The Multiple Adaptations of Multiple Imputation

Jerome P. Reiter and Trivellore E. Raghunathan*

Abstract

Multiple imputation was first conceived as a tool that statistical agencies could use to handle nonresponse in large sample, public use surveys. In the last two decades, the multiple imputation framework has been adapted for other statistical contexts. As examples, individual researchers use multiple imputation to handle missing data in small samples; statistical agencies disseminate multiply-imputed datasets for purposes of protecting data confidentiality; and, survey methodologists and epidemiologists use multiple imputation to correct for measurement errors. In some of these settings, Rubin's original rules for combining the point and variance estimates from the multiply-imputed datasets are not appropriate, because what is known—and therefore in the conditional expectations and variances used to derive inferential methods—differs from the missing data context. These applications require new combining rules and methods of inference. In fact, more than ten combining rules exist in the

*Jerome P. Reiter is Assistant Professor, Department of Statistical Science, Duke University, Durham, NC 27708-0251 (E-mail: *jerry@stat.duke.edu*). Trivellore E. Raghunathan is Professor, Department of Biostatistics and Institute for Social Research, University of Michigan, Ann Arbor, MI 48109 (E-mail: *teraghu@umich.edu*). This research was supported by the National Science Foundation grant ITR-0427889.

published literature. This article describes some of the main adaptations of the multiple imputation framework, namely missing data in large and small samples, data confidentiality, and measurement error. It reviews the combining rules for each setting and explains why they differ. Finally, it highlights research topics in extending the multiple imputation framework.

Key Words: Confidentiality; Measurement error; Missing data; Synthetic.

1 INTRODUCTION

Multiple imputation (Rubin, 1987) was first conceived as a tool that statistical agencies could use to handle nonresponse in large data sets that are disseminated to the public. The basic idea is for the statistical agency to simulate values for the missing data repeatedly by sampling from predictive distributions of the missing values. This creates multiple, completed datasets that are disseminated to the public. This has been done, for example, for public release files for the Fatality Analysis Reporting System (Heitjan and Little, 1991), the Consumer Expenditures Survey (Raghunathan and Paulin, 1998), the National Health and Nutrition Examination Survey (Schafer *et al.*, 1998), the Survey of Consumer Finances (Kennickell, 1998), and the National Health Interview Survey (Schenker *et al.*, 2006). See Rubin (1996) and Barnard and Meng (1999) for other examples of multiple imputation for missing data.

Multiple imputation is appealing for handling nonresponse in large datasets because it moves the missing data burden from data analysts to data producers, who typically have greater resources than analysts. When the imputation models meet certain conditions (Rubin, 1987, Chapter 4), analysts of the completed datasets can obtain

valid inferences using complete-data statistical methods and software. Specifically, the analyst computes point and variance estimates of interest with each dataset and combines these estimates using simple formulas developed by Rubin (1987). These formulas serve to propagate the uncertainty introduced by imputation through the analyst's inferences, enabling the analyst to focus on modeling issues rather than estimation technicalities.

In the last two decades, multiple imputation has evolved beyond the context of large sample survey nonresponse. Individual researchers now routinely use multiple imputation for missing data in small samples, as evidenced by the development of multiple imputation procedures for mainstream software like SAS, Stata, and S-Plus. Statistical agencies release multiply-imputed datasets to protect the confidentiality of survey respondents' identities or sensitive attributes in public-use files (Kennickell, 1998; Abowd and Woodcock, 2001). Survey methodologists and epidemiologists use multiple imputation to edit and correct for measurement errors (Ghosh-Dastidar and Schafer, 2003; Winkler, 2003; Cole *et al.*, 2006) or to recode variables due to changes in definitions (Clogg *et al.*, 1991; Schenker, 2003). In some of these settings, Rubin's (1987) rules for combining the point and variance estimates are not applicable, yielding confidence intervals without nominal coverage rates or significance tests without nominal levels. The original rules fail because what is considered known by the analyst, and therefore part of the conditional expectations and variances used to obtain the multiple imputation inferences, in these settings differs from the missing data setting. Consequentially, new adaptations of the multiple imputation framework have necessitated the development of new multiple imputation inferences. In fact, more than ten multiple imputation inference methods appear in the literature,

many published in the last five years.

This article summarizes some of the main adaptations of the multiple imputation framework and explains why different adaptations warrant different inferential methods. The rest of this paper is organized as follows. Section 2 reviews multiple imputation for missing data, including recent modifications. Section 3 reviews multiple imputation for data confidentiality, also known as synthetic data. Section 4 reviews multiple imputation for measurement error corrections, including a clarification of the appropriate combining rules in this context. In these sections, we write primarily from the perspective of a statistical agency releasing data to the public. Of course, this is only one area of application for multiple imputation. Section 5 cites examples of applications in other areas and suggests new applications.

2 MULTIPLE IMPUTATION FOR MISSING DATA

We begin this review with its original purpose: handling missing data in large samples. After summarizing Rubin's (1987) original methods, we discuss several adaptations including inference with small samples (Barnard and Rubin, 1999), significance tests of multi-component hypotheses (Li *et al.*, 1991b; Meng and Rubin, 1992), and nested imputation (Shen, 2000; Harel and Schafer, 2003; Rubin, 2003b). We do not cover conditional mean imputation (Schafer and Schenker, 2000), which is an approximation to multiple imputation. This section does not address practical issues like congeniality, specifying imputation models, and ignorability of the missing data. For excellent discussions of these issues, see Rubin (1987, 1996), Meng (1994), Schafer (1997), Little and Rubin (2002), Zhang (2003), Gelman *et al.* (2005), and Reiter *et al.* (2006).

2.1 Standard Multiple Imputation

For a finite population of size N, let $I_j = 1$ if unit j is selected in the survey, and $I_j = 0$ otherwise, where $j = 1, 2, \ldots, N$. Let $\mathbf{I} = (I_1, \ldots, I_N)$. Let $\mathbf{R}_j$ be a $p \times 1$ vector of response indicators, where $R_{jk} = 1$ if the response for unit j to survey item k is recorded, and $R_{jk} = 0$ otherwise. Let $\mathbf{R} = (\mathbf{R}_1, \ldots, \mathbf{R}_N)$. Let $\mathbf{Y}_{\text{inc}} = (\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$ be the $n \times p$ matrix of survey data for the n units with $I_j = 1$; $\mathbf{Y}_{\text{obs}}$ is the portion of $\mathbf{Y}_{\text{inc}}$ that is observed, and $\mathbf{Y}_{\text{mis}}$ is the portion of $\mathbf{Y}_{\text{inc}}$ that is missing due to nonresponse. Let $\mathbf{Y} = (\mathbf{Y}_{\text{inc}}, \mathbf{Y}_{\text{exc}})$ be the $N \times p$ matrix of survey data for all units in the population. Let $\mathbf{X}$ be the $N \times d$ matrix of design variables for all N units in the population, e.g. stratum or cluster indicators or size measures. We assume that such design information is known for all population units, for example from census records or the sampling frame(s). Missing values in $\mathbf{X}$ can be treated as part of $\mathbf{Y}_{\text{mis}}$. Finally, we write the observed data as $\mathbf{D} = (\mathbf{X}, \mathbf{Y}_{\text{obs}}, \mathbf{I}, \mathbf{R})$.

The agency fills in values for $\mathbf{Y}_{\text{mis}}$ with draws from the posterior predictive distribution of $(\mathbf{Y}_{\text{mis}} \mid \mathbf{D})$, or approximations of that distribution such as the sequential regression approach of Raghunathan *et al.* (2001). These draws are repeated independently m times to obtain m completed datasets, $\mathbf{D}^{(l)} = (\mathbf{D}, \mathbf{Y}_{\text{mis}}^{(l)})$ where $1 \leq l \leq m$, which are disseminated to the public. Multiple rather than single imputations are used so that analysts can estimate the variabilities due to imputing missing data.

2.1.1 Univariate Estimands: The Large Sample Case

From these imputed datasets the analyst seeks inferences about some estimand $Q = Q(\mathbf{X}, \mathbf{Y})$, for example a population mean or regression coefficient, where the notation $Q(\mathbf{X}, \mathbf{Y})$ indicates a function of $\mathbf{X}$ and $\mathbf{Y}$. In each imputed dataset, the analyst

estimates Q with some estimator $\hat{Q}$ and the variance of $\hat{Q}$ with some estimator $\hat{U}$. It is assumed that the analyst specifies $\hat{Q}$ and $\hat{U}$ by acting as if each $\mathbf{D}^{(l)}$ was in fact collected data from a random sample of $(\mathbf{X}, \mathbf{Y})$ based on the original sampling design $\mathbf{I}$, i.e., $\hat{Q}$ and $\hat{U}$ are complete-data estimators.

For $i = 1, \ldots, m$, let $Q^{(l)}$ and $U^{(l)}$ be respectively the values of $\hat{Q}$ and $\hat{U}$ in the completed dataset $\mathbf{D}^{(l)}$. Under assumptions described in Rubin (1987), the analyst can obtain valid inferences for scalar Q by combining the m replicates of $Q^{(l)}$ and $U^{(l)}$. Specifically, the following quantities typically arise in inferences:

$$\bar{Q}_m = \sum_{l=1}^{m} Q^{(l)}/m \tag{1}$$

$$B_m = \sum_{l=1}^{m} (Q^{(l)} - \bar{Q}_m)^2/(m-1) \tag{2}$$

$$\bar{U}_m = \sum_{l=1}^{m} U^{(l)}/m. \tag{3}$$

The analyst uses $\bar{Q}_m$ to estimate Q and $T_m = (1 + 1/m)B_m + \bar{U}_m$ to estimate $\mathrm{Var}(Q|\mathbf{D}^{(1)}, \ldots, \mathbf{D}^{(m)})$. Here, $\bar{U}_m$ estimates the variance if the data were complete, and $(1 + 1/m)B_m$ estimates the increase in variance because of the missing data.

When $m = \infty$, which is a useful case for motivating combining rules for other adaptations of multiple imputation, under the posited imputation model $\mathrm{Var}(Q|\mathbf{D})$ equals $\mathrm{Var}(E(Q|\mathbf{D}, \mathbf{Y}_{\mathrm{mis}})|\mathbf{D}) + E(\mathrm{Var}(Q|\mathbf{D}, \mathbf{Y}_{\mathrm{mis}})|\mathbf{D}) = B_\infty + \bar{U}_\infty$. This is because each $Q^{(l)}$ and $U^{(l)}$ is, respectively, a draw from the posterior distributions of $E(Q|\mathbf{D}, \mathbf{Y}_{\mathrm{mis}})$ and $\mathrm{Var}(Q|\mathbf{D}, \mathbf{Y}_{\mathrm{mis}})$, as discussed by Rubin (1987, Chapter 3).

When n is large and m is modest, inferences for Q are based on the t-distribution, $(\bar{Q}_m - Q) \sim t_{\nu_m}(0, T_m)$, with $\nu_m = (m-1)\left(1 + \bar{U}_m/((1+1/m)B_m)\right)^2$ degrees of freedom. This degrees of freedom is derived by matching the first two moments of $T_m/\mathrm{Var}(Q|\mathbf{D}^{(1)}, \ldots, \mathbf{D}^{(m)}, B_\infty)$ to the first two moments of a chi-squared distribution.

It has been shown (Wang and Robins, 1998; Robins and Wang, 2000; Nielsen, 2003; Kim *et al.*, 2006) that T_m can be biased. This bias is usually positive. While bias in T_m is clearly undesirable, Rubin (2003a) and others argue that it typically is not substantial enough to outweigh the benefits of using T_m—which is simple to compute—and multiple imputation more generally. The properties of confidence intervals for Q in genuine samples are more important than the asymptotic properties of T_m. Indeed, the primary purpose of estimating T_m lies with constructing confidence intervals for Q. Empirical evidence from genuine applications of the approach suggests that, for sensible complete-data inferences and imputation models, inferences based on T_m perform well for a variety of Qs (Rubin, 2003a).

2.1.2 Univariate Estimands: The Small Sample Case

Rubin's (1987) derivations assume that complete-data inferences about Q can be based on normal distributions. When n is small, however, t-distributions are more appropriate. Barnard and Rubin (1999) developed inferential methods that account for this difference. Their methods still use $\bar{Q}_m$ and T_m as the point and variance estimates, but the degrees of freedom change from ν_m to $\nu_m^* = (\nu_m^{-1} + \hat{\nu}_{\text{obs}}^{-1})^{-1}$, where $\hat{\nu}_{\text{obs}} = \nu_{\text{com}}(\bar{U}_m/T_m)(\nu_{\text{com}}+1)/(\nu_{\text{com}}+3)$, and ν_{com} is the degrees of freedom if the data were complete. The quantity $\hat{\nu}_{\text{obs}}$ is ν_{com} down-weighted by a multiplicative factor that equates the increase in variance due to missing data to a $(\bar{U}_m/T_m)\times 100\%$ reduction in effective sample size.

The quantity ν_m^* has several features that lead Barnard and Rubin (1999) to recommend its general use, regardless of the size of n. First, $\nu_m^* \leq \nu_{\text{com}}$, whereas ν_m can exceed ν_{com}. This property of ν_m^* is desirable, since the presence of missing

data should reduce the degrees of freedom rather than increase it. Second, $\nu_m^* < \nu_m$ with approximate equality when n is large, so that using ν_m^* instead of ν_m is slightly conservative in large samples. Third, ν_m^* is always between ν_{com} and ν_m, making it a compromise degrees of freedom.

2.1.3 Multi-component Estimands: The Large Sample Case

Using the m imputed datasets, the analyst seeks to test the null hypothesis $\mathbf{Q} = \mathbf{Q}_0$ for some k-component estimand $\mathbf{Q}$; for example, to test if k regression coefficients equal zero. Let $\bar{\mathbf{Q}}_m$, $\mathbf{B}_m$, and $\bar{\mathbf{U}}_m$ be the multivariate analogues of $\bar{Q}_m$, B_m, and $\bar{U}_m$. These are computed using k-dimensional estimates $\mathbf{Q}^{(l)}$ and $k \times k$ covariance matrices $\mathbf{U}^{(l)}$, where $1 \leq l \leq m$, in (1) – (3). It may appear reasonable to use a Wald test with statistic $(\bar{\mathbf{Q}}_m - \mathbf{Q}_0)^T \left((1 + 1/m)\mathbf{B}_m + \bar{\mathbf{U}}_m\right)^{-1} (\bar{\mathbf{Q}}_m - \mathbf{Q}_0)$. However, this test is unreliable when $k > m$ and m is moderate, as is frequently the case, because $\mathbf{B}_m$ can have large variability (Rubin, 1987; Li *et al.*, 1991b). Estimating $\mathbf{B}_m$ in such cases is akin to estimating a covariance matrix using fewer observations than there are dimensions. This difficulty is avoided by making m large.

To mitigate the effects of variability when m is moderate, Rubin (1987) proposed taking $\mathbf{B}_\infty = r_\infty \bar{\mathbf{U}}_\infty$, where r_∞ is a scalar. Equivalently, the percentage increases in variance due to nonresponse are equal for all components of $\mathbf{Q}$. Under this restriction, only one additional parameter, r_∞, is needed to estimate $\mathbf{B}_\infty$. Each diagonal element of $\mathbf{B}_m$ (after re-scaling) provides an estimate of r_∞. Hence, assuming $\mathbf{B}_\infty$ is proportional to $\bar{\mathbf{U}}_\infty$ turns the problem of having $m-1$ degrees of freedom to estimate $k(k+1)/2$ (possibly greater than m) parameters into the problem of having $k(m-1)$ degrees of freedom to estimate one parameter.

Using Rubin's proposal, the test statistic is $S_m = (\bar{\mathbf{Q}}_m - \mathbf{Q}_0)^T \bar{\mathbf{U}}_m^{-1} (\bar{\mathbf{Q}}_m - \mathbf{Q}_0)/(k(1+r_m))$ where $r_m = (1+1/m)\,\mathrm{tr}(\mathbf{B}_m \bar{\mathbf{U}}_m^{-1})/k$. The reference distribution for S_m is an approximate F-distribution, F_{k,v_w}, with $v_w = 4 + (t-4)(1+(1-2/t)/r_m)^2$ and $t = k(m-1) > 4$. When $t \leq 4$, we set $v_w = (k+1)\nu_m/2$. The p-value for testing $\mathbf{Q} = \mathbf{Q}_0$ is $\Pr(F_{k,v_w} > S_m)$. Simulations by Li *et al.* (1991b) suggest that for many practical situations with moderate m, this test has better properties than other tests.

The test statistic S_m has the familiar quadratic form of the Wald statistic, but with a correction factor $k(1+r_m)$ in the denominator. The factor of k is needed for a good F-approximation, which is derived by matching the first two moments of S_m. The factor of $1+r_m$ adjusts the quadratic form so that the test statistic is based on the appropriate estimate of variance rather than on $\bar{\mathbf{U}}_m$ alone. To see this, it is instructive to consider the case where Q is scalar, i.e. $k = 1$. Then, $(1+r_m) = T_m/\bar{U}_m$ estimates the percentage increase in the variance due to the missing data relative to the estimated complete-data variance, so that $(1+r_m)\bar{U}_m = T_m$ is the correct variance for the quadratic form. When $\mathbf{Q}$ is multivariate and proportionality holds, r_m estimates $(1+1/m)r_\infty$, so that $(1+r_m)$ can be interpreted as the average percentage increase in variance.

It may be cumbersome to work with $\bar{\mathbf{U}}_m$ for large k. Meng and Rubin (1992) developed an alternative significance test based on the log-likelihood ratio test statistics from the m imputed data sets, which are easily computed for common models like those from exponential families. Their strategy is to (i) find a statistic asymptotically equivalent to S_m based only on values of the Wald statistics calculated in each imputed dataset; (ii) use the asymptotic equivalence of Wald and likelihood ratio test statistics to define the likelihood ratio test statistic; and, (iii) use a reference F

distribution like the one for Wald tests. The key to this strategy is to approximate S_m and r_m without using $\bar{\mathbf{U}}_m$.

Let $\boldsymbol{\psi}$ be the vector of parameters in the analyst's model, and let $\boldsymbol{\psi}^{(l)}$ be the maximum likelihood estimate of $\boldsymbol{\psi}$ computed from $\mathbf{D}^{(l)}$, for $l = 1, \ldots, m$. Suppose the analyst is interested in a k-dimensional function, $\mathbf{Q}(\boldsymbol{\psi})$, and forms the hypothesis that $\mathbf{Q}(\boldsymbol{\psi}) = \mathbf{Q}_0$. Let $\boldsymbol{\psi}_0^{(l)}$ be the maximum likelihood estimate of $\boldsymbol{\psi}$ obtained from $\mathbf{D}^{(l)}$ subject to $\mathbf{Q}(\boldsymbol{\psi}) = \mathbf{Q}_0$. The log-likelihood ratio test statistic associated with $\mathbf{D}^{(l)}$ is $L^{(l)} = 2\log f(\mathbf{D}^{(l)}|\boldsymbol{\psi}^{(l)}) - 2\log f(\mathbf{D}^{(l)}|\boldsymbol{\psi}_0^{(l)})$. Let $\bar{L} = \sum_{l=1}^{m} L^{(l)}/m$; $\bar{\boldsymbol{\psi}} = \sum_{l=1}^{m} \boldsymbol{\psi}^{(l)}/m$; and, $\bar{\boldsymbol{\psi}}_0 = \sum_{l=1}^{m} \boldsymbol{\psi}_0^{(l)}/m$. Meng and Rubin (1992) also use the average of the log-likelihood ratio test statistics evaluated at $\bar{\boldsymbol{\psi}}$ and $\bar{\boldsymbol{\psi}}_0$, which we label as $\bar{L}_0 = (1/m)\sum_{l=1}^{m}(2\log f(\mathbf{D}^{(l)}|\bar{\boldsymbol{\psi}}) - 2\log f(\mathbf{D}^{(l)}|\bar{\boldsymbol{\psi}}_0))$.

The likelihood ratio test statistic is $\hat{S}_m = \bar{L}_0/(k(1+\hat{r}_m))$, where $\hat{r}_m = ((m+1)/t)(\bar{L} - \bar{L}_0)$ is asymptotically equivalent to r_m and $\bar{L}_0$ is asymptotically equivalent to $k(1+r_m)S_m$. The reference distribution for $\hat{S}_m$ is $F_{k,\hat{v}_w}$, where $\hat{v}_w$ is defined like v_w using $\hat{r}_m$ in place of r_m.

Because the likelihood ratio test is an asymptotic equivalent of the Wald test, it has similar properties to the Wald test when n is sufficiently large. Research comparing the properties of the two procedures for modest n is sparse. It also is possible to obtain inferences by combining only the p-values from Wald tests (Li *et al.*, 1991a). However, the performance of this method is unsatisfactory relative to other approaches (Meng and Rubin, 1992; Schafer, 1997).

2.1.4 Multi-component Estimands: The Small Sample Case

Tests of $\mathbf{Q} = \mathbf{Q}_0$ in small samples use the test statistic S_m. However, the denominator degrees of freedom v_w is not appropriate for small n. It is derived assuming that the reference distribution for the complete-data test is a χ^2 distribution, whereas for small samples it is an F-distribution. In fact, with small n, v_w can exceed v_{com}, which may result in a larger proportion of p-values below desired significance levels than would be expected under the null hypothesis for a test with valid frequentist properties.

Reiter (2007b) presents an alternative denominator degrees of freedom derived using a second-order Taylor series expansion and moment matching in the spirit of Barnard and Rubin (1999). A simplified approximation to this degrees of freedom is

$$v_{\text{fapp}} = 4 + \left(\frac{1}{v^*_{\text{com}} - 4(1+a)} + \frac{1}{t-4} \left(\frac{a^2(v^*_{\text{com}} - 2(1+a))}{(1+a)^2(v^*_{\text{com}} - 4(1+a))} \right) \right)^{-1} \quad (4)$$

where $v^*_{\text{com}} = v_{\text{com}}(v_{\text{com}}+1)/(v_{\text{com}}+3)$ and $a = r_m t/(t-2)$. A more complicated expression involving higher order terms is in Reiter (2007b). Note that $v_{\text{fapp}} \leq v_{\text{com}}$ with near equality for small fractions of missing information when t is large relative to v_{com}. Also, $v_{\text{fapp}} = v_w$ for infinite sample sizes, since in that case $(1+a)^2/a^2 = (1+(1-2/t)/r_m)^2$.

2.2 Nested Multiple Imputation

In some situations, it may be advantageous to generate different numbers of imputations for different variables. For example, imputers may want to generate relatively few values for variables that are computationally expensive to impute and many values for variables that are computationally inexpensive to impute. This approach was taken in the National Medical Expenditure Survey (Rubin, 2003b). As a related ex-

ample, when imputers seek to limit the total number of imputations, they may want to release few values for variables with low fractions of missing information—since the between imputation variance may be small for analyses involving these variables—and many values for variables with high fractions of missing information.

Using different numbers of imputations per variable is called nested multiple imputation (Shen, 2000) or two stage multiple imputation (Harel and Schafer, 2003). The nesting refers to the way in which imputations are generated; the data are not necessarily organized in a multi-level structure. To describe nested imputation, we use the setting of expensive and inexpensive imputations. Let $\mathbf{Y}_{\text{exp}}$ be the missing values that are expensive to impute, and let $\mathbf{Y}_{\text{inexp}}$ be the missing values that are inexpensive to impute. The imputer generates imputations in a two-step process. First, the imputer draws values of $\mathbf{Y}_{\text{exp}}^{(l)}$, for $l = 1, \ldots, m$, from the predictive distribution for $(\mathbf{Y}_{\text{exp}} \mid \mathbf{D})$, resulting in m partially completed datasets. Second, in each partially completed dataset, the imputer generates $\mathbf{Y}_{\text{inexp}}^{(l,1)}, \mathbf{Y}_{\text{inexp}}^{(l,2)}, \ldots, \mathbf{Y}_{\text{inexp}}^{(l,r)}$ by drawing from the predictive distribution of $(\mathbf{Y}_{\text{inexp}} \mid \mathbf{D}, \mathbf{Y}_{\text{exp}}^{(l)})$. The result is $M = mr$ completed datasets, $\mathbf{D}^{(l,i)} = (\mathbf{D}, \mathbf{Y}_{\text{exp}}^{(l)}, \mathbf{Y}_{\text{inexp}}^{(l,i)})$, where $l = 1, \ldots, m$ and $i = 1, \ldots, r$. Each dataset includes a label indicating its value of l; i.e., an indicator for its nest.

2.2.1 Univariate Estimands: The Large Sample Case

As shown in Shen (2000), analysts can obtain valid inferences from these released datasets by combining inferences from the individual datasets. As before, let $Q^{(l,i)}$ and $u^{(l,i)}$ be respectively the values of $\hat{Q}$ and $\hat{U}$ in dataset $\mathbf{D}^{(l,i)}$, where $1 \leq l \leq m$

and $1 \leq i \leq r$. Analogous to (1) – (3), we have

$$\bar{Q}_M = \sum_{l=1}^{m}\sum_{i=1}^{r} Q^{(l,i)}/(mr) = \sum_{l=1}^{m} \bar{Q}^{(l)}/m \tag{5}$$

$$\bar{W}_m = \sum_{l=1}^{m}\sum_{i=1}^{r} (Q^{(l,i)} - \bar{Q}^{(l)})^2/m(r-1) = \sum_{l=1}^{m} W^{(l)}/m \tag{6}$$

$$B_m = \sum_{l=1}^{m} (\bar{Q}^{(l)} - \bar{Q}_M)^2/(m-1) \tag{7}$$

$$\bar{U}_M = \sum_{l=1}^{m}\sum_{i=1}^{r} U^{(l,i)}/(mr). \tag{8}$$

Provided the complete-data inferences are valid from a frequentist perspective, and the imputations are proper, one can estimate Q with $\bar{Q}_M$. An estimate of $\mathrm{Var}(Q|\mathbf{D}^{(1,1)}, \ldots, \mathbf{D}^{(m,r)})$ is $T_M = (1 + 1/m)B_m + (1 - 1/r)\bar{W}_m + \bar{U}_M$. When n is large, inferences can be based on the t-distribution, $(Q - \bar{Q}_M) \sim t_{\nu_M}(0, T_M)$, with degrees of freedom,

$$\nu_M = \left(\frac{(1 + 1/m)B_m)^2}{(m-1)T_M^2} + \frac{((1 - 1/r)\bar{W}_m)^2}{m(r-1)T_M^2} \right)^{-1}. \tag{9}$$

To derive ν_M, match the first two moments of $T_M/\mathrm{Var}(Q|\mathbf{D}^{(1,1)}, \ldots, \mathbf{D}^{(m,r)}, B_\infty, \bar{W}_\infty)$ to those of a χ^2 distribution. An adjusted degrees of freedom for small n has not been developed for nested multiple imputation, although nested imputation is not particularly useful for small n since imputations are not computationally expensive.

The variance formula for T_M differs structurally from that for T_m because datasets within any nest l use the common set of imputed values $\mathbf{Y}_{\mathrm{exp}}^{(l)}$. To illustrate the difference, assume $m = r = \infty$. Then, $\bar{U}_\infty$ has the same interpretation as in standard multiple imputation: it estimates the complete-data variance associated with $Q(\mathbf{X}, \mathbf{Y}_{\mathrm{inc}})$. However, B_∞ has a different interpretation: it estimates the variance due to nonresponse in $\mathbf{Y}_{\mathrm{exp}}$ and part of the variance due to nonresponse in $\mathbf{Y}_{\mathrm{inexp}}$. This latter component is the variability of the $\bar{Q}^{(l)}$s across nests; i.e., the variance

of the within-nest expected values of the $Q^{(l,i)}$s. The variability of the $Q^{(l,i)}$s around their within-nest expected values is estimated by $\bar{w}_\infty$. Adding all sources together for $m = r = \infty$ gives

$$\begin{aligned}
\text{Var}(Q|\mathbf{D}) &= \text{Var}(E(E(Q|\mathbf{D}, \mathbf{Y}_{\text{exp}}, \mathbf{Y}_{\text{inexp}})|\mathbf{D}, \mathbf{Y}_{\text{exp}})|\mathbf{D}) \\
&+ E(\text{Var}(E(Q|\mathbf{D}, \mathbf{Y}_{\text{exp}}, \mathbf{Y}_{\text{inexp}})|\mathbf{D}, \mathbf{Y}_{\text{exp}})|\mathbf{D}) \\
&+ E(E(\text{Var}(Q|\mathbf{D}, \mathbf{Y}_{\text{exp}}, \mathbf{Y}_{\text{inexp}})|\mathbf{D}, \mathbf{Y}_{\text{exp}})|\mathbf{D}) = B_\infty + \bar{w}_\infty + \bar{U}_\infty. \quad (10)
\end{aligned}$$

For the more realistic setting of moderate m and r, we need to adjust for using only a finite number of imputations at each stage. In standard multiple imputation, we add B_m/m, which is the between-imputation variance divided by the number of imputed datasets. For nested imputation, we follow a similar strategy, but there are m imputed sets of $\mathbf{Y}_{\text{exp}}$ and $M = mr$ sets of $\mathbf{Y}_{\text{inexp}}$. Roughly speaking, the between-imputation variance associated with $\mathbf{Y}_{\text{exp}}$ should be divided by m, and the between-imputation variance associated with $\mathbf{Y}_{\text{inexp}}$ should be divided by M.

Given B_∞ and $\bar{W}_\infty$, for finite M the $\text{Var}(Q|\mathbf{D}^{(1,1)}, \ldots, \mathbf{D}^{(m,r)})$ is $(1 + 1/m)B_\infty + (1 + 1/M)\bar{W}_\infty + \bar{U}_\infty$ (Shen, 2000). We estimate B_∞ and $\bar{W}_\infty$ using an ANOVA decomposition. Here, B_m approximates $B_\infty + \bar{W}_\infty/r$, because each $\bar{Q}^{(l)}$ has between-nest (B_∞) and within-nest ($\bar{W}_\infty/r$) components of variance. And, $\bar{W}_m$ approximates $\bar{W}_\infty$. Plugging in the implied point estimates for B_∞ and $\bar{W}_\infty$ provides T_M.

2.2.2 Multi-component Estimands: Large Sample Case

Shen (2000) develops significance tests of k-dimensional multi-component hypotheses using strategies akin to those outlined in Section 2.1.3, with one key distinction: because there are two missing data variance components, the tests require two as-

sumptions of proportionality.

Let $\bar{\mathbf{Q}}_M, \bar{\mathbf{W}}_M, \mathbf{B}_M$, and $\bar{\mathbf{U}}_M$ be the multivariate analogues of the quantities in (5) – (8). To deal with high variability problems when m and r are modest relative to k, we assume that $\bar{\mathbf{W}}_\infty = r_\infty^{(w)}\bar{\mathbf{U}}_\infty$ and $\mathbf{B}_\infty = r_\infty^{(b)}\bar{\mathbf{U}}_\infty$. Equivalently, the fractions of missing information due to the missing $\mathbf{Y}_{\text{exp}}$ are the same for all components of $\mathbf{Q}$, and this is also the case—with a possibly different fraction—for $\mathbf{Y}_{\text{inexp}}$.

Under these assumptions, the Wald statistic is $S_M = (\bar{\mathbf{Q}}_M - \mathbf{Q}_0)^T\bar{\mathbf{U}}_M^{-1}(\bar{\mathbf{Q}}_M - \mathbf{Q}_0)/(k(1 + r_M^{(b)} + r_M^{(w)}))$, where $r_M^{(b)} = (1 + 1/m)\,\text{tr}(\mathbf{B}_M\bar{\mathbf{U}}_M^{-1})/k$ and $r_M^{(w)} = (1 - 1/r)\,\text{tr}(\bar{\mathbf{W}}_M\bar{\mathbf{U}}_M^{-1})/k$. The quantity $(1 + r_M^{(b)} + r_M^{(w)})$ adjusts the quadratic form so that it is based on a correct estimate of variance. The reference distribution for S_M derived by Shen (2000) is an approximate F-distribution, $F_{k,k\hat{\nu}_M}$, with

$$\hat{\nu}_M = \left(\frac{(r_M^{(b)})^2}{(m-1)(1 + r_M^{(b)} + r_M^{(w)})^2} + \frac{(r_M^{(w)})^2}{m(r-1)(1 + r_M^{(b)} + r_M^{(w)})^2}\right)^{-1}. \tag{11}$$

The denominator degrees of freedom is based on (9) with $(1+1/m)\mathbf{B}_m$, $(1-1/r)\bar{\mathbf{W}}_m$, and $\mathbf{T}_M$ replaced by their estimates under the above proportionality assumptions.

Shen's (2000) degrees of freedom were not derived by matching moments to an F-distribution, as was done for v_w in one-stage multiple imputation. For one-stage multiple imputation, Li *et al.* (1991b) found that tests based on v_w degrees of freedom performed better than tests based on $k\hat{v}_m$ degrees of freedom, where $\hat{v}_m = (m-1)(1+ r_m^{-1})^2$. For nested multiple imputation, Shen's (2000) degrees of freedom have not been compared to those based on the approach of Li *et al.* (1991b).

Shen (2000) derives a likelihood ratio test following the strategy outlined in Section 2.1.3. For each $\mathbf{D}^{(l,i)}$, let $\boldsymbol{\psi}_0^{(l,i)}$ and $\boldsymbol{\psi}^{(l,i)}$ be the maximum likelihood estimates of $\mathbf{Q}$ under the null and alternative hypotheses, respectively. Let $L^{(l,i)} =$

$2\log f(\mathbf{D}^{(l,i)}|\boldsymbol{\psi}^{(l,i)}) - 2\log f(\mathbf{D}^{(l,i)}|\boldsymbol{\psi}_0^{(l,i)})$ and $\bar{L}_M = \sum_{l=1}^m \sum_{i=1}^r L^{(l,i)}/(mr)$. Let $\bar{\boldsymbol{\psi}}^{(l)} = \sum_{i=1}^r \boldsymbol{\psi}^{(l,i)}/r$; $\bar{\boldsymbol{\psi}}_0^{(l)} = \sum_{i=1}^r \boldsymbol{\psi}_0^{(l,i)}/r$; and, $\bar{L}^{(l)} = (1/m)\sum_{i=1}^r (2\log f(\mathbf{D}^{(l,i)}|\bar{\boldsymbol{\psi}}^{(l)}) - 2\log f(\mathbf{D}^{(l,i)}|\bar{\boldsymbol{\psi}}_0^{(l)}))$. Shen (2000) also uses the average of the log-likelihood ratio test statistics evaluated at $\bar{\boldsymbol{\psi}}_M = \sum_{l=1}^m \boldsymbol{\psi}^{(l)}/m$ and $\bar{\boldsymbol{\psi}}_{0M} = \sum_{l=1}^m \boldsymbol{\psi}_0^{(l)}/m$, which we label as $\bar{L}_{0M} = (1/M)\sum_{l=1}^m \sum_{i=1}^r (2\log f(\mathbf{D}^{(l,i)}|\bar{\boldsymbol{\psi}}_M) - 2\log f(\mathbf{D}^{(l,i)}|\bar{\boldsymbol{\psi}}_{0M}))$.

The likelihood ratio test statistic is $\hat{S}_M = \bar{L}_{0M}/(k(1+\hat{r}_M^{(b)}+\hat{r}_M^{(w)}))$, where $\hat{r}_M^{(b)} = ((m+1)/t)\left(\sum_{l=1}^m \bar{L}^{(l)}/m - \bar{LR}_{0M}\right)$ and $\hat{r}_M^{(w)} = (1/k)\left(\bar{L}_M - \sum_{l=1}^m \bar{L}^{(l)}/m\right)$. The reference distribution for $\hat{S}_m$ is $F_{k,k\tilde{v}_M}$, where the $\tilde{v}_M$ is defined like $\hat{v}_M$: use $\hat{r}_M^{(b)}$ and $\hat{r}_M^{(w)}$ in place of $r_M^{(b)}$ and $r_M^{(w)}$.

3 MULTIPLE IMPUTATION FOR CONFIDENTIAL PUBLIC USE DATA

Many national statistical agencies, survey organizations, and researchers disseminate data to the public. Wide dissemination greatly benefits society, enabling broad subsets of the research community to access and analyze the collected data. Often, however, data disseminators cannot release data in their collected form, because doing so would reveal some survey respondents' identities or values of sensitive attributes. Failure to protect confidentiality can have serious repercussions for data disseminators. They may be in violation of laws passed to protect confidentiality, such as the recently enacted Health Insurance and Portability Act and Confidential Information Protection and Statistical Efficiency Act in the U.S. And, they may lose the trust of the public, so that potential respondents are less willing to give accurate answers or even participate in future surveys.

Data disseminators protect confidentiality by stripping unique identifiers like names, social security numbers, and addresses. However, these actions alone may not eliminate the risk of disclosures when quasi-identifiers—e.g., age, sex, race, and marital status—are released. These variables can be used to match units in the released data to other databases. Many data disseminators therefore alter values of quasi-identifiers, and possibly values of sensitive variables, before releasing the data. Common strategies include recoding variables, such as releasing ages or geographical variables in aggregated categories; reporting exact values only above or below certain thresholds, for example reporting all incomes above 100,000 as "100,000 or more"; swapping data values for selected records, for example switching the sexes of some men and women to discourage users from matching; and, adding noise to numerical data values to reduce the possibilities of exact matching or to distort the values of sensitive variables.

These methods can be applied to various degrees. Generally, increasing the amount of alteration decreases the risks of disclosures, but it also decreases the accuracy of inferences obtainable from the released data since these methods distort relationships among the variables. Unfortunately, it is difficult—and for some analyses impossible—for data users to determine how much their particular estimation has been compromised by the data alteration, because disseminators rarely release detailed information about the disclosure limitation strategy. Even when such information is available, adjusting for the data alteration may be beyond some users' statistical capabilities. For example, to properly analyze data that include additive random noise, users should apply measurement error models or the likelihood based approach of Little (1993), which are difficult to use for non-standard estimands.

Because of the inadequacies of standard disclosure limitation techniques, several

statistical agencies have decided to use, or are considering the use of, multiple imputation procedures to limit the risk of disclosing respondents' identities or sensitive attributes in public use data files. This idea, now called synthetic data, was first proposed by Rubin (1993). In his original approach, the data disseminator (i) randomly and independently samples units from the sampling frame to comprise each synthetic dataset, (ii) imputes the unknown data values for units in the synthetic samples using models fit with the original survey data, and (iii) releases multiple versions of these datasets to the public. These are called fully synthetic datasets. Some agencies use or are considering a variant of Rubin's approach: release multiply-imputed datasets comprising the units originally surveyed with only some collected values, such as sensitive values at high risk of disclosure or values of quasi-identifiers, replaced with multiple imputations. These are called partially synthetic datasets.

Releasing synthetic data can preserve confidentiality, since identification of units and their sensitive data can be difficult when some or all of the released data are not actual, collected values. Furthermore, using appropriate data generation and estimation methods based on the concepts of multiple imputation, analysts can make valid inferences for a variety of estimands using standard, complete-data statistical methods and software, at least for inferences congenial to the model used to generate the data. Provided the imputer releases some description of this model, analysts can determine whether or not their questions can be answered using the synthetic data. There are other benefits to using synthetic data, as well as limitations, most of which will not be described here. For further descriptions of fully synthetic data, see Rubin (1993), Raghunathan *et al.* (2003), Raghunathan (2003), and Reiter (2002, 2005a). For partially synthetic data, see Little (1993), Kennickell (1997), Abowd and

Woodcock (2001, 2004), Liu and Little (2002), Reiter (2003, 2004, 2005c), and Mitra and Reiter (2006).

As when imputing missing data, it is necessary to generate multiple copies of the synthetic datasets to enable analysts to estimate variances correctly. However, and perhaps surprisingly at first glance, the Rubin (1987) rules for combining the point and variance estimates do not work in the synthetic data contexts; in fact, they can result in severely biased estimates of variances. New combining rules are needed for each synthetic data strategy. In this section, we review these combining rules and explain why the rules differ across the different applications of multiple imputation.

3.1 Fully Synthetic Data

To construct fully synthetic data, the imputer follows a two-part process. First, the imputer imputes values of $\mathbf{Y}_{\text{exc}}$ to obtain a completed-data population, $(\mathbf{X}, \mathbf{Y}_{\text{com}}^{(l)})$. Imputations are generated from the predictive distribution of $(\mathbf{Y}|\mathbf{D})$, or some approximation to it. Second, the imputer takes a simple random sample of n_{syn} units from $(\mathbf{X}, \mathbf{Y}_{\text{com}}^{(l)})$, producing the synthetic dataset $\mathbf{d}^{(l)} = (\mathbf{X}, \mathbf{Y}_{\text{syn}}^{(l)})$. The lower case $\mathbf{d}$ distinguishes the use of imputed values as synthetic data from the use of imputed values to fill in missing data. The process is repeated independently m times to generate m different synthetic datasets, which are then released to the public. The imputer also could simulate $\mathbf{Y}$ for all N units, thereby avoiding releasing actual values of $\mathbf{Y}$.

In practice, it is not necessary to generate completed-data populations for constructing $\mathbf{Y}_{\text{syn}}^{(l)}$; the imputer need only generate values of $\mathbf{Y}$ for units in the synthetic samples. The formulation of completing the population, then sampling from it, aids in deriving the combining rules.

3.1.1 Univariate Estimands

The analyst specifies $\hat{Q}$ and $\hat{U}$ acting as if the synthetic data were in fact a simple random sample of $(\mathbf{X}, \mathbf{Y})$. The analyst need not worry about the original complex sampling design, which is one of the benefits of the fully synthetic data approach, because the design information is accounted for in the imputation stage (e.g., imputation models condition on stratum and cluster effects). As before, the analyst can use $\bar{Q}_m$ from (1) to estimate Q and $T_f = (1+1/m)B_m - \bar{U}_m$ to estimate $\text{Var}(Q|\mathbf{d}^{(1)}, \ldots, \mathbf{d}^{(m)})$, where B_m and $\bar{U}_m$ are defined in (2) and (3). Although it is possible for $T_f < 0$, negative values can be avoided by making m and n_{syn} large. A more complicated variance estimator that is always positive is described in Raghunathan *et al.* (2003). When $T_f > 0$, and n and n_{syn} are large, inferences for scalar Q can be based on a t-distribution with $\nu_f = (m-1)(1 - m\bar{U}_m/((m+1)B_m))^2$ degrees of freedom. A degrees of freedom for small n has not been derived, although the typical application for fully synthetic data is dissemination of survey data with large n.

Obviously, $T_f \neq T_m$: $\bar{U}_m$ is subtracted rather than added. This seemingly minor difference in T_f and T_m hides fundamental differences in the sources of variability estimated by B_m and $\bar{U}_m$. To illustrate these differences, we first take the case where $m = \infty$ and $n_{\text{syn}} = N$, so that each $\mathbf{d}^{(l)}$ is a completed population. In this case, each $U^{(l)} = 0$ because entire populations of values are released, so that $T_f = B_\infty$. The process of repeatedly completing populations and estimating Q is equivalent to simulating the posterior distribution of Q. Hence, when $n_{\text{syn}} = N$, B_∞ estimates $\text{Var}(Q|\mathbf{D})$. This differs completely from the standard missing data case, where B_∞ estimates the increase in variance due to nonresponse and $B_\infty + \bar{U}_\infty$ estimates $\text{Var}(Q|\mathbf{D})$.

Now consider the case when $m = \infty$ and $n_{\text{syn}} < N$ to motivate the subtraction in T_f. Each $Q^{(l)}$ is affected by two sources of variance: the variance due to imputing $\mathbf{Y}_{\text{exc}}$, which causes the values of $Q^{(l)}_{\text{com}} = Q(\mathbf{X}, \mathbf{Y}^{(l)}_{\text{com}})$ to differ across completed populations, and the variance due to sampling n_{syn} records from $(\mathbf{X}, \mathbf{Y}^{(l)}_{\text{com}})$. The first source is $\text{Var}(Q|\mathbf{D})$, since the infinite collection of $Q^{(l)}$s simulates the posterior distribution of Q. The second source is $\bar{U}_\infty$. Hence, $B_\infty = \text{Var}(Q|\mathbf{D}) + \bar{U}_\infty$, and $\text{Var}(Q|\mathbf{D}) = B_\infty - \bar{U}_\infty$. Since $\bar{Q}_\infty = Q_{\text{obs}}$, we have $\text{Var}(Q|\mathbf{d}^{(1)}, \ldots, \mathbf{d}^{(\infty)}) = \text{Var}(Q|\mathbf{D})$.

For moderate m, we replace B_∞ with B_m and U_∞ with $\bar{U}_m$, and add B_m/m to adjust for using only a finite number of synthetic datasets.

3.1.2 Multi-component Estimands

Significance tests for multi-component estimands are derived using the logic described in Section 2.1.3 (Reiter, 2005b). To minimize the impact of high variability in $\mathbf{B}_m$, the test statistics are derived under the assumption of equal fractions of missing information across all components of Q; i.e., $\mathbf{B}_\infty = r_\infty \mathbf{U}_\infty$. This assumption is generally reasonable in fully synthetic data, since all variables are imputed.

The Wald statistic is $S_f = (\bar{\mathbf{Q}}_m - \mathbf{Q}_0)^T \bar{\mathbf{U}}_m^{-1} (\bar{\mathbf{Q}}_m - \mathbf{Q}_0)/(k(r_f - 1))$, where $r_f = (1 + 1/m)\,\text{tr}(\mathbf{B}_m \bar{\mathbf{U}}_m^{-1})/k$. The reference distribution for S_f is an F-distribution, F_{k,v_f}, with $v_f = 4 + (t-4)(1 - (1 - 2/t)/r_f)^2$. The likelihood ratio test statistic is $\hat{S}_f = \bar{L}_0/(k(\hat{r}_f - 1))$, where $\hat{r}_f = ((m+1)/t)(\bar{L} - \bar{L}_0)$. The reference distribution for $\hat{S}_f$ is $F_{k,\hat{v}_f}$, where the $\hat{v}_f$ is defined as for v_f using $\hat{r}_f$.

The correction factor $r_f - 1$ serves a purpose akin to $1 + r_m$ in the missing data setting. It adjusts the quadratic form so that the test statistic is based on an appropriate estimate of the variance of $\bar{\mathbf{Q}}_m$. This is most easily seen with a scalar Q. Here,

$r_f - 1 = T_f/\bar{U}_m$, so that adding $r_f - 1$ appropriately adjusts the quadratic form to be based on T_f. The quantity $\bar{L}_0$ is an asymptotically equivalent replacement for the quadratic form in the Wald statistic for S_f, and $\hat{r}_f$ replaces r_f.

3.2 Partially Synthetic Data

Partially synthetic datasets look like datasets with multiple imputations for missing data. However, synthetic data imputations are replacements rather than completions, which leads to variance formulas that differ from T_m. We first describe partially synthetic data assuming no missing data, i.e. $\mathbf{Y}_{\text{inc}} = \mathbf{Y}_{\text{obs}}$, then broaden to the case when there are missing data.

To generate partially synthetic datasets when $\mathbf{Y}_{\text{inc}} = \mathbf{Y}_{\text{obs}}$, the imputer replaces selected values from the observed data with imputations. Let $Z_j = 1$ if unit j has any of its observed data replaced with synthetic values, and let $Z_j = 0$ for those units with all data left unchanged. Let $\mathbf{Z} = (Z_1, \ldots, Z_n)$. Let $\mathbf{Y}_{\text{rep}}^{(l)}$ be all the imputed (replaced) values in the lth synthetic dataset, and let $\mathbf{Y}_{\text{nrep}}$ be all unchanged (unreplaced) values of $\mathbf{Y}_{\text{obs}}$. The $\mathbf{Y}_{\text{rep}}^{(l)}$ are assumed to be generated from the predictive distribution of $(\mathbf{Y}_{\text{rep}} \mid \mathbf{D}, \mathbf{Z})$, or a close approximation of it. Each synthetic dataset, $\mathbf{d}^{(l)}$, comprises $(\mathbf{X}, \mathbf{Y}_{\text{rep}}^{(l)}, \mathbf{Y}_{\text{nrep}}, \mathbf{I}, \mathbf{Z})$. Imputations are made independently m times to yield m different partially synthetic data sets, which are released to the public.

3.2.1 Univariate Estimands

Inferences from partially synthetic datasets are based on quantities defined in (1) – (3). The analyst specifies $\hat{Q}$ and $\hat{U}$ acting as if each $\mathbf{d}^{(l)}$ was a random sample of $(\mathbf{X}, \mathbf{Y})$ collected with the original sampling design $\mathbf{I}$. As shown by Reiter (2003), the

analyst uses $\bar{Q}_m$ to estimate Q and $T_p = B_m/m + \bar{U}_m$ to estimate $\text{Var}(Q|\mathbf{d}^{(1)}, \ldots, \mathbf{d}^{(m)})$. Inferences are based on t-distributions with $\nu_p = (m-1)(1 + \bar{U}_m/(B_m/m))^2$ degrees of freedom. As for fully synthetic data, there is no adjusted degrees of freedom for small n, nor is one likely to be useful.

The formula for T_m includes $(1+1/m)B_m$, whereas the formula for T_p includes just B_m/m. This difference is explained by letting $m = \infty$. In the partially synthetic data context, $\bar{U}_\infty$ estimates the variance of the completed data, as is the case in standard multiple imputation. However, when $\mathbf{Y}_{\text{inc}} = \mathbf{Y}_{\text{obs}}$, the completed and observed data are identical, so that $\bar{U}_\infty$ by itself estimates $\text{Var}(Q|\mathbf{D})$. It is not necessary to add B_∞ as is done in the missing data case. For finite m, we replace $\bar{U}_\infty$ with $\bar{U}_m$, and add B_m/m for the additional variance due to using a finite number of imputations.

3.2.2 Multi-component Estimands

The logic for significance tests for multi-component estimands again parallels that summarized in Section 2.1.3 (Reiter, 2005b). The derivations of the test statistics use the assumption of equal fractions of missing information across all variables.

The Wald statistic for partially synthetic data is $S_p = (\bar{\mathbf{Q}}_m - \mathbf{Q}_0)^T \bar{\mathbf{U}}_m^{-1} (\bar{\mathbf{Q}}_m - \mathbf{Q}_0)/(k(1+r_p))$, where $r_p = (1/m)\,\text{tr}(\mathbf{B}_m \bar{\mathbf{U}}_m^{-1})/k$. The reference distribution for S_p is an F-distribution, F_{k,w_p}, with $\nu_p = 4 + (t-4)(1 + (1-2/t)/r_p)^2$, where $t = k(m-1)$. The likelihood ratio test statistic is $\hat{S}_p = \bar{L}_0/(k(1+\hat{r}_p))$, where $\hat{r}_p = (1/t)(\bar{L} - \bar{L}_0)$. The reference distribution for $\hat{S}_p$ is $F_{k,\hat{\nu}_p}$, where $\hat{\nu}_p$ is defined akin to ν_p using $\hat{r}_p$.

In this setting, $1+r_p$ can be interpreted as the average relative increase in variance across the components of $\mathbf{Q}$ from the partial synthesis. It adjusts the quadratic form so that the test statistic is based on an appropriate estimate of variance.

3.2.3 Imputation of Missing Data And Partially Synthetic Data

When some data are missing, it is logical to impute the missing and partially synthetic data simultaneously, possibly from different distributions since the replacement imputations should condition on $\mathbf{Z}$. Imputing the $\mathbf{Y}_{\text{mis}}$ and $\mathbf{Y}_{\text{rep}}$ simultaneously generates two sources of variability, in addition to the sampling variability in $\mathbf{D}$, that the analyst must account for to obtain valid inferences. Neither T_m nor T_p correctly estimate the total variation introduced by the dual use of multiple imputation. The bias of each can be illustrated with two simple examples. Suppose that only one value needs replacement, but there are hundreds of missing values to be imputed. Intuitively, the variance of the point estimator of Q should be well approximated by T_m, and T_p should underestimate the variance, as it is missing a B_m. On the other hand, suppose only one value is missing, but there are hundreds of values to be replaced. Then, the variance should be well approximated by T_p, and T_m should overestimate the variance, as it includes an extra B_m.

To allow analysts to estimate the total variability correctly, imputers can employ a three-step procedure for generating imputations (Reiter, 2004, 2007a). First, the imputer fills in $\mathbf{Y}_{\text{mis}}$ with draws from the predictive distribution for $(\mathbf{Y}_{\text{mis}} \mid \mathbf{D})$, resulting in m completed datasets, $\mathbf{D}^{(1)}, \ldots, \mathbf{D}^{(m)}$. Second, in each $\mathbf{D}^{(l)}$, the imputer selects the units whose values are to be replaced; i.e., those whose $Z_j = 1$. Third, in each $\mathbf{D}^{(l)}$, the imputer imputes values $\mathbf{Y}_{\text{rep}}^{(l,i)}$ for those units with $Z_j = 1$, using the predictive distribution for $(\mathbf{Y}_{\text{rep}} \mid \mathbf{D}^{(l)}, \mathbf{Z})$. This is repeated independently r times for $l = 1, \ldots, m$, so that a total of $M = mr$ datasets are generated. Each dataset, $\mathbf{d}^{(l,i)} = (\mathbf{X}, \mathbf{Y}_{\text{nrep}}, \mathbf{Y}_{\text{mis}}^{(l)}, \mathbf{Y}_{\text{rep}}^{(l,i)}, \mathbf{I}, \mathbf{R}, \mathbf{Z})$, includes a label indicating the l of the $\mathbf{D}^{(l)}$

from which it was drawn. These M datasets are released to the public.

This procedure is closely related to nested multiple imputation, and the methods for obtaining inferences use the quantities from Section 2.2.1. As described in Reiter (2004), the analyst can use $\bar{Q}_M$ to estimate Q, where $\bar{Q}_M$ is defined as in (5). An estimate of $\text{Var}(Q|\mathbf{d}^{(1,1)}, \ldots, \mathbf{d}^{(m,r)})$ is $T_{\text{MP}} = (1 + 1/m)B_m - \bar{W}_m/r + \bar{U}_M$, where B_m, $\bar{W}_m$, and $\bar{U}_M$ are as defined in (6) - (8). When n is large, inferences can be based on the t-distribution, $(Q - \bar{Q}_M) \sim t_{\nu_{\text{MP}}}(0, T_{\text{MP}})$, with degrees of freedom

$$\nu_{\text{MP}} = \left(\frac{((1 + 1/m)B_m)^2}{(m-1)T_M^2} + \frac{(\bar{W}_m/r)^2}{m(r-1)T_M^2} \right)^{-1}. \tag{12}$$

Significance tests for multi-component hypotheses have not yet been developed for this setting.

The behavior of T_{MP} and ν_{MP} in special cases is instructive. When r is large, $T_{\text{MP}} \approx T_m$. This is because each $\bar{Q}^{(l)} \approx Q^{(l)}$, which is the point estimate of Q based on $\mathbf{D}^{(l)}$, so that we obtain the results from analyzing $\mathbf{D}^{(l)}$. When the fraction of replaced values is small relative to the fraction of missing values, $\bar{W}_m$ is small relative to B_m, so that once again $T_{\text{MP}} \approx T_m$. In both these cases, ν_{MP} approximately equals ν_m, which is Rubin's (1987) degrees of freedom when imputing missing data only. When the fraction of missing values is small relative to the fraction of replaced values, $B_m \approx \bar{W}_m/r$, so that T_{MP} is approximately equal to T_p with M released datasets.

The distinction between T_M from Section 2.2.1 and T_{MP} mirrors the distinction between T_m and T_p: there is an extra $\bar{W}_m$ in T_M that is not present in T_{MP}. With T_M, the varying imputations within each nest fill in the inexpensive missing values, whereas with T_{MP}, the varying imputations within each nest replace existing values. This distinction results in the different formulas. To illustrate, consider the case with

$m = r = \infty$, so that $T_M = B_\infty + \bar{W}_\infty + \bar{U}_\infty$ and the $T_{\text{MP}} = B_\infty + \bar{U}_\infty$. In this case, T_M estimates $\text{Var}(Q|\mathbf{D})$, as explained in Section 2.2.1. The T_{MP} also estimates $\text{Var}(Q|\mathbf{D})$, since when $r = \infty$ the setting is equivalent to multiple imputation for missing data. The extra $\bar{W}_\infty$ is not needed in T_{MP} because the data associated with the replacements were observed, much like an extra B_∞ is not needed in T_p.

Given B_∞ and $\bar{W}_\infty$, for finite m and r, $\text{Var}(Q|\mathbf{d}^{(1,1)}, \ldots, \mathbf{d}^{(m,r)}) = (1+1/m)B_\infty + \bar{W}_\infty/M + \bar{U}_\infty$ (Reiter, 2004). As with nested imputation, B_m approximates $B_\infty + \bar{W}_\infty/r$. Plugging in the implied point estimates for B_∞ and $\bar{W}_\infty$ produces T_{MP}.

4 MULTIPLE IMPUTATION FOR MEASUREMENT ERROR/DATA EDITING

Many surveys contain non-sampling errors from sources other than nonresponse bias. For example, respondents can misunderstand questions or provide incorrect information; interviewers can affect respondents' answers; and, the recording process can generate errors. Statistical agencies routinely edit collected data, fixing obvious mistakes and inconsistencies. For some errors there is no deterministic fix-up, and the measurement error is treated as stochastic. Stochastic measurement error can be handled directly using measurement error models or Bayesian posterior simulation approaches. Similar issues plague observational data in many fields. For example, in medical studies, good measures of exposure may be available for only some records in the file, and other records have poorly measured exposures.

Agencies disseminating data to the public generally have more resources and knowledge to correct measurement errors than individual researchers. Thus, it is

prudent for the agency to make corrections for analysts before dissemination. The multiple imputation framework is well-suited for this task: the agency replaces values with stochastic measurement errors with draws from probability distributions designed to correct the errors, creating "ideal" datasets. Analysts of these datasets can use standard methods rather than measurement error techniques, since the adjustments for measurement error are automatically included in the ideal datasets. Releasing multiply-imputed ideal datasets enables analysts to incorporate the uncertainty due to simulation. For examples of the multiple imputation approach to data editing, see Winkler (2003) and Ghosh-Dastidar and Schafer (2003), who use multiple imputation to handle missing data and measurement error simultaneously. Individual researchers can follow similar strategies for measurement error correction, as is done in medical contexts by Raghunathan and Siscovick (1998), Yucel and Zaslavsky (2005), Cole *et al.* (2006), and Raghunathan (2006).

As with all applications of multiple imputation, the analyst estimates the parameters of interest and their associated measures of uncertainty in each ideal dataset. It is not obvious, however, how to combine the point and variance estimates: does Rubin's (1987) variance estimator T_m apply, or is a different combining rule needed? The answer depends on what distribution is used for imputations and what data are used for analyses, as we now illustrate.

For simplicity, suppose that the observed data comprise one variable $\mathbf{X}$ without measurement error and one variable $\mathbf{E}$ subject to measurement error. Let $\mathbf{Y}$ be the unknown, true values associated with $\mathbf{E}$. The observed data are $\mathbf{D} = (\mathbf{X}, \mathbf{E}, \mathbf{I})$. Assume there are no missing values in $\mathbf{D}$. Measurement error corrections utilize information about the relationship between $\mathbf{Y}$ and $\mathbf{E}$, for example from a validation

sample of records on which both the true and with-error values are measured. The validation sample could be records from an external file or records from $\mathbf{D}$.

Suppose that the validation sample is an external file, $\mathbf{D}_{\text{val}} = (\mathbf{X}_{\text{val}}, \mathbf{Y}_{\text{val}}, \mathbf{E}_{\text{val}})$. The imputer simulates values of $\mathbf{Y}$ in $\mathbf{D}$ by drawing from $f(\mathbf{Y}|\mathbf{D}, \mathbf{D}_{\text{val}})$ to obtain the ideal datasets $\mathbf{D}^{(l)} = (\mathbf{X}, \mathbf{Y}^{(l)}, \mathbf{E})$, for $l = 1, \ldots, m$. To analyze these data, one approach is to append $\mathbf{D}_{\text{val}}$ to each $\mathbf{D}^{(l)}$ without distinction, so that the analyses are based on $\mathbf{D}^{(*l)} = (\mathbf{D}^{(l)}, \mathbf{D}_{\text{val}})$, for $l = 1, \ldots, m$. This essentially treats the measurement error imputations as completions of missing values of $\mathbf{Y}$ in $(\mathbf{D}, \mathbf{D}_{\text{val}})$, so that Rubin's (1987) theory can apply (after adjustment of any original survey weights) and T_m is the appropriate variance estimator. In the public-use context, data producers can release each $\mathbf{D}^{(*l)}$ without $\mathbf{E}$, since analyses should be based on $\mathbf{Y}$. By similar logic, analyses with internal validation samples also use T_m.

On the other hand, suppose that $\mathbf{D}_{\text{val}}$ cannot be released to the public and is used solely for correcting measurement error. That is, analysis must be based on $\mathbf{D}^{(l)}$ for $l = 1, \ldots, m$. This does not fit cleanly into the missing data set-up, which calls the use of T_m into question. This is evident in the work of Rubin and Schenker (1987) and Clogg *et al.* (1991), who use multiple imputation to recode occupations in an example that fits this context. They find that T_m overestimates variances. In a way, releasing only the $\mathbf{D}^{(l)}$s is akin to synthesizing replacement values for $\mathbf{Y}$, as is done in the context of disclosure limitation. However, the parameter values of the imputation model are estimated from $\mathbf{D}_{\text{val}}$ rather than $\mathbf{D}$, which differs from partial synthesis. This suggests that inferential techniques other than those based on Rubin's (1987) rules or Reiter's (2003) rules are appropriate for this setting. As of this writing, the correct combining rules for this context have not been developed.

As suggested by Harel and Schafer (2003), versions of nested multiple imputation might be employed to handle missing data and editing simultaneously. The results in Section 2 suggest that T_M should be appropriate when the completed data include both the original and validation samples. It is not clear what variance is appropriate when the completed data include only the original sample.

5 OTHER APPLICATIONS AND OPEN RESEARCH TOPICS

The first four sections describe adaptations of multiple imputation within the context of an organization releasing survey data. By no means is multiple imputation limited to this context. Many problems fit into the incomplete data framework. As a small number of examples, multiple imputation is used to analyze data in coarse categories as occurs with age heaping (Heitjan and Rubin, 1990) or interval censored data (Pan, 2000); to estimate the distribution of times from HIV seroconversion to AIDS (Taylor *et al.*, 1990); to handle missing covariates in case-control studies involving cardiac arrest (Raghunathan and Siscovick, 1996); to integrate data from different sources into one file (Gelman *et al.*, 1998; Rassler, 2003); to estimate latent abilities in educational testing (Mislevy *et al.*, 1992); to reduce respondent burden by asking subsets of questions to different respondents (Thomas *et al.*, 2006); to impute missing outcomes in causal studies, for example to handle noncompliance in anthrax vaccine studies (Rubin, 2004); and, to select models in the presence of missing values (Yang *et al.*, 2005; Ibrahim *et al.*, 2005).

All multiple imputation analyses require combining the point and variance esti-

mates from the imputed datasets, but the rules for combining them depend on what is known in the conditioning, which in turn determines what the various multiple imputation quantities estimate. Using the proper conditioning results in variance formulas for the data confidentiality contexts that differ from the missing data context.

Many open research topics remain. These include (i) further investigating reference distributions for large-sample significance tests of multi-component hypotheses for nested multiple imputation, (ii) developing large-sample significance tests of multi-component significance tests for the version of nested multiple imputation presented in Section 3.2.3, (iii) developing small-sample degrees of freedom for the t-distributions used in inferences for nested multiple imputation and for multiple imputation for data confidentiality (although these techniques are most likely to be applied in large samples), and (iv) developing the correct combining rules for measurement error settings when only the original data are available for analysis. Additionally, it would be profitable to extend the two-stage procedure in nested multiple imputation to more than two stages. This could enable data disseminators to handle nonresponse, editing, and data confidentiality simultaneously in a principled manner.

As these research topics and existing applications indicate, even 20 years after Rubin's seminal book on multiple imputation, we can expect continued adaptation of multiple imputation to handle challenging statistical problems.

References

Abowd, J. M. and Woodcock, S. D. (2001). Disclosure limitation in longitudinal linked data. In P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes, eds., *Confidential-*

ity, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, 215–277. Amsterdam: North-Holland.

Abowd, J. M. and Woodcock, S. D. (2004). Multiply-imputing confidential characteristics and file links in longitudinal linked data. In J. Domingo-Ferrer and V. Torra, eds., *Privacy in Statistical Databases*, 290–297. New York: Springer-Verlag.

Barnard, J. and Meng, X. (1999). Applications of multiple imputation in medical studies: From AIDS to NHANES. *Statistical Methods in Medical Research* **8**, 17–36.

Barnard, J. and Rubin, D. B. (1999). Small-sample degrees of freedom with multiple-imputation. *Biometrika* **86**, 948–955.

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.

Cole, S. R., Chu, H., and Greenland, S. (2006). Multiple-imputation for measurement-error correction. *International Journal of Epidemiology* **35**, 1074–1081.

Gelman, A., King, G., and C., L. (1998). Not asked and not answered: Multiple imputation for multiple surveys. *Journal of the American Statistical Association* **93**, 846 – 857.

Gelman, A., Van Mechelen, I., Verbeke, G., Heitjan, D. F., and Meulders, M. (2005). Multiple imputation for model checking: Completed-data plots with missing and latent data. *Biometrics* **61**, 74 – 85.

Ghosh-Dastidar, B. and Schafer, J. L. (2003). Multiple edit/multiple imputation for multivariate continuous data. *Journal of the American Statistical Association* **98**, 807–817.

Harel, O. and Schafer, J. (2003). Multiple imputation in two stages. In *Proceedings of Federal Committee on Statistical Methodology 2003 Conference*.

Heitjan, D. F. and Little, R. J. A. (1991). Multiple imputation for the Fatal Accident Reporting System. *Applied Statistics* **40**, 13–29.

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.

Ibrahim, J. G., Chen, M. H., Lipsitz, S. R., and Herring, A. H. (2005). Missing data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association* **100**, 332 – 346.

Kennickell, A. B. (1997). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. In W. Alvey and B. Jamerson, eds., *Record Linkage Techniques, 1997*, 248–267. Washington, D.C.: National Academy Press.

Kennickell, A. B. (1998). Multiple imputation in Survey of Consumer Finances. In *Proceedings of the Section on Business and Economic Statistics of the American Statistical Association*, 11–20.

Kim, J. K., Brick, J. M., Fuller, W. A., and Kalton, G. (2006). On the bias of the multiple imputation variance estimator in complex sampling. *Journal of the Royal Statistical Society, Series B* **68**, 509–521.

Li, K. H., Raghunathan, T. E., Meng, X. L., and Rubin, D. B. (1991a). 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. (1991b). 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. (1993). Statistical analysis of masked data. *Journal of Official Statistics* **9**, 407–426.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data: Second Edition*. New York: John Wiley & Sons.

Liu, F. and Little, R. J. A. (2002). Selective multiple imputation of keys for statistical disclosure control in microdata. In *ASA Proceedings of the Joint Statistical Meetings*, 2133–2138.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (disc: P558-573). *Statistical Science* **9**, 538–558.

Meng, X. L. and Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* **79**, 103–111.

Mislevy, R. J., Beaton, A. E., Kaplan, B., and Sheehan, K. M. (1992). Population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement* **29**, 133–162.

Mitra, R. and Reiter, J. P. (2006). Adjusting survey weights when altering identifying

design variables via synthetic data. In J. Domingo-Ferrer and L. Franconi, eds., *Privacy in Statistical Databases*, 177–188. New York: Springer-Verlag.

Nielsen, S. F. (2003). Proper and improper multiple imputation. *International Statistical Review* **71**, 593–607.

Pan, W. (2000). A multiple imputation approach to Cox regression with interval-censored data. *Biometrics* **56**, 199–203.

Raghunathan, T. E. (2003). Evaluation of inferences from multiple synthetic data sets created using semiparametric approach. Report for the National Academy of Sciences Panel on Access to Confidential Research Data.

Raghunathan, T. E. (2006). Combining information from multiple surveys for assessing health disparities. *Allgemeines Statistisches Archiv* **90**, 515 – 526.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodology* **27**, 85–96.

Raghunathan, T. E. and Paulin, G. S. (1998). Multiple imputation of income in the Consumer Expenditure Survey: Evaluation of statistical inference. In *Proceedings of the Section on Business and Economic Statistics of the American Statistical Association*, 1–10.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics* **19**, 1–16.

Raghunathan, T. E. and Siscovick, D. S. (1996). A multiple-imputation analysis of

a case-control study of the risk of primary cardiac arrest among pharmacologically treated hypertensives. *Applied Statistics* **45**, 335–352.

Raghunathan, T. E. and Siscovick, D. S. (1998). Combining exposure information from multiple sources in the analysis of a case-control study. *Journal of Royal Statistical Society, Series D* **47**, 333–347.

Rassler, S. (2003). A non-iterative Bayesian approach to statistical matching. *Statistica Neerlandica* **57**, 58–74.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics* **18**, 531–544.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology* 181–189.

Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology* **30**, 235–242.

Reiter, J. P. (2005a). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A* **168**, 185–205.

Reiter, J. P. (2005b). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference* **131**, 365–377.

Reiter, J. P. (2005c). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics* **21**, 441–462.

Reiter, J. P. (2007a). Selecting the number of imputed datasets when using multiple imputation for missing data and disclosure limitation. *Statistics and Probability Letters* forthcoming.

Reiter, J. P. (2007b). Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. *Biometrika* forthcoming.

Reiter, J. P., Raghunathan, T. E., and Kinney, S. K. (2006). The importance of modeling the survey design in multiple imputation for missing data. *Survey Methodology* **32**, 143–150.

Robins, J. M. and Wang, N. (2000). Inference for imputation estimators. *Biometrika* **87**, 113–124.

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

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics* **9**, 462–468.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91**, 473–489.

Rubin, D. B. (2003a). Discussion on multiple imputation. *International Statistical Review* **71**, 619–625.

Rubin, D. B. (2003b). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica* **57**, 3–18.

Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics* **31**, 161–170.

Rubin, D. B. and Schenker, N. (1987). Interval estimation from multiply-imputed data: A case study using census agriculture industry codes. *Journal of Official Statistics* **3**, 375–387.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schafer, J. L., Ezzati-Rice, T. M., Johnson, W., Khare, M., Little, R. J. A., and Rubin, D. B. (1998). The NHANES III multiple imputation project. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 28–37.

Schafer, J. L. and Schenker, N. (2000). Inference with imputed conditional means. *Journal of the American Statistical Association* **95**, 144–154.

Schenker, N. (2003). Assessing variability due to race bridging: Application to census counts and vital rates for the year 2000. *Journal of the American Statistical Association* **98**, 818–828.

Schenker, N., Raghunathan, T. E., Chiu, P. L., Makuc, D. M., Zhang, G., and Cohen, A. J. (2006). Multiple imputation of missing income data in the National Health Interview Survey. *Journal of the American Statistical Association* **101**, 924–933.

Shen, Z. (2000). *Nested Multiple Imputation*. Ph.D. thesis, Harvard University, Dept. of Statistics.

Taylor, J. M., Munoz, A., Bass, S. M., Saah, A. J., Chmiel, J. S., and Kingsley, L. A. (1990). Estimating the distribution of times from HIV seroconversion to AIDS using multiple imputation: Multicentre AIDS Cohort Study. *Statistics in Medicine* **9**, 505 – 514.

Thomas, N., Raghunathan, T. E., N., S., Katzoff, M. J., and Johnson, C. L. (2006). An evaluation of matrix sampling methods using data from the National Health and Nutrition Examination Survey. *Survey Methodology* **32**, 217 – 232.

Wang, N. and Robins, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika* **85**, 935–948.

Winkler, W. E. (2003). A contingency-table model for imputing data satisfying analytic constraints. Tech. rep., Statistical Research Division, U.S. Bureau of the Census.

Yang, X., Belin, T. R., and Boscardin, J. (2005). Imputation and variable selection in linear regression models with missing covariates. *Biometrics* **61**, 498 – 506.

Yucel, R. M. and Zaslavsky, A. M. (2005). Imputation of binary treatment variables with measurement error in administrative data. *Journal of the American Statistical Association* **100**, 1123 – 1132.

Zhang, P. (2003). Multiple imputation: Theory and method. *International Statistical Review* **71**, 581–592.