Addressing Nonresponse for Categorical Data Items in Complex Surveys Using Full Information Maximum Likelihood

Susan L. Edwards[1], Marcus E. Berzofsky[1], Paul P. Biemer[1]
[1]RTI International, 3040 Cornwallis Rd., RTP, NC 27709

Abstract
This paper presents a comparison of two full information maximum likelihood (FIML) approaches for addressing nonresponse for categorical data items in complex surveys. Item nonresponse is an issue researchers using public use files for survey data often encounter. There are several techniques for dealing with item nonresponse in categorical data. A relatively new technique for handling missing data is FIML which incorporates all response patterns during the estimation process rather than ignoring cases with missing values. In 1982, Fuchs proposed a FIML method to handle nonresponse missing at random (MAR); Fay (1986) expanded the FIML method to handle nonresponse for both MAR and not missing at random (MNAR) by incorporating item nonresponse indicators into the model. The National Survey of Drug Use and Health (NSDUH) is an annual national and state level survey that collects information on the use of tobacco, alcohol, illicit drugs and mental health in the U.S. Using data from the NSDUH, these two FIML approaches are applied to a model of opioid use and depression using the software package LatentGOLD. The results of applying FIML to both independent and dependent variables will be compared.

Key Words: full information maximum likelihood, item nonresponse; NMAR; imputation

1. Background

1.1 Background
In survey data, missing data occurs when a respondent does not participate in the entire survey (unit nonresponse) or when the respondent does not respond to a survey question (item nonresponse). Item nonresponse can occur for a multitude of reasons including missing responses, inconsistent or invalid responses, and violations of skip patterns. (Lohr 2010; CBHSQ 2014)

Nonresponse is often classified according to one of three missing data mechanisms: missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). These missing data mechanisms define a mathematical relationship between the survey variables and the probability of nonresponse (Enders 2010). Originally defined by Rubin (1976), MCAR occurs when the missing does not depend on either the observed or unobserved data; MAR is less restrictive in that the missing depends only on the observed data; MNAR is the least restrictive mechanism where the missing depends on the unobserved data. Since failing to address missing data of any

classification can lead to biased and/or inefficient inference, many techniques have been developed to handle missing data.

Traditional methods for handling item nonresponse in categorical data analysis include listwise deletion, mean imputation, and hot deck imputation. Listwise deletion is sometimes referred to as complete case analysis and is implemented by default in several software packages. When item nonresponse is handled through listwise deletion, biased parameter estimates may result if the missing data does not follow a MCAR response mechanism and due to removing observations from the analysis, listwise deletion will always result in some loss of power (Graham 2009).

Imputation refers to the process of replacing missing data with plausible values. A benefit of imputation methods is that data are imputed prior to analysis and therefore a complete dataset is available for analyses. Mean imputation in categorical data analysis occurs by equating the effect for the missing value category to the average effect for the observed categories (Vermunt 2013). This is done by randomly assigning values to the missing observations based on the unweighted distribution of the observed data for that categorical variable and is sometimes suitable when the missing data response mechanism is MAR.

Another imputation method suitable for MAR missing data is hot deck imputation. Hot deck imputation assigns plausible values to respondents with missing data based on observed values from respondents with complete data. Hot deck methods preserve the univariate distribution of the data, but are not well suited for estimating measures of association and can yield biased estimates for correlations and regression coefficients (Enders 2010). When imputed values from these regression-based techniques are used in analysis, the resulting estimates are unbiased but the estimated variance is smaller than the true variance (Lohr 2010).

Two relatively newer methods for handling missing data are multiple imputation (MI) and full information maximum likelihood (FIML); asymptotically these methods yield similar results. A benefit of these methods is that they can restore the error variance lost from regression-based single imputation methods and maintain unbiased inferences on the estimates (Graham 2009; Little 2014). By default, these techniques assume a MAR response mechanism.

1.2 Full Information Maximum Likelihood (FIML)

When maximum likelihood estimation is a valid statistical technique, FIML methods can be used to fit a single hypothesized model. While not an imputation method, FIML makes use of all available data to maximize the log-likelihood by allowing some respondents to contribute more information than others (Enders 2010). For example, consider a dataset with four continuous variables observed for N respondents. The log-likelihood for respondent i when all four variables are observed is equal to

$$\log \mathcal{L}_i = \log P(x) f(y_{i1}|x) f(y_{i2}|x) f(y_{i3}|x) f(y_{i4}|x). \quad (1.1)$$

When the fourth variable is missing for respondent i, using a FIML approach the log-likelihood is reduced to include only the observed variables;

$$\log \mathcal{L}_i = \log P(x) f(y_{i1}|x) f(y_{i2}|x) f(y_{i3}|x). \quad (1.2)$$

The overall log-likelihood is then obtained by summing over the *N* respondents' likelihood functions. Therefore FIML techniques make use of all available information when estimating each case to account for item nonresponse.

For categorical data analysis, FIML approaches are similar to those developed to handle continuous data - partially observed information is used when fitting log-linear models. Take for example the following illustration presented by Vermunt (1997), suppose we have a dataset with four observed variables – *A, B, C, D* – with item nonresponse for *C* or *D*. In this case, there are four potential subgroups based on the potential nonresponse patterns. Subgroup *ABCD* consists of respondents with complete responses to all questions; subgroup *ABC* consists of respondents with complete responses to questions *A*, *B*, and *C* and nonresponse for question *D*; subgroup *ABD* consists of respondents with nonresponse for question *C* only; subgroup *AB* consists of respondents that responded to questions *A* and *B* but not questions *C* and *D*.

Assuming a multinomial sampling scheme for these four variables, estimation of a log-linear model with partially observed data involves maximizing the following incomplete data likelihood

$$\begin{aligned}\log \mathcal{L}_{(\pi,\theta)} = &\sum_{abcd} n_{abcd} \log \pi_{abcd}\theta_{ABCD|abcd} + \sum_{abc} n_{abc} \log \sum_{d} \pi_{abcd}\theta_{ABC|abcd} \\ &+ \sum_{abd} n_{abd} \log \sum_{c} \pi_{abcd}\theta_{ABD|abcd} \\ &+ \sum_{ab} n_{ab} \log \sum_{cd} \pi_{abcd}\theta_{AB|abcd}\end{aligned} \qquad (1.3)$$

where

$$\theta_{ABCD|abcd} + \theta_{ABC|abcd} + \theta_{ABD|abcd} + \theta_{AB|abcd} = 1.$$

Here n_{abcd}, n_{abc}, n_{abd}, and n_{ab} denotes the observed frequencies for each of the four subgroups. The structural probabilities are represented by πs and represent the probability of belonging to cell *A, B, C, D* in the joint distribution of the observed and unobserved variables. The response probabilities are represented by θs and contain the parameters associated with the response mechanism. Such that $\theta_{ABCD|abcd}, \theta_{ABC|abcd}, \theta_{ABD|abcd}$, and $\theta_{AB|abcd}$ denote the conditional probability of belonging to subgroup *ABCD*, *ABC*, *ABD*, and *AB*, respectively given that $A = a, B = b, C = c,$ and $D = d$.

Assuming the data follow a MAR (or MCAR) response mechanism, then the response probabilities are independent of the missing variables in the subgroup. In which case the likelihood can be factored into two components – one for the log-linear parameters and another for the response mechanism:

$$\log \mathcal{L}_{(\pi,\theta)} = \log \mathcal{L}_{(\pi)} + \log \mathcal{L}_{(\theta)} \qquad (1.4)$$

where

$$\begin{aligned}\log \mathcal{L}_{(\pi)} = &\sum_{abcd} n_{abcd} \log \pi_{abcd} + \sum_{abc} n_{abc} \log \sum_{d} \pi_{abcd} \\ &+ \sum_{abd} n_{abd} \log \sum_{c} \pi_{abcd} + \sum_{ab} n_{ab} \log \sum_{cd} \pi_{abcd}\end{aligned} \qquad (1.5)$$

and

$$\log \mathcal{L}_{(\theta)} = \sum_{abcd} n_{abcd} \log \theta_{ABCD|abcd} + \sum_{abc} n_{abc} \log \theta_{ABC|abc} + \sum_{abd} n_{abd} \log \theta_{ABD|abd} + \sum_{ab} n_{ab} \log \theta_{AB|ab} \,. \quad (1.6)$$

Under the MAR assumption, only the structural parameters need to be estimated since these two components can be maximized separately and the response mechanism is ignorable. Under the more restrictive MCAR assumption, the θs are assumed to be equal for every value of *A*, *B*, *C*, and *D*. (Vermunt 1997)

1.2.1 Fuchs Approach – Saturated MAR

In 1982, Fuchs extended the methodology of FIML to estimate the parameters of a saturated log-linear model using the Estimation-Maximization (EM) algorithm when item nonresponse is ignorable. In the estimation step, the conditional expected complete data likelihood for the structural probabilities is computed. Since the response mechanism is ignorable under the MAR assumption, only the following log-likelihood is considered:

$$\log \mathcal{L}^*_{(\pi)} = \sum_{abcd} \hat{n}_{abcd} \log \hat{\pi}_{abcd} \quad (1.7)$$

where

$$\hat{n}_{abcd} = n_{abcd} + n_{abc}\hat{\pi}_{d|abc} + n_{abd}\hat{\pi}_{c|abd} + n_{ab}\hat{\pi}_{cd|ab}.$$

In the maximization step, the likelihood from the estimation step is maximized to produce new estimates of $\hat{\pi}_{abcd}$ treating $\hat{n}_{abcd}$ as though they were observed frequencies in the next iteration of the estimation step. The chi-squared statistic of model fit resulting from the saturated MAR model jointly tests the MCAR assumption and the model fit. When the nonresponse mechanism is MNAR, this approach is not appropriate.

1.2.2 Fay Approach – Response Indicator

Fay (1986) further extended the methodology to model the response mechanism by using recursive causal log-linear models which treat the response indicators as dependent variables; thus providing a FIML technique that applies to data with non-ignorable item nonresponse and ignorable item nonresponse. In Fay's approach, response indicators are created for all variables with partially observed data and two log-linear models are fit. Fay stated that response indicators should never appear as independent variables in the logit equation for the structural model. First the structural variables are modeled and then the response mechanism is modeled by creating response indicators for all variables with item nonresponse.

In our four variable example from earlier, two response indicators are created for variables *C* and *D*. Response indicator *R* represents nonresponse for variable *C* and response indicator *S* represents nonresponse for variable *D*. *R* and *S* take the value of 1 if the variable of interest is observed and 2 if the variable of interest is missing.

$$r = \begin{cases} 1 & C \text{ observed} \\ 2 & C \text{ not observed} \end{cases} \text{ and } s = \begin{cases} 1 & D \text{ observed} \\ 2 & D \text{ not observed} \end{cases} \quad (1.8)$$

The inclusion of these two additional variables modifies the expected probabilities to:

$$\pi_{abcdrs} = \pi_{abcd}\pi_{rs|abcd} \quad (1.9)$$

where

$$\pi_{rs|abcd} = \Pr(R = r, S = s | A = a, B = b, C = c, D = d). \quad (1.10)$$

The $\pi_{rs|abcd}$ response probabilities can be modeled to represent MCAR, MAR, and MNAR response patterns by altering the dependencies of the response indicators. To model a MCAR response pattern then an interaction term between *R* and *S* is required. A MAR response pattern can be modeled using any variables without item nonresponse. When the response probabilities are influenced by their creation variable ($\pi_{rs|cd}$) then the response pattern is MNAR. Below are possible logit models for $\pi_{rs|abcd}$ under each of these mechanisms.

$$\begin{aligned}
&\text{MCAR:} && \pi_{rs|abcd} = \pi_{rs} = \frac{\exp(u_r^R + u_s^S + u_{rs}^{RS})}{\sum_{rs}\exp(u_r^R + u_s^S + u_{rs}^{RS})} \\
&\text{MAR:} && \pi_{rs|abcd} = \pi_{rs|ab} = \frac{\exp(u_r^R + u_s^S + u_{rs}^{RS} + u_{ra}^{RA} + u_{rb}^{RB} + u_{sa}^{SA} + u_{sb}^{SB})}{\sum_{rs}\exp(u_r^R + u_s^S + u_{rs}^{RS} + u_{ra}^{RA} + u_{rb}^{RB} + u_{sa}^{SA} + u_{sb}^{SB})} \\
&\text{MNAR:} && \pi_{rs|abcd} = \pi_{rs|cd} = \frac{\exp(u_r^R + u_s^S + u_{rs}^{RS} + u_{rd}^{RD} + u_{sc}^{SC})}{\sum_{rs}\exp(u_r^R + u_s^S + u_{rs}^{RS} + u_{rd}^{RD} + u_{sc}^{SC})}.
\end{aligned} \quad (1.11)$$

It should be noted that $\pi_{rs|abcd}$ can be further divided, such as $\pi_{r|abcd}\pi_{s|abcdr}$, in cases where such a response structure is expected as in the case with panel data. With the inclusion of the response indicators, the log-likelihood to be maximized takes on the form:

$$\begin{aligned}
\log \mathcal{L}_{(\pi)} = &\sum_{abcd} n_{abcd} \log \pi_{abcd}\pi_{11|abcd} + \sum_{abc} n_{abc} \log \sum_{d} \pi_{abcd}\pi_{12|abcd} \\
&+ \sum_{abd} n_{abd} \log \sum_{c} \pi_{abcd}\pi_{21|abcd} \\
&+ \sum_{ab} n_{ab} \log \sum_{cd} \pi_{abcd}\pi_{22|abcd}
\end{aligned} \quad (1.12)$$

where

$$\begin{aligned}
\pi_{11|abcd} &= \theta_{ABCD|abcd} \\
\pi_{12|abcd} &= \theta_{ABC|abcd} \\
\pi_{21|abcd} &= \theta_{ABD|abcd} \\
\pi_{22|abcd} &= \theta_{AB|abcd}.
\end{aligned}$$

In the estimation step, estimates of the observed frequencies are calculated by:

$$\begin{aligned}
\hat{n}_{abcd11} &= n_{abcd} \\
\hat{n}_{abcd12} &= n_{abc}\hat{\pi}_{d|abc12} \\
\hat{n}_{abcd21} &= n_{abd}\hat{\pi}_{c|abd21} \\
\hat{n}_{abcd22} &= n_{ab}\hat{\pi}_{cd|ab22}.
\end{aligned}$$

Note that these posterior probabilities are subgroup specific since each subgroup has different missing response patterns. Estimates of the observed frequencies are then used in the next maximization step to maximize the following log-likelihood:

$$\log \mathcal{L}^*_{(\pi)} = \sum_{abcdrs} \hat{n}_{abcdrs} \log \hat{\pi}_{abcd}\hat{\pi}_{rs|abcd}. \quad (1.13)$$

1.3 Purpose

This paper demonstrates how to apply Fay's FIML approach in LG. While applying Fuchs' approach in LG is somewhat straight forward, especially on the dependent

variable, Fay's method was never incorporated as an option in the LG software. However Fay's approach can be used in LG by an innovative and not well known work around using the syntax feature. Using a published model of youth opioid use from the National Survey of Drug Use and Health (NSDUH) the differences in estimates produced using FIML when MNAR and MAR response mechanisms are assumed are compared to determine if there are disadvantages to treating item nonresponse as MNAR when MAR can produce similar estimates.

The methods section contains a brief summary of current software packages capable of implementing FIML techniques on categorical data; along with the theory and LG syntax for fitting FIML models using Fuchs' and Fay's approaches under each response mechanism. In the results section, an overview of the NSDUH study with the model used by NSDUH analyst to explore past year opioid use among youths in the United States is presented with odds ratios and 95% confidence intervals from each of the missing data models fit with LG. The discussion section confirms the ability of Fay's approach to model MAR and MNAR response mechanisms, summarizes the differences between traditional methods and these newer FIML approaches and assesses the "cost" of these newer methods.

2. Methods – FIML Implementation

2.1 Software

Three programs that apply FIML approaches to handle missing data in structural equation modeling and latent class analysis are LEM, M-Plus, and LG. A brief comparison of how these programs handle nonresponse for categorical variables is presented in Table 1.

The freeware LEM package was developed by Jeroen K. Vermunt in 1997 to handle analysis of nominal, ordinal, and interval level categorical data. Parameters in LEM are estimated by means of maximum likelihood. Missing data are addressed through Fay's FIML approach. LEM can also use extended memory above 4GB of RAM. Despite these advantages, LEM is unable to account for complex survey designs during model estimation.

Another software program designed to handle analysis of continuous, ordinal, nominal, and count data is Mplus. Mplus was developed by Muthén & Muthén in 1998 as a latent variable modeling program; the most current version is 7.3. Mplus is able to account for complex survey designs during model estimation. Missing data are addressed through FIML or MI. The maximum likelihood estimation algorithm implemented in Mplus is based on the maximum-likelihood estimation of the saturated correlates model proposed by Graham (2003). Mplus appears to handle MNAR missing mechanisms for continuous variables in a FIML construct; for categorical MNAR response, MI is the suggested approach (Asparouhov & Muthén 2010).

The statistical software package LG was developed by Jeroen K. Vermunt to handle the analysis of categorical data; the most current version is 5.0. Missing data are addressed through Fuchs' FIML approach for dependent variables by default. The most current version of LG executes on a 32-bit platform and is able to access at most 4GB of RAM during model estimation. LG is capable of accounting for complex survey designs during model estimation through a survey option.

Due to the focus of LG on categorical data, its ability to account for complex survey designs, and its application of FIML, this paper focuses on handling missing data in the LG 5.0 software package with particular focus on highlighting a technique to use Fuchs' and Fay's FIML approaches to address item nonresponse through the syntax module. These features of LG have the potential to make the ability to apply FIML to variables with MNAR nonresponse more accessible to researchers using complex survey data.

Table 1: Comparison of Software Packages

Software Package	*Dependent Variables*	*Independent Variables*
LEM	FIML – Fay Multiple Imputation	FIML – Fay Multiple Imputation
MPlus	FIML – assumes continuous distribution Default: FIML – Fuchs	FIML – assumes continuous distribution Default: Mean Imputation
LG	Program:* FIML – Fay	Program: * FIML – Fuchs or Fay

* Requires Syntax Module to Implement

2.2 Modeling in LatentGold 5.0

2.2.1 Modeling Assuming MCAR

Fitting a model using cases where all data points are observed (ie. listwise deletion) is one of the easiest methods to implement to handle MCAR data. In LG 5.0, complete case analysis is requested in the options section of the syntax code with the keywords 'missing excludeall' (line 6 of *Exhibit 1*).

Let's modify the four variable situation detailed in equation 1.5 to contain five variables, *Y, A, B, C,* and *D*, where *Y* is a categorical dependent variable and *A, B, C,* and *D* are categorical independent variables. Similar to before, assume *Y, C,* and *D* suffer from item nonresponse. *Exhibit 1* illustrates a complete case MCAR model for this data using LG. In this case the MCAR complete case log-likelihood becomes

$$\log \mathcal{L}_{(\pi)} = \sum_y n_y \log \pi_{y|abcd}. \qquad (2.1)$$

```
1       options
2         algorithm
3          tolerance=1e-008 emtolerance=0.01 emiterations=250 nriterations=50;
4         startvalues
5          seed=0 sets=15 tolerance=1e-005 iterations=50;
6         missing excludeall;
7         output  parameters=last standarderrors;
8       variables
9         dependent Y nominal;
10        independent A nominal, B nominal, C nominal, D nominal;
11      equations
12        Y <- 1 + A + B + C + D;
```

Exhibit 1: Listwise (MCAR) Model Syntax in LatentGOLD 5.0

The LG syntax code consists of three sections - options, variables, and equations. The options section is used to set and turn off features available in LG. Bayes smoothers and Monte-Carlo methods, while not shown in *Exhibit 1*, can be set in the options section; see the LG technical manual for more guidance. The variables section is used to declare all dependent, independent and latent variables. Elements of a complex survey design would also be declare in the variables section. Finally the equations section contains any models

of interest. In *Exhibit 1*, a main effects logistic model for the dependent variable *Y* is defined with covariates *A*, *B*, *C*, and *D*.

2.2.2 Modeling Assuming MAR

Models that fit a MAR mechanism can be fit a variety of ways using either a saturated MAR (Fuchs) or response indicator (Fay) FIML approach. LG applies Fuchs approach to dependent variables by default. Since FIML techniques require complete independent variables during modeling, LG applies mean imputation on independent variables with missing data by default. A Fuchs-Mean model estimation can be requested by specifying 'missing includeall' in the options section of the syntax code (line 6 of *Exhibit 2*).

Applying Fuchs approach to the independent variables requires the use of quasi-latent variables. In LG, latent variables have the ability to be specified on either side of an equation. In our five variable example above, to fit a model for *Y* where *C* and *D* have been estimated using FIML techniques rather than mean imputation, independent variables *C* and *D* must be modeled in a latent framework. This is done by specifying latent variables (line 11) and equations (lines 13 and 14) in *Exhibit 3*. The use of the weight statement (w2~wei) on lines 13 and 14 preserves the observed values. Weight equations are always specified at the end of the equations section (see line 18). These FIML estimated values for C and D are then used in the regression formula on line 16 to model *Y* using Fuchs FIML approach. This process is repeated in an iterative EM fashion until convergence is reached for each model specified. In this case the estimated log-likelihood becomes

$$\log \mathcal{L}^*_{(\pi)} = \sum_{ycd} \hat{n}_{ycd} \log \hat{\pi}_{ycd} \qquad (2.2)$$

where

$$\hat{n}_{ycd} = n_{ycd} + n_{yc}\hat{\pi}_{d|yc} + n_{yd}\hat{\pi}_{c|yd} + n_{cd}\hat{\pi}_{y|cd} + n_{y}\hat{\pi}_{cd|y} + n_{c}\hat{\pi}_{yd|c} + n_{d}\hat{\pi}_{yc|d}.$$

```
1       options
2         algorithm
3          tolerance=1e-008 emtolerance=0.01 emiterations=250 nriterations=50;
4         startvalues
5          seed=0 sets=15 tolerance=1e-005 iterations=50;
6         missing includeall;
7         output  parameters=last standarderrors;
8       variables
9         dependent Y nominal;
10        independent A nominal, B nominal, C nominal, D nominal;
11      equations
12        Y <- 1 + A + B + C + D;
```

Exhibit 2: Fuchs-Mean (MAR) Model Syntax in LatentGOLD 5.0

To use Fay's approach in LG, response indicators must be added to the dataset for all variables with item nonresponse where Fay's FIML approach is desired. In *Exhibits 4 and 5*, only the dependent variable is modeled using Fay's approach; thus only the response indicator for the dependent variable (iY) is added to the variables section on line 9. In *Exhibit 4*, the missing values on independent variables *C* and *D* are imputed via the default mean imputation. In *Exhibit 5*, these values are estimated using Fuchs FIML approach. Notice on line 14 of *Exhibit 4* and line 18 of *Exhibit 5* that *iY* is dependent on the structural variables *A* and *B*. Under the MAR assumption, the model for *iY* can depend on any of the complete structural variables other than *Y*. Therefore there are several response pattern models that can be defined to model *iY*. The estimates of the structural model and the response models are influenced through the error terms, thus

various MAR response models should result in similar parameter and variance estimates for the model of *Y*.

```
options
  algorithm
   tolerance=1e-008 emtolerance=0.01 emiterations=250 nriterations=50;
  startvalues
   seed=0 sets=15 tolerance=1e-005 iterations=50 ;
  missing includeall;
  output  parameters=last standarderrors;
variables
  dependent Y nominal;
  independent A nominal, B nominal, C nominal, D nominal;
  latent   q_C nominal 2, q_D nominal 2;
equations
  q_C <- (w2~wei) C;
  q_D <- (w2~wei) D;

  Y <- 1 + A + B + q_C + q_D;

  w2 <- {1 0  0 1}:
```

Exhibit 3: Fuchs-Fuchs (MAR) Model Syntax in LatentGOLD 5.0

```
options
  algorithm
   tolerance=1e-008 emtolerance=0.01 emiterations=250 nriterations=50;
  startvalues
   seed=0 sets=15 tolerance=1e-005 iterations=50;
  missing includeall;
  output  parameters=last standarderrors;
variables
  dependent Y nominal, iY nominal;
  independent A nominal, B nominal, C nominal, D nominal;
equations
  Y <- 1 + A + B + C + D;

  iY <- 1 + A + B + A * B;
```

Exhibit 4: Fay-Mean (MAR) Model Syntax in LatentGOLD 5.0

```
options
  algorithm
   tolerance=1e-008 emtolerance=0.01 emiterations=250 nriterations=50;
  startvalues
   seed=0 sets=15 tolerance=1e-005 iterations=50;
  missing includeall;
  output  parameters=last standarderrors;
variables
  dependent Y nominal, iY nominal;
  independent A nominal, B nominal, C nominal, D nominal;
  latent   q_C nominal 2, q_D nominal 2;
equations
  q_C <- (w2~wei) C;
  q_D <- (w2~wei) D;

  Y <- 1 + A + B + q_C + q_D;

  iY <- 1 + A + B + A * B;
```

Exhibit 5: Fay-Fuchs (MAR) Model Syntax in LatentGOLD 5.0

Similar to Fuchs' approach, Fay's approach can also be applied to the independent variables through the use of quasi-latent variables. For the five variable example, consider response indicators *R, S,* and *T* which take on values of 1 when the variable is

observed and 2 otherwise for variables *Y, C,* and *D* respectively. In this case the estimated log-likelihood becomes

$$\log \mathcal{L}^{*}_{(\pi)} = \sum_{ycdrst} \hat{n}_{ycdrst} \log \hat{\pi}_{ycd} \hat{\pi}_{rst|ycd} \qquad (2.3)$$

where

$$\hat{n}_{ycd111} = n_{ycd} \qquad \hat{n}_{ycd211} = n_{cd}\hat{\pi}_{y|cd211}$$
$$\hat{n}_{ycd112} = n_{yc}\hat{\pi}_{d|yc112} \qquad \hat{n}_{ycd212} = n_{c}\hat{\pi}_{yd|c212}$$
$$\hat{n}_{ycd121} = n_{yd}\hat{\pi}_{c|yd121} \qquad \hat{n}_{ycd221} = n_{d}\hat{\pi}_{yc|d221}$$
$$\hat{n}_{ycd122} = n_{y}\hat{\pi}_{cd|y122} \qquad \hat{n}_{ycd222} = n_{\cdot}\hat{\pi}_{ycd|222}.$$

Models with more than one variable for Fay's method are more complicated. Every variable with missing for which Fay's method is desired must have a response indicator on the dataset. These response indicators are added to the dependent line of the variables section (line 9 of *Exhibit 6*). Next quasi-latent variables for each dependent variable and its response indicator that are needed on both sides of an equation must be specified on the latent line (line 11). Unless the joint distribution for the response indicators is known, each response indicator must be modeled separately. Appling this to our current example, equation 1.9 is modified as such

$$\pi_{ycdrst} = \pi_{ycd}\pi_{rst|ycd} = \pi_{ycd}\pi_{r|ycd}\pi_{s|ycdr}\pi_{t|ycdrs}. \qquad (2.4)$$

The equation section begins by estimating the quasi-latent independent variables using all complete independent data. Working from least amount of missing to most amount of missing, each variable is defined. Note that on line 14 the equation for the quasi-latent *D* variable contains the quasi-latent *C* variable. After line 13 all values for quasi-latent *C* are estimated. After all quasi-latent independent variables are estimated, the model of interest (*Y*) can be specified using all variables. After this, following Fay's instruction, the response indicators are modeled. In *Exhibit 6* starting on line 18, the missingness of *Y* is dependent on *A* and *B*; the missingness of *C* is dependent on *A*; and the missingness of *D* is dependent on *B*. Again, there are several MAR models that can be specified here. Lines 22 to 25 act to connect the quasi-latent variables to the observed data. When this set of equations are estimated at the same time using EM techniques, Fay's FIML approach is applied to both the independent and dependent variables with a MAR response mechanism.

2.2.3 Modeling Assuming MNAR

Extending Fay's MAR application to MNAR is straightforward. Under a MNAR response mechanism, the missingness of a variable depends on the variable itself. Consider the case of the Fay-Mean MAR application in *Exhibit 4*. To convert this model to a MNAR model, line 14 must be modified by adding *Y* to the dependent side. Since *Y* must now be used on both the independent and dependent side of separate equations, a quasi-latent variable for *Y* must be created. Therefore the MNAR code for a Fay-Mean model looks similar to *Exhibit 7*. Note this is the simplest MNAR response pattern, but other response patterns can be specified. Similar modifications allow Fay's FIML application to model MNAR for the dependent variables as well. Fay's method can also be mixed with Fuchs' method.

```
options
  algorithm
   tolerance=1e-008 emtolerance=0.01 emiterations=250 nriterations=50;
  startvalues
   seed=0 sets=15 tolerance=1e-005 iterations=50;
  missing includeall;
  output  parameters=last standarderrors;
variables
  dependent Y nominal, C nominal, D nominal,
            iY nominal, iC nominal, iD nominal;
  independent A nominal, B nominal;
  latent      q_C nominal 2, q_D nominal 2,
                q_iY nominal 2, q_iC nominal 2;
equations
  q_C <- 1 + A + B;
  q_D <- 1 + A + B + q_C;

  Y <- 1 + A + B + q_C + q_D;

  q_iY <- 1 + A + B;
  q_iC <- 1 + A + q_iY;
  iD <- 1 + B + q_iY + q_iC + q_iY*q_iC;

  iY <- (w2~wei) q_iY;
  iC <- (w2~wei) q_iC;
  C <- (w2~wei) q_C;
  D <- (w2~wei) q_D;

  w2 <- {1 0  0 1};
```

Exhibit 6: Fay-Fay (MAR) Model Syntax in LatentGOLD 5.0

```
options
  algorithm
   tolerance=1e-008 emtolerance=0.01 emiterations=250 nriterations=50 ;
  startvalues
   seed=0 sets=15 tolerance=1e-005 iterations=50 ;
  missing includeall;
  output  parameters=last standarderrors;
variables
  dependent Y nominal, iY nominal;
  independent A nominal, B nominal, C nominal, D nominal;
  latent  q_Y nominal 2;
equations
  q_Y <- 1 + A + B + C + D;

  iY <- 1 + q_Y;

  Y <- (w2~wei) q_Y;

  w2 <- {1 0  0 1};
```

Exhibit 7: Fay-Mean (MNAR) Model Syntax in LatentGOLD 5.0

3. Application Results and Discussion

3.1 The Illustrative Data

The National Survey of Drug Use and Health (NSDUH) is an annual national and state level survey that collects information on the use of tobacco, alcohol, illicit drugs and mental health in the United States. It is funded by the Substance Abuse and Mental Health Services Administration (SAMHSA), Center for Behavioral Health Statistics and Quality (CBHSQ). Several variables on the NSDUH undergo a weighted sequential hot deck imputation (WSHD); however the NSDUH public use data file contains many variables which undergo no imputation.

Recently Edlund et. al. (2015) used five years of NSDUH data to study the relationship of major depressive episodes (MDE) and nonmedical prescription opioid use (NMPOU) among adolescents between 12 and 17 years old. The authors fit a log-logistic regression model for past year opioid use with twelve covariates – past year MDE, age, gender, race, family income, rural/urban location, past year alcohol abuse/dependence, past year illicit drug abuse/dependence excluding opioid use, delinquency occurrences, most recent grades achieved in school, number of religious services attended in the past year, and family support.

For the purpose of this report, we used the model published by Edlund et. al. and four years of NSDUH public use data from 2010 to 2013 with 72,793 adolescent respondents as a starting point. Where possible, item nonresponse was reintroduced using imputation indicators for the variables that underwent WSHD imputation. The rural/urban location variable was not included on the NSDUH public use files; it was removed from our analysis.

Due to variable instability, the school performance grade variable was collapsed from a six level variable to a three level variable (1=A/B/C – average or attend school that does not give grades; 2=D/F - average; 3=dropped out). Limitations in LG required the use of a reduced model. The reduced model was selected using a forward selection approach using listwise deletion for handling missing data.

The final model for NMPOU used in this report contains the following eight covariates – past year MDE, age, gender, past year alcohol abuse/dependence, past year illicit drug abuse/dependence excluding opioid use, delinquency occurrences, most recent grades achieved in school (3 levels), and family support. The dependent variable contains about 2% item nonresponse; of the six covariates with item nonresponse rates range from around 1% to almost 8%.

To address the research questions of this paper, five types of models were fit each addressing nonresponse in a different manner. These models are listed in Table 2.

Table 2: Models Used to Explore Objectives

Model Type	*Dependent Variable*	*Independent Variables*	
1	Listwise	Listwise	**1 - MCAR - Listwise Deletion**
2	FIML MAR	Mean Imputation	**2a - Fuchs / Mean Imputation** **2b - Fay (MAR) /Mean Imputation**
3	FIML MAR	FIML MAR	**3a - Fuchs / Fuchs** **3b - Fay (MAR) / Fuchs** **3c - Fay (MAR) / Fay (MAR)**
4	FIML MNAR	Mean Imputation	**4 - Fay (MNAR) / Mean Imputation**
5	FIML MNAR	FIML MAR	**5a – Fay (MNAR) / Fuchs** **5b – Fay (MNAR) / Fay (MAR)**

3.2 MAR Models – Fuchs vs. Fay

The Fuchs' saturated MAR approach and Fay's response indicator approach resulted in the same model for past year NMPOU when the dependent variable missing mechanism was assumed to be MAR (model groups 2 and 3). The group 3 model that addressed independent nonresponse with Fay FIML MAR (3c) produced some estimates that

deviated from those observed with the other two models which used Fuchs FIML MAR to address independent nonresponse (3a/3b). While most of the odds ratio estimates produced by this model fit within the 95% confidence intervals produced by models 3a and 3b, two estimates fell outside the 95% confidence interval – adolescents between 12 and 13 years old and high school drop outs. This model had rank deficiency and boundary errors during model estimation and Bayes smoothers were required for the model to fit without errors.

Only one covariate changed interpretation depending on the model. Models 2a, 2b, 3a, and 3b all indicated the odds of past year NMPOU is similar between students with a D or F average and those with good grades; models 1 and 3c indicated otherwise. These differences were potentially due to the increased number of variables needed for the response indicator approach which may have contributed to potential model instability.

Odds ratio estimates and 95% confidence intervals for select parameters from select models are displayed in Figure 1.

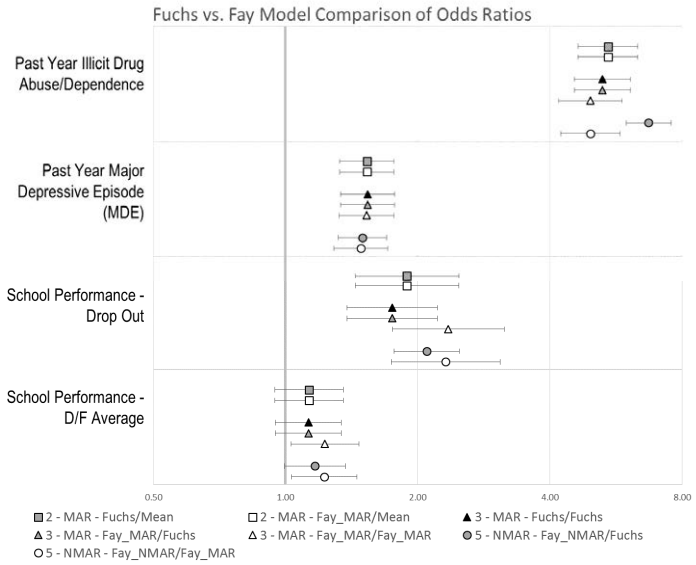

Figure 1: Odds Ratios for Past Year NMPOU among Adolescents Aged 12-17

3.3 MAR vs. MNAR Models

Odds ratio estimates from MNAR models tended to be near or higher than those from MAR models when the handling of the independent variables was held constant. When the independent variables were mean imputed, 8 of the 12 covariate levels produced larger estimates with FIML MNAR methods; similar results were observed for FIML MAR independent variable handling. Anywhere from 6 to 7 covariate levels changed by 0.10 points or less between the two methods.

The odds of past year NMPOU among high school drop outs compared to students with good grades (A, B, C average) increased the most; 0.58 points with mean imputation and 0.56 points with FIML MAR. This is expected, since this variable resulted in inconsistent estimates for models in the FIML MAR – FIML MAR model group 3. This may indicate that past year NMPOU use among high school drop outs follows a MNAR mechanism. It may also be a side effect of model stability.

As expected, the MNAR models for past year NMPOU resulted in models with larger variances compared to similar MAR models when the handling of the independent variables were held constant. Variance differences were observed by comparing widths of 95% confidence intervals for odds ratios for each model parameter. These ratios are displayed in Figure 2.

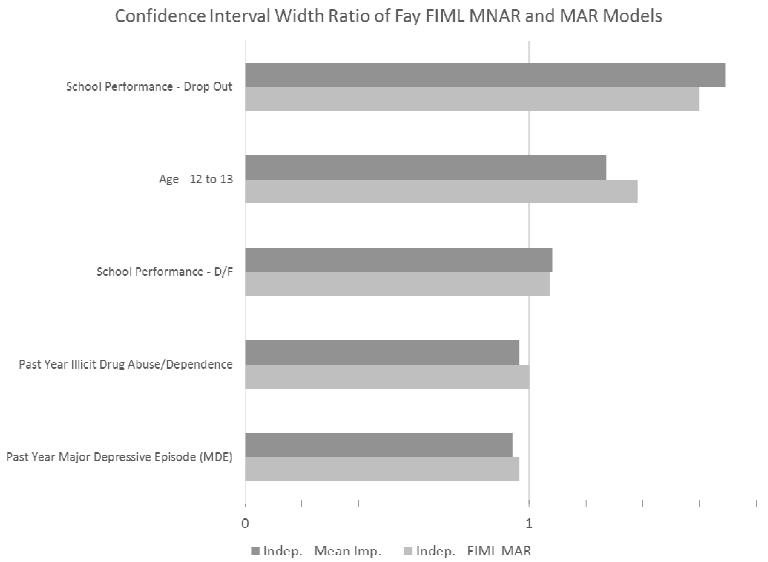

Figure 2: Ratio of Confidence Interval Width between Fay FIML MAR and MNAR Models

4. Conclusion

Implementing Fay's method in LG presented a few unique challenges. When Fay's method is desired on more than one variable in the model, the creation of a response indicator, two quasi-latent variables, and four equations for each additional variable with item nonresponse resulted in some models that were difficult to estimate. Currently LG executes on at most 4 GB of RAM. This restriction required the use of a more parsimonious model with a reduced set of covariates and the collapsing of 2 levels in the school performance variable.

From this analysis, the main advantage of treating item nonresponse as MNAR is the ability to identify variables that potentially have a MNAR nonresponse. This can only be assessed by comparing estimates from MAR and MNAR models; when estimates differ between the two methods, then a MNAR nonresponse may be present. The disadvantages of treating item nonresponse as MNAR is largely due to the increased complexity of the model.

In conclusion, these techniques are best used as a set. When FIML MAR and FIML MNAR models are fit to categorical data, information about the potential missing data mechanism can be gathered. Given the difficulty of fitting and writing the response indicator models compared to the saturated MAR approach, when a MAR mechanism is assumed Fuchs' approach should be used with the LG software. Only when a MNAR mechanism is suspected should Fay's approach be used and should be limited to only those variables suspected to have MNAR nonresponse. Other variables can be modeled with Fuchs' approach. Fitting several different models with various nonresponse

mechanisms is recommended to identify potential MNAR models and to identify when MNAR models produce untrustworthy estimates either by modeling fitting errors or through unrealistic parameter estimates.

Acknowledgements

The authors would like to thank NSF for sponsoring this research. However, we would like to note that the views expressed in this paper are those of the authors only and do not reflect the view or position of NSF, RTI, SAMHSA, or CBHSQ.

References

Asparouhov, T. & Muthén, B. (2010). Multiple Imputation with MPlus. MPlus Technical Report. http://www.statmodel.com

Center for Behavioral Health Statistics and Quality. (2014). *2012 National Survey on Drug Use and Health: Methodological Resource Book (2012 NSDUH Editing and Imputation Report).* Substance Abuse and Mental Health Services Administration, Rockville, MD.

Edlund M., Forman-Hoffman V., Winder C., Heller D., Kroutil L., Lipari R., & Colpe L. (2015). Opioid Abuse and Depression in Adolescents: Results from the National Survey on Drug Use and Health. *Science Direct*. Available online 22 April 2015. (http://www.sciencedirect.com/science/article/pii/S0376871615002057)

Enders C.K. (2010). *Applied Missing Data Analysis*. New York, NY: Guilford Press.

Fay, R.E. (1986). Causal Models for Patterns of Nonresponse. Journal of the American Statistical Association, 81(394), 354-365.

Fuchs, C. (1982). Maximum Likelihood Estimation and Model Selection in Contingency Tables with Missing Data. Journal of the American Statistical Association, 77(378), 270 – 278.

Graham, J.W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling,* 10(1), 80-100.

Graham, J.W. (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology,* 60, 549-576.

Little, T., Jorgensen, T., Lang, K., & Moore, W. (2014). On the Joys of Missing Data. *Journal of Pediatric Psychology,* 39(2), 151-162.

Lohr, S. L. (2010). *Sampling : design and analysis* (2nd ed.). Boston, Mass.: Brooks/Cole.

Muthén, L. K., & Muthén, B. O. (1998-2011). Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Rubin, D.B. (1976). Inference and Missing Data. *Biometrika,* 63(3), 581-592.

Vermunt, J. K., & Magidson, J. (2013). *Technical guide to Latent Gold 5.0: Basic, advanced, and syntax*. Belmont, MA: Statistical Innovations.

Vermunt, J.K. (1997). *Log-Linear Models for Event Histories.* London, UK: SAGE Publications, Inc.