads. Using model A for creating imputations (i.e., future predictions) yields a hit rate (agreements between predictions and outcomes) of $.6 \times .6 + .4 \times .4 = .52$, whereas using model B for predictions yields a hit rate of .6. This does not mean that model B is better than model A for handling missing values. Filling in missing values using model B yields the invalid statistical inference that in the future all coin tosses will be heads, clearly inconsistent for the estimand Q = fraction of tosses that are heads, whereas using model A yields consistent estimates for all such scientific estimands. The lesson is simple: Judging the quality of missing data procedures by their ability to recreate the individual missing values (according to hit-rate, mean square error, etc.) does not lead to choosing procedures that result in valid inference, which is our objective.

Statistical validity in our context is difficult because the answer that results from applying a complete-data analysis to an incomplete data set is generally invalid unless the complete-data analysis in the absence of missing data is valid—the ultimate user's responsibility, and the reasons for missing data are correctly modelled—the data-base con-

structor's responsibility. We can essentially never be sure that the data-base constructor's model is appropriate, but assuming it is, and assuming that the ultimate user is performing an analysis that would be valid if there were no missing data, we can expect that the ultimate user will obtain a valid inference.

Achievable Supplemental Objective. Assuming that the ultimate user's complete-data analysis is statistically valid for a scientific estimand, the answer that results from applying the same analysis method to an incomplete-data set remains statistically valid for the same scientific estimand assuming the truth of the data-base constructor's posited model for missing data.

I doubt that there is a much stronger objective regarding validity that we can achieve in this context where the ultimate user and the data-base constructor are distinct entities. Multiple imputation was designed to satisfy both achievable objectives by using the Bayesian and frequentist paradigms in complementary ways: the Bayesian model-based approach to *create* procedures, and the frequentist (randomization-based approach) to *evaluate* procedures.

2. REVIEW OF MULTIPLE IMPUTATION FRAMEWORK AND RESULTS

Multiple imputations for the set of missing values are multiple sets of plausible values; these can reflect uncertainty under one model for nonresponse and across several models. Each set of imputations is used to create a completed data set, each of which is to be analyzed using standard complete-data software to yield "completed-data" statistics, which are typically complete-data estimates, $\hat{Q}$, associated variance–covariance matrices, U, and p values. The complete-data statistics $\hat{Q}$ and U are general; for example, U may be obtained by mathematical analysis, linearization methods, balanced-repeated replication, the jackknife (see, e.g., Krewski and Rao 1981), the bootstrap (see, e.g., Efron 1994), or special routines for complex surveys such as SUDAAN or VPLX (see, e.g., Fay 1990). But no matter how $\hat{Q}$ and U are calculated with complete data, once missing data are filled in by imputation, they can be calculated as if the data set were complete.

2.1 Repeated Imputations

A theoretically fundamental form of multiple imputation is *repeated imputation* (Rubin 1987, pp. 75–76). Repeated imputations are draws from the posterior predictive distribution of the missing values under a specific model, that is, a particular Bayesian model for both the data and the missing-data mechanism. The m complete-data analyses corresponding to the m imputations under one model result in m repeated completed-data statistics, and these are combined to form one *repeated-imputation inference* that appropriately adjusts for nonresponse under the model used to create the repeated imputations. The values of the complete-data statistics $\hat{Q}$ and U calculated on the m completed data set are $\hat{Q}_{*1}, \ldots, \hat{Q}_{*m}$ and $U_{*1}, \ldots, U_{*m}$. The basic procedures for combining the m estimates $\{\hat{Q}_{*1}, \ldots, \hat{Q}_{*m}\}$, associated variance–covariance matrices $\{U_{*1}, \ldots U_{*m}\}$, and p values, that is, the final repeated-imputation inferences, are derived in chapter 3 in Rubin (1987) under the Bayesian paradigm for survey inference (introduced in chap. 2 of Rubin 1987), assuming that the multiple imputations are repeated imputations.

The key Bayesian motivation for multiple imputation is given by result 3.1 in Rubin (1987). Ignoring both technical details and indicator variables for sampling and response, the results and its consequences can be easily stated using the simplified notation that the complete-data are $Y = (Y_{\text{obs}}, Y_{\text{mis}})$, where Y_{obs} is observed and Y_{mis} is missing. Specifically, the basic result is

$$P(Q|Y_{\text{obs}}) = \int P(Q|Y_{\text{obs}}, Y_{\text{mis}})P(Y_{\text{mis}}|Y_{\text{obs}})\, dY_{\text{mis}},$$

or in words,

$$\begin{pmatrix} \text{actual posterior} \\ \text{distribution of } Q \end{pmatrix} = \text{AVE}\left[\begin{pmatrix} \text{complete-data posterior} \\ \text{distribution of } Q \end{pmatrix}\right],$$

where AVE[ ] refers to the average over the repeated imputations, which are draws from $p(Y_{\text{mis}}|Y_{\text{obs}})$, which is the posterior predictive distribution of missing data given the observed data. Two simple consequences follow (Rubin 1987, result 3.2). The first concerns the final estimate of Q:

$$E(Q|Y_{\text{obs}}) = E[E(Q|Y_{\text{obs}}, Y_{\text{mis}})|Y_{\text{obs}}],$$

or in words,

$$\begin{pmatrix} \text{Posterior mean} \\ \text{of } Q \end{pmatrix} = \text{AVE}\begin{bmatrix} \text{repeated complete-data} \\ \text{posterior means of } Q \end{bmatrix}.$$

The second concerns the final variance of Q:

$$V(Q|Y_{\text{obs}}) = E[V(Q|Y_{\text{obs}}, Y_{\text{mis}})|Y_{\text{obs}}] + V[E(Q|Y_{\text{obs}}, Y_{\text{mis}})|Y_{\text{obs}}],$$

or in words,

$$\begin{pmatrix} \text{Posterior} \\ \text{variance of } Q \end{pmatrix} = \text{AVE}\begin{bmatrix} \text{Repeated complete-data} \\ \text{variances of } Q \end{bmatrix} + \text{VAR}\begin{bmatrix} \text{repeated complete-data} \\ \text{posterior means of } Q \end{bmatrix},$$

where VAR refers to the variance over the repeated imputations. These simple relationships, which follow from standard probability calculations, underlie the repeated-imputation inferences recommended for practice.

2.2 Repeated-Imputation Inferences

The essential features of the repeated-imputation inference are the following. The repeated-imputation estimate is

$$\bar{Q}_m = \sum_{l=1}^{m} Q_{*l}/m, \tag{2.1}$$

and the associated variance–covariance of $\bar{Q}_m$ is

$$T_m = \bar{U}_m + \frac{m+1}{m} B_m, \tag{2.2}$$

where

$$\bar{U}_m = \sum_{l=1}^{m} U_{*l}/m = \text{ within-imputation variability}, \tag{2.3}$$

and

$$B_m = \sum_{l=1}^{m} (Q_{*l} - \bar{Q}_m)(Q_{*l} - \bar{Q}_m)'/(m-1)$$
$$= \text{ between-imputation variability}. \tag{2.4}$$

The large m repeated-imputation inference treats $(Q - \bar{Q}_m)$ as a normal random variable with variance–covariance matrix T_m; notationally, letting $m = \infty$ we have

$$(Q - \bar{Q}_\infty) \sim N(O, T_\infty), \tag{2.5}$$

where $T_\infty = \bar{U}_\infty + B_\infty$, and the eigenvalues of B_∞ relative to T_∞ measure the fractions of information missing about Q due to nonresponse.

The derivation of these expressions follows from the Bayesian perspective treating Q and $\hat{Q}$ as unobserved random variables with normal conditional distributions given the observed values $\{\hat{Q}_{*1}, \ldots, \hat{Q}_{*m}\}$ and $\{U_{*1}, \ldots U_{*m}\}$. For details, including specific small-m adjustments, see chapter 3 in Rubin (1987), and for more extensive results on p values, see Li, Raghunathan, and Rubin (1991), Li, Meng, Raghunathan, and Rubin (1991), and Meng and Rubin (1993).

2.3 Evaluating Repeated-Imputation Procedures Under the Randomization-Based Paradigm

The Bayesian paradigm, which is used to derive repeated-imputation inferences, is formally predicated on the correctness of all the model specifications. Although this paradigm is ideal for creating procedures, especially in complicated situations, its results cannot be unequivocally endorsed for routine practice because, in practice, we can never be sure any model assumptions are correct. Consequently, the Bayesianly-derived repeated-imputation procedures were evaluated in chapter 4 in Rubin (1987) under the randomization-based frequentist paradigm to investigate their sensitivity and robustness to model deviations and finite m. This paradigm extends that of Neyman (1934) to include a mechanism for nonresponse $\Pr(R|X, Y, I)$ in addition to the sampling mechanism $\Pr(I|X, Y)$, where I is the array of fully observed sampling indicators for which values of Y were included in the survey for observation, and R is the array of fully observed indicators for response (i.e., for which components of Y that were intended to be observed were observed). A component, Y_{ij}, is observed if both associated indicators, I_{ij} and R_{ij} are one, and is not observed if either is zero. This perspective is called the random-response randomization-based perspective.

2.4 Proper Multiple Imputation

A key concept underlying these randomization-based evaluations is that of *proper* multiple imputation, whose mathematical definition is purely frequentist, since it involves expectations given that the population values (X, Y) are fixed. The crucial result is that when (1) the multiple imputations are proper for $(\hat{Q}, U)$, and (2) the complete-data inference based on $(\hat{Q}, U)$ is randomization-valid for Q, then the large-m repeated-imputation inference given by (2.5) is randomization-valid for the scientific estimand Q, *no matter how complex the survey design.* Whether a multiple imputation procedure is proper depends, in general, on which complete-data estimates, $\hat{Q}$, and associated variance-covariance matrices, U, are being considered. The full definition is given in Rubin (1987, pp. 118–119); it is summarized here ignoring the more technical conditions in order to focus attention on three essential conditions.

The definition of a proper multiple imputation procedure treats (X, Y) *and* the intended sample (as indicated by I) as fixed [except for a minor technical condition—eq. (4.2.9) in Rubin 1987], and deals with the fixed but unknown values of the complete-data statistics $(\hat{Q}, \mathbf{U})$ in the sample as if they were estimands. That is, the randomization distribution critically involved in the definition of proper multiple imputation is generated by the response mechanism, in which X, Y, and I are fixed, and R is the random variable. Because the conditions for proper imputation involve large m, the simplified definition only involves expectations with respect to the response mechanism.

For proper imputation, the values of the complete-data statistics $\hat{Q}$ and U created by filling in the missing Y values, that is $\hat{Q}_{*l}$ and U_{*l}, must be approximately unbiased for their complete-data analog $\hat{Q}$ and U; that is, in terms of the large-m averages of $\hat{Q}_{*l}$ and U_{*l}:

$$E(\bar{Q}_\infty|X, Y, I) \doteq \hat{Q} \tag{2.6}$$

and

$$E(\bar{U}_\infty|X, Y, I) \doteq U. \tag{2.7}$$

Moreover, B_∞, which is the variance–covariance of the $\hat{Q}_{*l}$ across the m imputations, must be approximately unbiased for the randomization variance of $\bar{Q}_\infty$:

$$E(B_\infty|X, Y, I) \doteq \text{var}(\bar{Q}_\infty|X, Y, I). \tag{2.8}$$

Equation (2.6) for proper imputation is analogous to (1.1) for randomization validity: both require approximate unbiasedness of the estimate ($\bar{Q}_\infty$ or $\hat{Q}$) for its estimated ($\hat{Q}$ or Q) over its randomization distribution (induced by the response mechanism or the sampling mechanism). Equation (2.8) for proper imputation is analogous to (1.2) for randomization validity: both require approximately unbiased estimation by the ancillary statistic (B_∞ or U) for the variance of the estimate ($\bar{Q}_\infty$ or $\hat{Q}$) over its randomization distribution (induced by the response or sampling mechanism). Also, just as (1.1) and (1.2) together imply (at least in large-sample surveys) that randomization-valid inferences for Q can be based on the approximation

$$(\hat{Q}|X, Y) \sim N(Q, U),$$

(2.6) and (2.8) together imply that randomization-valid inferences for the complete-data statistics $\hat{Q}$ can be based on the approximation

$$(\bar{Q}_\infty|X,Y,I) \sim N(\hat{Q}, B_\infty),$$

where the randomization distributions are induced by the sampling and response mechanisms, respectively. The remaining condition for proper imputation has no direct analog in complete-data randomization validity: expression (2.7) implies that the complete-data ancillary statistic U, being treated as an ancillary complete-data estimand for the definition of proper imputation, is approximately unbiasedly estimated after imputation.

2.5 Conclusion Regarding Randomization Validity With Proper Multiple Imputation

The crucial result regarding the randomization validity of the large-m repeated-imputation inference, given by (2.5), averages over both the actual sampling mechanism and the posited response mechanism; it is simple and holds no matter how complex the survey design:

Result 4.1: If the complete-data inference is randomization-valid and the multiple-imputation procedure is proper, then the infinite-m repeated-imputation inference is randomization-valid under the posited response mechanism. (Rubin 1987, p. 119).

This result follows from combining the formal versions of (1.1), (1.2), (2.6), (2.7), and (2.8). Essentially, (1.1) and (2.6) imply that

$$E(\bar{Q}_\infty|X,Y) = E[E(\bar{Q}_\infty|X,Y,I)|X,Y] = E(\hat{Q}|X,Y) = Q,$$

and (1.2), (2.7), and (2.8) imply that

$$\begin{aligned}
E(T_\infty|X,Y) &= E(\bar{U}_\infty|X,Y) + E(B_\infty|X,Y) \\
&= E[E(\bar{U}_\infty|X,Y,I)|X,Y] \\
&\quad + E[E(B_\infty|X,Y,I)|X,Y] \\
&= E(U|X,Y) + E[\mathrm{var}(\bar{Q}_\infty|X,Y,I)|X,Y] \\
&= \mathrm{var}(\hat{Q}|X,Y) + E[\mathrm{var}(\bar{Q}_\infty|X,Y,I)|X,Y] \\
&= \mathrm{var}[E(\bar{Q}_\infty|X,Y,I)|X,Y] \\
&\quad + E[\mathrm{var}(\bar{Q}_\infty|X,Y,I)|X,Y] \\
&= \mathrm{var}(\bar{Q}_\infty|X,Y).
\end{aligned}$$

Thus approximately (2.5) follows, which is the conclusion of Result 4.1.

Rubin (1987, chap. 4) presented analytic results, simulation evaluations, and many examples of proper and improper multiple imputation methods, where the evaluations were all from the random-response randomization-based frequentist perspective. The trick in many of the examples of proper imputation was to get the variance condition (2.8) correct, and it was shown that when drawing imputations to approximate repetitions from a sensible Bayesian model, conditions (2.6)–(2.8) typically followed automatically. The more straightforward conditions, (2.6) and (2.7), typically were simple properties of any intelligent imputation scheme that tried to track the data. An example of a method that does not track the data is "fill in the mean," which although it may satisfy (2.6) for $\hat{Q} = \bar{y}$, fails to do so for $\hat{Q} = s^2$ or for $\hat{Q} =$ 25th percentile, or to satisfy (2.7) for $U = s^2/n$, etc. Hot-deck (Bootstrap) and random-draw regression methods tend to satisfy (2.6) and (2.7) but fail to satisfy (2.8) until a Bayesian, systematic between-imputation component of variability is added (e.g., via the Bayesian Bootstrap, Rubin 1981), to reflect uncertainty in the estimation of population parameters.

The view in 1987, which I still hold today, was summarized as follows.

Conclusion 4.1: If imputations are drawn to approximate repetitions from a Bayesian posterior distribution of Y_{mis} under the posited response mechanism and an appropriate model for the data, then in large samples the imputation method is proper. ... There is little doubt that if this conclusion were formalized in a particular way, exceptions to it could be found. Its usefulness is not as a general mathematical result, but rather as a guide to practice. Nevertheless, in order to understand why it may be expected to hold relatively generally, it is important to provide a general heuristic argument for it (Rubin 1987, pp. 125–126).

This heuristic argument treated the sample as the population with estimand $\hat{Q}$ (and U), where the resulting posterior distribution of $\hat{Q}$ was centered at $\bar{Q}_\infty$ with variance B_∞; assuming the Bayesian model appropriate [in the sense of satisfying (2.6) and (2.7)] and the samples large, standard arguments presented in chapter 2 of Rubin (1987) suggested that typically $(\hat{Q} - \bar{Q}_\infty)B_\infty^{-1/2}$ will have a sampling distribution (over the response mechanism) that is standard normal, thereby satisfying the basic conditions for proper multiple imputation.

2.6 Include All Variables in a Multiple Imputation Model To Make It Proper in General

The definition of proper concerns the situation where: "population" = complete-data sample, "estimands" = complete-data statistics $(\hat{Q}, U)$, "survey design" = the posited response mechanism, the criterion is valid frequency inference, and the method for creating inferences is Bayesian predictive inference using simulated values (i.e., multiple imputations). As with any finite population survey where valid frequency inference is desired from predictive procedures: (1) variables involved in the definition of estimands (i.e., $\hat{Q}, U$) should be predicted, and (2) variables involved in the survey design (i.e., the response mechanism) should be used as predictors. More explicitly, when $\hat{Q}$ or U involves some variable X, then leaving X out of the imputation scheme is improper and generally leads to biased estimation and invalid survey inference. For example, if X is correlated with Y but not used to multiply-impute Y, then the multiply-imputed data set will yield estimates of the XY correlation biased towards zero. In a complex survey, $\hat{Q}$, and especially U, depend on stratification and clustering indicators; consequently, in general these indicators need to be included as predictor variables in imputation models for the multiple imputation scheme to be proper. Minimally, major clustering and stratification indicators and sample de-

sign weights (or estimated propensity scores of being in the sample) should be included in imputation models. Ezzati-Rice, Johnson, Khare, Little, Rubin, and Schafer (1995) illustrates such efforts and the resulting valid inferences.

Since with public-use data sets it is always unclear what analyses the ultimate users will conduct, the range of statistics $(\hat{Q}, U)$ that might be used involves essentially any variable or combination of variables available in the data set, at least up to some level of interactions. Thus, the danger with an imputer's model is generally in leaving out predictors rather than including too many, and the advice has always been to include as many variables as possible when doing multiple imputation. The press to include all possibly relevant predictors is demanding in practice, but it is generally a worthy goal. For example, in the original prescription for the industry and occupation recoding project (Rubin 1983), thousands of logistic regressions were done, each with nearly 20 variables, and some with far fewer than 20 observations (e.g., 4), in order to preserve this theme of trying to include all variables that might be used to define statistics $\hat{Q}$ or U; this effort required the development of specialized but computationally efficient Bayesian logistic regression procedures for sparse data (Clogg, Rubin, Schenker, Schultz, and Weidman 1991). The possible lost precision when including unimportant predictors is usually viewed as a relatively small price to pay for the general validity of analyses of the resultant multiply-imputed data base.

2.7 Some Experience With Useful But Improper Multiple Imputation

In some cases, improper multiple imputations can still lead to confidence-valid repeated-imputation inferences. This issue will be discussed in more detail in Sections 3.5–3.8 in reply to a recent criticism of multiple imputation, but the issue has been previously considered. Rubin and Schenker (1987, sec. 7) explicitly consider the situation in the early industry and occupation example where some information used by the imputer (the original double-coded sample) is not available to the data analyst, and demonstrate the resulting potential conservative coverage. Also, the evaluations of the results of this project include cases where the data analyst uses variables not used by the imputer and, for this data set and practical analyses, find no deleterious consequences (Schenker, Treiman, and Weidman 1993; Treiman, Bielby, and Cheng 1989; Weld 1987). Careful and extensive evaluations of this general situation, involving variables omitted by the imputer, are also included in work conducted at ETS in the context of NAEP, which for a decade has created multiply-imputed public-use data bases (e.g., Mislevy, Johnson, and Muraki 1992).

Substantial empirical work, some given in the Appendix, supports the conclusion that, even if mildly important predictors are left out of the multiple imputation scheme, the repeated-imputation inferences are confidence-valid: with fractions of missing information typical in careful surveys, $m = 3$ or 5 works very well, with the complete-data procedure for small n typically breaking down before multiple imputation does. A heuristic reason for this robustness is that lack of model fit goes into residual variance, which in a Bayesian model inflates the between-imputation variance of draws (e.g., of regression coefficients), thereby leading to a large enough B_m to compensate for an omitted coefficient. Of course, this is an observation based on some experience, not a theorem, but a related theoretical result (Meng 1994, lem. 2) lends support to this observation.

Nevertheless, because problems can occur when the imputer's model leaves out important predictor variables, the data-base constructor must include a description of the imputation model with the multiply-imputed data base, so that ultimate users know which relationships among variables have been implicitly set to zero.

3. CURRENT ISSUES CONCERNING MULTIPLE IMPUTATION

There appear to be two distinct kinds of concerns about multiple imputation. The first type focuses on its implementation: operational difficulties for the data-base constructor and the ultimate user, as well as the acceptability of answers obtained partially through the use of simulation. The second type concerns the frequentist validity of repeated-imputation inferences when the multiple imputations are not proper, but appear "reasonable" in some sense.

3.1 Is Multiple Imputation Unprincipled or Unacceptable Because it Uses Simulation?

An early criticism, not much heard anymore but worthy of response, is that multiple imputation is theoretically unsatisfactory and practically unacceptable because it adds random noise to the data. In this context, it is critical to remember that multiple imputation does not pretend to *create* information through simulated values but simply to *represent* the observed information this way so as to make it amenable to valid analysis using complete-data tools. The extra noise created when using a finite number of imputations is the price to be paid for this luxury.

In response to this criticism, first appreciate that simulation methods are becoming more and more common and accepted in statistics. Consider jackknife and bootstrap methods for complete-data frequentist inference (e.g., Miller 1974; Efron and Tibsharani 1993), or data augmentation (Tanner and Wong 1987), the Gibbs sampler (e.g., Gelfand and Smith 1990; Gelman and Rubin 1992), and sampling importance resampling methods (Rubin 1983, 1987, 1988; Gelfand and Smith 1992) for complete-data Bayesian inference. These methods have now become accepted complete-data tools worthy of theoretical investigation and routine practical application.

Second, multiple imputation has a distinct advantage over such methods in principle, because with multiple imputation, the simulation is only being used to handle the missing information, with reliance for handling the rest of the information left to the complete-data method, be it analytic or simulation-based. Thus, the acceptable number of imputations can be *much* less than the acceptable number of simulations for a complete-data inference, at least assuming that the fraction of missing information, γ, is modest

Table 1. Approximate Factor for Inflating Normal Standard Errors in (2.5) to Reflect Finite Number of Imputations, m: $\sqrt{\nu/(\nu-2)}$, Where $\nu = (m-1)[1+(1+m^{-1})B_m/\bar{U}_m]^2$

	γ				
m	*10%*	*20%*	*30%*	*50%*	*70%*
3	1.01	1.03	1.07	1.22	1.53
5	1.00	1.01	1.03	1.08	1.17
10	1.00	1.01	1.01	1.03	1.06
20	1.00	1.00	1.01	1.01	1.03

(e.g., $< 30\%$) as commonly occurs in public-use surveys. More explicitly, few would recommend basing standard errors on fewer than 100 bootstrap or jackknife simulations, hundreds or thousands being more typical. In contrast, typically as few as five multiple imputations (or even three in some cases) is adequate under each model for nonresponse. Two simple calculations help to illustrate why only a few imputations can be adequate. First, the asymptotic efficiency of the repeated-imputation finite-m estimate relative to the infinite m estimate is $[1+(\gamma/m)]^{-1/2}$ in units of standard deviations, which is close to one with realistic fractions of missing information and modest m (Rubin 1987, table 4.1). Second, Table 1 displays an approximate factor for expanding standard errors in the infinite-m normal distribution (2.5) to reflect finite m. The table shows that the expansion of width of a confidence interval due to finite m is modest for most practical cases.

Finally, even when a particular multiple implementation method has deficiencies, it can only distort part of the inference in contrast to an incorrect complete-data analysis, which can distort the entire inference. For example, results in Heitjan and Rubin (1990) in a particular example suggest that doing some kind of multiple imputation, even if under a naive model, is far better inferentially than standard or sophisticated approaches with single imputation. In some vague sense, if a multiple imputation method is 20% deficient (80% okay) with 30% missing information, its total distortion is 20% of 30%, or 6%, implying that the repeated-imputation inference is 94% okay.

3.2 Is Multiple Imputation Too Much Work For The User?

My primary response to this question is: "Too much work relative to doing what?" Multiple imputation is intellectually trivial for the user. Running the identical complete-data software m times (e.g., 3, 5, or 10 times) and combining the results "by hand" is admittedly a burden, but is computationally trivial given appropriate macros (which are easy to write, e.g., in S-Plus; see Schafer 1995, or SAS, Freedman 1990). I believe it is substantially easier for the user, even without appropriate macros, than any other method that can validly address nonresponse in any generality. As repeatedly emphasized by many workers in this area, methods such as "fill in the mean and ignore," "available cases," "treat the data set as a two-way additive model and singly impute with zero interaction," etc., are not statistically valid in any generality, even for point estimation of a variety of estimands, such as means, variances, correlations, and are therefore not appropriate for public-use data bases.

3.3 Does a Multiply-Imputed Data Set Take Too Much Storage?

A multiply-imputed data set, in terms of needed storage locations, is $[1 + m \cdot (\%$ missing values overall$)]$ times as big as the original data set, typically a factor of two or less. For example, suppose the data set has 10,000 units; 20 background variables fully recorded; 20 "easy" survey questions, 5% missing; 30 "moderate" survey questions, 10% missing; 30 "difficult" survey questions, 30% missing: then the complete-data set = 1,000,000 items with 130,000 items missing. The associated multiply-imputed ($m = 5$) data set consists of the complete-data set of 870,000 data values and 130,000 pointers to the rows of the supplemental 130,000 $\times$ 5 matrix of imputations, for a total of 1,000,000 + 650,000 locations. Given the appropriate macros, we can unpack the multiple imputations to create five completed data sets only at the time of each of the five complete-data analyses, sequentially, in a manner transparent to the ultimate user, and using less than twice the storage needed for the original data set. Even with more missing values and more imputations per missing value, this issue should be easily handled with today's storage devices and simple and general macros, although it can be a burden without appropriate software. In situations with nonresponse confined to a few variables, an effective device can be to create a rectangular data set with m versions of these variables but one version of the fully observed variables.

3.4 Does It Take Too Much Work to Create Proper or Approximately Proper Multiple Imputations?

Again, my response to this question is "too much relative to what?" It certainly takes much more work than some methods that have no general validity. But multiple imputation takes little more work than other methods that attempt to address nonresponse validly and with some generality. Moreover, essentially all the extra work is needed from the data-base constructor, who may have the resources to do the job well, rather than the world of ultimate users with their varied and limited resources. In fact, some experience suggests that in practice it may be substantially easier to do model-based multiple imputation than to use previous approaches because we can apply powerful methods of direct and indirect simulation under full probability models (e.g., data augmentation, the Gibbs sampler) and let the computer do much of the work previously done by expensive and exhausting human iteration; consider, for example, the recent project dealing with nonmonotone missing data patterns in NHANES (Fahimi and Judkins 1993; Schafer et al., 1993; Ezzati-Rice et al. 1993; Johnson et al. 1993; and Little and Rubin 1993). For other examples dealing with the creation of multiple imputations and related issues, consider Kennickell (1991); Chand and Alexander (1994); Paulin and Ferraro (1994); and Eltinge, Yansaneh and Paulin (1994).

3.5 Can Repeated Imputations Under An Appropriate Bayesian Model Lead to Invalid Inferences?

Fay (1991, 1992, 1993; also see Kott 1992) claims that even when the model used to create repeated imputations is "appropriate" in some sense, the resulting repeated inferences can be invalid. I believe that this criticism is misguided for a variety of reasons, many of which have been exposed in the work and discussion of Meng (1994). Nevertheless, I will also briefly address the issue here because it has received attention, and because I believe my results, although less extensive and detailed than those of Meng (1994), will be more transparent to many readers.

The kernel of this criticism arises when an irrelevant predictor X of outcome Y *is not* used by the Bayesian multiple imputer to create repeated imputations, but *is* used by the ultimate analyst to define estimands (a case already introduced here in Sec. 2.7 because of historical discussion of it). More specifically, suppose X is dichotomous, (a, b), and Y is normal (0, 1) and independent of X in a population in which $X = a$ units and $X = b$ units are equally represented. Suppose a stratified random sample of size $2n$ is taken where there are n units with $X = a$ and n units with $X = b$, and further suppose that nonresponse is simply like another level of stratified random sampling that results in n_1 respondents and n_0 nonrespondents in both the $X = a$ sample and in the $X = b$ sample. The estimands are: $\bar{Y} = (\bar{Y}_a + \bar{Y}_b)/2$, the population mean value of Y; and $\bar{D} = (\bar{Y}_a - \bar{Y}_b)$, the population difference of means, which equals zero. The obvious complete-data estimators are $\bar{y} = (\bar{y}_a + \bar{y}_b)/2$ for $\bar{Y}$ and $\bar{d} = (\bar{y}_a - \bar{y}_b)$ for $\bar{D}$, with associated standard complete-data variance estimates $U_{\bar{y}}$ and $U_{\bar{d}}$, respectively, which result in randomization-valid complete-data inferences, at least for large n.

Now suppose repeated imputations for the $2n_0$ nonrespondents are generated using a fully exchangeable normal model based on the $2n_1$ respondents. That is, the imputations for both the $X = a$ and $X = b$ units will be centered at the observed grand mean $\bar{y}_{\text{obs}}$ rather than at the separate observed sample means $\bar{y}_{\text{obs},a}$ and $\bar{y}_{\text{obs},b}$. It is easy to show that the multiple imputation method is proper for $(\bar{y}, U_{\bar{y}})$, but it is improper for $(\bar{d}, U_{\bar{d}})$: (1), the expectation of $\bar{d}_\infty = (n_1/n)(\bar{y}_{\text{obs},a} - \bar{y}_{\text{obs},b})$ over the response mechanism, that is given (X, Y, I), does not equal $\bar{d}$, but $(n_1/n)\bar{d}$, thereby not satisfying (2.6) [nor (4.2.5) in Rubin 1987]; and (2) B_∞, the variance of the repeated values of $\bar{d}_{*l}$ across repeated imputations with fixed n_0, is greater than the variance of $\bar{d}_\infty$ over the response mechanism by the factor n/n_1, thereby not satisfying (2.8) [nor (4.2.6)–(4.2.7) in Rubin 1987].

3.6 Superefficient Imputations

In this example, the imputations are "superefficient" from the perspective of the data analyst interested in estimating $\bar{D}$ because the imputations use "extra" information, specifically the knowledge that the distribution of Y given $X = a$ is identical to the distribution of Y given $X = b$. For a more familiar example of superefficiency, if the data are normal with mean zero, then half the sample mean is a superefficient estimate of the population mean. The situation involving superefficient imputations is more subtle, however. Suppose that we have a multiply-imputed data set, but subsequently the data collector brings forth values of the missing data, thereby allowing us to calculate $\hat{Q}$ and U. Presumably, we would then be inclined to base our inferences for Q on $(\hat{Q}, U)$ and discard the imputations. If the imputations are superefficient, however, the standard complete-data procedure can be improved by using information in the imputations about Q beyond that in $\hat{Q}$, information supplied by the imputer (e.g., in the canonical example, the knowledge that $X = a$ units and $X = b$ units have the same population distribution of Y). The imputations are "strongly superefficient" if $\bar{Q}_\infty$ is at least as good an estimate as $\hat{Q}$ despite the existence of missing data, that is, despite the fact that $\bar{Q}_\infty$ is not identical to $\hat{Q}$ in the formal sense that

$$\text{var}(\bar{Q}_\infty - \hat{Q}|X, Y) > 0, \tag{3.1}$$

where with vector Q, ">" means at least one eigenvalue > 0.

More precisely, a multiple imputation procedure is strongly superefficient for the complete-data statistic $\hat{Q}$ if, first, $\bar{Q}_\infty$ and $\hat{Q}$ estimate the same estimand, that is, the procedure is "first-moment proper" for $\hat{Q}$,

$$E(\bar{Q}_\infty|X, Y) \doteq E(\hat{Q}|X, Y), \tag{3.2}$$

and second, $\bar{Q}_\infty$ has no larger variance than the complete-data estimate itself:

$$\text{var}(\bar{Q}_\infty|X, Y) \leq \text{var}(\hat{Q}|X, Y), \tag{3.3}$$

where with vector Q, (3.3) compares the generalized eigenvalues of the left side with respect to the right side. In the canonical example of Section 3.5, the imputations are strongly superefficient for $\hat{Q} = \bar{d}$ because $\bar{Q}_\infty = \bar{d}_\infty$ satisfies both (3.2) and (3.3).

The general definition of superefficiency concerns the existence of imputations that make $\bar{Q}_\infty$ informative about Q even with knowledge of $\hat{Q}$. Bayesian models can be superefficient when they incorporate appropriate smoothing information in their distributional assumptions. The resultant draws of Y_{mis} cannot be sharper than those from the parent distribution and still lead to valid inferences for a variety of estimands, but multiple imputations of Y_{mis} can be more efficient than the one true value of Y_{mis} because of their multiplicity. For instance, in the canonical example, suppose that the multiple imputation procedure drew the group difference effect from a normal distribution centered at $\frac{1}{2}(\bar{y}_{\text{obs},a} - \bar{y}_{\text{obs},b})$ rather than at $\bar{y}_{\text{obs},a} - \bar{y}_{\text{obs},b}$ (as when this effect is directly estimated from the data) or at zero (as with the strongly superefficient imputations of sec. 3.5). These imputations would effectively be additional data values, which could contribute to a better estimate of $\bar{D}$, even if the actual missing values were found. The general definition of superefficient imputations for $\hat{Q}$ replaces (3.3) with

$$\text{cov}(\bar{Q}_\infty, \hat{Q}|X, Y) < \text{var}(\hat{Q}|X, Y); \tag{3.4}$$

strong superefficiency implies superefficiency because (3.1) and (3.3) imply (3.4).

Table 2. Analysis of Results from Fay (1992)—Nominal 95% Intervals

	Multiple imputation ($m = 10$)			Rao and Shao	
	Statistics	Width	Estimated coverage	Width	Estimated coverage
Fay Table 1	Y	.24	95%	.26	95%
	Y_a	.33	97%	.33	95%
Fay Table 2	Y	.24	95%	.26	95%
	Y_a	.33	97%	.33	95%
	Y_s	.33	95%	.37	95%

NOTE: Y represents the sample mean; Y_a, sample mean for class a, not used in imputation; Y_s, sample mean in class s, used in imputation.

3.7 Confidence-Proper Multiple Imputation

We are now ready to provide an extended definition of proper imputation and state an extended result concerning frequency validity. Although the conditions and conclusion are similar to the major conclusions of Meng (1994), they are more direct and not as extensive since they avoid the issues of the ultimate user's incomplete-data procedure and congeniality between the imputer's and analyst's models. The definition of "confidence-proper" multiple imputation is still in terms of the complete-data statistics $(\hat{Q}, U)$, but involves averaging over both the response mechanism and the sampling mechanism and allows overestimation of between-imputation variability.

A multiple imputation procedure is *confidence-proper* for the complete-data statistics $(\hat{Q}, U)$ if the imputations are "first-moment proper" for $(\hat{Q}, U)$ in the sense of (3.2) and (3.5),

$$E(\bar{U}_\infty | X, Y) \doteq E(U | X, Y), \tag{3.5}$$

and if B_∞ conservatively estimates the "excess variance" of $\bar{Q}_\infty$ over $\hat{Q}$:

$$E(B_\infty | X, Y) \geq \text{var}(\bar{Q}_\infty | X, Y) - \text{var}(\hat{Q} | X, Y). \tag{3.6}$$

If a multiple imputation procedure is proper for $(\hat{Q}, U)$ it is confidence proper for $(\hat{Q}, U)$; (2.6) implies (3.2), (2.7) implies (3.5), and (2.6) with (2.8) implies (3.6) with equality. If a multiple imputation procedure is strongly superefficient for $\hat{Q}$ and first-moment proper for U, then it is confidence proper for $(\hat{Q}, U)$; (3.2) and (3.5) hold, and (3.3) implies that (3.6) holds for any B_∞. A superefficient multiple imputation procedure for $\hat{Q}$ is confidence proper for $(\hat{Q}, U)$ if it is first-moment proper for U, and if

$$E(B_\infty | X, Y) \geq \text{var}(\bar{Q}_\infty - \hat{Q} | X, Y); \tag{3.7}$$

(3.2) and (3.5) hold, and (3.40 implies that the right side of (3.7) is greater than the right side of (3.6), thereby satisfying (3.6). A "second-moment proper" imputation method (Meng 1995, p. 548) is defined by (3.2), (3.5), and equality in (3.7).

Analogous to Result 4.1 we have the following result:

Result on Confidence Validity. If the complete-data inference based on $(\hat{Q}, U)$ is confidence valid and the multiple imputation procedure is confidence proper for $(\hat{Q}, U)$, then the repeated-imputation inference is confidence valid with

$$E(\bar{Q}_\infty | X, Y) \doteq Q$$

and

$$E(T_\infty | X, Y) \geq \text{var}(\bar{Q}_\infty | X, Y).$$

The result follows because (3.2) and (1.1) imply that $\bar{Q}_\infty$ is approximately unbiased for Q, and (1.2), (3.5), and (3.6) together imply that $\bar{U}_\infty + B_\infty$ conservatively estimates $\text{var}(\bar{Q}_\infty | X, Y)$.

In the canonical example, the strong superefficiency in the imputer's model for $\bar{D}$ implies that the data analyst's resultant repeated-imputation interval for $\bar{D}$ will have at least nominal coverage and hence will be confidence-valid; whether it is superior or inferior to other valid procedures depends on its interval length and the lengths of intervals from other confidence-valid procedures.

The conclusion, however is as before: try to impute using a Bayesian or approximate Bayesian model that tracks the data and the posited response mechanism—if you do this and your complete-data inference is confidence-valid, the result will be confidence-valid repeated-imputation inferences *no matter how complex the survey design.*

3.8 Confidence Validity Versus Randomization Validity in Canonical Example

Fay (1991, 1992) effectively claims that (a) wider 95% confidence intervals with exact 95% (asymptotic) coverage are superior to (b) narrower 95% confidence intervals with at least 95% coverage. Specifically, in the discussion of tables 1 and 2 of Fay (1992), summarized here in Table 2 after a bit of analysis to produce approximate coverage, the claim is made that the Rao and Shao (1992) (RS) procedure, using single-imputation hot deck, which results in uniformly wider intervals but with asymptotic coverage equal to the confidence coefficient, is inferentially superior to the multiple-imputation version of the same procedure (MI), which results in uniformly narrower intervals with asymptotic coverage at least as great as the confidence coefficient. Both procedures as reported are confidence valid, and I believe many statisticians and scientists would agree with Neyman's criteria and prefer sharper intervals with at least 95% coverage rather than wider intervals with exact 95% coverage.

Fay (1993) repeats the same criticism as Fay (1992) in more extreme examples (e.g., with up to 80% nonresponse) and labels the confidence coverage of the repeated-imputation inference as "punishingly conservative." But from the analyst's perspective, punishingly conservative relative to what alternative procedure? Presumably relative to what would have happened if the imputer had done what the analyst expected, that is, had used the analyst's model for imputation rather than be superefficient. But that would have led to wider intervals with exactly nominal coverage—a valid procedure, but less preferred according to the Neyman definition and scientific criteria, than narrower intervals with greater coverage.

Of course, the confidence validity of the repeated-imputation inference does not mean it yields the best confidence-valid interval. By our mathematical analysis in this simple example we know that a shorter 95% confidence interval can be found with exact 95% coverage. Also, be-

cause the procedure is confidence valid but not randomization valid, inefficiencies can arise when combining various estimates using the assigned precisions as weights. But finding a randomization-valid procedure in general requires extra work beyond the use of standard complete-data methods, and is generally impossible for the ultimate user unless extra information is conveyed by the data-base constructor. Furthermore, this whole issue seems relatively unlikely to arise in practice because knowledge of population parameters by the data-base constructor must be unusual.

3.9 Reaching Correct Conclusions When Evaluating Multiple Imputation

Several points are critical in reaching correct conclusions concerning multiple imputation.

First, when evaluating repeated-imputation inferences by analysis or simulation, we need to monitor whether the complete-data inference with no missing data is valid: multiple imputation for missing data cannot fix problems with complete-data analyses (e.g., poor coverage properties of the normal approximation for the sample mean with rare binomial trials, where, for example, logit transforms can lead to more accurate complete-data inferences); Rubin and Schenker (1986) and Ezzati-Rice et al. (1995) provide examples of such evaluations. Also note when evaluating these procedures with the number of respondents fixed (e.g., as in sec. 4.3 and prob. 4-18, in Rubin, 1987) that the resultant answers are conditional on these quantities, which in practice are random. Moreover, when doing evaluations treating the number of respondents as random, the theoretical variances of unbiased estimators can be undefined, since, for any finite sample size, with positive probability, all units will be nonrespondents; in such cases, it makes more sense to report coverage properties of interval estimates, which are defined (no respondents implies zero coverage) and the objects of statistical inference anyway.

Also important in reaching correct conclusions about multiple imputation is the treatment of estimated sampling variances as ancillary statistics rather than as estimates of scientific estimands. For example, Fay (1992) treated the ratio of repeated-sampling covariances as an estimand, and thereby was led to misunderstand the effect of superefficient imputation on inference. This illustrates why it is important not to confuse scientific estimands and ancillaries. In particular, Fay (1992, sec. 3) states, in the context of the canonical example of Section 3.5:

> ... the design-based approach gives 19 times the covariance of multiple imputation ... such a limitation, if general, imposes severe restrictions on the validity of the multiple imputation inferences for complex applications, such as Clogg et al. (1991).

Consider the true sampling variance–covariance ellipsoid for $(\bar{y}_\infty, \bar{d}_\infty)$ under the exchangeable normal repeated-imputation scheme and the sampling ellipsoid for $(\bar{y}_\infty, \bar{d}_\infty)$ assigned to it by the repeated-imputation inference; both have zero correlation because $\bar{y}_{\infty,a} = (2\bar{y}_\infty + \bar{d}_\infty)$ and $\bar{y}_{\infty,b} = (2\bar{y}_\infty - \bar{d}_\infty)$ have equal variance. The repeated-imputation–assigned ellipsoid is outside because it touches the correct one at the two points along the $\bar{y}_\infty$ axis but is wider along the $\bar{d}_\infty$ axis. Using Fay's ratio of sampling covariances of $\bar{y}_{\infty,a}$ and $\bar{y}_{\infty,b}$ is equivalent to describing the difference between these two ellipsoids by the ratio of *differences* of variances (i.e., of eigenvalues) of $2\bar{y}_\infty$ and $\bar{d}_\infty$ in the two ellipsoids. The ratio of eigenvalues, or of variances in any direction, is relevant to inference, but the ratio of differences between eigenvalues, Fay's measure, is by itself, irrelevant.

4. COMPETING METHODS

If multiple imputations are proper (confidence proper) under the posited model for nonresponse, then using the repeated-imputation rules for combining complete-data statistics $(\hat{Q}, U)$ yields a randomization-valid (confidence-valid) final inference under the posited response mechanism, assuming that the complete-data inference was valid in the absence of nonresponse. And this holds no matter how complex the survey design. Moreover, the combining rules can be implemented using completely general software that is the same for all data sets and all complete-data analyses. Thus multiple imputation and the repeated-imputation combining rules satisfy both the basic objective and the supplemental achievable objective.

Are there competing methods that, in some cases at least, also satisfy these objectives? Yes, but such competitors appear to me in general to have substantially greater deficiencies for the intended situation with ultimate users distinct entities from database constructors. These competitors are single imputation, multiple imputation with some analysis for the ultimate user other than the repeated-imputation inference, and weighting methods.

4.1 Desiderata for Creating Imputations, Single or Multiple

If imputations are to be used, then the estimate will be the value of $\hat{Q}$ calculated on the imputed data, or the average of multiple values $\hat{Q}_{*l}, l = 1, 2, \ldots$. In broad generality, consistent estimation requires that the imputation method must be first-moment proper, in the sense of (3.2), for a variety of statistics $\hat{Q}$, for example $\hat{Q} =$ sample mean, sample variance, median, 25th percentile, factor loadings, and these quantities within strata, domains, subdomains, etc. For this to hold for each $\hat{Q}$ in such a range, the imputation method, single or multiple, must in general not only track the posited response mechanism but also must be a random draw method; otherwise, it cannot be first-moment proper for $\hat{Q} = \bar{y}, \hat{Q} = s^2, \hat{Q} =$ 25th percentile, etc.

Consequently, any imputation method that satisfies the validity objective in generality must not only reflect the underlying response mechanism but must also be a random draw method. Nonrandom draw methods can be applied in special cases but require special analysis techniques. The most careful work on this topic of deterministic imputation of which I am aware concerns imputing probabilities for missing dichotomous variables (Schenker 1989; Schafer and Schenker 1991), and this work reveals the substantial extra effort that is needed, even in a special situation.

When an imputation method is a random-draw method, then multiple draws will automatically provide the basis for improved efficiency of estimation and more accurate inference, and are no more difficult to obtain than a single random-draw imputation. Thus multiple imputation is more attractive than single imputation, and the larger m the better, no matter how variances are to be calculated from the multiply-imputed data set. Little (1988) provides additional discussion of desiderata for creating imputations, which is consistent with this position.

4.2 Imputation in Random Independent Replicates—An Alternative

Suppose the sampling mechanism is such that the primary sampling units can be randomly divided into K replicate groups, each with the same sample design. Then with complete data, $\hat{Q}$ can be calculated in each replicate, and a valid $(K-1)$ df estimate of the variance of the average of the K independent estimates, $\bar{Q} = \Sigma\hat{Q}/K$, can be found from their sample variance divided by K. This can be called the "random group estimator" (Wolter 1985).

This approach has been used with single imputation for missing data; I believe the method is appropriately attributed to Morris Hansen, but I cannot find the appropriate early reference (a relatively recent reference is Kalton 1983, pp. 112–123). Random-draw imputations are made in the K independent random replicates of the survey units, so that the variance of K values of $\hat{Q}$ on the imputed data is a $K-1$ df estimated variance of $\bar{Q}$ (or $\hat{Q}$ calculated on the full imputed data); this estimate reflects not only sampling variability but also increased variance due to imputation. In personal communications, Hansen realized the propriety of the use of multiple imputations within each independent replicate to reduce variance due to imputation, and realized the potential tremendous loss of efficiency by doing the imputations independently in each independent replicate. In Rubin (1990), when discussing a related approach with energy data (Burns 1990), I called the resulting estimate of uncertainty an estimate of "evaluation variance" in contrast to "inferential variance" because it evaluates the variability of the estimation procedure, perhaps including excessive variability due to an efficient procedure used to handle missing data.

Assuming the requisite richness of survey data to allow the independent replicate procedure to be applied and assuming that the imputation method is first-moment proper, Hansen's method almost satisfies the basic and validity objectives, without needing the second-moment conditions involved in proper or confidence-proper imputations; I say "almost" because the ultimate user must be willing to forgo variance estimation aspects of the complete-data analysis programs, and rely on the potentially far less efficient variance estimation via the replicates, which does not fully satisfy the basic objective. Nevertheless, the lack of need for second-moment conditions for valid variance estimation is a potential advantage relative to relying on the repeated-imputation inference. Some experience suggests, however, that these potential benefits often cannot be realized because two kinds of inefficiencies arise. First, because imputations are required to be independent within each of K replicates, there is $1/K$th the amount of data used for modeling imputations as actually available. Second, small K implies very poor variance estimation, and often the largest possible K is truly 1, so that actual independent replicates cannot be used when trying to apply the method.

I believe that Hansen agreed that the independent replicate approach was generally inadequate and that the Bayesian multiple imputation approach is necessary to handle missing data in surveys:

Olkin: Have you become involved with Bayesian statistics or other techniques developed within the last ten years?

Hansen: Not really. I guess I endorse and approve the kind of thinking that Don Rubin has been doing.

Olkin: With respect to missing observations?

Hansen: Yes, in missing observations. Sometimes it's necessary to do modeling in sample surveys, where probability sampling methods aren't applicable as in the case of the imputation for nonresponse. We certainly have been involved in such methods. In general, I can't say that we have been working in that area very much. However we are interested in the potential in that setting.

Olkin: Now, Morris to switch topics somewhat ... (Hansen 1987, p. 171)

4.3 Imputations in Hypothetical Independent Replicates—Another Alternative

One way to try to get around these inefficiencies is to try to do first-moment proper (multiple) imputation in K nonindependent samples, i.e., jackknife or bootstrap replicates (e.g., Burns 1990; Efron 1994). This is an interesting and useful idea, but it has limitations in our context. If the data-base constructor is to provide the imputations for the ultimate user, there must be a set of imputations for *each* of the K jackknife or bootstrap samples chosen by the data-base constructor, where K should be substantial for stable variance estimation (e.g., 100 or more). Moreover, if $K = 100$ replicate data sets are considered too many to provide, then the data-base constructor must include with the data base the software to be applied by the ultimate user to create the imputations on each of the ultimate user's jackknife or bootstrapped samples—in this case, superior imputations based on confidential or detailed information must be forgone. Also, as with independent replication, the basic objective is not fully satisfied for point or variance estimation, and more work is required of the ultimate user than with a multiply-imputed data set. Moreover, the variance estimation can be inaccurate inferentially, reflecting excessive procedural variance (see, e.g., Rao and Shao 1992, p. 813, and Burns 1990; incidentally, subsequently Burns found that multiple imputation worked well relative to replicate imputation, Burns 1991, 1993).

If neither the data-base constructor's bootstrap/jackknife imputations nor the data-base constructor's imputation software is delivered to the ultimate user, this approach effectively throws the entire problem into the ultimate user's lap, who may well do some sort of misguided imputation,

which is not even first-moment proper, take bootstrap or jackknife replicates and assume inferential validity despite badly biased estimates of scientific estimands (see, e.g., Rubin 1994, and Efron 1994, for differing views concerning the acceptability of such answers).

4.4 Other Imputation-Based Procedures

Rao and Shao (1992) provide a careful analysis of how to use the jackknife to adjust analyses when missing data have been singly imputed by a particular hot-deck procedure. This addresses an important problem because in current practice many public-use files have been singly imputed by the hot deck. But the ultimate user bears the burden of substantial extra work, because "special computations have to be performed to adjust the imputed values for each pseudo-replicate before applying the standard jackknife variance formula" (Rao and Shao 1992, p. 813), and new mathematical analysis and new software apparently must be developed for each new distinct situation (estimator $\times$ missing data pattern $\times$ survey design $\times$ imputation method). Consequently, this approach, at least at present, fails to satisfy the basic objective of relying only on complete-data analyses and general routines.

Fay's work is something of a moving target, with a variety of older and newer suggestions, which are described with little generality and under special assumptions (e.g., missing completely at random). For example, Fay (1996) seems now to accept multiple imputation as being superior to single imputation (and perhaps to standard weighting adjustments) but advocates creating "improper" multiple imputations and recommends analysis by weighting the data from the completed units in one analysis rather than using the repeated-imputed inference. Recommending creating "improper" multiple imputations is suggesting what we should *not* do, but it is not a prescription for doing anything in particular. Presumably, it refers to first-moment proper multiple imputation (because without this even point estimation can be badly biased) but without concern for the second-moment conditions (e.g., fixing parameters at point estimates rather than drawing them from their posterior distributions, as in Rubin 1987, ex. 4.1, prob. 13 in chap. 1, and prob. 46 in chap. 5). But this is not even defined in multistage complex surveys with clusters where valid imputation models need to be hierarchical, typically with levels of parametric structure: I know what it means to try to be proper in complex surveys by following a Bayesian analysis with variables for the survey structure included in the modelling, but I do not know what the advice to "not do this" means. Also consider the example in Rubin (1983, sec. 2.8, also described in Gelman, Carlin, Stern and Rubin 1995, chap. 15), which stimulated the methods in Clogg et al. (1991) and illustrates the need to be Bayesian and include variability in parameter estimation in order to obtain valid frequency inference.

Finally, consider the suggestion in Fay (1996) that the analysis of a multiply-imputed data set should proceed by replacing each incomplete unit with multiply-imputed versions of that unit's data with split weights. I considered and discarded this idea in Rubin (1977b; also see Rubin 1987, prob. 4-29, and the rejoinder in Meng 1994), because it seemed to have merit as a method of analysis only in simple cases (see, e.g., Little 1979). For valid analysis in general, I believe that such an approach requires extra routines for different complete-data analyses, and so fails to satisfy the basic objective. As a method for storing the multiply-imputed data sets, it can take substantially more memory than the standard form because all the observed data for units with some missing data are stored many times instead of just once. Nevertheless, I would certainly be interested in seeing any work that suggests I rejected this idea prematurely, and that in fact, it can be made to work for any posited response mechanism, complex survey, and complete-data analysis, with only the addition of completely general macros.

4.5 Conclusions Regarding Alternative Imputation Strategies

Given a situation with a single imputation method that is first-moment proper for many statistics, it is almost certainly a random-draw method, and then multiple imputations are easily created, and these are the basis of more accurate inference. Then the only reason not to create them and recommend to the ultimate user that the multiply-imputed data be analyzed using repeated-imputation combining rules is fear that the imputation method, although first-moment proper, is not fully proper for some analyses. If it is not proper but is confidence proper, the only legitimate fear is lost power and overcoverage, as due to superefficiency. But then another method is needed for the ultimate user to recover such superefficiency—I believe special methods for different situations. Are such special efforts needed? All realistic examples I know suggest that in practice the overcoverage is slight and a minor issue relative to omitted variables that can lead all methods astray because of biased estimation and undercoverage. General theory and examples suggest that second-moment properness of Bayesianly-motivated multiple imputation procedures typically follows automatically if the method is first-moment proper (see, e.g., Huber 1976, and results referenced in Rubin 1987, sec. 2.10). Nevertheless, more work on this issue is desirable and could make general theoretical contributions to understanding the robustness of Bayesian inference.

My conclusion when doing imputation is to do multiple imputation under carefully chosen models and use the repeated-imputation inference for analysis. Of course, more theoretical development is still desirable on such issues as: implicit imputation models that reflect both the uncertainty of parameter estimation and the uncertainty of the values to impute given a specific predictive fit (van Buuren, van Rijckevorsel, and Rubin 1993); models for sequential imputation (Kong, Liu, and Wong 1994; Liu and Chen 1995); the use of importance weights (Meng 1994); improved small m combining rules in especially difficult cases (Barnard 1995); and the development of realistic nonignorable models for particular settings.

4.6 Weighting Adjustments

Finally, consider weighting adjustments for nonresponse, which in principle, can be a very effective class of methods for obtaining approximately unbiased estimates. Each unit's weight is the inverse probability of observing that unit's pattern of missing data given (X, Y) information. If the patterns of missing data for the units are created by design, as with matrix sampling, these probabilities and thus the weights are known. When these patterns of missing data are affected by nonresponse, the nonresponse probabilities need to be estimated. Although this estimation can be undertaken by the data-base constructor, typically it is only done assuming the simplest case of nonresponse where the units are either respondents (with all of Y observed) or nonrespondents (with all of Y missing); in this case, the nonrespondents can be discarded, and (approximately) unbiased estimates can be obtained from the respondents and their weights, assuming they accurately reflect both the sampling and nonresponse mechanisms.

Several issues arise with the use of weighting adjustments. First, even in the simplest case of unit nonresponse, where the shared data base of respondents is fully observed, many ultimate users' complete-data analyses do not allow for sampling weights. Second, even with complete-data analyses that can deal with sampling weights, the construction of intervals and p-values that validly account for the fact that nonresponse adjustments in the weights are estimated from data are not immediate from complete-data analyses. Third, with general patterns of nonresponse, special analysis methods need to be developed and special software needs to be written—see Little 1988, sec. 5.1 for the case of monotone missing data, but attempting to do this in a manner that allows the use of standard complete-data software leads to ad hoc approaches such as "complete cases" and "available cases," which we have already rejected as unacceptable general solutions. These three issues imply that in general, weighting adjustments do not satisfy the objectives of allowing ultimate users to apply standard complete-data software to shared data bases to obtain valid inference.

A fourth issue with such weighting adjustments is that they are focused on unbiased estimation and are essentially blind to efficiency concerns. In most well-designed surveys, the planned pattern of missing data is such that efficient estimates are expected to result from standard weighted estimates. But nonrespondents do not necessarily create missing data in such a benign way, and so standard weighted estimates, even when approximately unbiased, can have excessive variability. Consider dealing with censored data by weighting—data beyond or approaching the censoring point have zero or very small probabilities of being observed, and so either cannot be dealt with by weighting or imply a few observations with dominant weights. Weighting by inverse probabilities cannot create estimates outside the convex hull of the observed data, and estimates involving weights near the boundary have extremely large variance.

For these reasons, weighting, although theoretically attractive in an asymptotic sense, has never really been claimed to be a complete practical solution to the problem of missing data in shared data bases; recall Hansen's (1987) comments reported in Section 4.2.

4.7 Concluding Comparative Comments

Multiple imputation is doing well, perhaps even flourishing, as documented by recent sessions at the annual meetings of the American Statistical Association and other professional associations (e.g., the International Statistical Institute, American Medical Informatics Association) and by the variety of recent publications documenting its applicability and extending its theory. It is even becoming so popular that the words "multiple imputation" can appear in the title of an article with no reference to a publication by me or any of my coauthors (e.g., James 1995). This change is occurring for two basic reasons. First multiple imputation is substantially easier for the ultimate user than any other current method that can satisfy the dual objectives of reliance only on complete-data methods and general validity of inference. And second, it is becoming relatively easy for the data collector to create multiply-imputed files using modern computing hardware and accompanying algorithmic developments for Bayesian models. Of course, the development of simply-used appropriate software for creating multiple imputations and analyzing multiply-imputed data is still badly needed, but fortunately progress is taking place in many places (e.g., Schafer 1996; Liu 1995; and van Buuren, van Mulligen, and Brand 1995). I expect that with the availability of this software, multiple imputation will become the standard method for handling missing data in public-use data sets.

As an anonymous referee of this paper wrote: "Multiple imputation is more flexible than replication and reweighting for the analysis of survey data when there are complex patterns of nonresponse. Case closed."

[Received August 1993. Revised June 1995.]

REFERENCES

Barnard, J. (1995), "Cross-Match Procedures for Multiple-Imputation Inference: Bayesian Theory and Frequentist Evaluation," unpublished doctoral thesis, University of Chicago, Dept. of Statistics.

Burns, E. M. (1990), "Multiple and Replicate Item Imputation in a Complex Sample Survey," in *Proceedings of the Bureau of the Census Annual Research Conference*.

——— (1991), "Multiple Imputation in the 1989 Commercial Buildings Energy Consumption Survey: Building Characteristics," CBECS Technical Note 86, U.S. Department of Energy.

——— (1993), "Assessment of Energy Use in Multibuilding Facilities," Report DOE/EIA-0555(93)/1, U.S. Department of Energy.

Chand, N., and Alexander, C. H. (1994), "Imputing Income For An N-Person Consumer Unit," Bureau of the Census paper presented at the American Statistical Association Annual Meeting, Toronto.

Clogg, C., Rubin, D. B., Schenker, N., Schultz, B., and Weidman, L. (1991), "Multiple Imputation of Industry and Occupation Codes in Census Public-Use Samples Using Bayesian Logistic Regression," *Journal of the American Statistical Association*, 86, 68–78.

Efron, B. (1994), "Missing Data, Imputation, and the Bootstrap" (with discussion), *Journal of the American Statistical Association*, 89, 463–478.

Efron, B., and Tibsharani, R. (1993), *An Introduction to the Bootstrap*, London: Chapman and Hall.

Eltinge, J. L., Yansaneh, I. S., and Paulin, G. D. (1994), "Assessment of Reported Differences Between Expenditures and Low Incomes in the

U.S. Consumer Expenditure Survey," paper presented at the American Statistical Association Annual Meeting, Toronto

Ezzati-Rice, T. M., Khare, M., and Schafer, J. L. (1993), "Multiple Imputation of Missing Data in NHANES III," paper presented at the American Statistical Association Annual Meeting, San Francisco.

Ezzati-Rice, T. M., Johnson, W., Khare, M., Little, R. J. A., Rubin, D. B., and Schafer, J. L. (1995), "A Simulation Study to Evaluate The Performance of Multiple Imputations in NCHS Health Examination Survey," in *Proceedings of the Bureau of the Census Eleventh Annual Research Conference*, pp. 257–266.

Fahimi, M., and Judkins, D. (1993), "Serial Imputation of NHANES III With Mixed Regression and Hot-Deck Techniques," paper presented at the American Statistical Association Annual Meeting, San Francisco,

Fay, R. E. (1990), "VPLX: Variance Estimation for Complex Surveys," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 266–271.

——— (1991), "A Design-Based Perspective on Missing Data Variance," in *Proceedings of the 1991 Annual Research Conference, U.S. Bureau of the Census,* pp. 429–440.

——— (1992), "When are Inferences From Multiple Imputation Valid?," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 227–232.

——— (1993), "Valid Inferences From Imputed Survey Data," paper presented at the Annual Meeting of the American Statistical Association, San Francisco.

——— (1996), "Alternative Paradigms for the Analysis of Imputed Survey Data," *Journal of the American Statistical Association*, this issue, 490–498.

Fisher, R. A. (1925), "Theory of Statistical Estimation," *Proceedings of the Cambridge Philosophical Society*, 22, 700–725.

——— (1934), Discussion of "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection," by J. Neyman, *Journal of the Royal Statistical Society*, Ser. A, 97, 614–619.

Freedman, V. (1990), "Using SAS to Perform Multiple Imputation," Discussion Paper Series UI-PSC-6, The Urban Institute, Washington, DC.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

——— (1992), "Bayesian Statistics Without Tears: A Sampling-Resampling Perspective," *The American Statistician*, 46, 84–88.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. B. (1995), *Bayesian Data Analysis*, New York: Chapman and Hall.

Gelman, A., and Rubin, D. B. (1992), "Inference From Iterative Simulation Using Multiple Sequences" (with discussion), *Statistical Science*, 7, 457–472.

Hansen, M. H. (1987), "A Conversation with Morris Hansen" (I. Olkin, interviewer), *Statistical Science*, 2, 162–179.

Heitjan, D. F., and Rubin, D. B. (1990), "Inference From Coarse Data via Multiple Imputation With Application to Age Heaping," *Journal of the American Statistical Association*, 85, 304–314.

Herzog, T., and Lancaster, C. (1980), "Multiple Imputation Modeling for Individual Social Security Benefit Amounts," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 398–403.

Huber, P. J. (1976), "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press, pp. 221–233.

James, I. R. (1995), "A Note on the Analysis of Censored Regression Data by Multiple Imputation," *Biometrics*, 51, 358–362.

Johnson, C. L., Curtin, L. R., Ezzati-Rice, T. M., Khare, M., and Murphy, R. S. (1993), "Single and Multiple Imputation: The NHANES Perspective," paper presented at the Annual Meeting of the American Statistical Association, San Francisco.

Kalton, G. (1983), *Compensating for Missing Survey Data*, Ann Arbor, MI: Institute of Social Research, University of Michigan.

Kennickell, A. B. (1991), "Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation," in *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 1–10.

Kong, A., Liu, J., and Wong, W. H. (1994), "Sequential Imputation and Bayesian Missing Data Problems," *Journal of the American Statistical Association*, 89, 278–288.

Kott, P. S. (1992), "A Note on a Counter-Example to Variance Estimation Using Multiple Imputation," technical report, U.S. National Agriculture Service.

Krewski, D., and Rao, J. N. K. (1981), "Inference From Stratified Samples: Properties of the Linearization, Jackknife, and Balanced Repeated Replication Methods," *The Annals of Statistics*, 9, 1010–1019.

Lehmann, E. L. (1959), *Testing Statistical Hypotheses*, New York: John Wiley.

Li, K. H., Meng, X. L., Raghunathan, T. E., and Rubin, D. B. (1991), "Significance Levels From Repeated p Values With Multiply-Imputed Data," *Statistica Sinica*, 1, 65–92.

Li, K. H., Raghunathan, T. E., and Rubin, D. B. (1991), "Large Sample Significance Levels from Multiply-Imputed Data Using Moment-Based Statistics and an F Reference Distribution," *Journal of the American Statistical Association*, 86, 1065–1073.

Little, R. J. A. (1979), "Maximum Likelihood for Multiple Regression With Missing Values: A Simulation Study," *Journal of the Royal Statistical Society*, Ser. B, 41, 76–87.

——— (1988), "Missing Data in Large Surveys" (with discussion), *Journal of Business and Economic Statistics*, 6, 287–301.

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, New York: John Wiley.

——— (1993), "Assessment of Trial Imputations for NHANES III," project report, Datametrics Research, Inc.

Liu, C. and Rubin, D. B. (1996), "M: Multiple Imputation System," report, Datametrics Research Inc.

Liu, J. S., and Chen, R. (1995), "Blind Deconvolution via Sequential Imputations," *Journal of the American Statistical Association*, 90, 567–576.

Meng, X. L. (1994), "Multiple Imputation With Uncongenial Sources of Input" (with discussion), *Statistical Science*, 9, 538–574.

Meng, X. L., and Rubin, D. B. (1992), "Performing Likelihood Ratio Tests With Multiply Imputed Data Sets," *Biometrika*, 79, 103–111.

Miller, R. G. (1974), "The Jackknife—A Review," *Biometrika*, 61, 1–17.

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

Neyman, J. (1934), "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection," *Journal of the Royal Statistical Society*, Ser. A, 97, 558–606.

Paulin, G. D., and Ferraro, D. L. (1994), "Do Expenditures Explain Income? A Study of Variables for Income Imputation," paper presented at the Annual Meeting of the American Statistical Association, Toronto.

Rao, J. N. K., and Shao, J. (1992), "Jackknife Variance Estimation With Survey Data Under Hot Deck Imputation," *Biometrika*, 79, 811–822.

Rubin, D. B. (1977a), "Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys," *Journal of the American Statistical Association*, 72, 538–543.

——— (1977b), "The Design of a General and Flexible System for Handling Non-Response in Sample Surveys," working document prepared for the U.S. Social Security Administration.

——— (1978), "Multiple Imputations in Sample Surveys—A Phenomenological Bayesian Approach to Nonresponse," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 20–34. Also in *Imputation and Editing of Faulty or Missing Survey Data*. U.S. Department of Commerce, pp. 1–23.

——— (1980), "Handling Nonresponse in Sample Surveys by Multiple Imputations," monograph, U.S. Department of Commerce, Bureau of the Census.

——— (1981), "The Bayesian Bootstrap," *The Annals of Statistics*, 9, 130–134.

——— (1983), "Progress Report on Project For Multiple Imputation of 1980 Codes," manuscript distributed to the U.S. Bureau of the Census, the U.S. National Science Foundation, and the Social Science Research Foundation.

——— (1984), "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician," *The Annals of Statistics*, 12, 1151–1172.

——— (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley.

——— (1988), "Using the SIR Algorithm to Simulate Posterior Distributions" (with discussion), in *Bayesian Statistics 3*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, New York: Oxford University Press, pp. 395–402.

——— (1990), "Imputation Procedures and Inferential Versus Evaluative Statistical Statements," in *Proceedings U.S. Census Bureau Sixth Annual Research Conference*, pp. 676–679.
——— (1993), "Satisfying Confidentiality Constraints Through the Use of Synthetic Multiply-Imputed Micro-Data," *Journal of Official Statistics*, 9, 461–468.
——— (1994), Comments on "Missing Data, Imputation, and the Bootstrap" by B. Efron, *Journal of the American Statistical Association*, 89, 485–8.
Rubin, D. B., and Schenker, N. (1986), "Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse," *Journal of the American Statistical Association*, 81, 366–374.
——— (1987), "Interval Estimation From Multiply Imputed Data: A Case Study Using Agriculture Industry Codes," *Journal of Official Statistics*, 3, 375–387.
Schafer, J. L. (1996), *Analysis of Incomplete Multivariate Data by Simulation*, New York: Chapman and Hall.
Schafer, J. L., and Schenker, N. (1991), "Variance Estimation With Imputed Means," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 696–701.
Schafer, J. L., Khare, M., Little, R. J. A., and Rubin, D. B. (1993), "Multiple Imputation of NHANES III," paper presented at the Annual Meeting of the American Statistical Association, San Francisco.
Schenker, N. (1989), "The Use of Imputed Probabilities for Missing Binary Data," in *Proceedings of the 5th Annual Research Conference, Bureau of the Census*, pp. 133–139.
Schenker, N., Treiman, D. J., and Weidman, L. (1993), "Analyses of Public Use Decennial Census Data With Multiply Imputed Industry and Occupation Codes," *Applied Statistics*, 42, 545–556.
Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data-Augmentation" (with discussion), *Journal of the American Statistical Association*, 82, 528–550.
Treiman, D. J., Bielby, W., and Cheng, M. (1989), "Evaluating a Multiple-Imputation Method for Recalibrating 1970 U.S. Census Detailed Industry Codes to the 1980 Standard," *Sociological Methodology*, 18, 309–345.
van Buuren, S., van Mulligen, E. M., and Brand, J. P. L. (1995), "Omgaan Met Ontbrekende Gegevens in Statistische Databases: Multiple Imputatie in HERMES," *Kwantitatieve Methoden, 50,* 503–504.
van Buuren, S., van Rijckevorsel, J. L. A., and Rubin, D. B. (1993), "Multiple Imputation by Splines," in *Bulletin of the International Statistical Institute, Contributed Papers II*, 503–504.
Weld, L. (1987), "Significance Levels from Public Use Data With Multiply-Imputed Industry Codes," unpublished doctoral thesis, Harvard University, Dept. of Statistics.
Wolter, K. M. (1984), *Introduction to Variance Estimation*, New York: Springer-Verlag.

BIBLIOGRAPHY: SOME OTHER WORK INVOLVING MULTIPLE IMPUTATION

Belin, T. R., Diffendal, G. J., Mack, S., Rubin, D. B., Schafer, J. L., and Zaslavsky, A. M. (1993), "Hierarchical Logistic Regression Models for Imputation of Unresolved Enumeration Status in Undercount Estimation" (with discussion), *Journal of the American Statistical Association*, 88, 1149–1166.
Belin, T. R., and Rubin, D. B. (1990), "Calibration of Errors in Computer Matching for Census Undercount" (with discussion), in *Proceedings of the Government Statistics Section, American Statistical Association*, pp. 124–131.
Bloxum, B., Pashley, P. J., Nicewander, W. A., and Yan, D. (1995), "Linking to a Large-Scale Assessment: An Empirical Evaluation," *Journal of Educational and Behavioral Statistics*, 20, 1–26.
Boshuizen, H. C., Izaks, G. J., van Buuren, S., and Ligthart, G. J. (1995), "Bloeddruk en Sterfte Bij Hoogbejaarden." TNO-rapport C95.014, ISBN 90-6743-377-2.
Brand, J., Gelsema, E. S., and van Buuren, S. (1995), "Verification of Multiple Imputation by Simulation," submitted to *SCAMC '95*.
Brand, J., van Buuren, S., van Mulligen, E. M., Timmers, T., and Gelsema, E. (1994), "Multiple Imputation as a Missing Data Machine." in *Proceedings of the Eighteenth Annual Symposium on Computer Applications in Medical Care (SCAMC)*, Philadelphia: Hanley and Belfus, pp. 303–307.
Brownstone, D. (1991), "Multiple Imputations for Linear Regression Models," Working Paper MBS 91-37, University of California, Irvine, Institute for Mathematical Behavioral Sciences.
Brownstone, D., and Valletta, R. (1996), "Modeling Earnings Measurement Error: A Multiple Imputation Approach," *The Review of Economics and Statistics*, In press.
Chao, M. T. (1994), "A Short Review of Recent Survey Methods for Nonresponse," *Journal of the Chinese Statistical Association*, 32, 169–177.
Chen, R., and Liu, J. S. (1994), "Predictive Updating Methods With Application to Bayesian Classification," *Journal of The Royal Statistical Society*, Ser. B, 58, 2.
Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B., and Weidman, L. (1991), "Multiple Imputation of Industry and Occupation Codes in Census Public-Use Samples Using Bayesian Logistic Regression," *Journal of the American Statistical Association*, 86, 68–78.
Dorey, F. J., Little, R. J. A., and Schenker, N. (1993), "Multiple Imputation for Threshold-Crossing Data With Interval-Censoring," *Statistics in Medicine*, 12, 1589–1603.
Ezzati-Rice, T. M., Khare, M., and Schafer, J. L. (1993), "Multiple Imputation of Missing Data in NHANES III," paper presented at the American Statistical Association Annual Meeting, San Francisco.
Fahimi, M., and Judkins, D. (1993), "Serial Imputation of NHANES III With Mixed Regression and Hot-Deck Techniques," paper presented at the American Statistical Association Annual Meeting, San Francisco.
Freedman, V., and Wolf, D. A. (1991), "Imputation of Mother's Marital Status in National Survey of Families and Households," Discussion Paper Series UI-PSC-8, The Urban Institute, Washington, DC.
Glynn, R., Laird, N., and Rubin, D. B. (1993), "The Performance of Mixture Models for Nonignorable Nonresponse With Follow Ups," *Journal of the American Statistical Association*, 88, 984–993.
Greenland, S., and Finkle, W. D. (1995), "A Critical Look at Basic Methods for Handling Missing Covariates in Epidemiologic Regression Analyses," *American Journal of Epidemiology, 142,* 1255–1264.
Journal of the American Statistical Association, 90, 54–63.
Heitjan, D. F. (1990), "Coping with Age Heaping and Digit Preference: A Multiple Imputation Approach," unpublished paper, Pennsylvania State University College of Medicine, Center for Biostatistics & Epidemiology.
Heitjan, D. F., and Landis, J. R. (1994), "Assessing Secular Trends in Blood Pressure: A Multiple-Imputation Approach," *Journal of the American Statistical Association*, 89, 750–759.
Heitjan, D. F., and Little, R. J. A. (1991), "Multiple Imputation for the Fatal Accident Reporting System," *Applied Statistics*, 40, 13–29.
Johnson, C. L., Curtin, L. R., Ezzati-Rice, T. M., Khare, M., and Murphy, R. S. (1993), "Single and Multiple Imputation: The NHANES Perspective," paper presented at the 1993 American Statistical Association Annual Meeting, San Francisco.
Johnson, E. G., and Zwick, R. (Eds.) (1988), *Focusing the New Design: The NAEP 1988 Technical Report*, Princeton, NJ: Educational Testing Service.
Kalleberg, A. L., Marsden, P. V., Aldrich, H. E., and Cassell, J. W. (1990), "Comparing Organizational Sampling Frames," *Administrative Science Quarterly*, 35, 658–688.
Kennickell, A. B. (1991), "Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 1–10.
Land, K. C., and McCall, P. L. (1993), "Estimating the Effect of Nonresponse in Sample Surveys: An Application of Rubin's Bayesian Method to the Estimation of Community Standards for Obscenity." *Sociological Methods and Research, 21,* 291–316.
Li, K. H. (1988), "Hypothesis Testing in Multiple Imputation," unpublished doctoral thesis, University of Chicago.
Little, R. J. A., and Rubin, D. B. (1989), "The Analysis of Social Science Data With Missing Values," *Sociological Methods and Research*, 18, 292–326. Also in *Modern Methods of Data Analysis* (1990), eds. S. Fox and J. S. Long, Newbury Park, CA: Sage Publications.
——— (1993), "Assessment of trial imputations for NHANES III," project report, Datametrics Research, Inc.
Liu, C. (1993), "Bartlett's Decomposition of the Posterior Distribution of the Covariance for Normal Monotone Ignorable Missing Data," *Journal of Multivariate Analysis*, 46, 198–206.
——— (1994), "Statistical Analysis Using the Multivariate *t* Distribution,"

unpublished doctoral thesis, Harvard University, Dept. of Statistics.
——— (1995), "Missing Data Imputation Using the Multivariate t Distribution," *Journal of Multivariate Analysis*, 53, 139–158.
Liu, J. S. (1994), "Fraction of Missing Information and Convergence Rate of Data Augmentation," in *Computationally Intensive Statistical Methods: Proceedings of the 26th Symposium Interface*, eds. J. Sall and A. Lehman, pp. 490–497.
Marsden, P. V., and Podolny, J. (1990), "Dynamic Analysis of Network Diffusion Processes," in *Social Networks Through Time*, eds. J. Weesie and H. Flap, Utrecht, The Netherlands: ISOR.
Meng, X. L. (1990), "Towards Complete Results for Some Incomplete-Data Problems," unpublished doctoral thesis, Harvard University, Dept. of Statistics.
Meng, X. L., and Rubin, D. B. (1990), "Likelihood Ratio Tests with Multiply-Imputed Data," in *Proceedings of the Statistical Computing Section, American Statistical Association*, pp. 78–82.
Meng, X. L., and Tu, X. M. (1993), "Correcting Reporting Delays in Surveillance Data by Multiple Imputation With Application to AIDS Surveillance," paper presented at the American Statistical Association Annual Meeting, San Francisco.
Mislevy, R. J. (1991), "Randomization-Based Inferences About Latent Variables From Complex Samples," *Psychometrika*, 56, 177–196.
Mislevy, R. J., Beaton, A. E., Kaplan, B., and Sheehan, K. M. (1992), "Estimating Population Characteristics From Sparse Matrix Samples of Item Responses," *Journal of Educational Measurement*, 29, 133–161.
Raghunathan, T. E. (1987), "Large Sample Significance Levels From Multiply-Imputed Data," unpublished doctoral thesis, Harvard University, Dept. of Statistics.
Raghunathan, T. E., and Grizzle, J. E. (1995), "A Split Question Survey Design," *The Journal of the American Statistical Association*, 90, 54–63.
Raghunathan, T. E., and Siscovick, D. S. (1996), "A Multiple Imputation Analysis of a Case-Control Study on the Risk of Primary Cardiac Arrest Among Pharmacologically Treated Hypertensives," *Applied Statistics*, 45, 3.
Reilly, M. (1993), "Data Analysis Using Hot-Deck Multiple Imputation," *The Statistician*, 42, 307–313.
Relles, D. A., and Stolzenberg, R. M. (1991), "An Assessment of the Consequences of Sample Censoring Bias in Graduate School Admission Test Validation," in *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 101–110.
Rubin, D. B. (1988), "Multiple Imputation for Data-Base Construction." *COMSTAT 88—Proceedings in Computational Statistics*, eds. D. Edwards and N. E. Raun, Heidelberg: Physica-Verlag, pp. 389–400.
——— (1988), "An Overview of Multiple Imputation," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 79–84.
——— (1991), "EM and Beyond," *Psychometrika*, 56, 241–254.
——— (1992), Comment on "Clinical Trials in Psychiatry: Should Protocol Deviations Censor Patient Data?" by Lavori, *Neuropsychopharmacology*, 6, 59–60.
Rubin, D. B., and Schafer, J. L. (1990), "Efficiently Creating Multiple Imputations for Incomplete Multivariate Normal Data," in *Proceedings of the Statistical Computing Section, American Statistical Association*, pp. 83–88.
——— (1988), "Imputation Strategies for Missing Values in the PES," *Survey Methodology*, 14, 209–221.
Rubin, D. B., Schafer, J. L., and Schenker, N. (1988), "Imputation Strategies for Estimating the Undercount" in *Bureau of the Census Fourth Annual Research Conference*, pp. 151–159.
Rubin, D. B., and Schenker, N. (1986), "Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse," *Journal of the American Statistical Association*, 81, 366–374.
——— (1987), "Interval Estimation from Multiply-Imputed Data: A Case Study Using Agriculture Industry Codes," *Journal of Official Statistics*, 3, 375–387.
——— (1991), "Multiple Imputation in Health-Care Data Bases: An Overview and Some Applications," *Statistics in Medicine*, 10, 585–598.
Rubin, D. B., Stern, H., and Vehovar, V. (1994), "Handling 'Don't Know' Survey Responses: The Case of the Slovenian Plebiscite," *Journal of the American Statistical Association*, 90, 822–828.
Rubin, D. B., and Zaslavsky, A. (1989), "An Overview of Representing Misenumerations in the Census Using Multiple Imputation," in *Proceedings of the Bureau of the Census Fifth Annual Research Conference*, pp. 109–117.
Schafer, J. L. (1991), "A Comparison of the Missing-Data Treatments in the Post-Enumeration Program," *Journal of Official Statistics*, 7, 475–498.
——— (1991), "Algorithms for Multiple Imputation and Posterior Simulation from Incomplete Multivariate Data With Ignorable Nonresponse," unpublished doctoral thesis, Harvard University, Dept. of Statistics.
Schafer, J. L., Khare, M., Little, R. J. A., and Rubin, D. B. (1993), "Multiple Imputation of NHANES III," paper presented at the American Statistical Association Annual Meeting, San Francisco.
Schenker, N., Treiman, D. J., and Weidman, L. (1988), "Evaluation of Multiply-Imputed Public-Use Tapes," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 85–92.
Schenker, N., Treiman, D. J., and Weidman, L. (1993), "Analysis of Public-Use Data With Multiply-Inputed Industry and Occupation Codes," *Applied Statistics*, 42, 545–556.
Schenker, N., and Welsh, A. H. (1988), "Asymptotic Results for Multiple Imputation," *The Annals of Statistics*, 16, 1550–1566.
Soldo, B. J., Wolf, D. A., and Freedman, V. A. (1990), "Coresidence With Older Mothers: The Children's Perspective," Urban Institute Report, Washington, DC.
Stein, M. L., Shen, X., and Styer, P. E. (1993), "Applications of a Simple Regression Model to Rain," *Canadian Journal of Statistics*, 21, 331–346.
Taylor, J. M. G., Muñoz, A., Bass, S. M., Saah, A., Chmiel, J. S., and Kingsley, L. A. (1990), "Estimating the Distribution of Times From HIV Seroconversion to AIDS Using Multiple Imputation," *Statistics in Medicine*, 9, 505–514.
Treiman, D. J., Bielby, W., and Cheng, M. (1989), "Evaluating a Multiple-Imputation Method for Recalibrating 1970 U.S. Census Detailed Industry Codes to the 1980 Standard," *Sociological Methodology*, 18, 309–345.
Tu, X. M., Meng, X. L., and Pagano, M. (1993a), "The AIDS Epidemic: Estimating Survival After AIDS Diagnosis From Surveillance Data," *Journal of the American Statistical Association*, 88, 26–36.
——— (1993b), "Survival Differences and Trends in Patients With AIDS in the United States," *Journal of Acquired Immune Deficiency Syndromes*, 6, 1150–1156.
van Buuren, S., van Mulligen, E. M., and Brand, J. P. L. (1994), "Routine Multiple Imputation in Statistical Data Bases," in *Proceedings of the Seventh International Working Conference on Scientific and Statistical Database Management*, eds. J. C. French and H. Hinterberger, Los Alamitos, CA: IEEE Computer Society Press, pp. 74–78.
Weld, L. H. (1987), "Significance Levels From Public-Use Data With Multiply-Imputed Industry Codes," unpublished doctoral thesis, Harvard University, Dept. of Statistics.
Williams, V. S. L., Billeaud, K., Davis, L. A., Thissen, D., and Sanford, E. (1995), "Projecting to the NAEP Scale: Results From the North Carolina End of Grade Testing Program," research report, National Institute of Statistical Sciences.
Zaslavsky, A. M. (1989), "Representing Census Undercount: A Comparison of Reweighting and Multiple Imputation Methods," unpublished doctoral thesis, Massachusetts Institute of Technology, Dept. of Mathematics.