WORKING PAPER

Multiple Imputation For Combined-Survey Estimation With Incomplete Regressors In One But Not Both Surveys

MICHAEL S. RENDALL, BONNIE GHOSH-DASTIDAR
MARGARET M. WEDEN AND ZAFAR NAZAROV

WR-887-1

October 2011

This paper series made possible by the NIA funded RAND Center for the Study of Aging (P30AG012815) and the NICHD funded RAND Population Research Center (R24HD050906).

RAND LABOR AND POPULATION

Multiple Imputation For Combined-Survey Estimation With Incomplete Regressors In One But Not Both Surveys

Michael S. Rendall,[1,2] Bonnie Ghosh-Dastidar, [2] Margaret M. Weden,[2]
and Zafar Nazarov[3,2]

[1] University of Maryland, College Park, [2] RAND Corporation, [3] Cornell University

October 24, 2011

Acknowledgements: This work was supported by grant R21-OH009320 from the National Institute of Occupational Safety and Health. The authors are grateful to Cha-Chi Fan for her assistance in compiling the SIPP data, and to Mark Handcock and Sanjay Chaudhuri for discussions about theoretical results on variance reduction and the model-fitting approach to testing for survey comparability. The paper also benefitted from comments received at presentations to the 2011 Population Association of America Annual Scientific Conference and to the Department of Statistics at Brigham Young University. Please address correspondence to Michael S. Rendall at mrendall@umd.edu.

Multiple Imputation For Combined-Survey Estimation

With Incomplete Regressors In One But Not Both Surveys

Abstract

Within-survey multiple imputation (MI) methods are adapted to pooled-survey regression estimation where one survey has a larger set of regressors but fewer observations than the other. This adaption is achieved through: (1) larger numbers of imputations to compensate for the higher fraction of missing values; (2) model-fit statistics to check the assumption that the two surveys sample from a common universe; and (3) specificying the analysis model completely from variables present in the survey with the larger set of regressors, thereby excluding variables never jointly observed. In contrast to the typical within-survey MI context, cross-survey missingness is monotonic and easily satisfies the Missing At Random (MAR) assumption needed for unbiased MI. Large efficiency gains in estimates of coefficients for variables in common between the surveys are demonstrated in an application to sociodemographic differences in the risk of experiencing a disabling occupational injury estimated from two nationally-representative panel surveys.

I. INTRODUCTION

Frequently a social scientist has a choice of more than one survey that he or she could use to analyze a given social phenomenon occurring at a given time. The survey with the best set of predictor variables will typically be chosen, as to do otherwise would risk introducing omitted variable bias. Standard methods for multivariate analysis rely on "rectangular" datasets (all predictor variables are present for all observations), and therefore an analysis that pooled observations across surveys would be rejected on the grounds that only predictor variables present in both surveys could be included. The problem of missing predictor variables and consequent non-rectangular datasets, however, is not unique to analysis with pooled surveys. It also frequently confronts a researcher using a single survey, due to survey item non-response (Allison 2002; Little and Rubin 2002). Standard analysis methods for rectangular datasets require that entire observations must be discarded if item non-response occurs for even one variable that belongs in the regression model, a practice sometimes referred to as "complete case analysis." In response to this apparently wasteful treatment of survey information, "missing data" methods of analysis that combine incomplete observations with complete observations have been developed and are now used widely in the social and health sciences (Schafer and Graham 2002; Raghunathan 2004).

Among missing-data methods, multiple imputation (MI, Rubin 1987) offers a flexible and statistically rigorous option. Little and Rubin (1989) argued for social scientists to consider the efficiency advantages of MI over complete-case analysis, and to consider the implementation advantages of MI over "direct methods" that combine separate likelihoods for incomplete observations and complete observations. These implementation advantages arise primarily from the separation of the imputation step from the target, post-imputation analysis. Successful early adoptions of MI in sociology and demography include studies by Freedman and Wolf (1995), Goldscheider et al (1999), and Sassler and McNally (2003).

In the present study, we extend these within-survey MI methods to the "cross-survey" multiple imputation of variables whose values are effectively missing for every observation in one survey but that are present in the other survey. Cross-survey MI was first proposed by Rubin (1986), but in the context of the “statistical matching” of surveys, each with one or more variables not present in the other. The resulting problems of post-imputation analysis with variables never jointly observed have discredited the statistical matching approach in the social sciences (Rodgers 1984, Ridder and Moffitt 2007), and have left it largely on the margins of statistics (but see also Rässler 2002 and Moriarity and Scheuren 2003). By insisting that the post-MI analysis be specified from variables completely present in one of the two surveys, and imputing variables only to the survey that we designate to be incomplete, we propose a form of cross-survey MI that avoids the “never jointly observed” problem. We address the remaining major challenge of cross-survey analysis, that of evaluating the appropriateness of pooling observations across surveys with differences in sampling design and survey instruments, with a model-fitting approach that compares pooled-survey models with and without regressors that indicate in which survey the observation is found.

The remainder of the paper is structured as follows. In section II immediately below, we describe the relationship of cross-survey MI to both within-survey MI and to direct methods for combining data sources using Maximum Likelihood and Generalized Method of Moments estimators. We describe adaptations of procedures and principles from both within-survey MI and direct data-combining methods that we suggest will implement cross-survey MI most effectively. To test the validity of the assumption that the two surveys sample from a common universe, we propose a model-fitting approach using pre-imputation data and the version of the analysis model that can be specified from variables in common between the two surveys. In section III, we demonstrate cross-survey MI in an application to sociodemographic differences in the risk of experiencing a disabling occupational injury estimated from two nationally-representative panel surveys. We show empirically that large efficiency gains in estimates of coefficients can be achieved for variables in common between the surveys. Our model-fit statistics support the assumption that the two surveys sample from a common social process. Discussion follows in section IV.

II. RELATIONSHIPS OF CROSS-SURVEY MI TO WITHIN-SURVEY MI AND TO DIRECT METHODS FOR COMBINING DATA SOURCES

As noted by Hellerstein and Imbens (1999), multiple imputation (MI) may be viewed as an alternative method to the imposing of moment restrictions from a larger data source to an equation estimated entirely from observations from a smaller data source. The rationale for using the smaller data source is found in the larger number of regressors that are available in the smaller than in the larger data source, allowing for a richer model specification with the smaller data source. Although the analysis may be conducted from the first data source (the smaller sample survey) alone, more efficient estimation can be achieved by additionally exploiting information in the second data source. Imbens and Lancaster (1994) reported large gains in efficiency by incorporating marginal moments from census data with sample-survey joint distributions using a Generalized Method of Moments estimator. Handcock, Rendall, and Cheadle (2005) developed a Constrained Maximum Likelihood estimator to similarly impose restrictions from population data on sample-survey data. Rendall et al (2008) showed that efficiency gains on a greater range of coefficients can be obtained by augmenting a small survey with both population data about bivariate relationships and additional surveys with data on a limited set of multivariate relationships. Hellerstein and Imbens (1999) considered the circumstance in which the data providing the moment restrictions are not from a population data source but instead from a large sample survey, and derived an expression for efficiency loss due to sampling error in the large survey. They considered in some detail the case in which the smaller and larger data sources do not sample from exactly the same universe, and described how an unbiased larger data source can be used to partially correct for bias in the smaller data source. Rendall, Handcock, and Jonsson (2009) considered the case in which the larger data source was biased, and described how improvements in mean square error can nevertheless be achieved by including the larger data source with bias correction in a combined analysis with the smaller data source.

Combined-data approaches based on imposing moment restrictions or constraints, however, quickly become computationally unwieldy as successively more covariate information is included from the larger data source. Moreover, it is often the case that more covariate information may be added only when the larger data source is no longer very large. In that case, more general methods to account for the sampling error introduced by both data sources are desirable. This argues for consideration of methods in which observations from the larger survey, and not merely aggregate moments computed from that survey, are combined with the observations from the smaller survey.

MI methods have been developed primarily for imputation from complete cases to incomplete cases in the same survey ("within-survey MI"). Compared to within-survey imputation, cross-survey imputation requires larger numbers of imputations to compensate for the higher fraction of missing values for variables missing entirely in the larger survey, and needs to account for any sample design and measurement differences between surveys. Cross-survey MI, however, has three desirable features not found in within-survey MI. Most obviously, major efficiency gains can be realized by pooling observations across surveys. This is analogous but additional to the efficiency gains obtained by combining incomplete and complete observations from the same survey in within-survey MI. Second, in cross-survey MI the variables to be imputed are "missing-by-design" (Raghunathan and Grizzle 1995), meaning that the reason an individual is "missing" a value for a variable of interest is that the survey did not ask the question. This is very different from values that are selectively missing due to item non-response or to dropout in a panel survey, and has the statistically important consequence that the "Missing at Random" (MAR) assumption will be more easily met in cross-survey MI than in within-survey MI. Third, the missing-by-design structure of the pooled observations used in cross-survey MI implies that missingness has a monotone rather than arbitrary pattern (Rubin 1987). An individual's being sampled in one survey and not the other means he or she will have missing values for all variables derived from questions not asked in that survey (but asked in the other). The resulting "monotone" missing pattern allows for sequential imputation with separate models for categorical and continuous variables. In within-survey MI the pattern of missingness is generally "arbitrary" (missing values on one variable does not invariably imply missing values on a second variable and vice versa). Consequently, a convenient parametric multivariate model needs to be assumed, which may reduce the flexibility in modeling of datasets with multiple types of variables (e.g. categorical, count, and continuous).

A cross-survey-MI setup that allows for the application of methods and results from within-survey MI and from direct methods of data combining

We describe data-combining in the context of two surveys, a "smaller survey" (Survey 1) that has fewer observations but more regressors and a "larger survey" (Survey 2) that has more observations but fewer regressors. We assume that Survey 1 has outcome variable Y and predictor variables X_1 and X_2 for a sample of size N_1, and that Survey 2 has outcome variable Y and predictor variable X_2 for a sample of size N_2, with $N_2 > N_1$. We assume that no values of Y, X_1, or X_2 are missing in Survey 1 and that no values of Y or X_2 are missing, but all values of X_1 are missing, in Survey 2. The goal of the analysis is to estimate the parameters and standard errors of a multivariate regression model that includes observations from both surveys and that is specified from the survey with the fullest set of available regressors. We consider the special case of a binary outcome variable Y and the logit model, $LOGIT[p] = \ln[p/(1-p)]$, for the regression:

$$LOGIT[\Pr\{Y \mid X_1, X_2)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \qquad (1)$$

Although X_1 and X_2 are predictor variables that will be assumed first to be scalar (single regressors), they may easily be generalized to vectors of regressors.

We first make the assumption that the surveys sample from a common universe using equivalent survey instruments. Under these conditions, it is clear that the methods of within-survey MI with complete and incomplete cases apply equivalently to cross-survey MI. Survey 1 provides the complete cases $\{Y_i, X_{1i}, X_{2i}\}_{i=1}^{N_1}$ and Survey 2 provides the incomplete cases $\{Y_j, X_{2j}\}_{j=1}^{N_2}$. Standard within-survey imputation methods (Rubin 1987) may then be applied to derive estimates of the parameters and standard errors of Model (1), as follows. An imputation model for $E[X_1 \mid X_2, Y]$ is first estimated as a regression of X_1 on X_2 and Y using Survey 1 observations only. Second, using the parameters estimated in the imputation model from the data of Survey 1, together with the values of X_2 and Y in Survey 2, a value for X_1 is imputed by randomly drawing m times for each of the N_2 observations in Survey 2. Third, each of these versions of Survey 2 observations containing a different randomly imputed value for X_1 is concatenated with the N_1 ('complete') observations in Survey 1 to create m 'completed' datasets each of size $N_1 + N_2$. Fourth, the analysis model (1) is estimated on each of the m completed datasets. Indexing the m datasets by k, the analysis model produces m unique realizations $\hat{\beta}_k = \{\hat{\beta}_{1k}, \hat{\beta}_{2k}\}$ of parameter vector $\beta = \{\beta_1, \beta_2\}$. Standard multiple imputation algorithms (Schafer 1997, p.109) are used to combine $\{\hat{\beta}_k\}_{k=1}^m$ to derive the final parameter estimates and their standard errors. The final parameter estimates are derived as simple averages over all m estimates $\hat{\beta} = \{(1/m)\sum_{k=1}^{m}\hat{\beta}_{1k}, (1/m)\sum_{k=1}^{m}\hat{\beta}_{2k}\}$. Let $\bar{U}$ represent the mean within-imputation variance $\bar{U} = (1/m)\sum_{k=1}^{m} Var(\hat{\beta}_k)$ and B represent the between-imputation variance $B = \frac{1}{m-1}\sum_{k=1}^{m}(\hat{\beta}_k - \hat{\beta})^2$. The final standard error estimates about the parameters are then derived as $S.E.(\hat{\beta}) = \sqrt{Var(\hat{\beta})} = \sqrt{\bar{U} + (1+1/m)B}$. The last term $(1+1/m)B$ represents the upward adjustment to the standard errors to account for the imputation of X_1 in the N_2 cases among the $N_1 + N_2$ cases pooled over Surveys 1 and 2 for the analysis.

Note that in the case that X_2 is a vector, sometimes in the practice of MI an imputation model to predict X_1 will be specified with regressors Y and a subset only of the variables in the X_2 vector. This is called an "uncongenial" method of imputation (Meng 1994) because variables in the analysis model are omitted from the imputation model. Uncongenial imputation has two theoretical drawbacks relevant to implementing and evaluating the performance of cross-survey MI. First, analytical expressions of variance reduction of Maximum Likelihood (ML) methods do not then apply directly to variance reduction in MI. Second, uncongenial imputation increases bias, as leaving out variables in the imputation model that are present in the analysis model may attenuate the analysis model regression coefficients (Schenker et al 2010). Uncongenial imputation is analogous in this way to imputation in which variables not used in the imputation model estimated from Survey 1 were actually not present in Survey 1. We discuss this below as the problem of having variables in the analysis model that are "never jointly observed" in Survey 1 or in Survey 2. This is a circumstance in which we advise against the use of cross-survey MI.

Theoretical results for variance reduction when combining complete and incomplete observations over estimation with only the complete observations were derived in the linear regression case by Little (1992), subsequently extended by White and Carlin (2010). A key parameter in evaluating gains to cross-survey MI is the "fraction missing." Following the terminology developed for within-survey MI in White and Carlin, we define the fraction with missing values of X_1 by $\pi = N_2 / (N_1 + N_2)$. This fraction missing can be considered either to be the fraction of incomplete cases in a single survey, or in our case to be the fraction of cases that come from the second of the two surveys.

The principal, or possibly only, variance reductions will then be in $Var(\beta_2)$, the parameter for which the regressor variable (X_2) is observed in both Survey 1 and Survey 2. Variance reduction about $Var(\beta_2)$ will depend not only on the fraction missing π but also on the correlation ρ_{12}^2 between X_2 and X_1, and on the partial correlation of Y and X_1 given X_2, $\rho_{1y.2}^2$. The expression for the proportion by which $Var(\beta_2)$ is reduced by adding observations from Survey 2 in the linear regression case is given by (White and Carlin 2010, p.2922):

$$\pi(1-\rho_{1y.2}^2)[1-\rho_{12}^2(1-2\rho_{1y.2}^2)] \qquad (2)$$

In the special case of no correlation between X_2 and X_1 and when X_1 has no association with of Y independent of variation in X_2, then $Var(\beta_2)$ reduces by the maximum amount, equal to the fraction of observations in Survey 2, π. In this case, however, we could estimate β_2 without the need for MI. Instead we would simply pool Survey 1 and Survey 2 observations and estimate $LOGIT[\Pr\{Y \mid X_1, X_2)] = \beta_0 + \beta_2 X_2$. In the more relevant case of $0 < \rho^2_{1y.2} < 1$ and $0 < \rho^2_{12} < 1$, cross-survey MI will always result in a reduction in $Var(\beta_2)$ in the linear regression case because both $(1-\rho^2_{1y.2})$ and $[1-\rho^2_{12}(1-2\rho^2_{1y.2})]$ will always be less than 1.

In the general case, reductions in $Var(\beta_1)$ will be negligible unless X_1 and X_2 are very highly correlated, given that the observations from Survey 2, in which X_1 is always missing, contribute no direct information about the relationship of X_1 to Y. This result of negligible reductions in coefficients about non-common variables between the two data sources was also found empirically by Imbens and Lancaster (1994), and by Handcock et al (2005) who referred to these β_1-type coefficients as being “indirectly constrained” only. White and Carlin (2010, p.2930) claim, moreover, that in the specific case of binary Y there is never any reduction in $Var(\beta_1)$ achieved through adding the observations from Survey 2. It is not clear, however, if this claim applies to the logistic model only or for all binary outcome models.

In addition to a “congenial” specification of the imputation equation, the number of imputations, m, needs to be sufficiently large for the MI variances $Var(\beta)$ to approach the variances of the Maximum Likelihood (ML) estimator. This approximation of the MI variances to those of ML as the number of imputations m becomes large follows from the expression for the ratio of the variance of the MI estimator to a corresponding ML estimator as $1 + f / m$, where f is “the fraction of missing information” for the parameter (Schafer 1997, p.110). The larger is f, the higher is the number of imputations needed to make MI nearly as efficient as ML, but given that f has an upper bound of 1 then f / m will always quickly converge to 0.

A recommended set of procedures for cross-survey multiple imputation

We propose three conditions and procedures to adapt within-survey MI methods successfully to cross-survey MI: (1) including in the analysis model only variables observed entirely within one of the two surveys; (2) use of sequential multiple imputation; and (3) testing for survey sampling and instrument differences using model-fit statistics calculated for an analysis model specified entirely from variables in common between the two surveys.

(1) Exclusion of Variables Never Jointly Observed

Special care must be taken in the specification of the variables to be included in the analysis and imputation models of a cross-survey MI study. We argue that a cross-survey imputation study should be designed such that the analysis model can be estimated with one of the surveys alone. This guarantees that "variables never jointly observed" (Rubin 1987) will be excluded. Violations of this exclusion condition discredited earlier attempts at cross-survey imputation and estimation conducted under the methodological heading of "statistical matching" (Rodgers 1984). A recent review and extension of that literature is found in D'Orazio, Di Zio, and Scanu (2006). The problem for the credibility of statistical matching techniques is in their handling of the variables not observed in common across the data sources. To give a simple illustration of the problem of variables never jointly observed, assume Survey 1 has outcome variable Y and predictor variable X_1 and Survey 2 has outcome variable Y and predictor variable X_2. The goal is to estimate $E[Y \mid X_1, X_2]$, for example in a multivariate regression model $Y = f(X_1, X_2)$. Without additional information on the joint distribution of Y, X_1, and X_2, estimation that combines observations across the two surveys leads to no additional knowledge about $Y \mid X_1$ that cannot be derived from estimation using observations from Survey 1 alone, and no additional knowledge about $Y \mid X_2$ that cannot be derived from estimation using observations from Survey 2 alone (Ridder and Moffitt 2007, pp.5491-5494).

Additional information may come from an auxiliary data source in which Y , X_1, and X_2 are all observed, though this auxiliary data source will often be for a sample drawn from a different universe such as a segment only of the population (Singh et al 1993). This difference in universe, may be appropriately handled by Bayesian methods that attach probability distributions to represent the degree of similarity of the joint distribution of Y , X_1, and X_2 in the auxiliary data to the true joint distribution in the target population. Such an approach was proposed by Rubin (1986) and was explored by Rässler (2002). No accepted methodology for implementing this, however, has taken hold in the social sciences. The analysis of Gelman, King, and Liu (1998a) is an extension of within-survey MI to a cross-survey context, but is not free of the "variables never jointly observed" problem. We consider their approach now in more detail. The setup of their problem is of 51 cross-sectional surveys (public opinion polls) conducted at various times preceding an election, and an analysis model predicting voter preference. Their cross-survey application of MI drew both on MI for within-survey non-response and on MI for split-survey "missing-by-design" data structures (Raghunathan and Grizzle 1995). Gelman et al developed and estimated a cross-survey multiple imputation model that combined the 51 surveys and that introduced an additional diagnostic layer to understand any unique "survey" effects though a Bayesian hierarchical model. Although they noted that 5 of the 51 surveys included all the questions used to construct their model variables (Gelman et al 1998b, p.870), a critical motivator of their study concerned variables not derived from the survey questions but instead from the period at which the poll was conducted. They handled this by including a variable for the date at which the survey was conducted (Gelman et al 1998a, p.850). This variable is in part a time-until-the-election variable. They noted also, however, that events such as a party convention were likely to affect on voter intentions separately from any overall time trend. A question for a key variable for their model, self-reported ideology (liberal, moderate, or conservative), was not asked in the poll conducted around the time of the party convention. Therefore party convention, political ideology, and voter intention were never jointly observed. The authors argued that this was not problematic for their analysis because "…public opinion shifts are generally uniform across the population…" (Gelman et al 1998a, p.855). This is a type of conditional independence assumption. The nature of this assumption, informed by previous studies and theory, is no stronger than those commonly used to identify statistical models in the social sciences. Nevertheless, given the history of skepticism about cross-survey imputation approach used in the statistical matching literature due to its invoking of conditional independence assumptions about variables never jointly observed, it seems useful to propose a context in which to evaluate and illustrate the utility of the cross-survey multiple imputation approach that does not include variables never jointly observed.

(2) Use of sequential imputation

Both continuous multivariate normal joint imputation methods and chained sequential imputation methods for deriving the joint distribution have been used in multiple imputation (Lee and Carlin 2010). We recommend that cross-survey MI take advantage of the monotone missing pattern (Rubin 1987) that comes with the "missing-by-design" structure of the incomplete data in the cross-survey context to conduct sequential imputation. In our example application below, only one variable has missing values, which can be considered the simplest case of monotone missingness. If X_1 is instead a vector of regressors, they are assumed still to be missing only in Survey 2, and therefore missing values for any one of the elements of X_1 implies missing values for all other elements of X_1, a monotone missingness pattern. If the missingness pattern were instead "arbitrary," for example if in Survey 1 some cases had missing values for X_1 and other cases had missing values for X_2, then a model for the joint imputation of X_1 and X_2 would be needed. This requires that the relationships between Y , X_1, and X_2 be estimated jointly through a parameterized multivariate distribution, and in the practice of multiple imputation this has meant imposing a multivariate normal distribution (Schafer 1997). If either or both X_1 and X_2 are categorical or count, the standard procedure is to first transform these variables using a continuous normal approximation.

Considerable work has been conducted on evaluating the biases induced by imposing a continuous normal approximation or partly categorical or count data (Raghunathan et al 2001; Allison 2009; Lee and Carlin 2010; White and Carlin 2010). When the categorical variable values have close to equal probabilities, the approximation is very good and results in almost no bias. When the categorical variable values have very disparate probabilities (e.g., a probability of less than .05 for a binary variable), the approximation is much worse and substantial bias may be introduced. This has led to the development of sequential regression methods of multiple imputation (Raghunathan et al 2001) that allow for categorical regression equations for categorical variables and linear regression equations for continuous variables. A theoretical concern with sequential imputation is that the distribution resulting from a sequence of sequential imputations may not converge to a true joint distribution. Simulation studies, however, have found this not to be a problem in practice (Raghunathan et al 2001; Lee and Carlin 2010). Moreover, the monotone missingness pattern of the cross-survey MI structure allows the joint distribution to be specified as a series of conditional distributions, whereas arbitrary missingness patterns typically used in simulation studies do not. For these reasons we recommend that cross-survey MI take advantage of the monotone missingness pattern of the missing-by-design structure of the data and use the sequential regression method of multiple imputation.

(3) A model-fitting approach to testing for sampling from the same universe

When the two surveys in the same period are both nationally-representative probability samples and have similar questions on a range of topics, statistical theory indicates that the observations from the two surveys may be considered as being sampled observations from a common population process.

In practice, two surveys will almost never sample from (or be designed to generalize to) exactly the same universe. Moreover, there will often be variations in variable definitions and in survey operations between the two surveys. To address this problem in cross-survey MI analysis, Gelman et al (1998a) proposed a Bayesian hierarchical model with random effects in which survey is a level in the hierarchy. Tighe et al (2010) similarly used a hierarchical approach in their pooled-survey analysis with variables common across all surveys that they pooled. Schenker, Raghunathan, and Bondarenko (2010) instead subdivided their two surveys' samples into subgroups that were identified through propensity-score analysis to have similar covariate distributions across the two surveys. Their analysis was a cross-survey MI approach to treating self-report error in the outcome variable. A disadvantage of this method is that smaller subsamples were then used to estimate the imputation models. Simpler imputation models had to be specified, possibly resulting in attenuation of the coefficients in the analysis model.

We propose a third approach that involves testing the reasonableness of the assumption that two surveys are independent realizations of a common social process based on the statistical theory of model fitting (e.g., Burnham and Anderson 2002; Weakliem 2004). We recommend the fitting of three sets of models: the first with no "survey" covariate; the second with a "survey" main effect only; and the third with a "survey" main effect plus full interactions between "survey" and the model covariates. For each, a penalized model-fit statistic such as the AIC or BIC should be estimated. We argue that this model-fitting diagnostics exercise is most appropriately conducted for an analysis model that can be estimated for both surveys with no imputation. This is because multiple correlations can be estimated and compared only for variables present in both surveys. Returning to the simple example above, and using S_2 as an indicator variable for the observation's being from Survey 2 $(S_2 = 1)$ versus in Survey 1 $(S_2 = 0)$, the models whose fit should be compared exclude X_1 because this is missing for every observation in Survey 2. The three models to be compared are:

$$LOGIT[\Pr\{Y \mid X_2\}] = \beta_0 + \beta_2 X_2 \qquad \text{(1a)}$$

$$LOGIT[\Pr\{Y \mid X_2, S_2)] = \beta_0 + \beta_2 X_2 + \beta_4 S_2 \tag{1b}$$

$$LOGIT[\Pr\{Y \mid X_2, S_2)] = \beta_0 + \beta_2 X_2 + \beta_4 S_2 + \beta_5 X_2 S_2 \tag{1c}$$

If Model (1b) has a smaller model-fit statistic than Model (1a), then the "intercept-shift" variable for Survey 2 is concluded to improve upon the model without it and the variable for S_2 (with its coefficient parameter β_4) should be added to the analysis model (1). If Model (1c) has a smaller model-fit statistic than Model (1b), then we would conclude that the observations from the surveys are not, after all, realizations of a common social process. Without a more complex model specification or partitioning of the sample space to account for the differences between the social processes revealed by the β_5 coefficient on the added term $X_2 S_2$, the estimates of the target model parameters $\{\beta_1, \beta_2\}$ may be biased. We return briefly to this issue in the Discussion section below.

III. APPLICATION TO SOCIODEMOGRAPHIC AND JOB CONTEXT DETERMINANTS OF OCCUPATIONAL INJURY

Overview

We discuss and illustrate all of the above-discussed factors for successful implementation of cross-survey MI in an application to sociodemographic differences in the risk of experiencing a seriously-disabling occupational injury. This is a very low-incidence event, motivating the pooling of observations across surveys to obtain sufficient individuals experiencing the event. Rarity of the outcome was the main motivating factor for Tighe et al's (2010) recent pooled-survey. In their case, however, the goal was simple prevalence estimation. Therefore no problem arose of predictor variables that were not present in one or more of the pooled surveys.

Our estimation combines two nationally representative, longitudinal surveys that we assume sample from a common social process. We conduct diagnostics under a model-fitting framework to test this assumption. Sociodemographic variables and the job-context variable of occupation are present in both surveys. Job satisfaction, which is highly predictive of occupational injury (dissatisfaction with the job is associated with a higher injury incidence), is present in only one of the surveys. We use cross-survey MI to impute the job satisfaction variable to the other survey and to account for the additional variability of the parameter estimates in the pooled-survey occupational injury model.

As noted earlier, standard multiple imputation methods based on imputing according to a joint multivariate normal distribution (e.g., Schafer 1997) have difficulty handling deviations from normality in low probability phenomena (Allison 2009). We address this problem by using a sequential multiple imputation method for monotone data structures (Raghunathan et al 2001), with a single binary logit equation to estimate an imputation model for job satisfaction.

Our example further allows us to take advantage of one of the strengths of multiple imputation over direct estimation methods, being the separation of the handling of missing data from the estimation of the analysis model. This allows for greater flexibility in the choice of analysis model. In our case, repeat observation of exposure to injury in one survey but not the other is handled easily by specifying a non-likelihood-based General Estimating Equation (GEE) estimator (Liang and Zeger 1986) in the analysis model. We are able to implement both the imputation and analysis steps using statistical package software, from SAS Version 9.2 (SAS Institute, no date).

Data and model for occupational injuries

The two surveys are the National Longitudinal Survey of Youth (NLSY79, Bureau of Labor Statistics, no date) and the Survey of Income and Program Participation (SIPP, U.S. Census Bureau, no date). The NLSY79 is of a single cohort aged 14 to 21 in 1979, whereas the SIPP consists of a series of panels of the U.S. population of all ages. Studies by Dembe, Erickson, and Delbos (2004) and Dong (2005) used NLSY79 data to show that both sociodemographic and job characteristics are important predictors of injuries. Previous work using the SIPP to study work-limiting conditions is found in studies by DeLeire (2000, 2001), who pooled the 1986 through 1993 panels of the SIPP to investigate the labor market consequences of disability. Of the two surveys, the NLSY79 allows for a fuller specification of a model predicting occupational injury, due to its inclusion of a variable for job satisfaction. Dembe et al (2004) classified this four-category variable into a dichotomous predictor for “dislikes job” and found this to be highly predictive of occupational injuries even after controlling for occupation and industry.

In accordance with the principle of avoiding variables never jointly observed, we restrict the model specification to one that may be estimated from the NLSY79 surveys alone. This narrows the study universe to individuals in the prime working ages, 24 to 43, as older working ages are not observed in the NSLY79. Moreover, we include SIPP panels initiated from 1986 through 2004 to match approximately the time period of 1988 through 2000 in which injury data are available in the NLSY79.

Occupational injuries are measured in different ways between the NLSY79 and the SIPP. In each NLSY79 wave, the individual is asked to report injuries experienced since the last wave. This is the most recent injury since the last wave (one year earlier for the 1988 to 1994 waves and two years earlier for the 1996, 1998, and 2000 waves). In the SIPP, the individual is asked about injuries only if he or she reports a current work limitation, and then only about the injury that resulted in the work limitation. To code the dependent variable of a “disabling work injury” similarly between the NLSY and SIPP, we consider in the NLSY only injuries reported in a survey wave in which the respondent also reports having a work limitation. That is, we infer the work limitation is due to the work injury based on their coincidence in timing and not on direct reporting of the link as in the SIPP. These definitions of work injury are clearly only approximately equivalent between the two surveys, and this is one of the reasons for using a model-fitting approach to evaluate survey comparability before conducting a pooled-survey analysis with imputation of the “dislikes job” variable from the NLSY79 to the SIPP.

A limited amount of wave-to-wave attrition occurred in the NLSY79. To keep the focus of the study on the novel, *cross-survey* multiple imputation of our study, we discard NLSY79 observations for which data are missing due to attrition or other non-response in any of the waves between 1988 and 2000.

Analysis Model and Imputation Strategy

A model for the hazard that individual *i* experiences a disabling injury in period *t*, designated by the (0,1) variable Y_{it}, is estimated for individuals who were working at the beginning of that period. A feature of our study that motivates the pooling of observations across surveys is the rarity of disabling occupational injuries. Previous, single-survey analyses of social differentials in occupational injury (e.g., Dembe et al 2004; Oh and Shin 2003) combined minor and major injuries in their outcome variable. Our stricter definition of occupational injury in our study than in previous, single-survey studies may yield greater insights into social disadvantage in the workplace than does a more all-encompassing definition of workplace injury.

The model is estimated using data during which the individual is aged 24 to 43, a range that is constrained by the ages of the NLSY79 cohort and the period 1988-2000 during which injuries were observed in that survey. The period of exposure that we model is 8 months. That is the maximum period length for which job variables can be observed at the beginning of the period of exposure to injury in the SIPP. The analysis model includes as regressors age (Age – 25), sociodemographic indicator variables for gender (Female), race/ethnicity (Hispanic and Black), and education (an indicator variable for having completed any years of college education), and work variables for occupation type (Manual and Service) and job satisfaction (Dislikes Job):

$$LOGIT[\Pr\{Y_{it}\}] = \beta_0 + \beta_1 Female_i + \beta_2 Hispanic_i + \beta_3 Black_i + \beta_4 CollegeEduc_{it} +$$

$$\beta_5 (Age - 25)_{it} + \beta_6 Manual_{it} + \beta_7 Service_{it} + \beta_8 Dislike_{it} \qquad (3)$$

In both the SIPP and the NLSY79, the individual is included in the universe of those exposed to injury in period *t* only if he or she is working at the beginning of that period. Occupation (observed in both surveys) and job satisfaction (observed only in the NLSY79) are also observed at the beginning of the period of exposure. In the SIPP there is a single period *t* for which exposure to injury is observed. In the NLSY79 there are up to 13 periods depending on the individual's age in 1988. The youngest individuals were aged 24 and were observed at age 26 in 1990 and at age 36 in 2000 (maximum of 11 periods of exposure to injury), the next youngest were aged 26 in 1989 and contributed a maximum of 12 periods of exposure. All other NLSY79 individuals were aged at least 26 in 1988 and each contributed a maximum of 13 periods of exposure up to 2000. The oldest were aged 31 in 1988 and contributed periods of exposure to injury up to age 43 in 2000. The NLSY79's observations therefore include repeat observations on each individual (from 1988 to 2000 inclusive), whereas the SIPP includes only one observation on each individual. Our estimation of the analysis model accounts for within-person correlation in the NLSY79 by using the method of GEE (Liang and Zeger 1986) in which an identifier for the individual is included. The SAS PROC GENMOD is use to implement this estimator, using unweighted observations.

The variable *Dislike* is missing for all cases in the SIPP. By using MI to produce "complete" SIPP observations with values for *Dislike*, we are able to estimate a model that includes all the relevant predictors found in the survey we would choose if we were limited to using only one of the two surveys. By using both surveys, we achieve increased statistical power to understand sociodemographic and occupational differences in experiencing work-limiting occupational injuries.

The imputation model for the *Dislike* job variable, combining occupation category across injury-exposed years, is:

$$LOGIT[\Pr\{Y_{it}\}] = \delta_0 + \delta_1 Female_i + \delta_2 Hispanic_i + \delta_3 Black_i + \delta_4 CollegeEduc_{it} +$$

$$\delta_5 (Age - 25)_{it} + \delta_6 ManualYears_{it} + \delta_7 ServiceYears_{it} + \delta_8 Injury_{it} \qquad (4)$$

This equation is estimated using all the NLSY79 observations, again unweighted.

For the SIPP, *Dislike* is missing for every observation and is imputed using the parameters estimated from the NLSY79 data in equation (4) together with the SIPP distribution of right-hand-side variables. Twenty versions of the datasets (the multiply-imputed data) are first created, and the same analytical model given by equation (3) is estimated on each dataset. Although 5 imputations are standard in within-survey imputation (Rubin 1987), the large fraction of observations with missing values in our cross-survey imputation case (just over half, being every SIPP observation) leads us to create instead 20 imputed datasets. The parameters and variances about those parameters for each of the 20 datasets are then combined using the standard MI algorithms given in section II above. We use the SAS PROC MI LOGISTIC, with the MONOTONE option to implement the logistic imputation model and the SAS PROC MIANALYZE to account for the additional variance due to imputation. We also used the IVEware software (Raghunathan, Solenberger, and van Hoewyk 2000) as a check on the PROC MI software and found identical results.

To evaluate the efficiency gains to including imputed data, estimation of the hazard of experiencing a work injury by sociodemographic characteristics is conducted on the NLSY79 data alone and on the ‘completed’ pooled dataset consisting of the NLSY79 pooled with the SIPP data that includes imputed values of *Dislike*.

Testing for data non-comparability in combining the NLSY and SIPP samples

When the two surveys running concurrently are both nationally representative and have similar questions on a range of topics, statistical theory indicates that the observations from the two surveys may be considered as being sampled observations from a common population process. In practice, two surveys will almost never sample from exactly the same universe. Moreover, there will often be variations in variable definitions and in survey operations between the two surveys. Differences that need to be addressed in combining the NLSY and SIPP samples include different over-sampling and sample weighting schemes between the two surveys, and differences in the NLSY and SIPP as survey samples and survey instruments. Oversampling of blacks and Hispanics is much greater in the NLSY than in the SIPP, and the sampling frames themselves imply differences, for example, in the likelihood that immigrants will be included in the survey. We address this oversampling in part by using unweighted regression that includes predictor variables encompassing differences in sampling schemes. By doing so, we assume that we are able to approximate a fully-specified model. In particular, predictor variables for black and for Hispanic are included with the expectation that their inclusion will obviate the need for including weights for differential selection into the NLSY79 and SIPP samples by race/ethnicity. Additionally, an indicator variable for SIPP survey allows for differences in the overall hazard that can be expected given the different ways that work injuries are identified between the NSLY79 and SIPP as described above.

We test the reasonableness of the assumption that the two surveys are independent realizations of a common social process based on the statistical framework of model fitting (e.g., Weakliem 2004). If inclusion of the survey variables (indicating the observation comes from one survey and not the other) improves the model fit when added to the regression specification, this indicates that the observations from the surveys are not after all realizations of a common social process. This test may be relaxed to allow for an intercept shift in which the relationships between the covariates and the outcome variable are equal across the surveys, but the overall level of the outcome variable is allowed to differ across the surveys (as in Tighe et al 2010). Accordingly we fit three sets of models corresponding to equations (1a), (1b), and (1c) of Section II above. The first model has no “survey” covariate; the second has a “survey” main effect only for presence in the SIPP data source; and the third has a “survey” main effect plus full interactions between the SIPP “survey” and the model covariates. We conduct this model-fitting diagnostics exercise for an analysis model that can be estimated for both surveys with no imputation. This is simply equation (3) but dropping the *Dislike* variable. We use the QIC model-fitting criterion of Pan (2001) that is designed for the GEE. It is a penalized quasi-likelihood criterion adapted from and analogous to the Akaike Information Criterion (AIC).

Results

Descriptive statistics for employed individuals in the NLSY79 and SIPP data are presented in Table 1. With the exception of the “Dislikes Job” variable that is entirely imputed for SIPP individuals, these are for observed data. Sociodemographic variables are age, gender, race/ethnicity (Hispanic, Black, and All Other), and education (any college years versus no college years). A total of 10,427 NLSY79 contributed any exposed periods (of 8 months length each) between the ages of 24 and 43, and these individuals contributed a total of 87,733 periods of exposure to disabling work injury. Pooling across the SIPP panels yields 113,160 8-month periods of exposure. The sample-weighted rate of injury (proportion of 8-month person-exposure intervals) is higher in the NLSY (0.25) than in the SIPP (0.20), a difference that is significant at the 0.05 level. This difference is substantively notable, indicating a 25% higher injury rate in the NLSY than in the SIPP.

The NLSY's greater oversampling of the lower-income population, who are at greater risk of occupational injury, additionally increases the observed (unweighted) number of injuries in that sample relative to the pooled SIPP panels sample. The weighted and unweighted injury characteristics of the NLSY relative to the SIPP together account for the slightly higher number of injuries in the NLSY sample (261) than in the larger SIPP sample (234).

[TABLE 1 ABOUT HERE]

The weighted distributions of sociodemographic characteristics are mostly quite similar between the NLSY and SIPP data. With the large sample sizes in each survey, however, statistically significant differences between the surveys are found for all variables. Age, gender, and occupation distributions are substantively similar. The SIPP's proportion of Hispanics is greater than in the NLSY, following the more recent sampling of the U.S. population in the SIPP (1986- to 2004-origin panels) than in the NLSY (1979-origin panel). Educational attainment is substantially greater in the SIPP. These substantively noteworthy differences across surveys in the regressor distributions provide additional reasons to use a statistical test of comparison. The substantively small differences that were seen to nevertheless be statistically significant makes it important that the statistical test of comparison include a penalty function, as we propose in our formal model-fitting approach.

[TABLE 2 ABOUT HERE]

The regressions of occupational injury on variables observed in both surveys are conducted for the two surveys separately and for the combined (pooled) samples (see Table 2, columns (1), (2), and (3)). The common variables consist of the sociodemographic variables of age, gender, education, and race/ethnicity, and additionally a single work-context variable for occupation group. The value of pooling across the surveys is already seen here. Standard errors about the coefficients are mostly similar between the NLSY and SIPP, the exception being the lower standard errors for black and Hispanic coefficients in the NLSY following the oversampling of those cases. The standard errors in the NLYS sample, however, are mostly around 1.3 times those of the pooled sample regression (see the ‘Ratio of S.E.s’ column (4)). This results in significant coefficients at the .05 level for Black and Hispanic workers (more likely to experience disabling injuries) and for college-educated workers (less likely to experience disabling injuries). In the NLSY, but not the SIPP, injuries become less likely with age between 24 and 43. This may be in part due to greater work experience with increasing age, a factor that was found by Oh and Shin (2003) to reduce injury. As expected, the occupational categories ‘manual’ and ‘service’ are found to elevate the risk of occupational injury compared with the reference category of other (‘white collar’) occupations.

The model fit statistic, the QIC, is slightly lower when adding the SIPP intercept (down from 6,718.5 to 6,715.9), but the QIC is not lowered further by adding interactions with the SIPP (6,716.4). The full set of interaction coefficients is given in Appendix Table A1. Of these interaction coefficients, only that for *Age*SIPP* is individually statistically significant. The assumption that the two surveys are realizations of the same social process may therefore be concluded to be reasonable, with only an intercept shift needed to account for survey differences.

Models including the subjective “Dislikes Job” variable are estimated for the NLSY data only and for the ‘completed’ pooled data consisting of the NLSY and the SIPP observations with multiply-imputed values of “Dislikes Job” (Table 2, Columns (5) and (6)). Recall that “Dislikes Job” is observed only in the NLSY. The coefficient for it in the NLSY model is statistically significant at the 0.01 level.

Including it substantially improves model fit over the NLSY-only model that excludes it (Column (1)): the QIC statistic falls from 3,472.4 to 3,443.8 when including “Dislikes Job.” Substantively, disliking one’s job is associated with a rise in risk of a disabling occupational injury, and this rise in risk is of a magnitude as large as that for working in one of the objectively higher-risk occupation categories (manual or service). A measure of the success of the imputation of the “Dislikes Job” variable to every observation in the SIPP is seen in the very similar magnitude of the coefficient between the observed and pooled models (Columns (5) and (6)). A similar proportion disliking their job was also found between the NLSY and the multiply-imputed SIPP dataset (Table 1).

Large efficiency gains in all the coefficients for variables observed in both the NLSY and SIPP are seen. The ratios of the standard errors in the NLSY-only estimation to those in the pooled estimation (see the ‘Ratio of S.E.s’ Column (7)) are similar to those for the NLSY versus pooled estimation of the model using common variables only (Column (4)). This indicates that the multiple correlations between the imputed variable and the observed-in-common variables have little effect on the variance reduction for the coefficients on the observed-in-common variables (see again Section II above). The standard error for the *Dislike* coefficient, however, is unchanged between the NLSY-only model and the pooled NLSY+SIPP model (S.E. ratio of 0.97). This result is consistent with the theoretical expectations and previous empirical findings described in Section II, and is intuitively unsurprising given that all the values of the *Dislike* variable in the SIPP are imputed. This, and the lack of any substantive change to the coefficient itself, shows that the multiple-imputation process succeeded in leaving the estimation of the parameter and standard error for the non-common variable unaffected by pooling the two surveys, while increasing substantially the efficiency of estimation of the parameters for the common variables. Thus a fuller regression model specified from the smaller NLSY survey full model was able to be estimated from 200,893 periods of exposure, yielding 495 injuries, by including observations from the larger SIPP survey. Job variables, both objective and subjective, are seen to be the strongest determinants of experiencing disabling work injuries. Sociodemographic variables, however, are also important. Statistically significant positive associations of Black and Hispanic and significant negative associations of having more education remain in the model that includes the observed job satisfaction variable in the NLSY and the imputed job satisfaction variable in the SIPP.

IV. DISCUSSION

Within-survey MI has become a common and widely-accepted practice for improving over complete-case analysis in sociology (e.g., Downey, von Hippel, and Broh 2004). Although it has been 25 years since Rubin (1986) proposed a form of cross-survey MI similar to that demonstrated here, we know of no successful adoptions of the method in sociology, and no more than occasional, experimental adoptions in the health and social sciences in general (e.g., Gelman et al 1998a; Schenker et al 2010).

The benefits of cross-survey MI, meanwhile, are potentially large. We demonstrated this in an empirical example in which we combined data from the National Longitudinal Survey of Youth (NLSY79) and the Survey of Income and Program Participation (SIPP) to analyze the sociodemographic and job-context associations with experiencing a serious occupational injury. We achieved large efficiency gains by pooling the complete cases of the NLSY with the incomplete cases of the SIPP after “completing” the latter by multiple imputation from an imputation model estimated with the NLSY cases. This allowed us to draw stronger conclusions about sociodemographic differences in risks of workplace injury. The magnitudes of the sociodemographic coefficients were themselves little changed after imputation, indicating that no apparent bias was introduced through the imputation process. The cross-survey multiple imputation strategy allowed us to introduce a larger set of regressors to include both objective (in the SIPP and NLSY) and subjective work characteristics (in the NLSY only). This fuller specification was possible even while retaining the major sample-size advantages of pooling observations across the surveys.

There are alternatives to cross-survey MI for combined-survey analysis. These are direct methods that involve forming a likelihood involving both complete and incomplete cases (see examples in Ridder and Moffitt 2007), and meta-analysis that combines a limited set of moments without incorporating the full covariance structure of each data source (e.g., Rao et al 2008). The computational demands on researchers of direct methods are much greater, requiring code to be built that is specific to each application (Schafer and Graham 2002). The combination of survey coefficients in meta-analysis involves implicit or explicit assumptions that differences in model specifications across studies can be ignored, whereas cross-survey MI accounts for differences in available variables across the surveys, imputing to allow for identical specifications once the surveys are pooled.

In thinking about what have been and continue to be barriers to the adoption of cross-survey MI, we consider the two largest of them to be (1) overcoming the problem of variables never jointly observed and (2) accounting for differences in survey sampling and measurement characteristics across the two (or more) surveys being pooled. The first of these problems is the easier one to solve, by simply specifying an analysis model that can be estimated with one of the surveys alone. This model is, in the best case, no worse than the model that may be specified in a regular analysis in which the researcher first chooses the best available data source and proceeds to estimate a model with that data source. With cross-survey analysis, this best case can be achieved when the outcome variable is measured similarly, and covering the same population, in the two surveys. In our example, the SIPP's population age and period coverage was broader than that for the NLSY, so that the NLSY population (defined by ages 25 to 43 over the period 1988 to 2000) was completely encompassed by the SIPP observations. This is likely to be the case often when combining a smaller-scale survey with a more general, larger survey.

The occupational-injury outcome in our application, however, was defined more restrictively in the SIPP to include only those occupational injuries resulting in a work limitation observed up to 8 months later. The consequences were first that only an analysis limited to "serious" occupational injuries could be conducted using the cross-survey MI analysis method, and second that the two surveys could not automatically be assumed to be measuring the same phenomenon. This is just one example of the kind of potential non-comparability problems that can occur when conducting cross-survey MI analyses. Because of the high likelihood that any two surveys of a given population will have differences in survey instruments, sampling schemes, and survey operations that could affect the character of responses, it is crucial to have a satisfactory method for evaluating the importance of such differences for the analysis of interest to the researcher. We argue here that the model-fitting diagnostics approach (Weakliem 2004) is an appropriate method for this purpose. The modeling-fitting approach penalizes the adding of complexity to a model specification, requiring that additional variables be sufficiently informative about the social process to justify their inclusion in the model. If differences in the two surveys do not cross this threshold of being sufficiently informative, we argue that this indicates that the two surveys can be reasonably assumed to sample the from same social process. This is consistent with the principles of model fit statistics, in which there is no assumption of one or other model being 'correct,' as there is in a strict hypothesis-testing framework.

We additionally argue that the model-fitting approach is best conducted in two steps: first adding only an intercept term for 'survey'; and second adding a full set of interactions of the covariates with 'survey.' If the model that includes the 'survey' intercept only improves in a model-fit sense over the model estimated with the pooled data but without the intercept term, this implies a difference in the overall level of the outcome variable between the surveys but relationships between the covariates and the outcome variable that are equal across the surveys. This is the circumstance modeled by Tighe et al (2010). This is unlikely to be a problem for the typical social science analysis, in which the goal is to understand relationships between covariates and the outcome variable. The post-imputation analysis model using the pooled surveys may then be estimated with an indicator variable for survey.

In the case that the best-fitting model includes interactions with 'survey,' this casts doubt on the appropriateness of pooling observations across the two surveys. Doing so may result in biased parameter estimates unless additional information, sometimes in the form of restrictions on the analysis model or data used to estimate that model, are introduced. One form of restriction is to partition the population into subgroups for which comparability can be established. This is the approach taken by Schenker et al (2010). As we noted earlier, a disadvantage of this method that it results in smaller subsamples being used. They found this to be a potential problem in finding sufficient data with which to estimate full (i.e., “congenial”) imputation models. Simpler imputation models had to be specified, which the authors suggested may have been responsible for attenuation of the coefficients in the analysis model. Estimating simpler imputation models is said to introduce a form of ‘uncongeniality’ in the terminology of MI (Meng 1994). We suggested that uncongenial imputation be considered also as analogous to imputation in which variables not used in the imputation model estimated from Survey 1 were actually not present in Survey 1, thereby introducing the equivalent of a “variables never jointly observed” situation. Insufficient attention to the seriousness of estimation with variables never jointly observed is possibly the biggest barrier to credible cross-survey multiple imputation, a theme brought out in comments on the Gelman et al (1998a) study (e.g., Judkins 1998), and in the more general review of methods for combining data by Ridder and Moffitt (2007).

The reason for the dangers of resorting to estimation with variables never jointly observed also likely arises from the perceived rationale for combining data as it originated in the statistical matching literature, which was to add variables to data sources without those variables. Schenker et al’s (2010) analysis is a variant on this in which cross-survey MI is used to add better-measured versions of variables from a smaller data source, where those variables are already present in less well-measured form in the larger data source. Significantly, they do not use pooled estimation after imputation, but instead estimation only with the larger data source augmented by the better-measured variables imputed from the smaller data source. We suggest that one of the roles of a pooled analysis is to assure the integrity of the imputation modeling process, as complete cases (from the smaller data source) and completed cases (from the larger data source) should be statistically equivalent. Analysis with observations from the larger data source only may be seen as an extreme alternative to complete-case analysis in which the analysis is conducted with incomplete cases only (after "completing" those missing values by multiple imputation).

An alternative way of thinking about the benefits to cross-survey MI that we argue for here is as a method not for adding variables but instead for adding *observations*. This presumes the existence of a data source with all the needed variables but that suffers from sample-size limitations. We suggest that the way that social science research is conducted makes this situation of sample-size limitations the norm rather than the exception. Given the choice between a first data source containing only a subset of variables considered important to a substantive model but with large numbers of observations (in the limit, a census) and a second data source containing most or all of the variables considered important to the substantive model but with a smaller sample size, the social scientist will usually opt for the smaller survey. He or she may even initiate a new survey data collection to obtain variables missing in existing data sources. Given the high costs of data collection, compromises leading to sample size limitations are the norm rather than the exception in social science analyses involving well-specified models. The situation in which sample size limitations can be mitigated by the adding of "incomplete" observations from a larger survey, that may then be "completed" by cross-survey MI, may then also be viewed as the norm rather than the exception in social science analysis. The present paper offers guidance of how this mitigation can be achieved using statistical package software with flexible specifications of both the analysis and imputation models, and provides an example of the substantial benefits that may be expected to result.

REFERENCES

Allison, P.D. (2002) Missing Data Newbury Park: Sage publications.

Allison, P.D. (2009) "Imputation of categorical variables with PROC MI." Paper 113-30, SUGI 30 Focus Session.

Bureau of Labor Statistics (no date) National Longitudinal Surveys www.bls.gov/nls/

Burnham, K.P., and D.R. Anderson (2002) Model Selection and Multimodel Inference: A Practical Information-theoretic Approach New York: Springer.

D'Orazio, M.B., M. Di Zio, and M. Scanu (2006) Statistical matching for categorical data: Displaying uncertainty and using logical constraints Journal of Official Statistics 22(1):137-157.

DeLeire, T. (2000) The wage and employment effects of the Americans with Disabilities Act Journal of Human Resources 35(4):693-715.

DeLeire, T. (2001) Changes in wage discrimination against people with disabilities: 1984-93 Journal of Human Resources 36(1):144-158.

Dembe, A.E., J.B. Erickson, and R. Delbos (2004) Predictors of work-related injuries and illnesses: National survey findings Journal of Occupational and Environmental Hygiene 1:542-550.

Dong, X.W. (2005) Long work hours, work scheduling, and work-related injuries among construction workers in the United States Scandanavian Journal of Work Environment and Health 31(5):329-335.

Downey, D.B., P.T. von Hippel, and B.A. Broh (2004) Are schools the great equalizer? Cognitive inequality during the summer months and the school year American Sociological Review 69(5):613-635.

Freedman, V., and D.A. Wolf (1995) A case study on the use of multiple imputation Demography 32:459-470.

Gelman, A., G. King, and C. Liu (1998a) Not asked and not answered: Multiple imputation for multiple surveys Journal of the American Statistical Association 94(443):846-857.

Gelman, A., G. King, and C. Liu (1998b) Rejoinder Journal of the American Statistical Association 94(443):869-874.

Goldscheider, F., C. Goldscheider, P. St. Clair, and J. Hodges (1999) Changes in returning home in the United States, 1925-1985 Social Forces 78(2):695-720.

Handcock, M.S., M.S. Rendall, and J.E. Cheadle (2005) Improved regression estimation of a multivariate relationship with population data on the bivariate relationship Sociological Methodology 35:291-334.

Hellerstein, J., and G.W. Imbens (1999) Imposing moment restrictions from auxiliary data by weighting Review of Economics and Statistics 81(1):1-14.

Imbens, G.W. and T. Lancaster (1994) Combining micro and macro data in microeconometric models Review of Economic Studies 61:655-680.

Judkins, D.R. (1998) Not asked and not answered: Multiple imputation for multiple surveys: Comment Journal of the American Statistical Association 94(443):861-864.

Lee. K.J., and J.B. Carlin (2010) Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation American Journal of Epidemiology 171:624-632.

Liang, K.Y., and S.L. Zeger (1986) Longitudinal data analysis using generalized linear models Biometrika 73:13-22.

Little, R.J.A. (1992) Regression with missing X's: A review Journal of the American Statistical Association 87(420):1227-1237.

Little, R.J.A., and D.B. Rubin (1989) The analysis of social science data with missing values Sociological Methods and Research 18:292-326.

Little, R.J.A., and D.B. Rubin (2002) Statistical Analysis with Missing Data (2nd Edition) Hoboken, NJ: Wiley.

Meng, X.L. (1994) Multiple-imputation inferences with uncongenial sources of input Statistical Science 9(4):538-573.

Moriarity, C., and F. Scheuren (2003) A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputation Journal of Business and Economic Statistics 21(1):65-73.

Oh, J.H., and E.H. Shin (2003) Inequalities in nonfatal work injury: The significance of race, human capital, and occupations Social Science and Medicine 57:2173-82.

Pan, W. (2001) Akaike's Information Criterion in generalized estimation equations Biometrics 57(1):120-125.

Raghunathan, T.E. (2004) What do we do with missing data? Some options for analysis of incomplete data Annual Review of Public Health 25:99-117.

Raghunathan, T.E., and J.E. Grizzle (1995) A split questionnaire survey design Journal of the American Statistical Association 94(447):896-908.

Raghunathan, T.E., J.M. Lepkowski, J. van Hoewyk, and P. Solenberger (2001) A multivariate technique for multiply imputing missing values using a sequence of regression models Survey Methodology 27(1):85-95.

Raghunathan, T.E., P. Solenberger, and J. van Hoewyk (2000) IVEware: Imputation and Variance Estimation Software http://www.isr.umich.edu/src/smp/ive/

Rao, S.R., B.I. Granbard, C.H. Schmid, S.C. Morton, T.A. Louis, A.M. Zaslavsky, and D.M. Finkelstein (2008) Meta-analysis of survey data: Application to health services research Health Services Outcomes Research Methods 8:98-114.

Rässler, S. (2002) Statistical matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches New York: Springer Verlag.

Rendall, M.S., R. Admiraal, A. DeRose, P. DiGiulio, M.S. Handcock, and F. Racioppi (2008) Population constraints on pooled surveys in demographic hazard modeling Statistical Methods and Applications 17(4):519-539.

Rendall, M.S., M.S. Handcock, and S.H. Jonsson (2009) Bayesian estimation of Hispanic fertility hazards from survey and population data Demography 46(1):65-84.

Ridder, G., and R.A. Moffitt (2007) "The econometrics of data combination," pp.5469-5547 in J.J. Heckman and E.E. Leamer (Eds.) Handbook of Econometrics Vol.6b. Amersterdam: North Holland.

Rodgers, W.L. (1984) An evaluation of statistical matching Journal of Business & Economic Statistics 2(1):91-102.

Rubin, D.B. (1986) Statistical matching using file concatenation with adjusted weights and multiple imputations Journal of Business and Economic Statistics 21(1):65-73.

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

SAS Institute (no date) SAS 9.2 Product Documentation http://support.sas.com/documentation/92/index.html.

Sassler S., and J. McNally (2003) Cohabiting couples' economic circumstances and union transitions: a re-examination using multiple imputation techniques Social Science Research 32(4):553-578.

Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data Boca Raton, FL: Chapman and Hall.

Schafer, J.L., and J.W. Graham (2002) Missing data: Our view of the state of the art Psychological Methods 7(2):147-177.

Schenker, N., T.E. Raghunathan, and I. Bondarenko (2010) Improving on analysis of self-reported data in a large-scale health survey by using information from an examination-based survey Statistics in Medicine 29:553-545.

Singh, A.C., H.J. Mantel, M.D. Kinack, and G. Rowe (1993) Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption Survey Methodology 19(1):59-79.

Tighe E., D. Livert, M. Barnett, and L. Saxe (2010) Cross-survey analysis to estimate low-incidence religious groups Sociological Methods and Research 39(1):56-82.

U.S. Census Bureau (no date) Survey of Income and Program Participation Washington, DC: US Census Bureau. www.census.gov/sipp/

Weakliem, W.L. (2004) Introduction to special issue on model selection Sociological Methods and Research 33(2):167-187.

White, I.R., and J.B. Carlin (2010) Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values Statistics in Medicine 29:2920-31.

Table 1 Sample Description, NLSY79 and SIPP ages 24 to 43
(percentage of weighted sample unless otherwise stated)

		observed	
	observed	plus imputed	~
	NLSY79	SIPP	
Injury^	0.25	0.20	*
Age (mean)	33.2	33.7	**
Female	45.6	47.0	*
Any College Education	51.3	58.0	**
Race/Ethnicity			**
Black	13.1	10.7	
Hispanic	7.2	9.8	
All Other	79.7	79.5	
Occupation			**
Manual	29.3	28.2	
Service	12.5	11.8	
All Other	58.2	60.0	
Dislikes Job~	8.5	8.4	
Year (mean)	1993.8	1994.8	**
Sample N (periods of injury exposure)	87,733	113,160	
Number of sample individuals	10,427	113,160	
Sample number of injuries	261	234	

Notes:
^ A work injury occurring in an 8 month period that results in a work limitation

Occupation and whether dislikes job in the NLSY79 are percentages of person-periods of exposure.

~ Imputed "Dislikes Job" variable for the SIPP

* $p < 0.05$, ** $p < 0.01$ in Chi-square test for difference between NLSY79 and SIPP

NLSY79 = National Longitudinal Survey of Youth, 1979 Cohort, 1988 to 2000 years
SIPP = Survey of Income Participation, 1986 to 2004 Panels

Table 2 Logistic regression of injury on sociodemographic and work characteristics in single-survey and pooled-survey samples from the NLSY79 and SIPP ages 24 to 43

	(1) observed			(2) observed			(3) observed pooled			(4) Ratio of S.E.s:
	NLSY79			SIPP			NLSY & SIPP			
	Estimate	S.E.	p value	Estimate	S.E.	p value	Estimate	S.E.	p value	(1)/(3)
Intercept	-6.13	0.223	0.000	-6.88	0.246	0.000	-6.55	0.162	0.000	1.37
Age	-0.04	0.014	0.012	0.02	0.011	0.085	0.00	0.008	0.623	1.72
Female	-0.03	0.137	0.824	0.13	0.156	0.413	0.04	0.103	0.687	1.34
Black	0.26	0.173	0.131	0.26	0.186	0.163	0.28	0.126	0.024	1.38
Hispanic	0.31	0.146	0.034	-0.04	0.217	0.851	0.26	0.115	0.023	1.28
Any College Education	-0.29	0.153	0.060	-0.56	0.170	0.001	-0.45	0.113	0.000	1.35
Manual Occupation	1.11	0.169	0.000	1.30	0.205	0.000	1.20	0.132	0.000	1.29
Service Occupation	0.85	0.195	0.000	1.01	0.215	0.000	0.92	0.145	0.000	1.35
Dislikes Job	-	-	-	-	-	-	-	-	-	-
SIPP	-	-	-	-	-	-	-	-	-	-
QIC	3,472.4			3,244.0			6,718.5			
QIC including SIPP intercept							6,715.9			
QIC including SIPP intercept and full interactions							6,716.4			
Sample N (periods of injury exposure)	87,733			113,160			200,893			
Number of sample individuals	10,427			113,160			123,587			
Sample number of injuries	261			234			495			

Notes: NLSY79 = National Longitudinal Survey of Youth, 1979 Cohort, 1988 to 2000 years
SIPP = Survey of Income Participation, 1986 to 2004 Panels
QIC = quasi-likelihood model fit statistic (Pan 2001)
Regressions are unweighted
Periods of injury exposure are 8 months long

Table 2 Logistic regression of injury on sociodemographic and work characteristics in single-survey and pooled-survey samples from the NLSY79 and SIPP ages 24 to 43 (Continued page)

	(5) observed NLSY79			(6) pooled observed NLSY and imputed SIPP			(7) Ratio of S.E.s:
	Estimate	S.E.	p value	Estimate	S.E.	p value	(5)/(6)
Intercept	-6.25	0.218	0.000	-6.55	0.164	0.000	1.33
Age	-0.03	0.014	0.022	0.00	0.009	0.988	1.59
Female	-0.06	0.142	0.649	0.02	0.103	0.883	1.38
Black	0.27	0.164	0.103	0.26	0.121	0.032	1.36
Hispanic	0.28	0.142	0.048	0.17	0.116	0.148	1.23
Any College Education	-0.30	0.146	0.039	-0.44	0.105	0.000	1.38
Manual Occupation	1.06	0.170	0.000	1.16	0.124	0.000	1.37
Service Occupation	0.83	0.192	0.000	0.90	0.141	0.000	1.36
Dislikes Job	0.93	0.153	0.000	0.92	0.158	0.000	0.97
SIPP				-0.20	0.094	0.031	-
QIC	3,443.8						
QIC including SIPP intercept							
QIC including SIPP intercept and full interactions							
Sample N (periods of injury exposure)	87,733			200,893			
Number of sample individuals	10,427			123,587			
Sample number of injuries	261			495			

Notes: NLSY79 = National Longitudinal Survey of Youth, 1979 Cohort, 1988 to 2000 years
SIPP = Survey of Income Participation, 1986 to 2004 Panels
QIC = quasi-likelihood model fit statistic (Pan 2001)
Regressions are unweighted
Periods of injury exposure are 8 months long

Table A1 Logistic regression of injury,SIPP only model including imputed 'Dislikes Job' and model of observed NLSY+SIPP common variables with full SIPP interactions

		imputed SIPP only			observed NLSY+SIPP	
	Estimate	S.E.	p-value	Estimate	S.E.	p-value
Intercept	-7.00	0.231	0.000	-6.13	0.223	0.000
Age	0.02	0.012	0.066	-0.04	0.014	0.012
Female	0.10	0.149	0.488	-0.03	0.137	0.824
Black	0.27	0.180	0.127	0.26	0.173	0.131
Hispanic	-0.07	0.217	0.745	0.31	0.146	0.034
Any College Education	-0.57	0.152	0.000	-0.29	0.153	0.060
Manual	1.26	0.181	0.000	1.11	0.169	0.000
Service	0.98	0.208	0.000	0.85	0.195	0.000
Dislikes Job	0.90	0.297	0.004	-	-	-
SIPP				-0.75	0.332	0.025
SIPP*Age				0.05	0.018	0.002
SIPP*Female				0.16	0.208	0.446
SIPP*Black				0.00	0.254	0.996
SIPP*Hispanic				-0.35	0.262	0.179
SIPP*Any College Education				-0.27	0.229	0.231
SIPP*Manual				0.19	0.266	0.483
SIPP*Service				0.15	0.290	0.602
QIC	-			6,716.4		
Sample N (injury exposure periods)	113,160			200,893		
Number of sample individuals	113,160			123,587		
Sample number of injuries	234			495		

Notes: NLSY79 = National Longitudinal Survey of Youth, 1979 Cohort, 1988 to 2000 years
SIPP = Survey of Income Participation, 1986 to 2004 Panels