# Multiple imputation when records used for imputation are not used or disseminated for analysis

BY JEROME P. REITER,

*Department of Statistical Science, Duke University, Durham, North Carolina 27708-0251, U.S.A.*

jerry@stat.duke.edu

SUMMARY

When some of the records used to estimate the imputation models in multiple imputation are not used or available for analysis, the usual multiple imputation variance estimator has positive bias. We present an alternative multiple imputation approach that enables unbiased estimation of variances and, hence, calibrated inferences in such contexts. First, using all records, the imputer samples $m$ values of the parameters of the imputation model. Secondly, for each parameter draw, the imputer simulates the missing values for all records $n$ times. From these $mn$ completed datasets, the imputer can analyze or disseminate the appropriate subset of records. We develop methods for interval estimation and significance testing for this approach. Methods are presented in the context of multiple imputation for measurement error.

*Some key words:* Combining data; Confidentiality; Measurement error; Missing data.

## 1. INTRODUCTION

The multiple imputation framework proposed by Rubin (1987) is useful for many

statistical problems, most notably to handle missing data and to correct measurement errors. In some scenarios, records used to estimate the imputation models are not used or available for analysis. This can arise, for example, when a statistical agency supplements the original data with administrative records to enable or improve estimation of imputation models. When confidentiality laws prevent the agency from sharing the supplementary data with outsiders, the agency might disseminate only the original data with multiple imputations. Alternatively, the supplementary data may not be representative of the intended target population, perhaps being records from subpopulations, from different time periods or from conveniently available databases. The agency may find it easier to disseminate or base inferences on the original data after using the combined data for imputations, especially for design-based estimation since combining the original and supplementary data complicates interpretation of sampling weights.

When records are used for imputation but not for analysis, Rubin's (1987) variance estimator can have positive bias, because of a mismatch in the conditioning used by the analyst and the imputer: the derivation of Rubin's (1987) variance estimator presumes that the analyst conditions on all records used in the imputation models, not just the available data. Unfortunately, the mismatch makes it difficult to find an alternative variance estimator that is generally unbiased.

This article presents a resolution to this dilemma based on a different approach for generating imputations, which leads to a different variance estimator for analysis from

that outlined by Rubin (1987). The approach is motivated with multiple imputation for measurement error, although it can be used in other contexts.

## 2. MULTIPLE IMPUTATION FOR MEASUREMENT ERROR

### 2.1. *General context*

In many datasets, some variables are measured with error. For example, survey respondents might provide incorrect information about their incomes, or medical patients' self-reported measurements of health variables like blood pressure or cholesterol levels may differ from clinical measurements. Occasionally, data analysts can obtain true values for these variables, or at least more accurate values than those collected initially, for some sampled or nonsampled units for which with-error measurements are also known. The data containing both gold-standard and with-error values are called validation data.

When validation data are available, inferences can be adjusted for measurement error via multiple imputation. We first append the records in the validation data to the original data when they are not in the original data, and treat the unknown gold-standard values in the original data as missing. Then, we complete the missing data with draws from a model relating the gold-standard and with-error values. Analyses are based on the combined data and Rubin's (1987) methods for combining the point and variance estimates from the multiple datasets. This and similar approaches have been used by Rubin & Schenker (1987), Clogg et al. (1991), Brownstone & Valleta

(1996), Raghunathan & Siscovick (1998), Ghosh-Dastidar & Schafer (2003), Yucel & Zaslavsky (2005), Cole et al. (2006), Durrant & Skinner (2006), Harel & Zhou (2006) and Schenker & Raghunathan (2007). Multiple imputation is especially appealing for statistical agencies disseminating data for public use, as they often have more resources, such as access to administrative databases or other surveys, for correcting measurement error than secondary data analysts.

We consider a scenario in which a statistical agency seeks to disseminate ideal datasets to the public without including the validation data used to estimate the imputation models. As a genuine example of these scenarios, Raghunathan (2006) and Schenker & Raghunathan (2007) use the National Health and Nutrition Examination Survey, which had both self-reported health measurements and clinical measurements from physical examinations for a sample of individuals, to multiply impute ideal values of health measurements for different sampled individuals in the larger National Health Interview Survey, which had only self-report measurements. They release only the multiply-imputed data from the National Health Interview Survey to the public.

To fix notation, let $Z$ represent gold-standard values, let $Y$ represent with-error measurements of those values, and let $X$ represent covariates measured without error. The agency owns two databases, an original survey $D_{\text{org}} = (X_{\text{org}}, Y_{\text{org}})$ with $s_{\text{org}}$ records, and validation data $D_{\text{val}} = (X_{\text{val}}, Y_{\text{val}}, Z_{\text{val}})$ with $s_{\text{val}}$ records. The records in $D_{\text{val}}$ are not in $D_{\text{org}}$. The gold-standard values $Z_{\text{org}}$ are not observed for the records in $D_{\text{org}}$. Variables in $D_{\text{val}}$ are on the same scales as the corresponding variables in $D_{\text{org}}$.

In what follows, we assume that there are no missing data in $D_{\text{org}}$ or $D_{\text{val}}$, although the approach can handle missing values.

### 2.2. *Inaccuracy of variance estimator when using one-stage multiple imputation*

We now illustrate via simulation that $T_m$, the usual multiple imputation variance estimator, can be badly biased when the validation data are used for imputation but not for analysis. For all records $j$, let $X_j \sim N(0,1)$ be a covariate, let $Z_j \sim N(X_j, 1)$ be the true value of the survey variable, and let $Y_j \sim N(Z_j, \tau^2)$ be the with-error measurement of $Z_j$, where $\tau \in (0.1, 1, 10)$. The original data, $D_{\text{org}} = (X_{\text{org}}, Y_{\text{org}})$, comprise $s_{\text{org}} = 200$ draws from these distributions. The validation data, $D_{\text{val}} = (X_{\text{val}}, Y_{\text{val}}, Z_{\text{val}})$, are drawn from the same distributions with sample size $s_{\text{val}}$ equal to either 100 or 2000. Using standard Bayesian methods with diffuse priors, we construct $m = 10$ idealized datasets by simulating values of $Z_{\text{org}}$ from

$$f(Z_{\text{org}}|D_{\text{org}}, D_{\text{val}}) = \int f(Z_{\text{org}}, \theta|D_{\text{org}}, D_{\text{val}})d\theta = \int f(Z_{\text{org}}|D_{\text{org}}, \theta)f(\theta|D_{\text{val}})d\theta, \quad (1)$$

where $\theta$ contains the coefficients of the model relating $Z_{\text{val}}$ to $(X_{\text{val}}, Y_{\text{val}})$. For $l = 1, \ldots, 10$, let $Z^{(l)}$ be the $l$th set of imputed values of $Z_{\text{org}}$, and let $D^{(l)} = (X_{\text{org}}, Z^{(l)})$ be the $l$th ideal dataset. We exclude $Y_{\text{org}}$ from $D^{(l)}$ to emphasize that inferences are based on the idealized rather than with-error values. We further suppose that the analyst bases inferences only on $(D^{(1)}, \ldots, D^{(10)})$; that is, $D_{\text{val}}$ is not available to the analyst.

We now briefly review Rubin's (1987) methods for inferences for multiply-imputed datasets; see Reiter & Raghunathan (2007) for more details. For $l = 1, \ldots, m$, let $q^{(l)}$

and $u^{(l)}$ be respectively the estimate of some population quantity $Q$ and the estimate of the variance of $q^{(l)}$ in completed dataset $D^{(l)}$. Analysts use $\bar{q}_m = \sum_{l=1}^{m} q^{(l)}/m$ to estimate $Q$, and use $T_m = (1 + 1/m)b_m + \bar{u}_m$ to estimate $\text{var}(\bar{q}_m)$, where $b_m = \sum_{l=1}^{m}(q^{(l)} - \bar{q}_m)^2/(m-1)$ and $\bar{u}_m = \sum_{l=1}^{m} u^{(l)}/m$. For large samples, inferences for $Q$ are obtained from the $t$-distribution, $(\bar{q}_m - Q) \sim t_{\nu_m}(0, T_m)$, where the degrees of freedom is $\nu_m = (m-1)\left[1 + \bar{u}_m / \left\{(1 + 1/m)b_m\right\}\right]^2$. A better degrees of freedom for small $s_{\text{org}}$ is presented by Barnard & Rubin (1999). Tests of significance for multicomponent null hypotheses are derived by Li et al. (1991a), Li et al. (1991b), Meng & Rubin (1992) and Reiter (2007).

Table 1 summarizes the properties of multiple imputation inferences for the population average of $Z$ obtained from 1000 simulations of each scenario. The averages of $\bar{q}_{10}$ are within simulation error of zero and so are not reported. In all settings except when $\tau = 0.1$, which represents very little measurement error, $T_m$ is positively biased and confidence interval coverages exceed 95%. Similar results were obtained for a subdomain mean and tail probability of the distribution of $Z$, and for the coefficients in the regression of $Z$ on $X$ and of $X$ on $Z$.

Since $T_m$ is positively biased, we also examine the performance of $T_p = \bar{u}_m + b_m/m$, the variance estimator for partially synthetic data (Little, 1993; Abowd & Woodcock, 2004; Reiter, 2003, 2005b). The aim of partially synthetic data is to protect confidentiality in public-use data by replacing values of sensitive variables with multiple imputations. Results using $T_p$ are quite poor for $\tau \in \{1, 10\}$: $T_p$ is at

least 25% too small when $s_{\text{val}} = 100$ and at least 25% too large when $s_{\text{val}} = 2000$. Thus, $T_p$ is not appropriate either.

The bias in $T_m$ can be illustrated analytically with a simple example. Let $D_{\text{org}}$ and $D_{\text{val}}$ be simple random samples without covariates $X$. In both $D_{\text{org}}$ and $D_{\text{val}}$, let $f(Z_j|Y_j) = N(Y_j\beta, \sigma^2)$ for all records $j$. Let $s_{\text{val}}$ be large enough that $\beta$ and $\sigma^2$ are measured with negligible variance. Finally, let $m = \infty$ so that $T_m = \bar{u}_\infty + b_\infty$. When only the idealized versions of $D_{\text{org}}$ are analyzed, the estimate of the population mean of $Z$ is $\bar{q}_\infty = \lim_{m\to\infty} \sum_{l=1}^{m} \bar{Z}^{(l)}/m = \bar{Y}_{\text{org}}\beta$, where $\bar{Y}_{\text{org}}$ is the mean of $Y_{\text{org}}$. Using repeated sampling arguments, we have $\text{var}(\bar{q}_\infty) = \beta^2\text{var}(\bar{Y}_{\text{org}})$. However, the expectation of $T_m$ is something different. For any $u^{(l)} = \sum_j (Z_j^{(l)} - \bar{Z}^{(l)})^2/\{s(s-1)\}$, we have $E(u^{(l)}) \doteq \beta^2\text{var}(\bar{Y}_{\text{org}}) + \sigma^2/s$. We also have $E(b_\infty) = \sigma^2/s$. Hence, in this case, $E(T_m) > \text{var}(\bar{q}_\infty)$. One might think that the bias can be avoided by using $\bar{u}_m - b_m$ to estimate $\text{var}(\bar{q}_\infty)$. While this works here, it fails when $\theta$ is not known with very high precision. In fact, when $s_{\text{val}} < s_{\text{org}}$, as is typical, the variability in $q^{(l)}$ induced by sampling values of $\theta$ can exceed $\bar{u}_m$, so that $\bar{u}_m - b_m < 0$.

## 3. THE TWO-STAGE IMPUTATION APPROACH

### 3.1. *Summary of two-stage approach*

The inadequacy of $T_m$ and other combinations of $\bar{u}_m$ and $b_m$ suggests that alternatives to (1) may be useful for generating idealized datasets. We propose imputation in two stages. First, the imputer samples $m$ values of $\theta$ from $f(\theta|D_{\text{val}})$. Secondly,

for each $\theta^{(l)}$, for $l = 1, \ldots, m$, the imputer simulates $n$ versions of $Z^{(l)}$ by drawing $n$ times from $f(Z_{\text{org}}|D_{\text{org}}, \theta^{(l)})$. Let $Z^{(l,i)}$ be the $i$th copy in the $l$th nest of the simulated $Z_{\text{org}}$, and let $D^{(l,i)} = (X_{\text{org}}, Z^{(l,i)})$. The imputer releases the collection of datasets, $D^* = \{D^{(l,i)} : l = 1, \ldots, m; i = 1, \ldots, n\}$, without $D_{\text{val}}$. Each $D^{(l,i)}$ includes an index of its nest $l$. Related two-stage imputation schemes have been proposed for handling missing data, see Rubin (2003), Schafer & Harel (2002), and an unpublished 2000 Harvard University Ph.D. thesis by Z. Shen, and for imputing confidential and missing data simultaneously (Reiter, 2004).

Analysts can obtain valid inferences from $D^*$ by combining quantities computed from each $D^{(l,i)}$. For $l = 1, \ldots, m$ and $i = 1, \ldots, n$, let $q^{(l,i)}$ and $u^{(l,i)}$ be respectively the estimate of $Q$ and the estimated variance computed with $D^{(l,i)}$. The following quantities are used for inferences:

$$\bar{q}_M = \sum_{l=1}^{m} \sum_{i=1}^{n} q^{(l,i)}/(mn) = \sum_{l=1}^{m} \bar{q}_n^{(l)}/m, \tag{2}$$

$$\bar{w}_M = \sum_{l=1}^{m} \sum_{i=1}^{n} (q^{(l,i)} - \bar{q}_n^{(l)})^2/\{m(n-1)\} = \sum_{l=1}^{m} w_n^{(l)}/m, \tag{3}$$

$$b_M = \sum_{l=1}^{m} (\bar{q}_n^{(l)} - \bar{q}_M)^2/(m-1), \tag{4}$$

$$\bar{u}_M = \sum_{l=1}^{m} \sum_{i=1}^{n} u^{(l,i)}/(mn). \tag{5}$$

The analyst can use $\bar{q}_M$ to estimate $Q$ and $T_M = \bar{u}_M - \bar{w}_M + (1+1/m)b_M - \bar{w}_M/n$ to estimate the variance of $\bar{q}_M$. When $s_{\text{org}}$ is large, inferences are based on a $t$-

distribution, $(\bar{q}_M - Q) \sim t_{\nu_M}(0, T_M)$. The degrees of freedom, $\nu_M$, equal

$$\nu_M = \left( \frac{\{(1+1/m)b_M\}^2}{(m-1)T_M^2} + \frac{\{(1+1/n)\bar{w}_M\}^2}{\{m(n-1)\}\, T_M^2} \right)^{-1}. \tag{6}$$

It is possible that $T_M < 0$, particularly for small $m$ and $n$. Instead, analysts can use the always positive variance estimator, $\tilde{T}_M = \lambda T_M + (1-\lambda)(1+1/m)b_M$, where $\lambda = 1$ when $T_M > 0$ and $\lambda = 0$ otherwise. Motivation for this estimator is provided in the next section. Generally, negative values of $T_M$ can be avoided by making $m$ and $n$ large. When $T_M < 0$, inferences are based on a $t$-distribution with $(m-1)$ degrees of freedom, which comes from using only the first term and $\tilde{T}_M$ in (6).

### 3.2. *Derivation of inferences for the two-stage approach*

To derive the inferential methods, we follow a Bayesian paradigm, as done by Rubin (1987, Ch. 3) for standard multiple imputation. Let $Q^{(\theta)}$ be the estimate of $Q$ if the true $\theta$ was known and used by the analyst to impute $Z_{\text{org}}$, and let $V^{(\theta)}$ be the estimate of $\text{var}(Q|Q^{(\theta)}, D_{\text{org}})$. Let $Q^{(l)}$ be the estimate of $Q$ if $\theta^{(l)}$ was known and used by the analyst to impute $Z_{\text{org}}$, and let $\bar{Q}_M = \sum_{l=1}^m Q^{(l)}/m$. Let $B_\infty = \lim \sum_{l=1}^m (Q^{(l)} - \bar{Q}_M)^2/(m-1)$ as $m \to \infty$, let $W_\infty^{(l)} = \lim \sum_{i=1}^n (q^{(l,i)} - Q^{(l)})^2/(n-1)$ as $n \to \infty$, and let $\bar{W}_\infty = \lim \sum_{l=1}^m W_\infty^{(l)}/m$ as $m \to \infty$. Finally, let $Q^* = \{Q^{(1)}, \ldots, Q^{(m)}\}$ and $W_\infty^* = \{W_\infty^{(1)}, \ldots, W_\infty^{(m)}\}$.

The derivation proceeds by using $Q^{(\theta)}$ to estimate $Q$, $\bar{Q}_M$ to estimate $Q^{(\theta)}$ and $\bar{q}_M$

to estimate $\bar{Q}_M$. To be specific, the analyst seeks $f(Q|D^*)$, which is obtained from

$$\begin{aligned} f(Q|D^*) &= \int f(Q|D^*, Q^{(\theta)}, Q^*, V^{(\theta)}, B_\infty, W^*_\infty) f(Q^{(\theta)}|D^*, Q^*, V^{(\theta)}, B_\infty, W^*_\infty) \quad (7) \\ &\times \; f(Q^*|D^*, V^{(\theta)}, B_\infty, W^*_\infty) f(V^{(\theta)}, B_\infty, W^*_\infty|D^*) dQ^{(\theta)} dQ^* dV^{(\theta)} dB_\infty dW^*_\infty. \end{aligned}$$

We assume that the sample sizes are large enough to permit normal approximations for the distributions involving $Q, Q^{(\theta)}$ and each $Q^{(l)}$. Thus, we require only the first two moments for each of these distributions, which we derive using standard large-sample Bayesian arguments. Diffuse priors are assumed for all parameters.

To begin, all quantities associated with the imputations are irrelevant for inference about $Q$ given $Q^{(\theta)}$ and $V^{(\theta)}$. Hence, $f(Q|D^*, Q^{(\theta)}, Q^*, V^{(\theta)}, B_\infty, W^*_\infty) = f(Q|D_{\text{org}}, Q^{(\theta)}, V^{(\theta)})$. We assume that

$$Q|D_{\text{org}}, Q^{(\theta)}, V^{(\theta)} \sim N(Q^{(\theta)}, V^{(\theta)}). \quad (8)$$

The imputed values in $D^*$ and all elements in $W^*_\infty$ are irrelevant for inference about $Q^{(\theta)}$ given $Q^*$ and $B_\infty$. Hence, $f(Q^{(\theta)}|D^*, Q^*, B_\infty, W^*_\infty) = f(Q^{(\theta)}|D_{\text{org}}, Q^*, B_\infty)$. Since each $Q^{(l)}$ is a draw from the posterior distribution of $Q^{(\theta)}$, which we assume is normally distributed, we have

$$Q^{(\theta)}|D_{\text{org}}, Q^*, B_\infty \;\sim\; N\{\bar{Q}_M, (1 + 1/m)B_\infty\}. \quad (9)$$

Only $q^{(l,i)}$, where $1 \leq i \leq n$, from the $l$th nest and $W^{(l)}_\infty$ are relevant for inferences about each $Q^{(l)}$. We assume that the sampling distribution for $q^{(l,i)}$ is

$$q^{(l,i)}|D_{\text{org}}, Q^{(l)}, W^*_\infty \;\sim\; N(Q^{(l)}, W^{(l)}_\infty), \quad (10)$$

so that

$$Q^{(l)}|D^*, W^*_\infty \sim N(\bar{q}^{(l)}_n, W^{(l)}_\infty/n), \tag{11}$$

$$\bar{Q}_M|D^*, W^*_\infty \sim N\{\bar{q}_M, \bar{W}_\infty/(mn)\}. \tag{12}$$

Having derived the distributions of the mean parameters in (7), we now turn to the variance parameters. To estimate $f(V^{(\theta)}|D^*, B_\infty, W^*_\infty)$, we first define $V^{(l,i)} = \text{var}(Q|D^{(l,i)}, Q^{(\theta)} = Q^{(l)}, B_\infty, W^*_\infty)$. Given only one dataset $D^{(l,i)}$, the analyst would use $u^{(l,i)}$ to estimate $\text{var}(Q|D^{(l,i)}, B_\infty, W^*_\infty)$. Relating these quantities and using an iterated variance computation, we have

$$u^{(l,i)} = E(V^{(l,i)}|D^{(l,i)}, B_\infty, W^*_\infty) + \text{var}(Q^{(l)}|D^{(l,i)}, B_\infty, W^*_\infty).$$

Rewriting this as an expression for $V^{(l,i)}$, we have $E(V^{(l,i)}|D^{(l,i)}, B_\infty, W^*_\infty) = u^{(l,i)} - W^{(l)}_\infty$. We assume that the sampling distribution of $V^{(l,i)}$ has mean $V^{(\theta)}$, so that

$$\begin{aligned} E(V^{(\theta)}|D^*, B_\infty, W^*_\infty) &= E\{E(V^{(\theta)}|D^*, B_\infty, W^*_\infty, Q^*)|D^*, B_\infty, W^*_\infty\} \\ &= E\left\{\sum_{l,i} V^{(l,i)}/(mn)|D^*, B_\infty, W^*_\infty\right\} = \bar{u}_M - \bar{W}_\infty. \end{aligned}$$

Finally, we assume that the variance in the sampling distribution of $V^{(l,i)}$ is of lower order than $V^{(\theta)}$. This implies negligible sampling variance in $\bar{u}_M$, which typically is reasonable in multiple imputation contexts with large sample sizes (Rubin, 1987, Ch. 3). Thus, we write $f(V^{(\theta)}|D^*, B_\infty, W^*_\infty)$ as a distribution concentrated at $\bar{u}_M - \bar{W}_\infty$ with negligible variance.

To obtain the conditional distributions of $B_\infty$ and each $W^{(l)}_\infty$, we use an analysis-

of-variance set-up. From (10), we have

$$\frac{(n-1)w_n^{(l)}}{W_\infty^{(l)}}\bigg| D^* \sim \chi^2_{n-1}. \tag{13}$$

From (9) – (13) and the simplifying assumption that $W_\infty^{(l)} = \bar{W}_\infty$ for all $l$, we have

$$\frac{(m-1)b_M}{B_\infty + \bar{W}_\infty/n}\bigg| D^*, \bar{W}_\infty \sim \chi^2_{m-1}, \tag{14}$$

$$\frac{m(n-1)\bar{w}_M}{\bar{W}_\infty}\bigg| D^* \sim \chi^2_{m(n-1)}. \tag{15}$$

We now have all distributions needed for inference about $Q$. Integrating over the distributions in (8), (9) and (12), we have

$$Q|D^*, V^{(\theta)}, B_\infty, W^*_\infty \sim N\{\bar{q}_M, V^{(\theta)} + (1+1/m)B_\infty + \bar{W}_\infty/(mn)\}. \tag{16}$$

To obtain $f(Q|D^*)$, we should integrate (16) with respect to the distributions of $V^{(\theta)}$, $B_\infty$ and $W^*_\infty$. Although this integration can be carried out numerically, we desire a straightforward approximation that can be easily computed by analysts using $D^*$. For large $m$ and $n$, we can approximate $f(Q|D^*)$ by a normal distribution with mean $E(Q|D^*) = \bar{q}_M$. If we suppress the conditioning on $D^*$, the variance is

$$\begin{aligned}
\mathrm{var}(Q|D^*) &= E\{\mathrm{var}(Q|Q^{(\theta)})\} + \mathrm{var}\{E(Q|Q^{(\theta)})\} \\
&= E\{E(V^{(\theta)}|Q^*)\} + E\{\mathrm{var}(Q^{(\theta)}|Q^*)\} + \mathrm{var}\{E(Q^{(\theta)}|Q^*)\} \\
&= \bar{u}_M - E(\bar{W}_\infty) + E\{(1+1/m)B_\infty\} + \bar{w}_M/(mn). \qquad (17)
\end{aligned}$$

Based on (14) and (15), we approximate the expectations in (17) as $E(\bar{W}_\infty|D^*) \doteq \bar{w}_M$ and $E(B_\infty|D^*) \doteq b_M - \bar{w}_M/n$. Substituting these into (17), we have

$$\mathrm{var}(Q|D^*) \doteq \bar{u}_M - \bar{w}_M + (1+1/m)(b_M - \bar{w}_M/n) + \bar{w}_M/(mn) = T_M.$$

For modest $m$ and $n$, we use a $t$-distribution, $(\bar{q}_M - Q) \sim t_{\nu_M}(0, T_M)$. The degrees of freedom, $\nu_M$, is derived by matching the first two moments of $(\nu_M T_M)/\{\bar{u}_M - \bar{W}_\infty + (1 + 1/m)B_\infty + \bar{W}_\infty/(mn)\}$ to those of a $\chi^2_{\nu_M}$ distribution. The derivation is presented in the Appendix.

When $T_M < 0$ we estimate $\text{var}(Q|D^*)$ with $(1 + 1/m)b_M$. This is motivated as follows. We might observe $T_M < 0$ because $\bar{u}_M - \bar{w}_M < 0$. This is most likely to happen when $\bar{W}_\infty$ is close to $\bar{U}_\infty \doteq \bar{u}_M$, and variability in $\bar{w}_M$ is relatively large. Hence, we set $\bar{U}_\infty - \bar{W}_\infty$ to zero and drop $\bar{u}_M - \bar{w}_M$ from the variance estimator. We also might observe $T_M < 0$ because $b_M - \bar{w}_M/n < 0$. Hence, we drop $\bar{w}_M/n$ from the variance estimator. When $\bar{U}_\infty - \bar{W}_\infty$ is small, $(1 + 1/m)b_M$ should be positively biased for $\text{var}(Q|D^*)$, since $(1 + 1/m)b_M$ in expectation equals $(1 + 1/m)B_\infty + (1 + 1/m)\bar{W}_\infty/n$. It is possible to use other conservative variance estimators, such as $\max(0, \bar{u}_M - \bar{w}_M) + (1 + 1/m)b_M$, when $T_M < 0$. In the simulations described in §3.3, $\tilde{T}_M$ had up to 40% smaller bias than this more conservative estimator for the few settings where negative variances were a nontrivial occurrence.

### 3.3. *Illustrative simulations of the two-stage approach*

To illustrate the performance of the approach, we adapt the simulation of §2.2. All values of one variable are measured with error in $D_{\text{org}}$, and both with-error and gold-standard values are available in $D_{\text{val}}$. The combinations of $m$ and $n$ include $(m, n) = (12, 3), (6, 6)$ and $(3, 12)$. With the three levels of $\tau$ and two levels of $s_{\text{val}}$, there are eighteen simulation scenarios. We obtain inferences about the mean of $Z$,

the mean of $Z$ for observations with $X < 1$, the percentage of observations with $Z > 1$, truly about 24%, and the coefficients in the regressions of $Z$ on $X$ and of $X$ on $Z$. We also obtain inferences for these quantities using the true values $(X_{\text{org}}, Z_{\text{org}})$ for the records in $D_{\text{org}}$. The simulation is repeated 1000 times.

For all ninety estimands, Fig. 1 displays the ratio of the simulated average $T_M$ over the corresponding simulated $\text{var}(\bar{q}_M)$ that it is meant to estimate. The ratios are generally near one, and the simulated averages of $T_M$ are rarely inaccurate by more than 10% in either direction. In contrast, the simulated averages of $T_m$ for the one-stage imputation procedure in §2.2 are inaccurate by 40% or more when $\tau \in \{1, 10\}$.

In scenarios with $\tau < 10$, few values of $T_M$ are negative. When $s_{\text{val}} = 2000$ and $\tau = 10$, however, $T_M < 0$ in about 35% of the simulations for the coefficient in the regression of $Z$ on $X$. In this case, $\bar{u}_M$ and $\bar{w}_M$ are nearly equal in expectation, but $\bar{w}_M$ has higher-order variability than $\bar{u}_M$. This creates a relatively high chance that $\bar{w}_M > \bar{u}_M$, which leads to negative variance estimates. Using $\tilde{T}_M$ results in overestimation of variances in these settings by factors of between two and three, depending on $m$ and $n$.

Fig. 2 displays 95% confidence interval coverage rates for all ninety estimands. For the most part, coverage rates for the imputed data are in line with those from the true data and near the nominal 95% mark. All but one of the incidences in which coverages dip below 90% occur when $(m, n) = (3, 12)$ and $\tau \in \{1, 10\}$. This suggests

that, when entire variables are subject to measurement error, imputers should use large $m$ to obtain coverage rates that are close to nominal. The average coverage rate is 94.0% when $m = 12$ and 93.3% when $m = 6$, suggesting that, for fixed $M$, allocating replicates to the first stage, i.e. making $m$ large, may result in better calibrated inferences than allocating them to the second stage, i.e. making $n$ large.

Analysts pay a price in terms of efficiency for not using or having $D_{\text{val}}$ for analysis. To illustrate this efficiency loss, we compared variances based on $D^*$ to those based the combined data, $(D_{\text{org}}, D_{\text{val}})$. Standard multiple imputation was used to create 36 completed versions of the combined data. When $\tau = 1$ and $s_{\text{val}} = 100$, the simulated variances of the five estimands were 1.15 to 1.33 times higher when using the two-stage imputations with $(m, n) = (12, 3)$ than when using the multiply-imputed versions of the combined data. Variances were 5.7 to 8.8 times higher when $s_{\text{val}} = 2000$.

Although not shown here, the simulation was repeated with $(m, n) = (10, 10)$. The 95% confidence intervals were well calibrated, with simulated coverage rates between 93.0% and 95.5%, with coverage rates within $\pm$ 1% of those based on the true data.

## 4. LARGE SAMPLE SIGNIFICANCE TESTS

### 4.1. *Summary of tests*

In addition to interval estimation for scalar quantities, analysts of $D^*$ may seek to test the null hypothesis $Q = Q_0$ for some $d$-component estimand $Q$, to test if $d$ regression coefficients equal zero, for example. We now derive a large-sample significance

test for multicomponent null hypotheses for the two-stage multiple imputation approach. In each $D^{(l,i)}$, let $q^{(l,i)}$ be the estimate of $Q$ and let $u^{(l,i)}$ be the estimate of the $d \times d$ covariance matrix associated with $q^{(l,i)}$. We use multivariate analogues of (2) – (5) and $T_M$ for inferences; for example, we use $b_M = \sum_{l=1}^{m}(\bar{q}_n^{(l)} - \bar{q}_M)(\bar{q}_n^{(l)} - \bar{q}_M)'/(m-1)$.

Given the normal distributions used in inferences about $Q$, it may appear reasonable to use a Wald test with statistic $\Delta = (\bar{q}_M - Q_0)'T_M^{-1}(\bar{q}_M - Q_0)$. However, this test is unreliable when $d$ is large and $m$ and $n$ are moderate, as may be the case when statistical agencies release ideal datasets to the public, because $b_M$ or $\bar{w}_M$ can have large variability. Estimating $b_M$ or $\bar{w}_M$ in such cases is akin to estimating a covariance matrix using few observations compared to the number of dimensions. This same problem arises in multiple imputation for missing data (Rubin, 1987). The instability in $T_M$ can be avoided by making $m$ and $n$ large.

For modest $m$ and $n$, analysts can use the test statistic

$$S \;=\; (\bar{q}_M - Q_0)'\bar{u}_M^{-1}(\bar{q}_M - Q_0)/\{d(1 + r_M^{(b)} - r_M^{(w)})\},$$

where $r_M^{(b)} = (1+1/m)\mathrm{tr}(b_M\bar{u}_M^{-1})/d$ and $r_M^{(w)} = (1+1/n)\mathrm{tr}(\bar{w}_M\bar{u}_M^{-1})/d$. This statistic is derived assuming that $B_\infty = r_\infty^{(b)}\bar{u}_M$ and that $W_\infty^{(l)} = r_\infty^{(w)}\bar{u}_M$ for all $l$, thereby reducing the number of unknown parameters in $B_\infty$ and each $W_\infty^{(l)}$. These proportionality assumptions may not be met in practice, especially when only some variables are subject to imputation for measurement error. However, similar assumptions have been used in other multiple imputation contexts with good results; see Li et al. (1991a), Li et al. (1991b), Meng & Rubin (1992), Reiter (2005a, 2007) and Z. Shen's

thesis. Simulation studies by those authors show that tests based on proportionality assumptions, even when they are not met, tend to have better properties than tests based on the equivalents of $\Delta$.

The reference distribution for $S$ is an $F_{d,v_s}$ distribution, where

$$v_s = 4 + \left(1 + \frac{r_M^{(b)} c_b}{c_b - 2} - \frac{r_M^{(w)} c_w}{c_w - 2}\right)^2 \Big/ \left\{ \frac{(r_M^{(b)} c_b)^2}{(c_b-2)^2(c_b-4)} + \frac{(r_M^{(w)} c_w)^2}{(c_w-2)^2(c_w-4)} \right\}, \quad (18)$$

for $c_b > 4$ and $c_w > 4$, and $c_b = d(m-1)$ and $c_w = dm(n-1)$. The $p$-value for testing $Q = Q_0$ equals $\mathrm{pr}(F_{d,v_s} > S)$. When $c_b \leq 4$ or $c_w \leq 4$, $v_s$ is not defined. This occurs only for small $d$ when $m = 2$, a combination that is not likely to arise in practice because of the benefits of making $m$ large for inferences for scalar $Q$.

### 4.2. *Derivation of test*

Let $T_\infty = \bar{U}_\infty - \bar{W}_\infty + (1 + 1/m)B_\infty + \bar{W}_\infty/(mn)$. Extending the derivations in §3.2 to multivariate $Q$, we have $(Q|D^*, T_\infty) \sim N(\bar{q}_M, T_\infty)$ and

$$(m-1)b_M(B_\infty + \bar{W}_\infty/n)^{-1}|D^*, \bar{W}_\infty \sim W(m-1, I), \quad (19)$$

$$m(n-1)\bar{w}_M \bar{W}_\infty^{-1}|D^* \sim W\{m(n-1), I\}. \quad (20)$$

Conditional on $T_\infty$, the $p$-value for testing $Q = Q_0$ equals $\mathrm{pr}\{\chi^2_d > (\bar{q}_M - Q_0)' T_\infty^{-1}(\bar{q}_M - Q_0)|D^*, T_\infty\}$. Since $T_\infty$ is not known, we should average over its distribution to obtain the $p$-value conditional only on $D^*$. If we assume that $B_\infty = r_\infty^{(b)}\bar{U}_\infty$, that $W_\infty^{(l)} = r_\infty^{(w)}\bar{U}_\infty$ for all $l$, and that $\bar{U}_\infty \doteq \bar{u}_M$, this requires averaging over the distributions of $r_\infty^{(b)}$ and $r_\infty^{(w)}$. If we write $T_\infty = \bar{u}_M[1 + (1 + 1/m)r_\infty^{(b)} - \{1 - 1/(mn)\}r_\infty^{(w)}]$, the $p$-value

equals

$$\int \mathrm{pr}\left[\chi^2_d > \frac{(\bar{q}_M - Q_0)'\bar{u}_M^{-1}(\bar{q}_M - Q_0)}{1 + (1+1/m)r_\infty^{(b)} - \{1 - 1/(mn)\}r_\infty^{(w)}}\middle| D^*, r_\infty^{(b)}, r_\infty^{(w)}\right]$$
$$\times \mathrm{pr}(r_\infty^{(b)}|D^*, r_\infty^{(w)})\mathrm{pr}(r_\infty^{(w)}|D^*)dr_\infty^{(b)}dr_\infty^{(w)}$$
$$= \int \mathrm{pr}\left[(\chi^2_d/d)\,\frac{1 + (1+1/m)r_\infty^{(b)} - \{1 - 1/(mn)\}r_\infty^{(w)}}{(1 + r_m^{(b)} - r_M^{(w)})} > S\middle| D^*, r_\infty^{(b)}, r_\infty^{(w)}\right] \tag{21}$$
$$\times \mathrm{pr}(r_\infty^{(b)}|D^*, r_\infty^{(w)})\mathrm{pr}(r_\infty^{(w)}|D^*)dr_\infty^{(b)}dr_\infty^{(w)}.$$

The conditional distributions of $r_\infty^{(b)}$ and $r_\infty^{(w)}$ can be obtained from (19) and (20) if we assume that $B_\infty = r_\infty^{(b)}\bar{U}_\infty$ and that $W_\infty^{(l)} = r_\infty^{(w)}\bar{U}_\infty$ for all $l$. Applying multivariate normal theory, we have

$$\frac{d(m-1)\mathrm{tr}(b_M\bar{u}_M^{-1})/d}{r_\infty^{(b)} + r_\infty^{(w)}/n}\middle| D^*, r_\infty^{(w)} \sim \chi^2_{d(m-1)}, \tag{22}$$
$$\frac{dm(n-1)\mathrm{tr}(\bar{w}_M\bar{u}_M^{-1})/d}{r_\infty^{(w)}}\middle| D^* \sim \chi^2_{dm(n-1)}. \tag{23}$$

Substituting the implied distributions for $r_\infty^{(b)} + r_\infty^{(w)}/n$ in (22) and for $r_\infty^{(w)}$ in (23) into (21), after some algebra we have

$$\mathrm{pr}\left\{(\chi^2_d/d)\frac{1 + c_b r_m^{(b)}/\chi^2_{c_b} - c_w r_M^{(w)}/\chi^2_{c_w}}{1 + r_m^{(b)} - r_M^{(w)}} > S\right\}. \tag{24}$$

Following the approach of Li et al. (1991b), we approximate the random variable in (24) as proportional to a $F$-distributed random variable, $\delta F_{d,v_s}$. The approximation is obtained by matching the first two moments of $\delta F_{d,v_s}$ to those of the left-hand side of the inequality in (24). This yields the expression in (18) for $v_s$ and $\delta = \{(v_s - 2)/v_s\}\{1 + c_b r_m^{(b)}/(c_b - 2) - c_w r_M^{(w)}/(c_w - 2)\}/(1 + r_m^{(b)} - r_M^{(w)})$. The derivation is presented in the Appendix. When $c_b$ and $c_w$ are sufficiently large, $\delta \simeq 1$, and the approximate $p$-value equals $\mathrm{pr}(F_{d,v_s} > S)$.

### 4.3. *Illustrative simulations of significance tests*

This section illustrates the test based on $S$, showing its advantages over the test based on $\text{pr}(\chi^2_d > \Delta)$. The original data, $D_{\text{org}}$, comprise one variable, which for any record $j$ we label as $Z_{j0}$, measured with error and $d$ variables measured without error, where $d = 5$ or $d = 20$. For all records $j$, $Z_{j0} \sim N(2, 1)$ and $Z_{jk} \sim N(1, 1)$, for $k = 1, \ldots, d$. The with-error measurements are $Y_{j0} \sim N(Z_{j0}, 1)$. To ensure large sample size for the Wald approximation, $D_{\text{org}}$ comprises $s_{\text{org}} = 1000$ observations. The validation sample, $D_{\text{val}}$, comprises independent observations, $(Y_{j0}, Z_{j0}, \ldots, Z_{jd})$ where $j = 1, \ldots, s_{\text{val}} = 200$. The two-stage multiple imputation proceeds with $(m, n) = (3, 12)$ and $(12, 3)$. There are two true null hypotheses of interest: that all coefficients equal zero in the regression of $Z_0$ on $(Z_1, \ldots, Z_d)$, and that all coefficients equal zero in the regression of $Z_1$ on $(Z_0, Z_2, \ldots, Z_d)$. The simulation is repeated 1000 times.

Table 2 displays the simulated significance levels of the tests for nominal levels of $\alpha = 0.10, \alpha = 0.05$, and $\alpha = 0.01$. The $p$-values for the test based on $S$ are reasonably well calibrated. They tend to be better calibrated when $m = 12$ than when $m = 3$, providing further evidence that increasing $m$ can improve inferences. On the whole, the simulated levels based on $S$ are closer to the corresponding nominal levels than those based on $\Delta$. Additionally, in some settings $\Delta < 0$ for large percentages of simulations; for example, $\Delta < 0$ in 78% of the simulations for the scenario with $d = 5$, $m = 3$, $n = 12$, and $Z_0$ as a dependent variable. In contrast, $S$ is always positive in these simulations. Using $(1 + 1/m)b_M$ in place of $T_M$ when $\Delta < 0$ results

in even higher simulated significance levels than those shown here.

## 5. CONCLUDING REMARKS

In addition to measurement-error contexts, this two-stage imputation approach can be useful in missing-data contexts. For example, a statistical agency may seek to improve precision by supplementing the original data with complete records from other sources, especially when the original data have high fractions of missing information. The agency can apply the procedures of §3 with a slight modification: the parameters of the imputation model are estimated with the combined data rather than just the supplementary data. A related application is to use two-stage multiple imputation on an entire dataset to handle missing values, but release only a subsample of records to the public to protect confidentiality.

The two-stage imputation approach can be used when agencies seek to combine information from two data sources, often called data fusion (Rassler, 2003). Suppose that a large dataset contains only the variables $X$ and a small dataset contains the same variables plus some others, $Z$. It may be desired to disseminate multiply imputed, ideal datasets $(X, Z^*)$ for analysis. The methods in §3 apply directly: simply exclude the with-error variables in the conditioning arguments.

The two-stage approach can also be used to adjust for changes in the codings or definitions of variables over time. As an example, consider a modification of race bridging (Schenker, 2003; Schenker & Parker, 2003). In the 2000 U.S. census, individuals were allowed to select more than one race from five categories. However,

some state-level data sources do not allow for multiple-race reporting and use only four categories. These state agencies plan to phase in the federal standard over time. Fortunately, there exists a validation sample for bridging the inconsistencies in race reporting in the intervening years: the National Health Interview Survey allows multiple-race reporting but also asks respondents to select one primary race. Thus, using the relationships between the multiple-race and single-race variables in the Survey, we can impute multiple-race reporting for the state-level data sources using the two-stage imputation procedure.


## ACKNOWLEDGEMENT

This research was supported by a grant from the U.S. National Science Foundation.


## APPENDIX

### *Derivations of degrees of freedom*

*Scalar estimands.* Inferences from $D^*$ are made using a $t$-distribution. A key step is to approximate the distribution of

$$\frac{\nu_M T_M}{\bar{u}_M - \bar{W}_\infty + (1+1/m)B_\infty + \bar{W}_\infty/(mn)} | D^* \tag{A1}$$

as a $\chi^2_{\nu_M}$ distribution. We determine $\nu_M$ by matching the mean and variance of the chi-squared distribution to those of (A1).

Let $\alpha = (B_\infty + \bar{W}_\infty/n)/b_M$ and $\gamma = \bar{W}_\infty/\bar{w}_M$. Then, $\alpha^{-1} \mid D^*, \bar{W}_\infty$ and $\gamma^{-1} \mid D^*$ have mean-square distributions, i.e. the distribution of $x/k$ when $x$ is a random

variable with a $\chi^2_k$ distribution (Rubin, 1987, p. 91), with degrees of freedom $m-1$ and $m(n-1)$, respectively. Let $f = (1+1/m)b_M/\bar{u}_M$, and $g = (1+1/n)\bar{w}_M/\bar{u}_M$. We write (A1) as

$$\frac{T_M}{\bar{u}_M + (1+1/m)(B_\infty + \bar{W}_\infty/n) - (1+1/n)\bar{W}_\infty} = \frac{\bar{u}_M(1+f-g)}{\bar{u}_M(1+\alpha f - \gamma g)}. \tag{A2}$$

For the expectation of (A2), we use an iterated expectation and first-order Taylor series expansions in $\alpha^{-1}$ and $\gamma^{-1}$ around their expectations, which equal one, to obtain

$$E\left\{E\left(\frac{1+f-g}{1+\alpha f-\gamma g} \mid D^*, \bar{W}_\infty\right) \mid D^*\right\} \simeq E\left(\frac{1+f-g}{1+f-\gamma g} \mid D^*\right) \simeq 1.$$

For the variance of (A2), we use the conditional variance representation

$$E\left\{\mathrm{var}\left(\frac{1+f-g}{1+\alpha f-\gamma g} \mid D^*, \bar{W}_\infty\right) \mid D^*\right\} + \mathrm{var}\left\{E\left(\frac{1+f-g}{1+\alpha f-\gamma g} \mid D^*, \bar{W}_\infty\right) \mid D^*\right\}.$$

For the interior variance and expectation, we use a first-order Taylor series expansion in $\alpha^{-1}$ around one. Since $\mathrm{var}(\alpha^{-1} \mid D^*, \bar{W}_\infty) = 2/(m-1)$, this expression equals approximately

$$E\left\{\frac{2(1+f-g)^2 f^2}{(m-1)(1+f-\gamma g)^4} \mid D^*\right\} + \mathrm{var}\left(\frac{1+f-g}{1+f-\gamma g} \mid D^*\right). \tag{A3}$$

We now use first-order Taylor series expansions in $\gamma^{-1}$ around one to determine the components of (A3). The first term in (A3) is

$$E\left\{\frac{2(1+f-g)^2 f^2}{(m-1)(1+f-\gamma g)^4} \mid D^*\right\} \simeq \frac{2f^2}{(m-1)(1+f-g)^2}. \tag{A4}$$

Since $\mathrm{var}(\gamma^{-1} \mid D^*) = 2/\{m(n-1)\}$, the second term in (A3) is

$$\mathrm{var}\left(\frac{1+f-g}{1+f-\gamma g} \mid D^*\right) \simeq \frac{2g^2}{m(n-1)(1+f-g)^2}. \tag{A5}$$

If we combine (A4) and (A5), the variance of (A2) equals approximately

$$\frac{2f^2}{(m-1)(1+f-g)^2} + \frac{2g^2}{m(n-1)(1+f-g)^2}. \tag{A6}$$

Since a mean-square random variable has variance equal to 2 divided by its degrees of freedom, we conclude that $\nu_M$ equals the expression in (6).

*Significance tests.* Significance tests of the null hypothesis $Q = Q_0$ with $D^*$ are made using an $F$-distribution. Here, we approximate

$$\left(\frac{\chi_d^2}{d}\right)\left[\frac{1+(1+1/m)r_\infty^{(b)} - \{1-1/(mn)\}r_\infty^{(w)}}{1+r_M^{(b)} - r_M^{(w)}}\right] \tag{A7}$$

as proportional to an $F_{d,v_s}$ distribution. The $v_s$ is determined by matching the mean and variance of the $F$-distribution to the mean and variance of (A7).

It is useful to rewrite the second part of the random variable as

$$\frac{1+(1+1/m)(r_\infty^{(b)} + r_\infty^{(w)}/n) - (1+1/n)r_\infty^{(w)}}{1+r_M^{(b)} - r_M^{(w)}} = \frac{1+c_b r_M^{(b)}/\chi_{c_b}^2 - c_w r_M^{(w)}/\chi_{c_w}^2}{1+r_M^{(b)} - r_M^{(w)}}, \tag{A8}$$

where the equality results from (22) and (23).

For the first two moments of (A8), we use an iterated expectation, conditioning on $(D^*, r_\infty^{(w)})$ for the innermost expectation. Since $E(1/\chi_t^2) = 1/(t-2)$, we have

$$E\left(\frac{1+c_b r_M^{(b)}/\chi_{c_b}^2 - c_w r_M^{(w)}/\chi_{c_w}^2}{1+r_M^{(b)} - r_M^{(w)}} | D^*\right) = \frac{1+c_b r_M^{(b)}/(c_b-2) - c_w r_M^{(w)}/(c_w-2)}{1+r_M^{(b)} - r_M^{(w)}}. \tag{A9}$$

Since $E(\chi_d^2/d) = 1$, the first moment of (A7) equals the expression in (A9).

For the second moment of (A8), we add its variance to the square of (A9). Since

$\mathrm{var}(1/\chi^2_t) = 2/\{(t-2)^2(t-4)\}$, we have

$$E\left\{\left(\frac{1+c_b r_M^{(b)}/\chi^2_{c_b} - c_w r_M^{(w)}/\chi^2_{c_w}}{1+r_M^{(b)}-r_M^{(w)}}\right)^2\right\} = \frac{2(c_b r_M^{(b)})^2/\{(c_b-2)^2(c_b-4)\}}{(1+r_M^{(b)}-r_M^{(w)})^2}$$

$$+\frac{2(c_w r_M^{(w)})^2/\{(c_w-2)^2(c_w-4)\}}{(1+r_M^{(b)}-r_M^{(w)})^2} + \left(\frac{1+c_b r_M^{(b)}/(c_b-2) - c_w r_M^{(w)}/(c_w-2)}{1+r_M^{(b)}-r_M^{(w)}}\right)^2. \quad \text{(A10)}$$

Since $E\{(\chi^2_d/d)^2\} = d(d+2)/d^2$, the second moment of (A7) equals the expression in (A10) multiplied by $d(d+2)/d^2$.

We set the two moments of (A7) equal to the first two moments of $\delta F_{d,v_s}$, which are $E(\delta F_{d,v_s}) = \delta v_s/(v_s-2)$, and $E(\delta^2 F^2_{d,v_s}) = \delta^2(v_s/d)^2 d(d+2)/\{(v_s-2)(v_s-4)\}$. Solving yields $\delta = \{(v_s-2)/v_s\}\{1+c_b r_M^{(b)}/(c_b-2) - c_w r_M^{(w)}/(c_w-2)\}/(1+r_M^{(b)}-r_M^{(w)})$ and the expression in (18) for $v_s$.

# References


ABOWD, J. M. & WOODCOCK, S. D. (2004). Multiply-imputing confidential characteristics and file links in longitudinal linked data. In *Privacy in Statistical Databases*, Ed. J. Domingo-Ferrer & V. Torra, pp. 290–7. New York: Springer-Verlag.

BARNARD, J. & RUBIN, D. B. (1999). Small-sample degrees of freedom with multiple-imputation. *Biometrika* **86**, 948–55.

BROWNSTONE, D. & VALLETTA, R. G. (1996). Modeling earnings measurement error: A multiple imputation approach. *Rev. Econ. Statist.* **78**, 705–17.

CLOGG, C. C., RUBIN, D. B., SCHENKER, N., SCHULTZ, B. & WEIDMAN, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *J. Am. Statist. Assoc.* **86**, 68–78.

COLE, S. R., CHU, H. & GREENLAND, S. (2006). Multiple-imputation for measurement-error correction. *Int. J. Epidemiol.* **35**, 1074–81.

DURRANT, G. B. & SKINNER, C. (2006). Using missing data methods to correct for measurement error in a distribution function. *Survey Methodol.* **32**, 25–36.

GHOSH-DASTIDAR, B. & SCHAFER, J. L. (2003). Multiple edit/multiple imputation for multivariate continuous data. *J. Am. Statist. Assoc.* **98**, 807–17.

HAREL, O. & ZHOU, X. H. (2006). Multiple imputation for correcting verification bias. *Statist. Med.* **25**, 3769–86.

LI, K. H., MENG, X. L., RAGHUNATHAN, T. E. & RUBIN, D. B. (1991a). Significance levels from repeated $p$-values with multiply-imputed data. *Statist. Sinica* **1**, 65–92.

LI, K. H., RAGHUNATHAN, T. E. & RUBIN, D. B. (1991b). Large-sample significance levels from multiply-imputed data using moment-based statistics and an F reference distribution. *J. Am. Statist. Assoc.* **86**, 1065–73.

LITTLE, R. J. A. (1993). Statistical analysis of masked data. *J. Offic. Statist.* **9**, 407–26.

MENG, X. L. & RUBIN, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* **79**, 103–11.

RAGHUNATHAN, T. E. (2006). Combining information from multiple surveys for assessing health disparities. *Allgemeines Statist. Archiv.* **90**, 515–26.

RAGHUNATHAN, T. E., LEPKOWSKI, J. M., VAN HOEWYK, J. & SOLENBERGER, P. (2001). A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodol.* **27**, 85–96.

RAGHUNATHAN, T. E. & SISCOVICK, D. S. (1998). Combining exposure information from multiple sources in the analysis of a case-control study. *Statistician* **47**, 333–47.

RASSLER, S. (2003). A non-iterative Bayesian approach to statistical matching. *Statist. Neer.* **57**, 58–74.

REITER, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodol.* **29**, 181–9.

REITER, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodol.* **30**, 235–42.

REITER, J. P. (2005a). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *J. Statist. Plan. Infer.* **131**, 365–77.

REITER, J. P. (2005b). Using CART to generate partially synthetic, public use microdata. *J. Offic. Statist.* **21**, 441–62.

REITER, J. P. (2007). Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. *Biometrika* **94**, 502–8.

REITER, J. P. & RAGHUNATHAN, T. E. (2007). The multiple adaptations of multiple imputation. *J. Am. Statist. Assoc.* **102**, 1462–71.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons.

RUBIN, D. B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statist. Neer.* **57**, 3–18.

RUBIN, D. B. & SCHENKER, N. (1987). Interval estimation from multiply-imputed data: A case study using census agriculture industry codes. *J. Offic. Statist.* **3**, 375–87.

SCHAFER, J. & HAREL, O. (2002). Multiple imputation in two stages. In *ASA Proceedings of the Joint Statistical Meetings*, pp. 1359–63, Alexandria: VA, American Statistical Association.

SCHENKER, N. (2003). Assessing variability due to race bridging: Application to census counts and vital rates for the year 2000. *J. Am. Statist. Assoc.* **98**, 818–28.

SCHENKER, N. & PARKER, J. D. (2003). From single-race reporting to multiple-race reporting: using imputation methods to bridge the transition. *Statist. Med.* **22**, 1571–87.

SCHENKER, N. & RAGHUNATHAN, T. E. (2007). Combining information from multiple surveys to enhance estimation of measures of health. *Statist. Med.* **26**, 1802–11.

YUCEL, R. M. & ZASLAVSKY, A. M. (2005). Imputation of binary treatment variables with measurement error in administrative data. *J. Am. Statist. Assoc.* **100**, 1123–32.

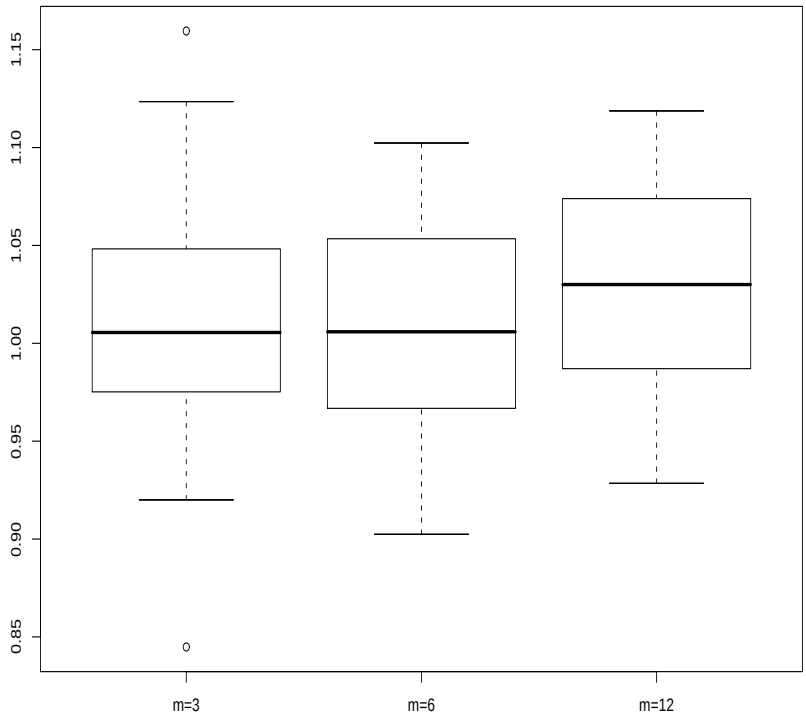


Figure 1: Simulation study. Box plots of ratios of simulated average $T_M$ to simulated $\mathrm{var}(\bar{q}_M)$, for $m = 3$, $m = 6$ and $m = 12$.

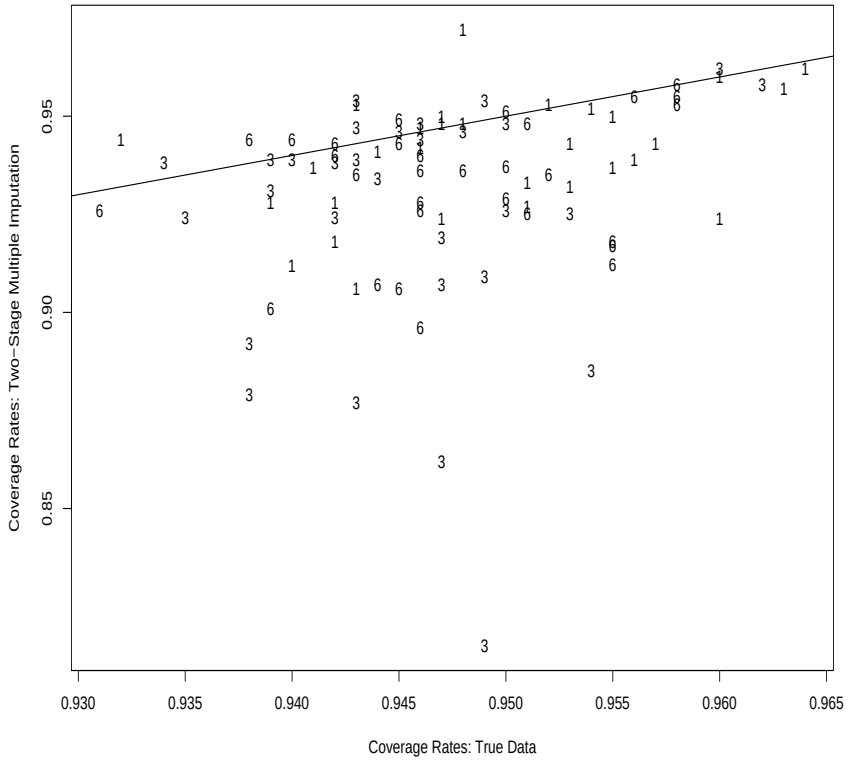


Figure 2: Simulated 95% confidence interval coverage rates for two-stage multiple imputation for measurement error. Rates for $m = 3$ are displayed with '3', for $m = 6$ with '6', and for $m = 12$ with '1'. The horizontal and vertical axes show the rates when using the true values and the imputed data, respectively.

Table 1: *Illustration of the bias in $T_m$ and conservative 95% confidence interval coverage rates when using standard multiple imputation for measurement error.*

| Setting | $\text{var}(\bar{q}_{10})$ | Avg. $T_{10}$ | 95% CI Cov. |
|---|---|---|---|
| $s_{\text{val}} = 100, \tau = 0.1$ | 0.010 | 0.010 | 94.0% |
| $s_{\text{val}} = 100, \tau = 1$ | 0.013 | 0.019 | 98.3% |
| $s_{\text{val}} = 100, \tau = 10$ | 0.015 | 0.027 | 99.1% |
| $s_{\text{val}} = 2000, \tau = 0.1$ | 0.011 | 0.010 | 94.5% |
| $s_{\text{val}} = 2000, \tau = 1$ | 0.008 | 0.013 | 99.0% |
| $s_{\text{val}} = 2000, \tau = 10$ | 0.006 | 0.016 | 99.7% |

Note: Avg., average; 95% CI Cov., 95% confidence interval coverage rate.

Table 2: *Simulated significance levels for two-stage multiple imputation approach.*

| | | Based on $S$ | | | Based on $\Delta$ | | |
|---|---|---|---|---|---|---|---|
| | True significance level | 10% | 5% | 1% | 10% | 5% | 1% |
| | | $Z_0$ is dependent variable | | | | | |
| $(m, n) = (12, 3)$ | $d = 5$ | 8.3 | 4.6 | 0.9 | 27.3 | 22.9 | 17.5 |
| | $d = 20$ | 11.2 | 6.9 | 1.8 | 29.2 | 27.5 | 25.2 |
| $(m, n) = (3, 12)$ | $d = 5$ | 7.7 | 4.6 | 0.6 | 4.7 | 4.5 | 3.7 |
| | $d = 20$ | 10.8 | 5.8 | 2.1 | 17.6 | 16.8 | 15.7 |
| | | $Z_0$ is independent variable | | | | | |
| $(m, n) = (12, 3)$ | $d = 5$ | 10.2 | 5.9 | 1.3 | 18.5 | 12.5 | 5.9 |
| | $d = 20$ | 11.2 | 5.7 | 2.6 | 17.6 | 16.8 | 15.7 |
| $(m, n) = (3, 12)$ | $d = 5$ | 14.1 | 9.7 | 3.6 | 7.6 | 4.4 | 2.4 |
| | $d = 20$ | 13.6 | 8.7 | 2.5 | 11.0 | 6.7 | 2.8 |