Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2014, Article ID 368791, 9 pages
http://dx.doi.org/10.1155/2014/368791

Research Article

Working with Missing Data: Imputation of Nonresponse Items in Categorical Survey Data with a Non-Monotone Missing Pattern

Machelle D. Wilson[1] and Kerstin Lueck[2,3]

[1]*Department of Public Health Sciences, Division of Biostatistics, University of California, Davis, Davis, CA 95616, USA*
[2]*Social Psychology, The University of Adelaide, Adelaide, SA 5005, Australia*
[3]*Department of Integration and Conflict, Max Planck Institute, 06017 Halle, Germany*

Correspondence should be addressed to Machelle D. Wilson; mdwilson@phs.ucdavis.edu

Received 9 June 2014; Accepted 16 October 2014; Published 7 December 2014

Academic Editor: Jin Liang

The imputation of missing data is often a crucial step in the analysis of survey data. This study reviews typical problems with missing data and discusses a method for the imputation of missing survey data with a large number of categorical variables which do not have a monotone missing pattern. We develop a method for constructing a monotone missing pattern that allows for imputation of categorical data in data sets with a large number of variables using a model-based MCMC approach. We report the results of imputing the missing data from a case study, using educational, sociopsychological, and socioeconomic data from the National Latino and Asian American Study (NLAAS). We report the results of multiply imputed data on a substantive logistic regression analysis predicting socioeconomic success from several educational, sociopsychological, and familial variables. We compare the results of conducting inference using a single imputed data set to those using a combined test over several imputations. Findings indicate that, for all variables in the model, all of the single tests were consistent with the combined test.

1. Introduction

The problem of bias due to missing data has received a good deal of attention over the last 20 years and the correction of bias due to item and unit nonresponse remains an important problem for investigators using survey data [1–9]. For data missing because of item nonresponse, imputation of the missing data is often the best solution. However, methods for imputing categorical data are still experimental in some software releases. Many software packages will automatically remove cases with missing values from the analysis, greatly reducing the sample size, often causing a drastic loss of information. Additionally, if the data are not *missing completely at random*, removing cases with missing items will result in biased parameter estimates in subsequent analyses.

Durrant [10] conducted an extensive review of various imputation methods. She showed that parameter estimates can vary considerably with different methods and noted their advantages and disadvantages. She noted that with the regression model approach problems arise with failures in model assumptions, but in the case where model assumptions hold the modelling approach works well. Regression modelling is superior to mean imputation or similar methods in that it makes use of the information in the entire sample to impute the missing value for each observation. However, as with mean imputation, regression imputation leads to underestimation of the variability in the data, because predicted values from the regression model are treated in the substantive analysis as if they were random observations from the sample population, leading to biased estimates of the population variance and subsequently the standard errors used to conduct inference about the model parameters. There does not appear to be a consensus as to the best method, as much depends on the nature of the data and

the missing data process. However, in many situations, the use of model-based approaches such as Markov chain Monte Carlo methods (MCMC) is superior. These methods define a model or distribution for the missing data (the missing data model) and sample from this distribution to impute the missing values [9], hence simulating or mimicking random sampling from the population of interest. These methods have been shown via theory and simulation to converge to the true target distribution. Sampling from the distribution of the missing variable reduces the underestimation of the population variance, as well as the standard errors of the parameter estimates for the substantive model, which would result from most other single-imputation methods such as mean imputation or regression imputation alone. Often the model-based approaches are combined with regression modelling to perform the imputation so as to also more fully exploit the information in the data. For categorical data, logistic regression is a natural choice and has the advantage of accurately modelling the distribution of the missing data given the observed data. The parameters are easily estimated via the incomplete observed data [11, 12]. However, the model-based approaches in conjunction with logistic regression become problematic for data sets with a large number of variables. This is because the relationships between the variables are modelled via cross-tabulation and the size of the contingency table grows exponentially with additional variables, often creating a situation which exceeds the limitations of the software package [13]. The limitation can be circumvented if the missing data pattern is monotone. However, assessing the existence of a monotone missing pattern is equally problematic for a large number of variables.

Furthermore, for the typical researcher in education and the social sciences, ease of implementation of the imputation so as to devote time and energy to the substantive model is of more interest than programming advanced statistical algorithms. Hence, our goal here is to provide a simple method for exploiting the modelling capabilities of SAS or other software packages and circumventing the difficulties for the case of a large number of categorical variables with a nonmonotone missing data pattern.

In this paper we will review the degrees of randomness and the implications for imputation. We will discuss MCMC implementation of multivariate normal data and MCMC combined with logistic regression for imputation of categorical data, the monotone missing requirement, creation of a monotone missing pattern, and how to perform multiple imputations using this method. Finally, we will apply the method to data drawn from the National Latino and Asian American Study (NLAAS) and show the use of multiple imputations for an example substantive model using these data.

1.1. Degrees of Randomness in Missing Data. A missing data process is said to be *missing completely at random* (*MCAR*) if the probability that a subject is missing a variable is completely independent of the value of the variable and of the values of any other variables. For example, missing data from lost survey pages are *MCAR* because, presumably, the probability that a page was lost is not in any way related to the value of any of the variables measured on that page, nor to any other possible variables related to the missing data. To restate, a missing variable is said to be *MCAR* if the probability that the variable is missing from a subject is neither related to the value of the missing variable nor to the value of any other variable for that subject.

A missing variable is said to be *missing at random* (*MAR*) when the probability that a variable is missing from a given subject depends on the value of another variable for that subject but not on the value of the missing variable itself. For example, if we have a variable *income* that is more likely to be missing from respondents with higher levels of education but is no more likely to be missing for higher or lower incomes, then both the missing and the observed income values in a survey are a random sample of the population of *income* at a given level of *education* but are not a random sample of the population of all possible *income*. That is, the conditional distribution of *income* given *education* is unbiased and is representative of the distribution of *income* for any given level of *education*. Hence, the income variables are *missing at random* given the education level. However, the sample of *income* may be biased for the unconditional distribution of *income* because any relationship between *income* and *education* will cause bias in the observed sample. For example, if higher education is correlated with higher income, then the sample mean of all incomes in the data set will be biased downwards because the higher incomes associated with higher education were more likely to be missing.

If the probability that a particular question is not answered is dependent on the answer itself, then the missing data process is *nonrandom* and the resulting bias in the parameter estimates cannot be corrected without information from outside the sample. For example, if low-income respondents are more likely to refuse to answer a question about their income level, then the estimates for *income* will be biased, since lower levels of income were more likely to be excluded. If we have other sources of information about income, we may be able to correct this, but the sample itself is biased and by itself will produce biased parameter estimates. Similarly, the model parameter estimates containing statistics based on the sample values for *income* may also be biased. This issue is similar to bias that can result from unit nonresponse to surveys and similar corrective measures may be possible. In any case, certainly, avoidance of both unit and item nonresponse is the best solution, when possible [6–8, 14].

For most survey data, including the National Latino and Asian American Study, we cannot assume that the missing data are *MCAR*. Some respondents may be more likely to refuse to answer certain questions depending on their understanding of the question, their education level, their cultural identity, or other characteristics. However, it can usually be argued for surveys with a large number of variables that the missing data in the survey can be assumed to be *MAR* because we have a large number of variables with which to model the missing data process. That is, the larger the number of variables we have, the more likely that there is a variable (or a combination of variables) in our data set for

which the conditional distribution of the missing variable is unbiased [11, 15, 16].

It should be noted here that if the number of continuous variables in the data set is small, we are more likely to encounter problems with the MAR assumption. Under the MAR assumption, unbiased estimates of the missing data values can be obtained by conditioning the imputed value on the observed variables that model the missing data process. Imputation of missing data using small data sets can increase the risk of violating the MAR assumption in that the missingness may depend on a variable we did not include in the imputation model. If auxiliary variables (variables not intended for the substantive model, but may be correlated with those that are) can be collected, it is expedient to do so [17, 18]. Care should be taken and expert knowledge employed to consider possible relationships between variables and the probability of missingness when building the imputation model.

1.2. Imputation of Missing Data Using Bayesian Methods. One useful approach to imputation is to use a Bayesian model-based method. In this method, a posterior distribution for the parameters of the missing data distribution given the observed data is obtained using Bayes' Rule. The *posterior distribution* is based on the maximum likelihood estimates for the population parameters for the data and a prior distribution for the parameters. The prior distribution for the parameters models our uncertainty about the true parameter values. In general, the posterior distribution of the parameter vector, θ, of the distribution of a given random variable, Y, is expressed as

$$\Pr\{\theta \mid Y\} = \frac{L\{Y \mid \theta\}\, p(\theta)}{\sum_{\theta} L\{Y \mid \theta\}\, p(\theta)_j}, \tag{1}$$

where $p(\theta)$ is the prior distribution of θ, $L\{Y \mid \theta\}$ is the likelihood function for the data, and the denominator is the sum (or integral for continuous sets) over all possible values of θ. The denominator is a normalizing constant (once the data have been observed) that ensures that the posterior distribution is a valid probability distribution [19]. In the missing data context, once we have obtained the posterior distribution of the parameters of the missing data distribution, we sample the missing data population parameter values from their posterior distribution using simulation, and then we impute the missing data by sampling via simulation of the missing data values from their distribution given the previously sampled parameter values and the observed data. Finding the posterior distribution often involves the use of MCMC methods, most often the Gibbs sampler. Note that, in the Bayesian framework, the parameters are random variables and come from a probability distribution. This probability distribution is meant to model our uncertainty in the model parameter values. In practice, if we have no prior information about the distribution of the parameters, we can specify a non-informative prior. With non-informative prior, the Bayesian point estimate for the parameters will be equal to the maximum likelihood estimator in most situations. The benefit of the Bayesian approach, especially for the imputation of missing values, is that, unlike with regression imputation, we are not limited to point estimation in building the imputation model but can model our uncertainty in the parameter estimates for each imputed variable by sampling them from their posterior distribution for each imputation and thus each imputation uses a different parameter estimate in the imputation model to impute the missing value, thus more realistically incorporating variability due to uncertainty in the parameter estimates into the imputed data set.

In non-model-based approaches, such as mean imputation, hot-deck imputation, and regression imputation, the variability in the imputed data will be less than the variability in the population. That is, variance estimates based on the complete data (the observed data with imputed values replacing the missing data) will be biased downwards because the imputed data does not contain information about uncertainty in the parameter estimates used in the imputation [11, 20–22]. Likewise, standard errors based on the variance estimates will be biased downwards, possibly affecting inference in the substantive model. Hence, with the Bayesian approach, the downward bias of the population variance estimates for the complete data set is reduced because we are modelling the uncertainty in the imputation model parameters.

1.3. Imputation of Multivariate Normal Data. In the method we discuss here we need to first impute any multivariate normal data that we may have in our survey before imputing the categorical data. Having at least a few complete variables (either observed or imputed) will help us in establishing a monotone missing pattern for the categorical data. Some software packages, such as SAS, can perform model-based imputations, such as the Bayesian method described above, within a canned procedure so that the investigator does not need to have mastered advanced statistical computing.

In the context of missing multivariate normal data we aim to sample from the posterior distribution, $p(\theta \mid Y_{\text{obs}}, Y_{\text{mis}})$, to obtain estimates for the population parameters, the mean vector, $\boldsymbol{\mu}$, and the covariance matrix $\boldsymbol{\Sigma}$. Once we have population parameter estimates we sample from the distribution of the missing data, given the parameters and the observed data, $p(Y_{\text{mis}} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, Y_{\text{obs}})$, in order to impute the missing data. These methods produce a Markov chain whose stationary distribution is the target distribution. For data from an approximately multivariate normal distribution, the imputation process involves the following steps.

(1) *The Imputation Step*. At step t, given the current estimates for the mean vector, $\boldsymbol{\mu}^{\mathbf{t}} = (\mu_1, \mu_2, \mu_3, \ldots, \mu_p)$, and covariance matrix, $\boldsymbol{\Sigma}^{\mathbf{t}}$, the I-step simulates the missing values for each missing observation independently. That is, if the variables with missing values for the ith observation are denoted by $Y_{i(\text{mis})}$ and the variables with observed values by $Y_{i(\text{obs})}$, then the I-step draws independent values for $Y_{i(\text{mis})}$ from the current conditional distribution for $Y_{i(\text{mis})}$ given $Y_{i(\text{obs})}$ and $\boldsymbol{\theta}^{\mathbf{t}} = (\boldsymbol{\mu}^{\mathbf{t}}, \boldsymbol{\Sigma}^{\mathbf{t}})$.

(2) *The Posterior Step*. Given a complete sample, that is, with all missing values provisionally imputed,

the P-step simulates the mean vector and covariance matrix from their respective posterior distributions. These new estimates are then used in the next I-step, $t + 1$.

These two steps constitute a Markov chain whose equilibrium distribution is the true distribution of $Y_{i(\text{mis})}$ given $Y_{i(\text{obs})}$ and the parameter estimates that have been simulated from their respective posterior distributions. That is, with a current parameter estimate $\boldsymbol{\theta}^{\mathbf{t}}$ at the tth iteration, the I-step draws $Y_{\text{mis}}^{(t+1)}$ from $p(Y_{\text{mis}} \mid Y_{\text{obs}}, \boldsymbol{\theta}^{\mathbf{t}})$ and the P-step draws $\boldsymbol{\theta}^{\mathbf{t+1}}$ from $p(\boldsymbol{\theta} \mid Y_{\text{obs}}, Y_{\text{mis}}^{t+1})$. This produces a Markov chain $, \ldots, (Y_{\text{mis}}^{B+1}, \boldsymbol{\theta}^{\mathbf{B+1}}), (Y_{\text{mis}}^{B+2}, \boldsymbol{\theta}^{\mathbf{B+2}}), \ldots$, where B is a large number such that the chain has converged to the target distribution $p(Y_{\text{mis}} \mid Y_{\text{obs}}, \boldsymbol{\theta})$. Once the chain has converged, each simulation is an independent draw from this distribution, the values of which are then used to impute the missing data [11].

1.4. Imputation of Categorical Data. For ordinal or nominal data, we can use a logistic regression model to impute missing data once a monotone missing data pattern had been established and the posterior distribution of the parameters for the regression imputation model has been found. Once a model has been fitted, the missing values can be imputed using the predicted values from the model [4, 23].

For a missing binary class variable Y_j with possible outcomes 0 and 1, we fit a logistic regression model using the observed data for Y_j and its covariates $X_1, X_2, X_3, \ldots, X_k$ and a vector of β sampled from their posterior distribution. We have

$$\text{logit}\left(\mu_j\right) = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \cdots + \beta_k X_{kj}, \tag{2}$$

where $X_{1j}, X_{2j}, X_{3j}, \ldots, X_{kj}$ are the covariates for Y_j, where

$$\mu_j = \Pr\left(Y_j = 1 \mid X_{1j}, X_{2j}, \ldots, X_{kj}\right) \tag{3}$$

and where

$$\text{logit}\left(\mu_j\right) = \log\left(\frac{\mu_j}{1 - \mu_j}\right). \tag{4}$$

The imputed values are simulated algorithmically using the following steps.

(1) At step t, randomly draw a new parameter vector, $\boldsymbol{\beta}^t = (\beta_1, \beta_2, \ldots, \beta_k)$, from the current posterior predictive distribution, where $\boldsymbol{\beta}^t = \boldsymbol{\beta}^{\mathbf{t-1}} + \mathbf{V}'_{\mathbf{hj}}\mathbf{Z}$, where $\mathbf{V}'_{\mathbf{hj}}$ is the upper triangular matrix of the Cholesky decomposition and $\mathbf{Z}$ is a vector of $k + 1$ independent random normal variates. The posterior predictive distribution is updated at each step, t, given the observed data and the imputed data from the last step.

(2) For each observation with missing Y_j given covariates $X_{1j}, X_{2j}, X_{3j}, \ldots, X_{kj}$ and $\boldsymbol{\beta}^{\mathbf{t}}$ find the expected probability that $Y_j = 1$ given by p_j.

(3) Draw u, a uniform (0, 1) random variable. If $u < p_j$, impute $Y_j = 1$; else impute $Y_j = 0$.

The above algorithm produces a Markov chain whose stationary distribution converges to the true distribution of Y. The imputed Y_j are our best estimates of the true value of the missing variable for each observation given the observed data and the covariates. Furthermore, sampling from the posterior distribution of the parameter vector, $\boldsymbol{\beta}$, models our uncertainty in the imputation model parameters, thus providing a more realistic variability in the imputed data. This algorithm can be extended for ordinal or categorical variables with more than two categories.

1.5. The Monotone Missing Pattern Requirement. For the imputation of categorical data, if the missing data pattern is non-monotone, this can cause difficulties in the imputation in a variety of situations [12, 24–26]. A data set is said to have a monotone missing pattern when it is possible to arrange the variables in order such that if an individual is missing variable Y_j then that individual is also missing all subsequent variables $Y_k, k > j$. The data set below has a monotone missing pattern:

Obs	Y_1	Y_2	Y_3	Y_4
1	2.1	0	10.2	0.5
2	3.0	1	11.0	·
3	1.9	1	·	·
4	3.2	·	·	·

(5)

Because survey data are often categorical with a large number of variables, finding such an ordering of the variables may be prohibitively time consuming or impossible. However, if we can achieve a monotone missing data pattern, we can use the automated model-based capabilities of many software packages to impute our missing data and avoid many of the pitfalls of other types of imputation.

1.5.1. Creating a Monotone Missing Pattern. To construct a monotone missing data pattern, the first step is to use the model-based approach described above for multivariate normal data in the data set to simultaneously impute the missing data for the continuous variables. Once this step is completed, we can impute the incomplete categorical variables, one at a time, using the model-based approach for categorical data described above, which requires a monotone missing pattern for implementation. Once a variable is complete it can be used in the imputation of the next variable. Hence, at each step in the process only one variable is incomplete, creating by default a monotone missing data pattern. We recommend starting with the variable with the fewest missing values and ending with the variable with the most missing values. This procedure is repeated until all variables are complete.

1.6. A Word about Multiple Imputations. Some investigators argue that multiple imputation is necessary to obtain unbiased estimates of the standard errors and hence for conducting inference [1, 4, 5, 14, 27]. However, most multiple imputation procedures work in tandem with the procedure for the substantive analysis. For example, SAS's *Proc MI* works in tandem with *Proc MIanalyze*, which performs the substantive analysis after each of the multiple imputations. Because we must first build a monotone missing pattern,

we must first impute each missing variable for each case before building the substantive logistic regression model and we cannot exploit the "multiple" aspects of *Proc MI* and other similar software implementations. Furthermore, we must impute the normal variables using different methods than the categorical variables and hence these need to be imputed in a separate step. Furthermore, in our case study as well as in many studies involving ordinal data, we construct indices from item responses to measure constructs such as socioeconomic status, family cohesion, and language proficiency. We need to have complete data to build these indices and building constructs is not an imbedded part of any software implementation. Multiple imputation procedures work in conjunction with the substantive analysis by repeating the imputation several (up to 10 usually) times, each time estimating the parameters of the substantive model and their standard errors. Less biased estimates of the standard errors can then be obtained based on changes in the parameter estimates across different imputations. Inference about the parameters is then conducted using this improved standard error.

This cannot be implemented within the canned procedure if the data need to be imputed variable by variable. Hence confidence intervals and P values could retain some downward bias when performing single imputation even with the model-based approach.

We can, of course, perform our own procedure for the imputation and the data analysis several times and calculate the variance of the different parameter estimates of interest across different analyses with different imputations. Here we show an example of this procedure which involved imputing the data, calculating indices based on the complete data, fitting the substantive logistic regression model, repeating the entire process several times, and calculating the total variance as a weighted sum of the within- and between-imputation variance estimates. The resulting standard error could then be used to conduct inference. While it may sound tedious, in practice once the code is written it is quite simple and straightforward.

Let m be the number of imputations performed, producing m different point estimates for the parameters and their standard errors. The combined point estimate for the parameter, θ, is given by the mean over all imputations:

$$\overline{\theta} = \frac{1}{m}\sum_{i=1}^{m}\widehat{\theta}_i, \tag{6}$$

where $\widehat{\theta}_i$ is the estimate from the ith imputation. Let W_i be the variance estimate from the ith imputation; then the within-imputation variance is given by the mean over all m imputations:

$$\overline{W} = \frac{1}{m}\sum_{i=1}^{m}W_i. \tag{7}$$

The between-imputation variance is given by

$$B = \frac{1}{m-1}\sum_{i=1}^{m}\left(\widehat{\theta}_i - \overline{\theta}\right)^2. \tag{8}$$

The estimate for the total variance for $\widehat{\theta}$ is given by

$$T = \overline{W} + \left(1 + \frac{1}{m}\right)B. \tag{9}$$

The statistic has a t distribution given by

$$T^{-1/2}\left(\theta - \overline{\theta}\right) \sim t\left(\nu_m\right), \tag{10}$$

where θ is the value of the parameter under the null hypothesis and the degrees of freedom are given by

$$\nu_m = (m-1)\left[1 + \frac{\overline{W}}{(1+1/m)\,B}\right]^2. \tag{11}$$

Inference can then be conducted via the construction of confidence intervals or hypothesis testing [4, 28].

1.7. Assessing Imputation. Assessing how well an imputation worked is somewhat problematic. If the *MCAR* assumption holds, then we would expect only small changes in the means of the different variables and no change in the basic shape of the distribution for quantitative variables. Hence, small changes in the histograms of variables between the incomplete and complete data (before and after imputation) indicate *MCAR*. For quantitative *MAR* data, we would expect small changes in the shape of the conditional distribution of Y given each level of X or combination of X's. If there are a large number of variables it may be impossible to check. However, it can be instructive to plot histograms of the incomplete data for the missing variable by different categories of a few categorical variables and compare these histograms to the same for the complete data. If there are no drastic changes, then this is evidence for the data being at least *MAR*. For categorical data, assessing the imputation is even harder. In practice, usually we can only examine summary statistics and look for any problematic data or data patterns. This is a difficult theoretical problem and until it is resolved by theorists, the investigator must rely on common sense and reasonable care with checking of assumptions. If, in expert opinion and experience, respondents are likely to refuse to answer certain types of questions based on the answer to the question itself, and no number of other participant characteristics can be used to model this probability of refusal, then methods to correct this bias using information from outside the sample are indicated.

In general, though, in the assessment of imputation using model-based approaches, if the algorithm converges and produces no anomalous values, then we have no reason to question the results, as MCMC methods have been shown by strong theory and simulation to produce samples from the target distribution.

2. Methods

We tested the method on a case study using the National Latino and Asian American Study (NLAAS). The NLAAS core sampling procedure resulted in a nationally representative sample of 4649 Latino and Asian Americans and

immigrants who resided in the contiguous United States. Regarding the Latino sample, there were 577 Cubans, 868 Mexicans, and 614 other Latinos. The subcategory "other Latinos" included immigrants from Colombia, the Dominican Republic, Ecuador, El Salvador, Guatemala, Honduras, Nicaragua, and Peru. The Asian sample consisted of 600 Chinese, 508 Filipino, and 520 Vietnamese participants and 467 other Asians. The subcategory "other Asians" consisted of Koreans, Japanese, Asian Indians, and individuals of other Asian backgrounds.

The NLAAS data set had one of the most comprehensive and advanced designs ever developed. A detailed description of the NLAAS methods of data collection has been previously documented [29–32].

The sampling techniques consisted of three major approaches. First, core and secondary sampling units were selected according to probability proportionate to size, from which household members in the continental United States were sampled. While primary sampling units were defined as metropolitan statistical areas, secondary sampling units were formed from contiguous groupings of census blocks. Second, high-density supplemental sampling was applied, using a greater than 5% density criterion, in which Asian and Latino groups were oversampled. Asian and Latino individuals who did not belong to the target groups under which these geographical areas were classified were still eligible to take part in NLAAS. For example, Vietnamese individuals living in a Chinese high-density census block were eligible. Third, secondary respondents were recruited from households in which one eligible participant had already been recruited and interviewed. Secondary respondents sampling was used to further increase the number of study participants. In all three sampling procedures explained above weighting corrections were applied to take into account joint probabilities of selection.

The NLAAS instruments were available in Cantonese, Mandarin, Tagalog, Vietnamese, Spanish, and English. They were translated using standard translation as well as back-translation techniques. All participants received an introductory letter and the study brochure in their preferred language. Those who gave their consent to take part in the study were screened and interviewed by professionals who had linguistic and cultural backgrounds similar to those of the sample population. Interviews were conducted with computer-assisted interviewing software in the preferred language of the participants. Face-to-face interviews with the participants were administered in the core and high-density samples. Exceptions were made when respondents specifically requested a telephone interview or when face-to-face interviewing was prohibitive. The average length of each interview was 2.4 hours. As a measure of quality control, a randomly selected sample of participants with completed interviews was contacted to validate the data.

Written consent was obtained for all study participants, protocols, and procedures. Human subject approval was given by the Cambridge Health Alliance, Harvard University, the University of Michigan, and the University of Washington.

2.1. Imputing Missing Values for the NLAAS Data Set. All but a few variables in the NLAAS data set had missing observations. We selected a total of 75 out of approximately 3000 items available in the NLAAS dataset for the imputation, with 68 items of interest in the substantive model. The 75 items include both single variables such as sex, race, participant's education, spouse's education, mother's education, father's education, child labor, economic resources, and multivariable constructs such as social networks, family cohesion, language preference, ethnic or native language proficiency, English language proficiency, and socioeconomic success. Out of the 75 items used for imputation, 5 were either approximately normal or could be normalized using transformations. The variable SE2 (spouse's education) was transformed using the square root transformation and EM2 (child labor/age at employment) was transformed using exponentiation to 3/2. The remaining 70 variables were binary or ordinal with 2 to 5 categories. Additionally, four variables had no missing data. Hence we had 9 variables with which to build the first imputation model and 75 variables with which to build the final imputation model. The extent of the missing data can be visualized in the histogram and box-plot shown in Figures 1 and 2.

We used the SAS procedure *Proc MI* to perform the imputations using the MCMC model-based method described above. First the continuous variables were normalized by a suitable transformation, if necessary. Then the multivariate normal imputation, as described above, was performed on these. Next, the categorical variable with the fewest missing values was imputed using all completely observed variables and the normal variables imputed in the first stage of imputation. We used the *monotone discrim* model for binary and 3-category variables and *monotone logistic* for variables with more than 3 categories. These methods implement the methods model-based MCMC procedure described and require the monotone missing pattern.

Once all missing values were imputed, we developed indices to measure various abstract concepts such as *English language proficiency*, *ethnic or native language proficiency*, *language preference*, *social networks*, and *family cohesion*. We developed a model to predict socioeconomic success in Latino and Asian immigrants based on constructs such as *language preference* and *proficiency*, *economic resources*, *social networks*, *family cohesion*, and *child labor*. The constructs were built from responses to items in the survey using the complete data.

For illustration of the multiple imputation procedure, we performed the procedure 10 times for data used in a substantive model to predict socioeconomic success based on several constructs. We estimated the total variance and performed the t-test for the null hypothesis that the true parameter value is zero versus the alternative that is it not equal to zero for each parameter in the model.

To assess the imputation, we checked for extreme or nonsensical values after each imputation and graphed histograms of the continuous variables. Examples of the histograms are shown in Figures 3, 4, 5, and 6. All tables and figures were produced in SPSS.

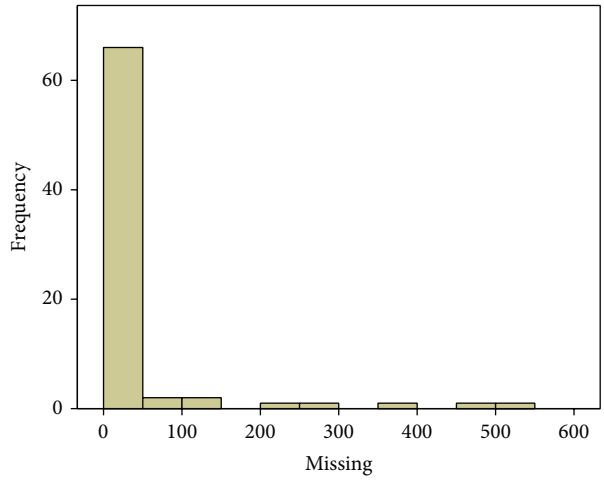

Figure 1: Histogram of the number of missing observations in 75 items used from the NLAAS data.

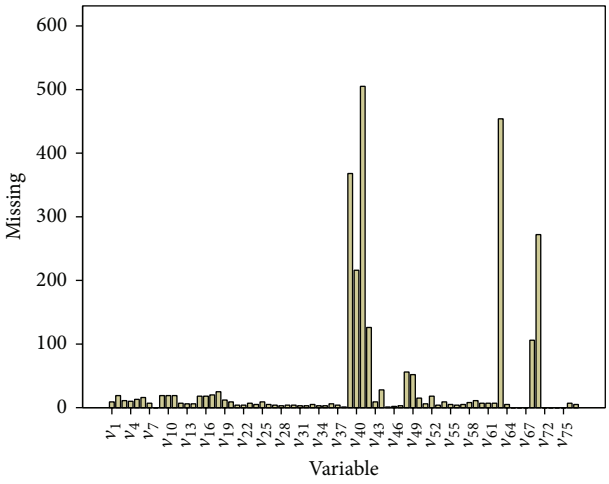

Figure 2: Bar chart of the extent of missingness showing each variable.

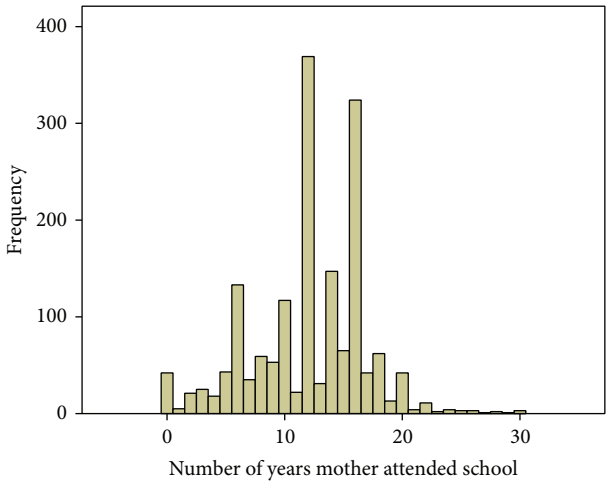

Figure 3: Histograms of raw data for the number of years the participant's mother attended school.

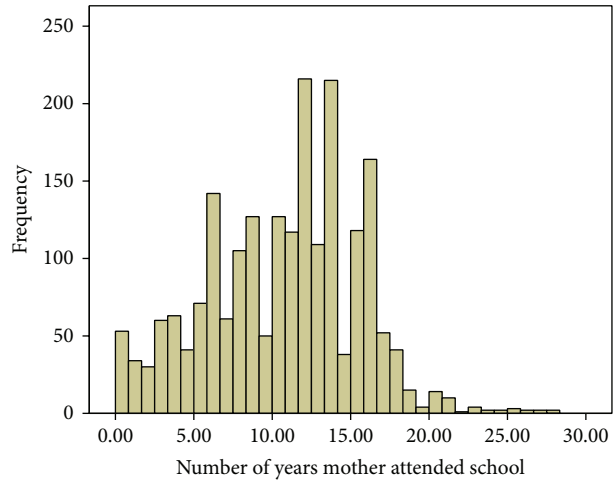

Figure 4: Histogram after imputation of missing values of the number of years the participant's mother attended school.

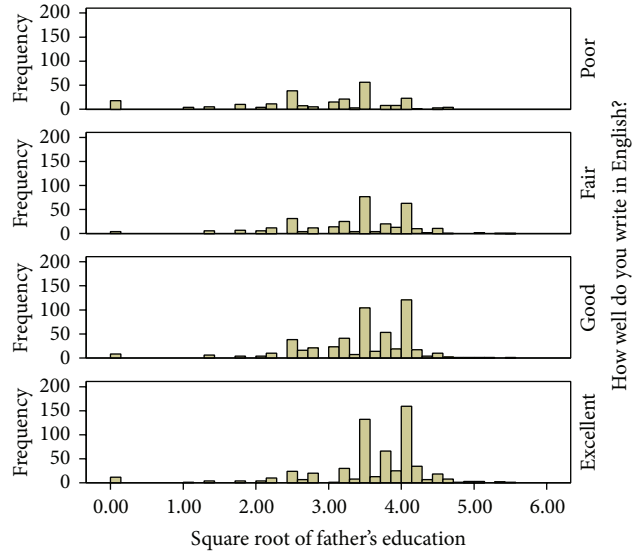

Figure 5: Histograms of raw data of the square root of father's education by writing proficiency in English.

3. Results and Discussion

3.1. Imputation. For this study, the imputation appeared to work well. There were no problems with convergence and no implausible values were observed. Figure 3 shows the small changes in the distribution of the variable *Mother's Ed*, the number of years of education for the respondent's mother. These results are typical of the quantitative variables imputed. Figures 5 and 6 show the histograms for a continuous variable which is the square root of the number of years of education for the participant's father. The square root was necessary to achieve approximate normality. The small changes in the conditional distribution lend evidence that the missing data process for this variable is *MAR*.

Table 1: Results of logistic regression substantive analyses on 10 imputed data sets. The far right column reports the percentage of the imputed data sets which resulted in rejecting the null hypothesis that the parameter is equal to zero versus the two-sided alternative. The other columns represent the parameter estimates and t values for the t-test described in Section 1.6. $N = 4649$, $t_9^*(0.05) = 2.26$.

Index	$\widehat{\theta}$	W	B	T	t	% Reject H_o
Child labor	$-4.40E-01$	$1.17E-02$	$3.15E-03$	$1.51E-02$	-3.57	100
English language proficiency	$1.98E-01$	$1.65E-04$	$4.00E-04$	$6.05E-04$	8.07	100
Parents' education	$6.83E-02$	$8.86E-05$	$2.59E-04$	$3.73E-04$	3.54	100
Race	$-3.67E-01$	$1.26E-03$	$3.72E-04$	$1.67E-03$	-8.98	100
Sex	$2.25E-01$	$1.11E-03$	$3.13E-03$	$4.55E-03$	3.33	100

Dependent variable: socioeconomic success (0, 1).

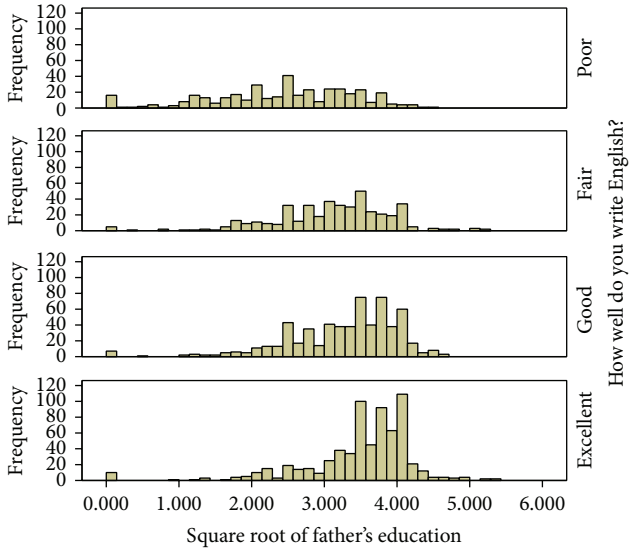

Figure 6: Histograms of complete data of the square root of father's education by writing proficiency in English.

3.2. Multiple Imputation. The results for the multiple imputations are shown in Table 1. Findings indicate that, for all variables in the model, all of the single tests were consistent with the combined test.

Our approach represents a simple and very effective method for imputation of survey data, which are often ordinal or nominal. Our method combines the capability of modelling the missing data distribution of the automated model-based procedures, such as Bayesian MCMC methods, commonly available in many software packages, while circumventing the current limitations in many of these packages for the imputation of a large number of categorical variables.

Conflict of Interests

The authors declare there is no conflict of interests regarding the publication of this paper.

Acknowledgment

The project described was supported in part by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), through Grant no. UL1 TR000002.

References

[1] P. D. Allison, "Multiple imputation for missing data: a cautionary tale," *Sociological Methods & Research*, vol. 28, no. 3, pp. 301–309, 2000.

[2] P. D. Allison, *Missing Data*, Sage, Thousand Oaks, Calif, USA, 2001.

[3] P. A. Gimotty and M. B. Brown, "Imputation procedures for categorical data: their effects on the goodness-of-fit chi-square statistic," *Communications in Statistics: Simulation and Computation*, vol. 19, no. 2, pp. 681–703, 1990.

[4] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York, NY, USA, 1987.

[5] D. B. Rubin, "Multiple imputation after 18+ years," *Journal of the American Statistical Association*, vol. 91, no. 434, pp. 473–489, 1996.

[6] U. Kohler, "Surveys from inside: an assessment of unit nonresponse bias with internal criteria," *Survey Research Methods*, vol. 1, no. 2, pp. 55–67, 2007.

[7] T. W. Smith, "Survey non-response procedures in cross-national perspective: the 2005 ISSP non-response survey," *Survey Research Methods*, vol. 1, no. 1, pp. 45–54, 2007.

[8] A. W. Hoogendoorn and J. Daalmans, "Nonresponse in the recruitment of an internet panel based on probability sampling," *Survey Research Methods*, vol. 3, no. 2, pp. 59–72, 2009.

[9] G. Chen and T. Åstebro, "How to deal with missing categorical data: test of a simple Bayesian method," *Organizational Research Methods*, vol. 6, no. 3, pp. 309–327, 2003.

[10] G. B. Durrant, "Imputation methods for handling item–nonresponse in practice: methodological issues and recent debates," *International Journal of Social Research Methodology*, vol. 12, no. 4, pp. 293–304, 2009.

[11] J. L. Schafer, *Analysis of Incomplete Multivariate Data*, Chapman & Hall, New York, NY, USA, 1997.

[12] J. K. Vermunt, J. R. van Ginkel, L. A. van der Ark, and K. Sijtsma, "Multiple imputation of incomplete categorical data using latent class analysis," *Sociological Methodology*, vol. 38, no. 1, pp. 369–397, 2008.

[13] N. J. Horton and K. P. Kleinman, "Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models," *The American Statistician*, vol. 61, no. 1, pp. 79–90, 2007.

[14] R. M. Groves, "Non-response rates and nonresponse bias in household surveys," *Public Opinion Quarterly*, vol. 70, no. 5, pp. 646–675, 2006.

[15] J. L. Schafer, "Multiple imputation: a primer," *Statistical Methods in Medical Research*, vol. 8, no. 1, pp. 3–15, 1999.

[16] S. van Buuren, H. C. Boshuizen, and D. L. Knook, "Multiple imputation of missing blood pressure covariates in survival analysis," *Statistics in Medicine*, vol. 18, pp. 681–694, 1999.

[17] P. D. Allison, "Handling missing data by maximum likelihood," SAS Global Forum: Statistics Data Analysis 312-2012, 2012, http://www.statisticalhorizons.com/wp-content/uploads/MissingDataByML.pdf.

[18] "Statistical Computing Seminars: Multiple Imputation in Stata," Institute for Digital Research and Education, UCLA, http://www.ats.ucla.edu/stat/stata/seminars/missing_data/mi_in_stata_pt1.htm.

[19] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, Chapman & Hall, New York, NY, USA, 1995.

[20] K. H. Li, "Imputation using Markov chains," *Journal of Statistical Computation and Simulation*, vol. 30, no. 1, pp. 57–79, 1988.

[21] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, John Wiley & Sons, New York, NY, USA, 2002.

[22] C. Liu, "Bartlett's decomposition of the posterior distribution of the covariance for normal monotone ignorable missing data," *Journal of Multivariate Analysis*, vol. 46, no. 2, pp. 198–206, 1993.

[23] M. A. Tanner and W. H. Wong, "The calculation of posterior distributions by data augmentation," *Journal of the American Statistical Association*, vol. 82, pp. 528–540, 1987.

[24] P. Minini and M. Chavance, "Sensitivity analysis of longitudinal binary data with non-monotone missing values," *Biostatistics*, vol. 5, no. 4, pp. 531–544, 2004.

[25] P. J. Lavrakas, *Encyclopedia of Survey Research Methods*, Sage, Thousand Oaks, Calif, USA, 2008.

[26] D. Zhang, *A Monte Carlo Investigation of Robustness to Non-normal Incomplete Data of Multilevel Modeling*, Texas A&M University Press, College Station, Tex, USA, 2005.

[27] K. Lueck, *Evaluationsmethoden der Bildungsforschung*, LISUM, Ludwigsfelde, Germany, 2006.

[28] J. Barnard and D. B. Rubin, "Small-sample degrees of freedom with multiple imputation," *Biometrika*, vol. 86, no. 4, pp. 948–955, 1999.

[29] M. Alegría, D. Takeuchi, G. Canino et al., "Considering context, place and culture: the National Latino and Asian American Study," *International Journal of Methods in Psychiatric Research*, vol. 13, no. 4, pp. 208–220, 2004.

[30] S. G. Heeringa, J. Wagner, M. Torres, N. Duan, T. Adams, and P. Berglund, "Sample designs and sampling methods of the Collaborative Psychiatric Epidemology Studies (CPES)," *International Journal of Methods in Psychiatric Research*, vol. 13, no. 4, pp. 221–240, 2004.

[31] K. Lueck and M. Wilson, "Acculturative stress in Asian immigrants: the impact of social and linguistic factors," *International Journal of Intercultural Relations*, vol. 34, no. 1, pp. 47–57, 2010.

[32] K. Lueck and M. Wilson, "Acculturative stress in Latino Immigrants: the impact of social, socio-psychological and migration-related factors," *International Journal of Intercultural Relations*, vol. 35, no. 2, pp. 186–195, 2011.