A Comparison of Methods for Creating Multiple Imputations of Nominal Variables

Kyle M. Lang[1]

Wei Wu[2]

[1]Institute for Measurement, Methodology, Analysis & Policy at Texas Tech University

[2]University of Kansas Department of Psychology

This work was partially supported by NSF grant No. 1053160 to W. Wu and T. Little. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies. The authors wish to thank Dr. Karrie Shogren for sharing the data used in the real-data resampling experiment reported below. Correspondence concerning this article should be addressed to Kyle Lang (email: kyle.lang@ttu.edu)

Abstract

Many variables that are analyzed by social scientists are nominal in nature. When missing data occur on these variables, optimal recovery of the analysis model's parameters is a challenging endeavor. One of the most popular methods to deal with missing nominal data is multiple imputation (MI). This study uses a Monte Carlo simulation study and a real-data resampling study to evaluate the capabilities of five MI methods that are often recommended for treating incomplete nominal variables. Practical recommendations are provided based on the findings.

Keywords: Missing data, Nominal variables, Multiple imputation, General location model, Multiple imputation with chained equations

A Comparison of Methods for Creating Multiple Imputations of Nominal Variables

Incomplete categorical variables present an intriguing challenge for missing data analysts, because much of the theory underpinning modern missing data analysis assumes continuously distributed data. As a result, missing data methods that work well for normally distributed variables can produce nonsensical results, or entirely break down, when naively applied to categorically distributed missing data (Enders, 2010; Little & Rubin, 2002; Schafer, 1997). Unfortunately, many interesting phenomena in social and behavioral research are best represented via discrete variables. For example, choice of political candidate, the number of cigarettes smoked in a day, and responses to Likert-type items all represent categorical variables. Therefore, missing data researchers have long sought principled methods to treat categorical nonresponse (i.e., methods that model the nonresponse via appropriate, discrete distributions).

The majority of the solutions presented in the literature pertain to the treatment of binary or ordinal variables. In the case of ordinal variables, it can often suffice to naively impute the missing data under the assumption of normality, if the discrete nature of the measurement is unimportant (e.g., when the variables will be averaged into a scale score; Wu, Jia, & Enders, in press). When the discrete measurement level must be maintained, there are powerful solutions based on thresholding the latent variable produced by a probit regression model (Enders & Keller, 2014). Monte Carlo simulations suggest that imputation methods based on logistic or probit regression equations perform well when imputing binary data (Brand, 1999; Enders & Keller, 2014). However, treating incomplete nominal variables (i.e., unordered categorical variables) with more than two response levels (e.g., race/ethnicity, political affiliation) has received relatively little attention in the missing data literature. Treatments for such incomplete nominal variables are the focus of this study. Although binary variables are also nominal, we will hereafter use the terms “nominal variables” and “unordered factors” to refer to unordered

categorical variables with three or more possible response levels.

Unique Challenges of Nominal Missing Data

Nominal-level responses can be modeled as realizations of a multinomial trial. Thus, missing data on these variables follows a convenient functional form, namely the multinomial distribution. However, the way that these variables are represented in data leads to inherent difficulties with treatment of their missing values. Nominal variables can either be represented by a set of $K - 1$ dummy codes (where K is the number of response categories) or as a single unordered factor with K levels. Data analysts must be wary of naively treating either of these representations with imputation methods that were not designed for nominal data.

The chance of falling into any one of a nominal variable's K response categories, versus the reference category, is given by a binomial probability. Yet, simply imputing each nominal variable's dummy-coded representation as though each code is an independent binary item can lead to impossible results (i.e., subjects endorsing multiple response categories; Allison, 2002). Likewise, if the nominal variable is represented as an unordered factor, applying normal-theory imputation models will produce quantitative, decimal-valued imputations for variables that contain only qualitative, integer-valued labels. Thus, a more nuanced approach is needed.

Because nominal variables are distributed in the same way as a series of binary variables, a natural approach to treating missingness on nominal variables is to extend methods designed for binary variables. However, the increased computational burden induced by the exponential growth of the contingency table that occurs when adding response levels can cause methods that work well for binary missing data to fail when applied to nominal missing data. Although some authors have acknowledge the additional computation burden inherent in treating nominal variables (e.g., Belin, Hu, Young, & Grusky, 1999; Schafer, 1997), we were unable to locate any systematic examination of missing data methods specifically targeting incomplete nominal

variables. To address this gap in the literature, the study reported here compares several recommended methods for creating multiple imputations (MIs) of incomplete nominal variables.

Extant Approaches

Imputation methods that are tailored for nominal data can be generally classified into one of three types: (a) ad hoc, rounding-based techniques, (b) principled, parametric methods that model the missing data with generalized linear models (GLMs), and (c) methods based on nonparametric predictive algorithms. A short overview of these approaches is given below.

Ad hoc approaches. The most convenient approach entails transforming the nominal variables into dummy codes and naively imputing these codes under the multivariate normal model. In the case of nominal variables (as opposed to certain instances of ordinal variables), these decimal-valued imputations cannot be left unaltered because the numeric values of the original dummy codes merely represent qualitative labels. Thus, the imputed values must be rounded back to meaningful, integer-valued labels. Several different rounding schemes have been proposed to transform normal-theory imputations into binary values (e.g., Bernaards, Belin, & Schafer, 2007; Yucel, He, & Zaslavsky, 2008, 2011). The most natural approach prescribes naively assigning a label of 1 to imputed values greater than or equal to .5 and assigning a label of 0 to imputed values less than .5 (Allison, 2002; Honaker & King, 2010; Schafer, 1997).

When treating dummy-coded nominal variables, as opposed to binary variables, this naïve rounding procedure also requires modification. Simply rounding the imputations for each dummy code, independently, can lead to impossible solutions (e.g., participants endorsing more than one response category). To circumvent such possibilities, Allison (2002) suggested a ranking-based method (hereafter RANK) that extends the idea of naïve rounding to the multinomial context. RANK is implemented by first imputing all $K - 1$ dummy codes under a normal-theory imputation model. The reference group's imputation is then computed as one

minus the sum of these $K - 1$ normal-theory imputations. These steps produce K decimal-valued imputations, one for each possible response level of the incomplete nominal variable. Finally, an imputed set of dummy codes is constructed by assigning a value of 1 to the response level with the largest imputation and assigning a value of 0 to all other response levels. In the case of incomplete binary variables, this process reduces to naïve rounding of the decimal-valued imputations. Allison (2002) offered an intuitive justification for this approach but did not present any empirical evidence of its performance.

Principled approaches. The most expedient principled approach is *multiple imputation with chained equations* (MICE; also known as sequential regression imputation or fully conditional specification) employing multinomial logistic regression (MNR) as the elementary imputation method (hereafter MICE-MNR). The MICE approach entails creating the MIs by sequentially imputing each incomplete variable with a univariate predictive model (in this case, a multinomial logistic regression). Brand (1999) developed one such approach that is now employed as the default method for imputing unordered factors in the R package **mice** (van Buuren & Groothuis-Oudshoorn, 2011). Two joint modeling approaches can also be applied in certain circumstances. If all of the incomplete variables are categorical, and all predictors of the nonresponse mechanism are also categorical (so that all variables related to the missing data process can be collapsed into a contingency table), the missing data can be imputed via a fully saturated multinomial model or a simpler loglinear model in which certain marginal associations are not modeled (Schafer, 1997). Yet, when the data analyst has continuous covariate data available, imputing under the saturated multinomial model or loglinear model can exclude useful auxiliary information because these models cannot incorporate continuous variables. The R package **mix** (Schafer, 2013) can produce MIs from multinomial and loglinear models.

When the incomplete data set contains both categorical and continuous variables,

imputations can be created with a joint modeling approach based on the general location model (GLOC). GLOC was originally developed by Olkin and Tate (1961) for inferential modeling of mixed categorical and continuous items. In this original implementation, the categorical data were quantified by a marginal, saturated multinomial model and the continuous variables were assigned a conditional multivariate normal model. Under GLOC, the means of each conditional multivariate normal model is given by the expected proportion of their respective cell in the contingency table, but each normal model shares a common covariance matrix. In the case of binary incomplete variables, GLOC reduces to simple linear discriminant analysis.

Little and Schluchter (1985) extended the original Olkin and Tate (1961) formulation of GLOC by using a loglinear model to describe the categorical data. The Little and Schluchter (1985) GLOC allowed for structural models of the nominal variables. They also developed methods for applying GLOC to the task of missing data imputation. A Gibbs sampling scheme for creating MIs from the Little and Schluchter (1985) GLOC was described by Schafer (1997). Although MI under GLOC is mathematically appealing due to its reliance on a joint probability model to describe the missing data, high computational cost limits its applicability (Allison, 2002; Little & Rubin, 2002; Schafer, 1997). Belin et al. (1999) employed GLOC-based MI to treat the missing data in a typical cancer study and showed that GLOC can be computationally infeasible and breaks down with real-world-sized problems.

Nonparametric approaches. Because they are not subject to any distributional assumption, nonparametric approaches are conceptually well-suited to imputing nominal missing data. Many nonparametric methods are donor-based techniques in which the missing data are replaced by observed data from a matched *donor* case. *Predictive mean matching* (PMM) is a popular donor-based method that has been advocated for possible use with nominal variables (van Buuren, 2012). In PMM the donor and recipient cases are matched based on their respective

predicted outcome values (i.e., their $\hat{Y}_i$ values) from a univariate regression equation in which the incomplete variable acts as the dependent variable. Marshall, Altman, and Holder (2010) and Marshall, Altman, Royston, and Holder (2010) both suggested that MICE with PMM as the elementary imputation method (hereafter MICE-PMM) can outperform rounded normal-theory MI when imputing nominal data in prognostic modeling applications. However, they noted that researchers should be wary of applying PMM with small sample sizes due to the possibly restricted size of the donor pool. Andridge and Little (2010) also raised this point by noting that the validity of PMM (or any donor-based imputation method) is reliant upon the existence of an adequate number of matched respondents to replace the missing data of the nonresponents. In the case of incomplete nominal variables, this limitation is exacerbated by the exponential growth of the contingency table's size as the number of response categories increases (Agresti, 2007).

Decision tree modeling is another nonparametric method that can be used to create imputations that automatically preserve the distributional properties of the incomplete data (Borgoni & Berrington, 2013). The flexibility and simplicity of decision tree models has led some authors to highlight them as ideal candidates for imputation engines (e.g., Harrell, 2001; Hastie, Tibshirani, & Friedman, 2009). Burgette and Reiter (2010), Drechsler and Reiter (2011), Reiter (2005), and Wallace, Anderson, and Mazumdar (2010) all explored the capability of MICE with *classification and regression trees* (CART) as the elementary imputation method (hereafter MICE-CART). Their applications were able to naturally accommodate mixed continuous and categorical missing data, and they all found that MICE-CART led to mostly unbiased estimates of parameters but tended to induce under-coverage for the associated CIs.

The Current Project

This project compares several currently advocated methods for imputing nominal missing

data to offer guidance to researchers faced with nominal incomplete data. Practicality in substantive research domains was a paramount concern in choosing the imputation methods to examine in this study. Therefore, the methods were chosen to satisfy two requirements: (1) Each method needed to simultaneously treat mixed variable types (i.e., both continuous and categorical data), and (2) each method needed to produce legal imputations for the nominal variables (i.e., qualitative, integer-valued labels rather than quantitative, real-valued numbers). Based on these considerations, five MI approaches were selected for comparison: three MICE approaches (MICE-MNR, MICE-CART, and MICE-PMM), and two joint modeling approaches (GLOC and RANK). Each of these five techniques were used for the imputation phase of a standard MI analysis (see Enders, 2010, Chapters 7 & 8 for a detailed introduction to the three phases of MI-based analyses). These five techniques were compared based on their ability to recover the population-level regression coefficients of MNR and multiple linear regression (MLR) models in which the imputed variables are included as outcomes and predictors, respectively. A Monte Carlo simulation study was used to assess each method's performance under well-controlled conditions with simple data-generating models. A real-data resampling study was also conducted to further elucidate the imputation methods' strengths and weaknesses in a more ecologically valid situation with more complex population-level models.

In this context, the following hypotheses are posed:

1) Because MICE-MNR employs the correct distribution for the missing data, we expect it to produce the most accurate results—especially when imputing outcome variables.
2) Consistent with previous findings regarding the performance of naïve rounding approaches, we expect RANK to produce the least accurate results.
3) Due to its high degree of flexibility and the strong prediction performance of its underlying algorithm, we expect that MICE-CART will produce generally accurate

results but will not perform as well as MICE-MNR.

4) Because the imputation models employed in the simulation study will be small, the computational limitations of GLOC should not be an issue. Thus, GLOC is expected to perform well in the simulation study due to its use of a saturated joint model for the nonresponse.
5) Because PMM can only perform optimally with large donor pools, we expect MICE-PMM to perform disproportionately worse (relative to the other techniques studied) with smaller sample sizes and larger proportions of missing data.
6) Because we have employed discriminative analysis models, we expect the choice of imputation method to have a greater impact when imputing outcomes than when imputing predictors.

Monte Carlo Simulation Study

Methods

The Monte Carlo simulation study was used to assess the performance of the different imputation methods under well-controlled conditions. The techniques were evaluated based on how well they facilitate recovery of the true values of the regression coefficients in MNR and MLR models.

Simulation parameters. Three simulation parameters were varied: total sample size (N = {250, 500, 1000}), proportion of missing data (PM = {0.1, 0.2, 0.3, 0.4, 0.5}), and the number of nominal response levels (K = {3, 5, 7, 10}). We did not consider sample sizes lower than N = 250 because categorical data analyses tend to require large samples (Agresti, 2007). These factors were fully crossed with the five MI methods to produce a full-factorial design with 3(N) × 5(PM) × 4(K) × 5(*MI Method*) = 300 crossed conditions.

Software. All of the MICE-based imputations were created with the R package **mice** (van Buuren & Groothuis-Oudshoorn, 2011). The MICE-based conditions each employed 10 rounds of fully conditional updates. For MICE-PMM and MICE-CART, both the continuous and nominal missing data were imputed with the same elementary imputation method. For MICE-MNR, however, the continuous missing data were imputed with normal-theory Bayesian regression. RANK was implemented with the R package **Amelia II** (Honaker, King, & Blackwell, 2011), and an adaptive ridge prior, with a maximum value of $N/10$, was used to stabilize the fitted imputation models. This prior was imposed by setting `autopri = 0.1` in the **Amelia II** software (see, Honaker et al., 2011 for more details of this feature). GLOC was implemented with the R package **mix** (Schafer, 2010). For GLOC, estimates from the expectation-maximization algorithm were used as the starting values for the data augmentation algorithm. Each imputation was generated from its own Markov Chain that was run for 1000 iterations before drawing the imputations (see Schafer, 1997, Chapter 9 for details on parameterizing the GLOC algorithm).

Data generation. The population values employed in the Monte Carlo simulation were chosen arbitrarily. Three analysis variables were simulated for the current study: a nominal outcome variable Y with K response levels, a normally distributed predictor variable $X \sim \mathrm{N}(0.25, 1.0)$, and a normally distributed outcome variable $Z \sim \mathrm{N}(Y\boldsymbol{\zeta}, 1.0)$. These three analysis variables were related via two different population models that were used for data generation:

$$\operatorname{logit}\left(Y\right) = \mathbf{1}_n\boldsymbol{\alpha} + X\boldsymbol{\beta}, \tag{1}$$

$$Z = Y\boldsymbol{\zeta} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\mathbf{1}_n$ is an N-dimensional column vector of ones, $\boldsymbol{\alpha}$ is an K-dimensional row vector of intercept terms, $\boldsymbol{\beta}$ is a K-dimensional row vector of multinomial regression coefficients, $\mathbf{Y}$ is an

$N \times (K - 1)$ matrix containing the dummy-coded representation of Y, $\boldsymbol{\zeta}$ is a $(K - 1)$-dimensional column vector of linear regression coefficients, and $\boldsymbol{\varepsilon} \sim \mathrm{N}(0, 1)$ is an N-dimensional column vector of residual error terms. Each multinomial regression coefficient was set to $\beta_k = \ln(2)$ so that the population-level odds of endorsing the kth level of Y, versus the reference level, doubled for every unit increase in X. The linear regression coefficients were specified according to the rule: $\zeta_k = k - (K / 2)$ for $k = 1, 2, \ldots, K - 1$ (e.g., taking $K = 5$ would produce $\boldsymbol{\zeta} = \{-2, -1, 1, 2\}^T$). The multinomial intercept terms were specified as an exponentially decreasing function of the number of response categories according to the rule: $\alpha_k = \ln(3)(1/2)^{(k-1)}$ for $k = 1, 2, \ldots, K$ (e.g., taking $K = 5$ would produce $\boldsymbol{\alpha} \approx \{1.1, 0.55, 0.27, 0.14, 0.07\}$). This rule was chosen to reflect a situation in which most participants endorse one of the initial response categories and few participants endorsed the higher response levels (e.g., in many elections, most people will vote for one of a few front-running candidates but a small subset of voters will endorse several additional "fringe" candidates). Given these population models, the data were simulated in a multi-stage process. First, the normally distributed predictor values X were used to compute the probabilities of each observation falling into category k as follows:

$$P(y_n = k \mid x_n) = \frac{\exp\left(\alpha_k + \beta_k x_n\right)}{1 + \sum_{k=1}^{(K-1)} \exp\left(\alpha_k + \beta_k x_n\right)}. \tag{3}$$

This process produced an $N \times K$ matrix $\mathbf{P}$ of model-implied probabilities. The nominal outcomes in Y were simulated by looping over the rows of $\mathbf{P}$ and submitted the elements of each row as a vector of success probabilities to the R function `rmultinom`. The continuous outcome Z was simulated by dummy coding Y to produce $\mathbf{Y}$ which was then plugged into Equation 2.

The simulated data needed to include an incomplete continuous variable in addition to the incomplete nominal response Y, because the ability to simultaneously treat mixed variable types

was a critical characteristic of the imputation methods examined in this study. We simulated a nuisance covariate $X_{cov} \sim \mathrm{N}(0, 1)$ to include an incomplete continuous variable during the missing data imputation. In order to maintain adequate experimental control, the continuous variables that entered the analysis models (i.e., X and Z) were left complete so that any contamination of the fitted model parameters could be linked directly to the imputation methods' treatments of the nominal data. We also simulated a continuous auxiliary variable $X_{aux} \sim \mathrm{N}(0, 1)$ to act as a predictor of the nonresponse process. Each simulated dataset, therefore, contained one nominal variable (Y) and four continuous variables (X, Z, X_{cov}, X_{aux}).

Missing data imposition. Missing data were imposed on Y and X_{cov} according to a *missing at random* (MAR) mechanism. The MAR missingness on both variables was predicted by X_{aux} according to an ordinary probit regression model by applying the following formulas variable-wise:

$$\begin{aligned} R_n^{(1)} &= \mathrm{I}\left[\Phi\left(x_{aux,n};\, \bar{x}_{aux},\, \sigma_{x_{aux}}\right) \le PM\right], \\ R_n^{(2)} &= \mathrm{I}\left[\Phi\left(x_{aux,n};\, \bar{x}_{aux},\, \sigma_{x_{aux}}\right) \ge (1-PM)\right], \end{aligned} \tag{4}$$

where $R_n^{(1)}$ and $R_n^{(2)}$ are nonresponse indicator variables that take a value of 1 when an observation is missing and 0 otherwise, $\Phi(\cdot\,;\, \bar{X}_{aux},\, \sigma_{X_{aux}})$ is the normal cumulative distribution function with mean $\bar{X}_{aux}$ and standard deviation $\sigma_{X_{aux}}$, $\mathrm{I}[\cdot]$ is an indicator function that returns a 1 when its argument is true and 0 otherwise, and PM is the proportion of nonresponse to introduce. Once $R_n^{(1)}$ and $R_n^{(2)}$ had been created as above, the elements of the Y for which $R_n^{(1)} = 1$ were set to missing data and elements of X_{cov} were set to missing where $R_n^{(2)} = 1$. Thus, observations with low values of X_{aux} were missing Y data, and observations with high values of X_{aux} were missing X_{cov} data.

Outcome measures. The relative performance of the different imputation methods was assessed by how well they could facilitate recovering the true regression coefficients of the MNR model given by Equation 1 and the MLR model given by Equation 2. To quantify how well these regression coefficients were recovered, we employed two outcome measures: *percentage relative bias* (PRB) and *confidence interval coverage* (CIC):

$$\text{PRB} = 100\left(\frac{\bar{\hat{\theta}} - \theta}{\theta}\right),$$

$$\text{CIC} = \sum_{r=1}^{R} \text{I}\left(\theta \in CI_r\right),$$

where R indexes Monte Carlo replications, θ is the true value of the parameter, $\bar{\hat{\theta}}$ is the mean of the estimated parameter's R Monte Carlo replicates, and CI_r is the estimated confidence interval for the rth replication. For the current study, conditions with |PRB| > 10 were considered to have unacceptably large degrees of bias.

Procedure. This study employed $R = 500$ Monte Carlo replications. Within each of these replications, for each number of nominal response levels $K \in \{3, 5, 7, 10\}$, a complete data set of size $N \in \{250, 500, 1000\}$ was simulated. The procedure described above was then used to impose MAR missingness at rates of $PM \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. These missing data were then imputed 100 times with each of the five imputation methods described above. These sets of imputed data were then analyzed via the models given by Equations 1 and 2, and their fitted regression coefficients were pooled via *Rubin's Rules* (Rubin, 1987). Finally, after completing all 500 replications, PRB and CIC values were computed for the fitted regression coefficients.

Results

MNR analysis models. Table 1 contains the average PRB values for the slope coefficients of the MNR analysis models in which the imputed factors were predicted by the

normally distributed predictor variable X (see Equation 2). Table 2 contains the average CIC rates of these coefficients. These models each produced $K - 1$ estimated regression slopes, so the values reported in Tables 1 and 2 are averages taken over these $K - 1$ coefficients.

MICE-MNR, MICE-CART, RANK, and GLOC all led to models with bias levels that stayed mostly within the acceptable range. The only violations of this pattern occurred for $PM = 0.5$ where MICE-CART tended to exhibit an unacceptably large negative bias and RANK exhibited a problematic degree of negative bias in conditions with many nominal response categories (i.e., $K = 10$) or with large sample sizes (i.e., $N = 1000$). MICE-MNR only strayed into the realm of unacceptable bias when $N = 250$, $K = 10$, and $PM = 0.5$ (i.e., the most degenerate condition tested). MICE-PMM stood apart as exhibiting much larger biases than the other four imputation methods with bias levels entering the unacceptable range in most of the conditions.

In terms of CIC rates, only GLOC produced CIs with coverage rates that were consistently close to nominal. The remaining four methods generally exhibited a trend of decreasing coverage rates as nonresponse rates increased. MICE-CART, MICE-PMM, and RANK also tended to produce lower coverage rates with increasing sample sizes. This sample size dependence was most pronounced for MICE-PMM which exhibited approximately nominal coverage rates at $N = 250$, but extremely underestimated coverage rates for $N = 1000$. After GLOC, RANK produced the second most stable, and appropriate, CI coverage rates.

MLR analysis models. Table 3 contains the average PRB values for the slope coefficients of the MLR analysis models in which the imputed factors were dummy-coded and used to predict the normally distributed outcome variable Z (see Equation 3). Table 4 contains the average CIC rates of these coefficients. These models also produced $K - 1$ estimated regression slopes (i.e., one slope for each dummy code), so the values reported in Tables 3 and 4 are averages taken over these $K - 1$ coefficients.

When the imputed factors entered the analysis models as dummy coded predictors, rather than outcomes, the patterns of bias changed considerably from those reported above. MICE-MNR and GLOC both exhibited negligible degrees of bias across all conditions. MICE-CART and MICE-PMM exhibited satisfactory bias levels for $K \geq 5$. MICE-CART induced less bias than MICE-PMM did. For $K = 3$ MICE-PMM led to unacceptably large levels of bias for $PM \geq 0.3$ as did MICE-CART for $PM = 0.5$. Unlike its strong performance for the MNR analysis models, RANK was clearly the weakest method, in terms of coefficient bias, when applied in the context of MLR models. Although RANK led to lower bias rates than MICE-PMM for $K = 3$, when $K \geq 5$, RANK frequently led to unacceptably large negative biases, and the degree of this bias increased with rising rates of nonresponse and increasing numbers of response categories.

In terms of CI coverage rates, GLOC again provided the strongest results, producing coverage rates that were near nominal for all conditions. MICE-PMM produced good coverage rates for $N = 250$, but it led to considerably deflated coverage rates when $N \geq 500$ and $K = 3$ as well as when $N = 1000$ and $PM \geq 0.4$. MICE-CART and RANK both produced unstable coverage rates. The coverage rates of MICE-CART presented as a nonlinear function of the nonresponse rate whereby increasing rates of missingness led to decreasing coverage rates up to $PM = 0.4$ with a sudden jump in coverage rates for $PM = 0.5$. At lower sample sizes (i.e., $N \leq 500$) RANK tended to produce CIs with inflated coverage rates, and the degree of this inflation became worse as the number of nominal response categories increased. Yet, for $N = 1000$ RANK produced deflated coverage rates for $K \geq 5$ and $PM \geq 0.4$. After GLOC, MICE-MNR produced the second most stable coverage rates for the MLR analysis models. MICE-MNR *did* lead to decreasing coverage rates as nonresponse rates increased, but the degree of this deflation was minor except under high missing data rates (i.e., $PM \geq 0.4$) and many nominal response categories (i.e., $K \geq 7$).

Real-Data Resampling Experiment

Methods

The Monte Carlo simulation study reported above offered compelling insight into the relative performance of the MI methods examined in this study, but it prioritized strong experimental control over ecological validity. This focus on experimental control somewhat limits the practical generalizability of the simulation study's findings. To address this limitation, we also performed a real-data resampling experiment in which a more realistic set of analysis models were fit to data originally collected by Wehmeyer, Palmer, Lee, Williams-Diehm, and Shogren (2011) and Wehmeyer, Palmer, Shogren, Williams-Diehm, and Soukup (2013) to evaluate the efficacy of several self-determination interventions.

Population data. The original data set contain data from $N = 782$ participants collected over three waves of measurement. Items collected included the Arc's Self Determination Scale (Wehmeyer & Kelchner, 1995), information on post-high school employment and academic outcomes, and a large battery of demographic information. Full details of the original data set can be found in Shogren, Wehmeyer, Palmer, Rifenbark, and Little (2015). For the current study, five Wave 1 variables were examined: Student gender (*Gender*, binary), Student primary disability label (*Disability*, nominal), Student primary course of study (*Course*, nominal), Student age (*Age*, continuous), and Student mean self-determination score (*SDS*, continuous). This restricted data set was further subset by excluding any incomplete cases as well as factor levels with very low endorsement rates. This process produced a final, population data set with $N = 535$ observations of the five previously described variables. Table 5 contains the summary statistics of these population data.

Analysis models. To parallel the structure of the Monte Carlo simulation study, two analysis models were employed, a MNR model and a MLR model:

$$\text{logit}\left(Course\right) = \boldsymbol{\alpha} + \boldsymbol{\beta}_1 SDS + \boldsymbol{\beta}_2 Age + \boldsymbol{\beta}_3 Autism + \boldsymbol{\beta}_4 BD + \boldsymbol{\beta}_5 LD + \boldsymbol{\beta}_6 ADHD\,, \quad (5)$$

$$SDS = \alpha + \beta_1 Gender + \beta_2 Age + \beta_3 Autism + \beta_4 BD + \beta_5 LD + \beta_6 ADHD\,. \quad (6)$$

Table 6 contains the parameter estimates produced by fitting these models to the full, population data set.

Outcome measures. The resampling study had two limitations that drove our choice of outcome measure. (1) Bootstrapped estimates are correlated, so they violate the independence of replications assumption required to make valid Monte Carlo inferences about parameter variability. (2) The small "population" sample size (i.e., $N = 535$) limited our ability to interpret the absolute bias in parameter estimates, because the population estimates themselves were subject to large amounts of sampling error. Limitation 1 narrowed the universe of useful outcome measures to different types of parameter bias. Limitation 2 led us to focus our discussion on the relative performance of the imputation methods in lieu of scrutinizing their absolute performance. We, therefore, chose the raw bias in the analysis models' regression coefficients as the outcome measure in the resampling study. The "population" values used to calculate bias were taken to be the coefficients estimated from the full, cleaned sample described above.

Procedure. After subsetting and cleaning the data as described above, the cleaned population data were resampled with replacement (i.e., bootstrapped) 500 times. Each of these subsamples contained $N_{rs} = 300$ observations. For each subsample, 20% MAR missingness was imposed using the same procedure employed in the Monte Carlo simulation study. *Age* acted as the nonresponse predictor such that younger participants were missing *SDS* and *Disability* while older participants were missing *Gender* and *Course*. These missing data were imputed 100 times with each of the five MI methods employed in the Monte Carlo simulation study.

All software settings were equivalent between the studies with two exceptions. First, the

MICE-based conditions used 20 rounds of fully conditional updates in the resampling study to facilitate imputation model convergence with the noisy, real data. Second, due to the size of the contingency table and sparse cells therein, the GLOC model had to be restricted to achieve reliable convergence. This restricted GLOC modeled the categorical variables with a loglinear model that forced marginal independence between *Gender*, *Course*, and *Disability* and only allowed two-way interactions between these variables when modeling *Age* and *SDS* (see Schafer, 1997, Chapter 9 for details on restricting GLOC). After the missing data were imputed, the analysis models described by Equations 5 and 6 were fit to the imputed data sets and their coefficients were pooled with Rubin's Rules.

Results

Figure 1 contains bar plots of the raw bias in the regression coefficients associated with Disability in the MNR analysis model. Figure 2 shows analogous barplots of the predictors in the MLR analysis model as well as Age and SDS in the MNR analysis model. These figures show that the patterns of performance were largely similar between the resampling and simulation studies. The GLOC remained a strong performer across conditions, despite the considerable restrictions placed on the model to facilitate convergence. When imputing the predictors of the MLR analysis models, MICE-MNR was also a consistently strong performer while RANK and MICE-PMM led to the largest biases in this context (though not every coefficient was biased).

When imputing the outcome of the MNR analysis models, the performance picture was not as clear. The coefficients associated with continuous predictors (i.e., *SDS* and *Age*) were sporadically biased. Each imputation method performed well for some of effects and performed poorly for others. No definitive pattern was discernable for the continuous predictors of the MNR analysis models. For the nominal predictors of the MNR analysis models (i.e., *Disability*), the GLOC was again a consistently strong performer, and MICE-MNR showed the most

consistently biased results. The remaining imputation methods performed similarly to one another, tended to produce somewhat more bias than GLOC but less than MICE-MNR, overall. The poor performance of MICE-MNR represented a clear difference from the results of the simulation study wherein MICE-MNR produced consistently unbiased results. The simulation study did not, however, include any categorical predictors in the MNR analysis models. So, the simulation study contained no condition that was directly analogous to the Disability coefficients in the resampling study. Future work should extend the simple models used in our simulation to include both continuous and categorical predictor variables to elucidate potential limitations of MICE-MNR with categorical predictors.

The large biases seen for the "*BD:Life Skills*" and "*ADHD:Life Skills*" coefficients in Figure 1 were caused by low cell counts (i.e., the *BD* × *Life Skills* and *ADHD* × *Life Skills* cells contained only one observation in the population contingency table because life skills training is only provided to students with significant intellectual disabilities). There is very little, if any, information in the observed data to suggesting plausible values for these effects, so the estimates are expected to be biased. MICE-MNR appears to be especially sensitive to low cell counts. This pattern highlights an inherent difficulty of missing data analysis with categorical data. Low observed cell counts are already a difficult problem for researchers who wish to employ categorical data analytic methods (Agresti, 2007), but compounding this sparsity with additional nonresponse makes the challenge substantially worse. Thus, initial data screening is that much more important when attempting to impute missing nominal variables.

General Discussion

The results of this study make one point clear: there is no "magic bullet" technique for imputing incomplete nominal variables. None of the methods examined here will outperform the others in all circumstances, and the preferred method will differ depending on whether the

incomplete nominal variables enter the analysis model as outcomes or as predictors and on how many nominal variables enter the imputation model. Thus, few of our original hypotheses were unambiguously resolved. Hypothesis 1 was partially supported in that MICE-MNR consistently produced unbiased point estimates of regression coefficients for both MNR and MLR analysis models. In the simulation study, it was the second best all-around performer (after GLOC) for imputing nominal predictors in MLR models. Yet, MICE-MNR tended to produce substantially shrunken CI coverage rates when imputing the outcomes of MNR models and the resampling study suggested the possibility of issues when using MICE-MNR to impute categorical predictors of nominal outcomes. So, Hypothesis 1 was not fully supported, but MICE-MNR was one of the strongest performers for the MLR analysis models in the resampling study.

Hypothesis 2 was clearly supported for the case of imputing nominal predictors in MLR models where RANK led to very poor results, but the exact opposite of the hypothesized pattern was observed when imputing the outcomes of MNR models where RANK was the second most preferred method (after GLOC). An intuitive explanation of this disparity is difficult to construct, so future research should work to elucidate those conditions under which RANK will perform well and those in which it will fail.

Hypothesis 3 was not supported by this study. MICE-CART was one of the weakest performers for both MNR and MLR analysis models. Although it did produce unbiased point estimates of regression coefficients for most conditions, it produced very unstable confidence intervals with a strong tendency to undercover the true parameter values. This result is consistent with previous findings (e.g., Drechsler & Reiter, 2011; Reiter, 2005).

Hypothesis 4 was resoundingly supported, at least in terms of the simulation study's results. For the simulation study, GLOC was the universally strongest performer; it produced unbiased regression coefficients with nominal confidence interval coverage rates in every

simulation condition. The GLOC was also the strongest, relative, performer in the resampling study. Yet, the simplistic structure of the models employed in the simulation study could be hiding a fatal flaw in GLOC. Previous work (e.g., Belin et al., 1999) has shown that GLOC will produce poor results except with very small models (i.e., those with few incomplete categorical variable). To achieve convergence in the resampling study, the GLOC model had to be substantially restricted. The data set used for the resampling study was not particularly complex and the contingency table describing its categorical variables was not very large. So, it seems likely that the types of missing data problems that applied researchers will face in their day-to-day work could require so many restrictions on the GLOC model that the strong performance seen in this study is completely negated.

Surprisingly, Hypothesis 5 was not supported by this study. Although MICE-PMM did perform quite poorly for the MNR analysis models and for MLR analysis models with small numbers of nominal response categories, this poor performance was not differentially exacerbated by small samples or high nonresponse rates. Counter-intuitively, MICE-PMM was actually one of the strongest performers when imputing the nominal predictors of MLR models under small samples with high nonresponse rates. This performance could be driven by the simplicity of the matching algorithm PMM uses to find donor observations. This simplicity may mitigate the sensitivity to deflated sample sizes, especially when considering that even at the highest nonresponse rate and lower samples size tested here, MICE-PMM would still have a pool of at least $N = 125$ relatively homogeneous observations from which to draw donor cases.

Finally, Hypothesis 6 was supported. With the notable exception of RANK, all of the tested imputation methods had a general tendency toward stronger performance when imputing the predictors of MLR models than they did when imputing the outcomes of MNR models. This pattern was evident in both the simulation and resampling studies.

Practical Guidance for Applied Researchers

Methods that do not work. For models with simple, linear systematic components like the analysis models employed in this study, MICE-CART should be generally avoided. Although MICE-CART did lead to minimal bias, it never outperformed MICE-MNR. MICE-CART also produced unstable confidence interval coverage rates, particularly for MLR models. These two limitations suggest that MICE-MNR would be preferred to MICE-CART in any of the scenarios examined here. The CART algorithm is a very flexible, nonparametric modeling scheme, so MICE-CART might perform better in situations where its "automatic interaction detection" could come into play (see Borgoni & Berrington, 2013, for details on CART's merits in this context).

MICE-PMM should not be used to impute the outcomes of MNR models or the predictors of MLR models when those predictors have few nominal response levels (i.e., $K < 4$ or 5). Applying MICE-PMM to either of these cases is expected to introduce large amounts of bias into the fitted regression coefficients and to seriously deflate confidence interval coverage rates. Likewise, RANK should not be used to impute the nominal predictors in MLR models, under any circumstances. Doing so is expected to introduce considerable bias into the fitted regression coefficients and to lead to very unstable CI coverage. If accurate standard errors are important, MICE-MNR should not be used to impute the outcomes of MNR models, but it can be used for this purpose when unbiased estimates of the regression coefficients are the primary concern and deflated standard errors are not problematic.

Methods that work well. GLOC was clearly the strongest performer in the simulation and resampling studies. It led to unbiased fitted regression coefficients whose CIs consistently achieved nominal coverage rates for both MNR and MLR analysis models. Even though GLOC is known to be highly sensitive to imputation model complexity, the very strong performance in our studies suggests that it may be worth considering if convergence can be achieved without

placing unrealistic constraints on the model.

MICE-MNR led to unbiased regression coefficients in both MNR and MLR analysis models. Due to this lack of bias and the good CI coverage rates for low to moderate missing data rates (i.e., $PM < 0.4$), MICE-MNR is generally recommended for use when imputing the nominal predictors of MLR models. MICE-MNR can also be recommended when the incomplete nominal variables enter the analysis model as outcomes and deflated standard errors are not a concern (e.g., when hypothesis test are conducted via nested model $\Delta\chi^2$ tests). The resampling study did suggest the possibility of issues when using MICE-MNR to imputed nominal predictors in MNR analysis models. Our results do not provide enough information to make firm recommendations in this context, but researchers are encouraged to proceed with caution until future studies can more fully elucidate these potential problems.

When fitting such MNR models in situations where accurate standard error estimates are required, RANK is recommended over MICE-MNR. RANK induced minimal bias in the regression coefficients of the MNR analysis models, and it produced the second best confidence interval coverage rates after GLOC. Although RANK may lead to shrunken confidence intervals with large sample sizes (i.e., $N \geq 1000$), it still represents the best all-around performer when imputing the outcome of a MNR model.

Finally, MICE-PMM can fill a niche role when imputing the nominal predictors of a MLR model. When the predictors had a relatively large number of categories (i.e., $K \geq 5$) MICE-PMM performed well, and it maintained better confidence interval coverage rates than MICE-MNR under high missing data rates, particularly when sample sizes were small to moderate (i.e., $N \leq 500$). Thus, MICE-PMM may be preferred to MICE-MNR when the number of nominal response levels is large and the proportion of missing data is high. This finding is comforting because such circumstances will often cause convergence difficulties for MICE-MNR.

Limitations & Future Directions

The results of this study offer clear practical guidance, but this guidance must be interpreted in light of certain study limitations. From the standpoint of advising applied researchers, the largest limitation of the simulation study was the small number of variables included in the analysis models. By including the real-data resampling experiment, we have made progress towards addressing this shortcoming in the simulation study, but the results of the resampling study are not as generalizable as those from a full-fledged Monte Carlo simulation would be. Without independent replications and the ability to control the population characteristics, the resampling study can only offer supporting evidence to reinforce the findings of the simulation study (which it clearly did), but future work would benefit from exploring more complex data structures within a proper Monte Carlo simulation framework. Some key complications that should be included in future Monte Carlo simulations include: (1) incorporating categorical predictors in the analysis models, (2) examining mixes of continuous, nominal, and ordinal variables, (3) including more variables in the imputation and analysis models, and (4) examining a wider range of sample sizes.

References

Agresti, A. (2007). *An introduction to categorical data analysis.* Hoboken, NJ: Wiley-Interscience. doi: 10.1002/0470114754

Allison, P. D. (2002). *Missing data.* Thousand Oaks, CA: Sage Publications. doi: 10.4135/9780857020994.n4

Andridge, R. R., & Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, *78*(1), 40–64. doi: 10.1111/j.1751-5823.2010.00103.x

Belin, T. R., Hu, M.-Y., Young, A. S., & Grusky, O. (1999). Performance of a general location model with an ignorable missing-data assumption in a multivariate mental health services study. *Statistics in Medicine, 18*, 3123–3135. doi: 10.1002/(SICI)1097-0258(19991130)18:22<3123::AID-SIM277>3.0.CO;2-2

Bernaards, C. A., Belin, T. R., & Schafer, J. L. (2007). Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine, 26*, 1368–1382. doi: 10.1002/sim.2619

Borgoni, R., & Berrington, A. (2013). Evaluating a sequential tree-based procedure for multivariate imputation of complex missing data structures. *Quality & Quantity, 47*(4), 1991–2008. doi: 10.1007/s11135-011-9638-3

Brand, J. P. L. (1999). *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets*. (Doctoral thesis, Erasmus University, Rotterdam, The Netherlands). Retrieved from http://repub.eur.nl/pub/19790/990408_BRAND,%20Jacob%20Pieter%20Laurens.pdf

Burgette, L. F., & Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology, 172*(9), 1070–1076. doi: 10.1093/aje/kwq260

Drechsler, J., & Reiter, J. P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics and Data Analysis, 55*(12), 3232–3243. doi: 10.1016/j.csda.2011.06.006

Enders, C. K. (2010). *Applied missing data analysis.* New York: The Guilford Press.

Enders, C. K., & Keller, B. T. (2014, May). *A latent variable chained equations approach for multilevel multiple imputation.* Paper presented at the annual Modern Modeling Methods (M^3) conference, Storrs, CT.

Harrell, F. E., Jr. (2001). *Regression modeling strategies*. New York: Springer-Verlag. doi: 10.1007/978-1-4757-3462-1

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). New York: Springer. doi: 10.1007/978-0-387-84858-7

Honaker, J., & King, G. (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science, 54*(2), 561–581. doi: 10.1111/j.1540-5907.2010.00447.x

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software, 45*(7), 1–47.

Little, R. J. A., & Schluchter, M. D. (1985). Maximum likelihood estimation for mixed continuous and categorical variables with missing data. *Biometrika, 72*, 497–512. doi: 10.2307/2336722

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc. doi: 10.1002/9781119013563

Marshall, A., Altman, D. G., & Holder, R. L. (2010). Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: A resampling study. *BMC Medical Research Methodology*, *10*(1), 112–122. doi: 10.1186/1471-2288-10-112

Marshall, A., Altman, D. G., Royston, P., & Holder, R. L. (2010). Comparison of techniques for handling missing covariate data within prognostic modeling studies: A simulation study. *BMC Medical Research Methodology*, *10*(1), 7–23. doi: 10.1186/1471-2288-10-7

Olkin, I., & Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *The Annals of Mathematical Statistics, 32*(2), 448–465. doi: 10.1214/aoms/1177705052

Reiter, J. P. (2005). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics, 21*(3), 7–30.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in survey.* New York: Wiley.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* Boca Raton, FL: Chapman & Hall/CRC. doi: 10.1201/9781439821862

Schafer, J. L. (2013). Estimation/multiple imputation for mixed categorical and continuous data (R package version 1.0-8) [Computer software]. Retrieved from http://sites.stat.psu.edu/~jls/misoftwa.html

Shogren, K. A., Wehmeyer, M. L., Palmer, S. B., Rifenbark, G. G., & Little, T. D. (2015). Relationships between self-determination and postschool outcomes for youth with disabilities. *The Journal of Special Education, 48*(4), 256–267.

van Buuren, S. (2012). *Flexible imputation of missing data.* Boca Raton, FL: Chapman & Hall/CRC. doi: 10.1201/b11826

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67.

Wallace, M. L., Anderson, S. J., & Mazumdar, S. (2010). A stochastic multiple imputation algorithm for missing covariate data in tree-structured survival analysis. *Statistics in Medicine*, *29*(29), 3004–3016. doi: 10.1002/sim.4079

Wehmeyer, M. L., & Kelchner, K. (1995). *The Arc's Self-Determination Scale*. Arlington, TX: The Arc National Headquarters.

Wehmeyer, M. L., Palmer, S. B., Lee, Y., Williams-Diehm, K., & Shogren, K. (2011). A randomized-trial evaluation of the effect of Whose Future is it Anyway? On self-determination. *Career Development for Exceptional Individuals, 34*(1), 45–56.

Wehmeyer, M. L., Palmer, S. B., Shogren, K., Williams-Diehm, K., & Soukup, J. H. (2013). Establishing a causal relationship between intervention to promote self-determination and enhanced student self-determination. *The Journal of Special Education, 46*(4), 195–210.

Wu, W., Jia, F., & Enders, C. K. (in press). A comparison of imputation strategies to ordinal missing data. *Multivariate Behavioral Research.*

Yucel, R. M., He, Y., & Zaslavsky, A. M. (2008). Using calibration to improve rounding in imputation. *The American Statistician, 62*, 1–5. doi: 10.1198/000313008X300912

Yucel, R. M., He, Y., & Zaslavsky, A. M. (2011). Gaussian-based routines to impute categorical variables in health surveys. *Statistics in Medicine, 30*, 3447–3460. doi: 10.1002/sim.4355

Table 1

Percentage Relative Bias for the Multinomial Logistic Regression Slope Coefficients (Averaged Over K – 1 Coefficients)

K	PM	N = 250					N = 500					N = 1000				
		MNR	CART	PMM	RANK	GLOC	MNR	CART	PMM	RANK	GLOC	MNR	CART	PMM	RANK	GLOC
3	*0.1*	0.56	-0.54	-2.12	-0.95	0.42	0.30	-0.60	-2.63	-1.42	0.16	-0.10	-0.70	-3.05	-1.87	-0.21
3	*0.2*	0.39	-1.63	-4.55	-2.51	0.15	-0.07	-1.80	-5.35	-3.27	-0.33	-0.24	-1.94	-5.76	-3.62	-0.42
3	*0.3*	0.35	-3.03	-6.60	-3.43	0.25	0.01	-2.72	-7.55	-4.49	-0.29	-0.33	-2.71	-8.10	-5.21	-0.54
3	*0.4*	0.29	-3.74	-8.20	-3.95	0.78	-0.09	-3.26	-9.43	-5.54	-0.18	-0.46	-3.71	**-10.19**	-6.48	-0.58
3	*0.5*	-3.56	-9.84	**-11.78**	-6.02	-1.68	-2.14	**-11.54**	**-11.99**	-7.37	-1.88	-1.41	**-13.45**	**-12.20**	-8.12	-1.77
5	*0.1*	2.65	0.07	-3.14	0.00	2.02	1.24	-0.60	-4.29	-1.23	0.88	0.67	-1.01	-4.84	-1.78	0.49
5	*0.2*	2.02	-2.74	-8.69	-3.03	0.64	1.19	-2.47	-9.36	-3.67	0.45	0.81	-2.17	-9.64	-3.99	0.32
5	*0.3*	2.34	-5.75	**-12.91**	-4.97	0.27	1.49	-4.19	**-13.68**	-5.72	0.24	0.81	-3.99	**-14.08**	-6.26	-0.03
5	*0.4*	2.12	-9.74	**-16.83**	-7.42	-0.90	1.36	-7.38	**-17.65**	-8.00	-0.31	0.74	-6.32	**-18.05**	-8.42	-0.23
5	*0.5*	2.87	**-13.70**	**-19.73**	-8.88	-1.36	1.48	**-14.01**	**-20.79**	-9.82	-1.09	0.89	**-15.02**	**-21.56**	**-10.30**	-0.66
7	*0.1*	2.17	-0.32	-3.95	-0.52	1.35	1.37	-0.75	-4.76	-1.41	0.84	0.75	-0.66	-5.21	-1.94	0.50
7	*0.2*	2.98	-2.33	-9.08	-2.32	1.33	2.08	-2.13	-9.78	-3.42	1.10	1.11	-2.14	**-10.28**	-4.24	0.62
7	*0.3*	4.06	-4.17	**-13.91**	-4.48	1.22	2.80	-3.55	**-14.26**	-5.44	1.28	1.34	-3.73	**-15.04**	-6.66	0.51
7	*0.4*	5.14	-7.22	**-18.23**	-6.18	1.16	3.11	-6.27	**-18.87**	-7.69	0.83	1.56	-5.84	**-19.46**	-8.99	0.28
7	*0.5*	6.84	**-11.97**	**-22.58**	-8.17	1.74	3.58	**-12.43**	**-22.97**	-9.77	0.63	1.40	**-13.16**	**-23.63**	**-11.31**	-0.26
10	*0.1*	5.07	2.59	-1.80	1.65	4.18	2.42	0.25	-3.81	-0.71	1.92	1.78	0.23	-4.41	-1.37	1.46
10	*0.2*	6.51	1.29	-7.14	-0.62	4.42	3.15	-1.04	-9.08	-3.00	2.03	2.30	-0.96	-9.54	-3.72	1.71
10	*0.3*	8.01	-1.19	**-12.10**	-2.94	4.58	3.46	-2.95	**-14.29**	-5.87	1.79	2.64	-2.80	**-14.71**	-6.41	1.61
10	*0.4*	8.74	-4.39	**-17.74**	-4.79	3.61	3.54	-6.65	**-19.30**	-8.60	1.04	2.62	-4.57	**-19.40**	-9.26	1.12
10	*0.5*	**10.28**	**-10.90**	**-22.90**	-7.89	1.85	3.58	**-11.69**	**-23.86**	**-11.49**	-0.15	2.44	**-11.33**	**-23.93**	**-12.24**	0.27

Note: K = Number of multinomial response categories, PM = proportion of missing data, MNR = MICE-MNR imputation, CART = MICE-CART imputation, PMM = MICE-PMM imputation, RANK = Allison (2002) ranking approach, GLOC = general location model imputation. Bold-faced entries exceed the threshold of unacceptably large bias (i.e.,| PRB| > 10).

Table 2

Confidence Interval Coverage for the Multinomial Logistic Regression Slope Coefficients (Averaged Over K – 1 Coefficients)

K	PM	N = 250					N = 500					N = 1000				
		MNR	CART	PMM	RANK	GLOC	MNR	CART	PMM	RANK	GLOC	MNR	CART	PMM	RANK	GLOC
3	*0.1*	0.945	0.935	0.948	0.949	0.951	0.939	0.926	0.933	0.945	0.943	0.941	0.920	0.928	0.937	0.945
3	*0.2*	0.933	0.921	0.940	0.953	0.949	0.941	0.927	**0.911**	0.948	0.951	0.937	**0.911**	**0.884**	0.935	0.951
3	*0.3*	0.931	**0.913**	0.936	0.955	0.957	0.923	**0.903**	**0.885**	0.949	0.958	**0.917**	**0.881**	**0.808**	**0.914**	0.953
3	*0.4*	**0.891**	**0.901**	0.925	0.946	0.951	**0.905**	**0.907**	**0.874**	0.936	0.961	**0.890**	**0.891**	**0.747**	**0.909**	0.950
3	*0.5*	**0.873**	0.926	0.923	0.944	0.939	**0.859**	**0.919**	**0.849**	0.921	0.950	**0.855**	**0.909**	**0.706**	**0.869**	0.947
5	*0.1*	0.952	0.957	0.965	0.957	0.953	0.948	0.946	0.953	0.955	0.950	0.949	0.941	0.942	0.953	0.951
5	*0.2*	0.947	0.945	0.956	0.958	0.955	0.940	0.937	0.934	0.956	0.953	0.936	**0.917**	**0.878**	0.944	0.946
5	*0.3*	0.924	0.932	0.948	0.956	0.951	0.923	0.924	**0.898**	0.945	0.952	0.924	**0.907**	**0.779**	0.933	0.944
5	*0.4*	**0.910**	0.926	0.936	0.957	0.955	**0.909**	**0.910**	**0.838**	0.945	0.953	**0.898**	**0.892**	**0.637**	**0.914**	0.948
5	*0.5*	**0.905**	0.939	0.928	0.956	0.950	**0.895**	**0.916**	**0.779**	0.939	0.952	**0.876**	**0.880**	**0.512**	**0.886**	0.945
7	*0.1*	0.945	0.947	0.955	0.954	0.949	0.944	0.945	0.956	0.957	0.948	0.952	0.953	0.953	0.961	0.957
7	*0.2*	0.936	0.942	0.955	0.953	0.947	0.938	0.935	0.941	0.955	0.947	0.943	0.937	**0.915**	0.953	0.948
7	*0.3*	0.929	0.935	0.952	0.957	0.949	0.931	0.931	0.927	0.955	0.949	0.931	**0.917**	**0.846**	0.946	0.952
7	*0.4*	0.922	0.933	0.949	0.958	0.951	**0.913**	0.923	**0.901**	0.951	0.952	0.921	**0.915**	**0.723**	0.931	0.951
7	*0.5*	**0.902**	0.938	0.940	0.957	0.939	**0.901**	0.930	**0.854**	0.949	0.955	**0.898**	**0.909**	**0.584**	**0.912**	0.949
10	*0.1*	0.935	0.941	0.958	0.949	0.939	0.944	0.941	0.959	0.955	0.947	0.945	0.947	0.958	0.955	0.950
10	*0.2*	0.929	0.938	0.963	0.956	0.944	0.940	0.945	0.960	0.959	0.949	0.940	0.941	0.946	0.960	0.951
10	*0.3*	**0.918**	0.940	0.966	0.959	0.943	0.932	0.943	0.949	0.964	0.952	0.932	0.933	**0.895**	0.954	0.951
10	*0.4*	**0.911**	0.941	0.965	0.964	0.946	**0.919**	0.936	0.932	0.962	0.951	0.922	0.931	**0.835**	0.945	0.952
10	*0.5*	**0.908**	0.949	0.959	0.965	0.940	**0.908**	0.948	**0.909**	0.955	0.949	**0.912**	0.943	**0.731**	0.936	0.953

Note: K = Number of multinomial response categories, PM = proportion of missing data, MNR = MICE-MNR imputation, CART = MICE-CART imputation, PMM = MICE-PMM imputation, RANK = Allison (2002) ranking approach, GLOC = general location model imputation. Bold-faced entries have CI coverage rates lower than 92%.

Table 3

Percentage Relative Bias for the Multiple Linear Regression Slope Coefficients (Averaged Over K – 1 Coefficients)

K	PM	N = 250					N = 500					N = 1000				
		MNR	CART	PMM	RANK	GLOC	MNR	CART	PMM	RANK	GLOC	MNR	CART	PMM	RANK	GLOC
3	*0.1*	0.33	-0.42	-3.69	-1.99	-0.09	-0.56	-1.08	-4.28	-2.90	-0.78	-0.33	-1.03	-3.95	-2.67	-0.43
3	*0.2*	0.47	-0.53	-7.63	-4.25	-0.36	-0.47	-1.38	-7.82	-5.05	-0.91	-0.23	-1.62	-7.48	-4.90	-0.49
3	*0.3*	-0.02	-2.57	**-12.31**	-6.91	-1.45	-0.61	-2.29	**-11.89**	-7.49	-1.40	-0.35	-2.55	**-11.35**	-7.31	-0.79
3	*0.4*	0.37	-2.61	**-16.92**	-8.83	-1.75	-0.68	-4.43	**-16.26**	-9.88	-1.83	-0.25	-3.74	**-15.27**	-9.49	-0.90
3	*0.5*	-0.26	**-13.22**	**-25.59**	**-10.66**	-3.93	-1.09	**-14.67**	**-24.31**	**-12.00**	-3.21	-0.32	**-15.74**	**-22.57**	**-11.57**	-1.51
5	*0.1*	-0.89	-1.00	-1.67	-3.91	-0.92	-0.51	-0.55	-1.17	-3.65	-0.53	-0.26	-0.23	-0.81	-3.51	-0.27
5	*0.2*	-0.63	-0.83	-2.33	-6.75	-0.71	-0.43	-0.32	-1.71	-6.77	-0.40	-0.05	0.25	-1.19	-6.57	-0.03
5	*0.3*	-0.87	-0.93	-3.53	-9.83	-0.83	-0.60	-0.19	-2.58	-9.94	-0.52	-0.20	-0.06	-2.00	-9.95	-0.14
5	*0.4*	-1.44	-2.22	-5.28	**-12.98**	-1.14	-0.94	-1.23	-3.79	**-13.22**	-0.69	-0.34	-0.56	-2.88	**-13.19**	-0.25
5	*0.5*	-2.39	-5.90	-9.19	**-15.95**	-1.00	-1.32	-4.31	-6.66	**-16.27**	-0.21	-0.47	-3.49	-5.11	**-16.34**	0.07
7	*0.1*	0.36	0.52	0.13	-2.88	0.52	-0.15	-0.10	-0.33	-3.55	-0.06	-0.21	-0.13	-0.44	-3.73	-0.19
7	*0.2*	0.31	0.70	-0.08	-6.06	0.75	-0.10	0.07	-0.52	-6.83	0.08	-0.19	-0.13	-0.67	-7.18	-0.10
7	*0.3*	-0.34	0.31	-0.91	-9.63	0.53	-0.36	-0.03	-0.93	**-10.34**	0.08	-0.42	-0.20	-1.04	**-10.80**	-0.21
7	*0.4*	-1.00	0.30	-1.68	**-12.62**	0.67	-0.42	0.35	-1.24	**-13.56**	0.37	-0.55	-0.17	-1.34	**-14.32**	-0.19
7	*0.5*	-2.77	-1.79	-4.04	**-16.16**	0.66	-1.17	-0.96	-2.47	**-16.98**	0.50	-0.90	-0.31	-2.15	**-17.89**	-0.04
10	*0.1*	-0.09	0.00	-0.01	-3.51	0.18	0.05	0.11	0.05	-3.61	0.18	0.03	0.09	0.00	-3.76	0.10
10	*0.2*	-0.47	0.06	-0.14	-7.02	0.22	-0.12	0.13	0.02	-7.28	0.22	-0.05	0.09	-0.05	-7.55	0.14
10	*0.3*	-1.04	-0.20	-0.36	**-10.46**	0.19	-0.45	0.02	-0.16	**-11.00**	0.19	-0.29	-0.02	-0.17	**-11.42**	0.10
10	*0.4*	-2.21	-0.57	-0.79	**-13.78**	0.20	-0.91	-0.12	-0.41	**-14.64**	0.21	-0.44	0.07	-0.25	**-15.13**	0.18
10	*0.5*	-4.48	-1.19	-1.21	**-16.89**	0.09	-2.05	-0.80	-0.64	**-18.12**	0.32	-1.03	-0.36	-0.37	**-18.91**	0.23

Note: K = Number of multinomial response categories, PM = proportion of missing data, MNR = MICE-MNR imputation, CART = MICE-CART imputation, PMM = MICE-PMM imputation, RANK = Allison (2002) ranking approach, GLOC = general location model imputation. Bold-faced entries exceed the threshold of unacceptably large bias (i.e., |PRB| > 10).

Table 4

Confidence Interval Coverage for the Multiple Linear Regression Slope Coefficients (Averaged Over K – 1 Coefficients)

K	PM	N = 250					N = 500					N = 1000				
		MNR	CART	PMM	RANK	GLOC	MNR	CART	PMM	RANK	GLOC	MNR	CART	PMM	RANK	GLOC
3	*0.1*	0.952	0.958	0.962	0.968	0.957	0.952	0.944	0.955	0.962	0.953	0.938	0.933	0.935	0.950	0.938
3	*0.2*	0.952	0.952	0.952	0.972	0.961	0.954	0.937	0.947	0.970	0.958	0.938	**0.901**	**0.896**	0.957	0.946
3	*0.3*	0.937	0.924	0.940	0.961	0.950	0.941	**0.916**	**0.916**	0.969	0.947	0.926	**0.891**	**0.856**	0.961	0.939
3	*0.4*	0.924	**0.919**	0.923	0.975	0.956	0.938	**0.913**	**0.864**	0.972	0.958	0.926	**0.895**	**0.767**	0.958	0.948
3	*0.5*	0.930	0.969	0.931	0.979	0.960	0.939	0.977	**0.850**	0.977	0.961	0.926	0.980	**0.709**	0.952	0.956
5	*0.1*	0.951	0.954	0.954	0.971	0.954	0.950	0.950	0.953	0.963	0.953	0.950	0.945	0.951	0.960	0.952
5	*0.2*	0.948	0.941	0.954	0.977	0.958	0.952	0.946	0.948	0.967	0.959	0.942	0.929	0.942	0.951	0.952
5	*0.3*	0.940	0.938	0.947	0.975	0.956	0.948	0.932	0.940	0.963	0.959	0.939	**0.917**	0.926	0.939	0.954
5	*0.4*	0.930	0.920	0.942	0.977	0.959	0.926	**0.918**	0.924	0.951	0.955	0.934	**0.911**	**0.903**	0.923	0.954
5	*0.5*	0.923	0.947	0.946	0.979	0.961	0.920	0.971	**0.913**	0.945	0.957	0.926	0.984	**0.895**	**0.895**	0.964
7	*0.1*	0.949	0.947	0.957	0.975	0.950	0.936	0.936	0.941	0.965	0.938	0.953	0.947	0.951	0.962	0.956
7	*0.2*	0.946	0.943	0.954	0.982	0.951	0.928	0.922	0.936	0.969	0.940	0.945	0.933	0.936	0.952	0.950
7	*0.3*	0.933	0.929	0.947	0.984	0.952	0.927	**0.917**	0.932	0.967	0.940	0.933	**0.915**	0.923	0.924	0.949
7	*0.4*	**0.918**	**0.909**	0.939	0.980	0.948	**0.917**	**0.915**	0.928	0.956	0.941	0.922	**0.911**	**0.907**	**0.889**	0.953
7	*0.5*	**0.915**	0.934	0.943	0.978	0.948	**0.897**	0.948	0.922	0.945	0.948	**0.907**	0.970	**0.896**	**0.849**	0.942
10	*0.1*	0.957	0.949	0.958	0.991	0.957	0.947	0.945	0.951	0.987	0.949	0.952	0.941	0.950	0.982	0.951
10	*0.2*	0.949	0.940	0.952	0.997	0.953	0.949	0.940	0.956	0.989	0.956	0.947	0.933	0.949	0.966	0.954
10	*0.3*	0.945	0.938	0.949	0.997	0.952	0.943	0.931	0.945	0.983	0.957	0.944	**0.918**	0.945	0.935	0.955
10	*0.4*	0.931	**0.916**	0.939	0.990	0.949	0.929	0.929	0.938	0.975	0.954	0.928	**0.914**	0.929	**0.894**	0.954
10	*0.5*	0.942	0.936	0.942	0.993	0.949	0.921	0.960	0.934	0.963	0.956	**0.915**	0.975	**0.919**	**0.857**	0.953

Note: K = Number of multinomial response categories, PM = proportion of missing data, MNR = MICE-MNR imputation, CART = MICE-CART imputation, PMM = MICE-PMM imputation, RANK = Allison (2002) ranking approach, GLOC = general location model imputation. Bold-faced entries have CI coverage rates lower than 92%, and shaded entries have CI coverage rates higher than 98%.

Table 5

Summary Statistics of Population Data for Real-Data Resampling Experiment

Categorical Variables			
Variable	*Level*	*Count*	*Proportion*
Gender	Male	324	0.61
	Female	211	0.39
Disability	ID	185	0.35
	Autism	35	0.07
	BD	56	0.10
	LD	229	0.43
	ADHD	30	0.06
Course	Vocational	221	0.41
	College Prep	119	0.22
	Life Skills	109	0.20
	General Diploma	86	0.16
Continuous Variables			
	Variable	*Mean*	*SD*
	SDS	2.64	0.55
	Age	16.98	1.41

Note: ID = Intellectual disability, BD = Behavioral disturbance, LD = Learning disability, ADHD = Attention deficit hyperactivity disorder, SDS = Self-determination scale

Table 6

Parameter estimates produced by fitting the real-data resampling study's analysis models to the population data

Effect	*Coefficient*	*SE*
Multiple Linear Regression		
Intercept	2.1	0.32
Autism	-0.16	0.1
BD	0.21	0.08
LD	0.25	0.06
ADHD	0.33	0.11
Gender	0.12	0.05
Age	0.02	0.02
Multinomial Logistic Regression		
Intercept: Vocational	-0.1	1.92
Intercept: College Prep	10.93	2.61
Intercept: Life Skills	-4.02	2.19
SDS: Vocational	-0.32	0.25
SDS: College Prep	0.23	0.3
SDS: Life Skills	-0.74	0.3
Age: Vocational	0.1	0.11
Age: College Prep	-0.76	0.15
Age: Life Skills	0.4	0.12
Autism: Vocational	-0.25	0.53
Autism: College Prep	1.38	0.68
Autism: Life Skills	-1.43	0.6
BD: Vocational	0.3	0.45
BD: College Prep	1.42	0.56
BD: Life Skills	-2.99	1.08
LD: Vocational	0.54	0.31
LD: College Prep	1.42	0.45
LD: Life Skills	-1.74	0.42
ADHD: Vocational	-0.34	0.6
ADHD: College Prep	1.62	0.64
ADHD: Life Skills	-2.41	1.12

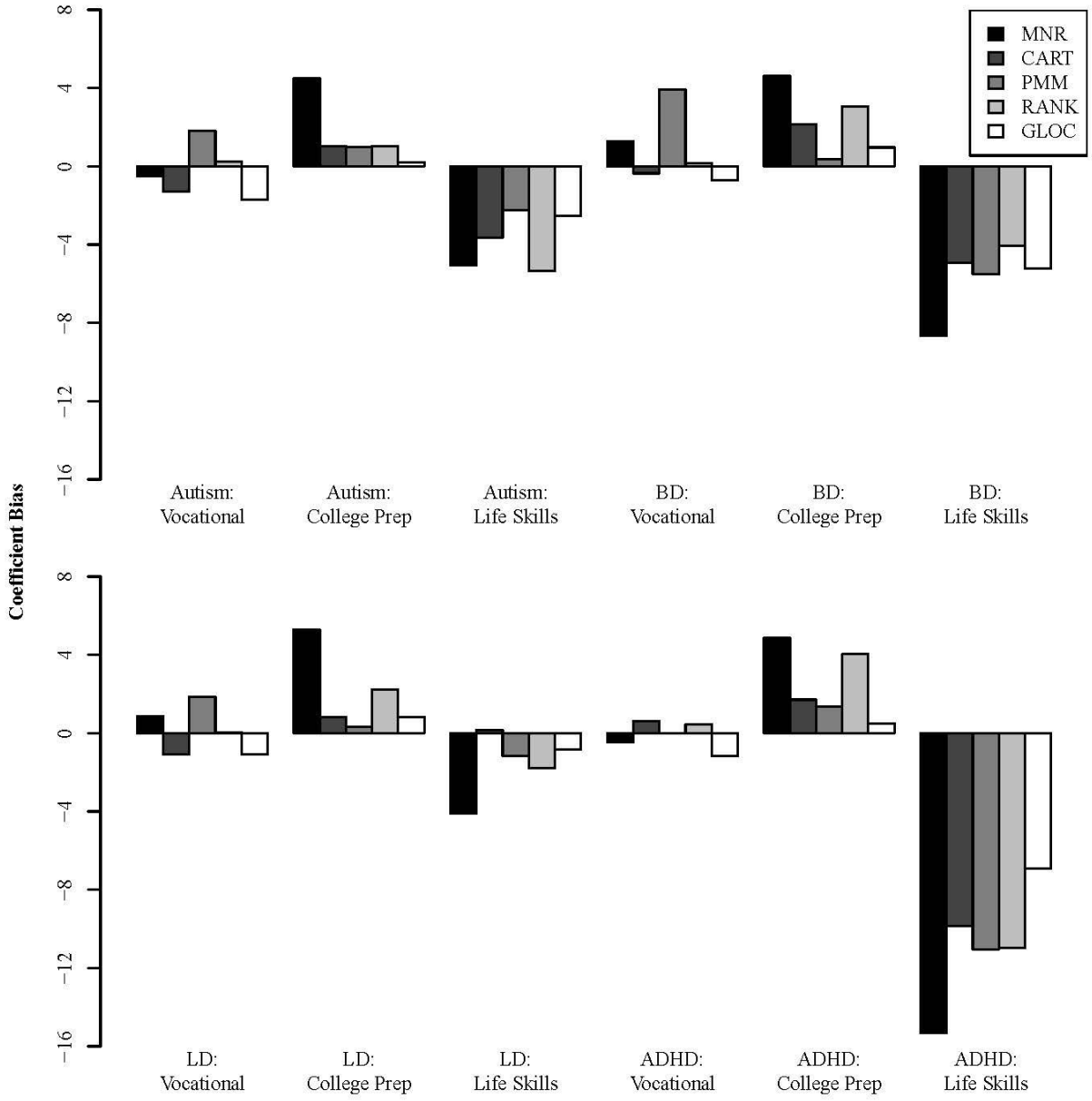

Figure 1

Bias in Disability coefficients in the real-data resampling experiment's multinomial logistic regression model

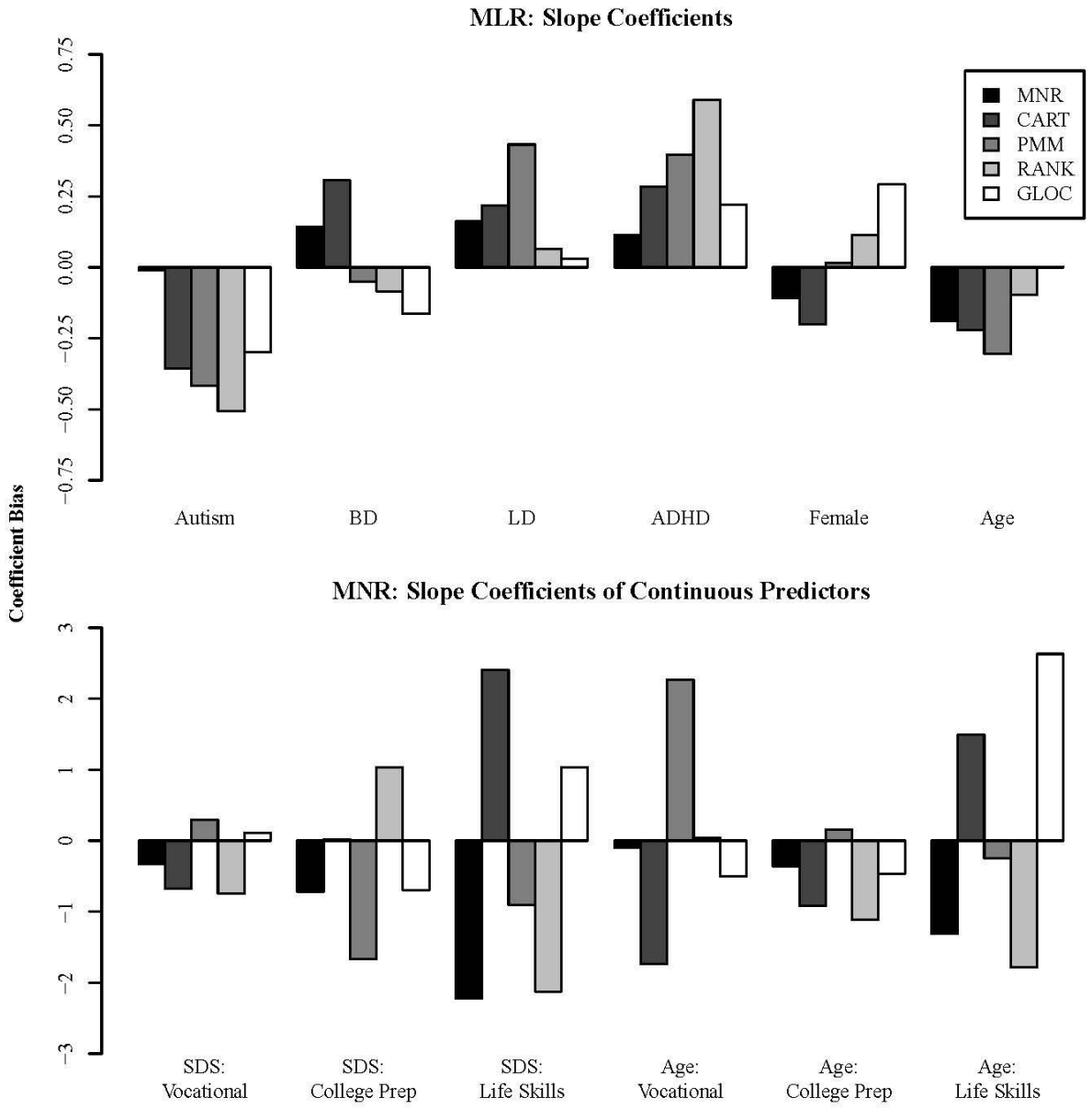

Figure 2

Bias in the coefficients of SDS and Age in the real-data resampling experiment's multinomial logistic regression model and in all predictors of the multiple linear regression analysis model