



Addressing Nonresponse for Categorical Data Items Using Full Information Maximum Likelihood with Latent GOLD 5.0

Susan L. Edwards, Marcus E. Berzofsky, and
Paul P. Biemer

RTI Press publication MR-0038-1809

RTI International is an independent, nonprofit research organization dedicated to improving the human condition. The RTI Press mission is to disseminate information about RTI research, analytic tools, and technical expertise to a national and international audience. RTI Press publications are peer-reviewed by at least two independent substantive experts and one or more Press editors.

Suggested Citation

Edwards, S. L., Berzofsky, M. E., and Biemer, P. P. (2018). *Addressing Nonresponse for Categorical Data Items Using Full Information Maximum Likelihood with Latent GOLD 5.0*. RTI Press Publication No. MR-0038-1809. Research Triangle Park, NC: RTI Press. <https://doi.org/10.3768/rtipress.2018.mr.0038.1809>

This publication is part of the
RTI Press Methods Report series.

RTI International
3040 East Cornwallis Road
PO Box 12194
Research Triangle Park, NC
27709-2194 USA

Tel: +1.919.541.6000
E-mail: rtipress@rti.org
Website: www.rti.org

©2018 RTI International. RTI International is a registered trademark and a trade name of Research Triangle Institute. The RTI logo is a registered trademark of Research Triangle Institute.



This work is distributed under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 license (CC BY-NC-ND), a copy of which is available at <https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>

<https://doi.org/10.3768/rtipress.2018.mr.0038.1809>

www.rti.org/rtipress

Contents

About the Authors	i
Acknowledgments	ii
Abstract	ii
Introduction	1
Missing Data Mechanisms	1
Full Information Maximum Likelihood (FIML)	2
Software Packages	3
Methods and Results: Implementing the Full Information Maximum Likelihood (FIML) Technique	3
Modeling Assuming Data Missing Completely at Random (MCAR)	3
Modeling Assuming Data Missing at Random (MAR)	6
Modeling Assuming Data Missing Not at Random (MNAR)	8
Discussion	9
References	11

About the Authors

Susan L. Edwards, MS, is a research statistician in the International Statistics Program at RTI International.

Marcus E. Berzofsky, PhD, is a senior research statistician in the Social Statistics Program at RTI International.

Paul P. Biemer, PhD, is Distinguished Fellow in Statistics at RTI International.

Acknowledgments

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 or RTI.

Abstract

Full information maximum likelihood (FIML) is an important approach for compensating for nonresponse in data analysis. Unfortunately, only a few software packages implement FIML and even fewer have the capability to compensate for missing not at random (MNAR) nonresponse. One of these packages is Statistical Innovations' Latent GOLD; however, the user documentation for Latent GOLD provides no mention of this capability. The purpose of this paper is to provide guidance for fitting MNAR FIML models for categorical data items using the Latent GOLD 5.0 software. By way of comparison, we also provide guidance on fitting FIML models for nonresponse missing at random (MAR) using the method of Fuchs (1982) as well as the approach by Fay (1986), who incorporated item nonresponse indicators within a structural modeling framework. We compare implementation and results for both FIML for MAR and MNAR nonresponse models for independent and dependent variables. Recommendations for future applications of FIML using Latent GOLD are provided.

Introduction

Nonresponse can occur for several reasons including refusals, inconsistent or invalid responses, “don’t knows,” and inadvertent skips. In the statistical literature, nonresponse is assumed to arise from three types of stochastic mechanisms (see, for example, Rubin, 1976). Missing completely at random (MCAR) occurs when the nonresponse is completely unrelated to independent or dependent variables (i.e., not dependent on any attributes); missing at random (MAR) occurs when the nonresponse is related to the observed information but not the dependent variable(s); and missing not at random (MNAR) occurs when the nonresponse is not MCAR or MAR.

Regardless of the underlying mechanism, ignoring missing data in a data analysis can lead to biased and/or inefficient inference. Fortunately, many techniques have been developed to handle missing data. Some common techniques are listwise deletion and single imputation including mean imputation, hot deck imputation, and predictive mean matching. Multiple imputation and full information maximum likelihood (FIML) are among the newer techniques that can yield unbiased inferences and valid standard error estimates. Asymptotically, multiple imputation can yield results that are identical to those obtained using FIML (Graham, 2009; Little & Rubin, 2002).

Several statistical software packages have incorporated methods for addressing nonresponse. Some of these are LEM, MPlus, SAS, Stata, R, and Latent GOLD (LG). Implementing more standard approaches for handling missing data is quite often straightforward in these software packages; for instance, regression analyses using SAS and Stata employ listwise deletion by default. However, implementing more complex missing data techniques may require a more intimate knowledge of the software and the data.

This report provides a primer for applying FIML to compensate for missing data in log-linear models for users of LG. While not detailed in this report, the techniques presented can apply to latent class models (see the supplementary article materials published with Edwards, Berzofsky, & Biemer, 2017 for LG

code applying these techniques to Markov latent class models).

Two methods for handling missing data in LG are well documented and specified as user options: complete case analysis (i.e., listwise deletion) and FIML using Fuchs’ (1982) method, which assumes MAR (Vermunt & Magidson, 2005, 2016). The latter method can only be applied for missing data in the dependent variables and automatically invokes mean imputation to address any missing data in the independent variables. Although Fuchs’ approach works well for MAR data, the method by Fay (1986) is more general; it can be applied for both MAR and MNAR nonresponse by incorporating and modeling item nonresponse indicators. Unfortunately, Fay’s method is not a user option in LG; however, it is not difficult to set up the LG software to force-fit Fay’s method for handling missing data, as we show for MAR nonresponse in the section Modeling Assuming Data Missing at Random (MAR). The section Modeling Assuming Data Missing Not at Random (MNAR) describes in detail how to model item missing data for all variables in categorical data analysis using LG models when the missing data mechanism is MNAR. Doing so allows us to model both MAR and MNAR missingness for both dependent and independent variables. This makes LG unique since Fay’s method is not available in any other data analysis software packages to our knowledge.

The rest of this introduction discusses missing data mechanisms, FIML assumptions, and available software packages in more depth. The Methods and Results section details the LG syntax for fitting FIML models using two different approaches (Fuchs’ and Fay’s) under each response mechanism for both dependent and independent variables. The discussion summarizes these various methods and discusses potential difficulties with implementing each approach.

Missing Data Mechanisms

Every imputation method makes assumptions about the causal nature of the missing data; this is known as the *missing data mechanism*. 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). Originally defined by Rubin (1976), MCAR occurs when the missingness does not depend on either the observed or unobserved data; MAR is a less restrictive assumption in that the missingness depends only on the observed data; and MNAR is the least restrictive mechanism regarding modeling assumptions, where the missingness depends on both the observed and unobserved data.

Under MAR and MNAR the respondents and nonrespondents may differ on the outcome variable of interest. Under a MAR mechanism, the missing outcomes are explained by other observed independent variables, so the response mechanism is assumed to be ignorable conditional on these observed independent variables. Under a MNAR approach, the missing data mechanism interacts with the outcome variable. This interaction can be expressed using a response indicator to incorporate information regarding the response mechanism into the imputation model (Rubin, 1976). Thus, to model MNAR data, the observed data and the response mechanism must be modeled jointly to account for observed and unobserved influences on the missingness.

Current approaches to modeling MNAR data can be classified into two types: selection models and pattern mixture models (Heckman, 1976; Little, 1993). Selection models require fitting a two-part model—with one model for the outcome variable and another model for the response mechanism given the outcome variable. Pattern mixture models form subgroups of cases that share the same missing data pattern and then fit a model for each pattern. The overall model is then estimated using a weighted average of the individual models; standard error estimates for a pattern mixture model are often estimated using the delta method (Enders, 2010). Formulas for these models are presented in Table 1, where Y is the incomplete outcome variable of interest and R is a response indicator of Y .

These MNAR techniques require strong assumptions about the data in order for the models to be

identifiable (that is, to contain reliable estimates). Selection models assume that the underlying response probabilities and the incomplete outcome variable follow a bivariate normal distribution. Pattern mixture models require the researcher to specify values for the inestimable parameters and have largely been used in conducting sensitivity analyses (Enders, 2010). When these assumptions are met, and the missing data mechanism is MNAR, simulation studies have shown that these methods tend to inflate the variance estimates (Fay, 1986). Our report addresses MNAR missing data through a selection model approach. For more information on pattern mixture modeling, refer to Little (1993) and references citing Little (1993).

Full Information Maximum Likelihood (FIML)

When the general conditions for estimation are satisfied, FIML methods can be used to fit a structural (or substantive) model for the outcome variable and nonresponse model simultaneously. While not an imputation method, FIML makes use of all available data (including partially observed data) to maximize the log-likelihood of this joint model (Enders, 2010). A considerable amount of work has been done around applying FIML under a MAR response mechanism and modeling MNAR data with a continuous outcome (Anderson, 1957; Arbuckle, 1996; Graham, 2003; Little & Rubin, 2002).

For analysis of categorical dependent variables, FIML approaches are similar to those developed to handle continuous data in that partially observed information is used when maximizing a log-linear likelihood function. The primary difference is in the assumption about the sampling distribution: continuous data analysis assumes normality, and categorical data analysis assumes a multinomial sampling distribution (Vermunt, 1997).

Table 1. Missing not at random (MNAR) modeling approaches

MNAR Model:	
$P(Y,R)$ = joint distribution of observed data and response pattern	
Approach 1: Selection Model $P(Y,R) = P(Y)P(R Y)$	Approach 2: Pattern Mixture Model $P(Y,R) = P(R)P(Y R)$

P stands for probability function, Y is the incomplete outcome variable of interest, and R is a response indicator of Y .

If the data follow a MAR (or MCAR) response mechanism, this implies the response probabilities are independent of the missing variables. In this case, the likelihood can be factored into two components—one for the log-linear parameters and another for the response mechanism:

$$\log \mathcal{L}_{(\pi, \Theta)} = \log \mathcal{L}_{(\pi)} + \log \mathcal{L}_{(\Theta)} \quad (1)$$

where the structural probabilities are represented by π and the response probabilities are represented by Θ . 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.

In 1982, Fuchs extended the methodology of FIML to estimate the parameters of a saturated log-linear model using the estimation-maximization algorithm when item nonresponse is ignorable. 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 because, in that case, the log likelihood does not factor, as shown in equation (1).

Fay extended the FIML 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 either nonignorable nonresponse (MNAR) and ignorable nonresponse (MAR or MCAR). In Fay's approach, which uses a selection model, response indicators are created for all variables with partially observed data, and outcome and nonresponse models are fit using their joint likelihood.

Software Packages

Many programs are capable of applying FIML approaches to handle MAR missing data. A few of these programs include LEM, MPlus, SAS, Stata, R, and Latent GOLD (LG). While all of these packages are widely used, LG is specifically designed to analyze nominal, ordinal, and interval-level categorical data by assuming a multinomial sampling distribution with the capability to account for complex survey designs. LG fits all models as log-linear models,

which can be conceptualized as mixture models. In LG, missing data are addressed by default through Fuchs' FIML approach for dependent variables and a stochastic mean imputation for independent variables (Vermunt & Magidson, 2016). With some adjustments that are not documented in the 5.0 user manual, LG can accommodate both Fuchs' and Fay's FIML approaches on all variables with nonresponse in the log-linear model.

MPlus is a popular package used in structural equation modeling to analyze continuous, ordinal, nominal, and count data; MPlus can address a complex survey design and apply FIML to missing data. MPlus assumes a normal sampling distribution to implement FIML on categorical variables (Muthén & Muthén, 1998–2011). Due to the distributional assumptions, these models are not fit as log-linear models. While this approach may be logical for binary, ordinal, or interval-level categorical variables, its effectiveness when used on nominal categorical variables is unclear. We are unaware whether MPlus can apply FIML in a log-linear model. Although MPlus handles MNAR missing mechanisms for continuous variables in a FIML construct, for categorical MNAR response, multiple imputation is the suggested approach (Asparouhov & Muthén, 2010).

This report focuses on using LG 5.0 to account for missing data, 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 can make applying FIML to variables with MAR and MNAR nonresponse more accessible to researchers, as shown in the following section.

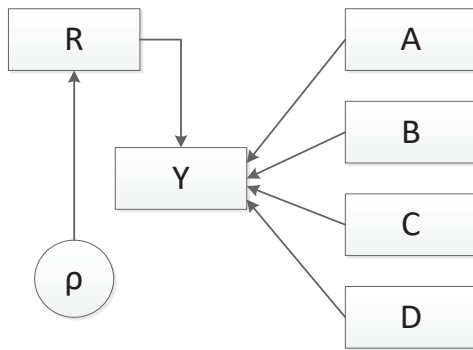
Methods and Results: Implementing the Full Information Maximum Likelihood (FIML) Technique

Modeling Assuming Data Missing Completely at Random (MCAR)

Fitting a model using cases where all data points are observed (i.e., listwise deletion or complete case analysis) is one of the easiest methods to implement

to account for MCAR data. A path diagram of an MCAR model with four complete independent variables and one dependent variable with missing values is shown in Figure 1.

Figure 1. Missing completely at random (MCAR) model path diagram



In LG 5.0, complete case analysis is requested in the options section of the syntax code with the keyword “missing excludeall” (line 6 of Figure 2). In Figure 2, a main effects multinomial-logistic model for the dependent variable *Y* is defined with covariates *A*, *B*, *C*, and *D* using only observations with complete data for variables *Y*, *A*, *B*, *C*, and *D*. In Figure 2, all variables are multinomial. If *Y* is a binary variable, a log-linear logistic model is fit; if *Y* has more than two levels, a log-linear model is fit. In later figures we illustrate applications of FIML using binary variables for simplicity; these models can easily be extended to higher level multinomial variables, either ordinal or nominal.

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. In the options section of LG 5.0, Bayes smoothers can be set to prevent boundary solutions, Monte-Carlo simulation requested, and power calculation methods invoked, although this is not shown in Figure 2. Refer to the LG technical manual for more guidance on how to implement these options. The variables section is used to declare all dependent, independent, and latent variables. Latent variables are unobserved variables that are inferred from other observed variables; they can be either continuous or categorical. Elements of a complex survey design would also be set in the variables section. Finally, the equations section contains any models of interest.

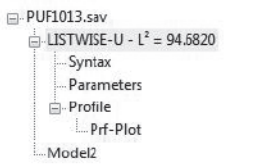
Once executed, LG creates several windows of output—Model Summary Output, Syntax, Parameters, and Profile. The Model Summary Output provides information on the model—the number of observations used to fit the model, the number of parameters in the model, seed values, and fit statistics—and can be viewed by clicking on the model name. Any warning or error messages regarding model estimation are listed in this window. The Parameters window contains model estimates for every model specified in the equations section of the syntax. Figures 3 and 4 contain example screenshots of the Model Summary Output and Parameters windows, respectively. For more information on the output windows refer to the LG users guide.

Figure 2. Listwise (MCAR) model syntax in Latent GOLD 5.0

```

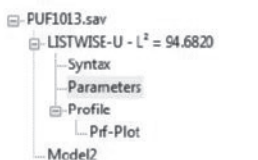
1  options
2      algorithm
3          tolerance=1e-008 emtolerance=0.01 emiterations=250 nriterations=50;
4      startvalues
5          seed=0 sets=15 tolerance=1e-005 iterations=50;
6      missing  excludeall;
7      output   parameters=last standarderrors;
8  variables
9      dependent Y nominal;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


Figure 3. Example model summary output from Latent GOLD



Syntax (1) Model	
Number of cases	68137
Number of parameters (Npar)	7
Random Seed	82000
Best Start Seed	82000
Chi-squared Statistics	
Degrees of freedom (df)	29
L-squared (L ²)	94.6820
X-squared	109.4965
Cressie-Read	103.5004
BIC (based on L ²)	-228.0670
AIC (based on L ²)	36.6820
AIC3 (based on L ²)	7.6820
CAIC (based on L ²)	-257.0670
SABIC (based on L ²)	-135.9043
Dissimilarity Index	0.0048
Log-likelihood Statistics	
Log-likelihood (LL)	-11417.8744
Log-prior	0.0000
Log-posterior	-11417.8744
BIC (based on LL)	22913.6537
AIC (based on LL)	22849.7488
AIC3 (based on LL)	22856.7488
CAIC (based on LL)	22920.6537
SABIC (based on LL)	22891.4075
Classification Statistics	
Classification errors	
Reduction of errors (Lambda)	
Entropy R-squared	
Standard R-squared	
Classification log-likelihood	-11417.8744
Entropy	0.0000
CLC	22835.7488
AWE	23012.5586
ICL-BIC	22913.6537

Figure 4. Example parameters output from Latent GOLD



Regression Parameters					
	term	coef	s.e.	z-value	p-value
anIYR2(0)	← 1	3.0675	0.0372	82.5483	2.0e-1482
anIYR2(1)	← 1	0.0000	.	.	.
anIYR2(0)	← irsex(1)	0.4322	0.0380	11.3635	6.4e-30
anIYR2(1)	← irsex(1)	0.0000	.	.	.
anIYR2(0)	← irsex(2)	0.0000	.	.	.
anIYR2(1)	← irsex(2)	0.0000	.	.	.
anIYR2(0)	← CATAG7(1)	1.1894	0.0576	20.6646	7.2e-95
anIYR2(1)	← CATAG7(1)	0.0000	.	.	.
anIYR2(0)	← CATAG7(2)	0.3841	0.0413	9.3067	1.3e-20
anIYR2(1)	← CATAG7(2)	0.0000	.	.	.
anIYR2(0)	← CATAG7(3)	0.0000	.	.	.
anIYR2(1)	← CATAG7(3)	0.0000	.	.	.

Modeling Assuming Data Missing at Random (MAR)

Models that implement a MAR mechanism can be fit a variety of ways using either a saturated MAR (Fuchs') or response indicator (Fay's) FIML approach. These methods are detailed in Table 2. Mean imputation is included as an option for the independent variables since it is the default method for handling item-nonresponse in independent variables in LG. From here on, we refer to models by model type, which is an abbreviation of the missing data approach applied to the dependent and independent variables, separated by a hyphen.

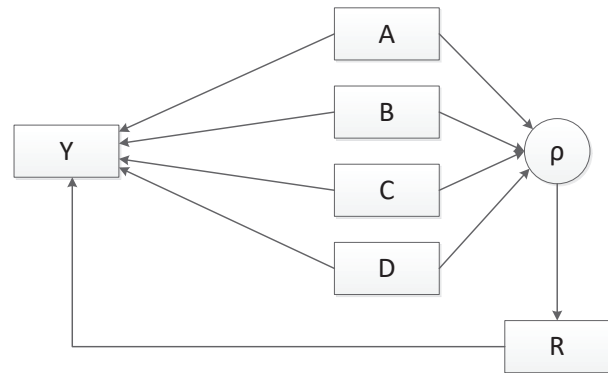
Table 2. Modeling approaches assuming MAR

Model Type	Dependent Variable	Independent Variables
1—Fuchs-Mean	FIML MAR via Saturated MAR (Fuchs)	Mean Imputation
2—Fay-Mean (MAR)	FIML MAR via Response Indicator (Fay)	Mean Imputation
3—Fuchs-Fuchs	FIML MAR via Saturated MAR (Fuchs)	FIML MAR via Saturated MAR (Fuchs)
4—Fay-Fuchs (MAR)	FIML MAR via Response Indicator (Fay)	FIML MAR via Saturated MAR (Fuchs)
5—Fay-Fay (MAR)	FIML MAR via Response Indicator (Fay)	FIML MAR via Response Indicator (Fay)

FIML = full information maximum likelihood; MAR = missing at random.

A path diagram showing the relationship between an outcome variable, its response indicator, and four independent variables in a MAR model is displayed in Figure 5. As in Figure 1, the four independent variables are complete, and the outcome variable suffers from item nonresponse. To model the missing at random response pattern, the response mechanism, ρ , is dependent on the four independent variables. In this single model case with missing only in the dependent variable, the FIML model estimates are equivalent to those from a MCAR model.

Figure 5. Missing at random (MAR) model path diagram



LG applies Fuchs' approach to dependent variables by default. Since Fuchs' FIML approach requires complete independent variables during modeling, LG by default applies mean imputation on independent variables with missing data. A Fuchs-Mean model estimation can be requested by specifying "missing includeall" in the options section of the syntax code (line 6 of Figure 6). We use this scenario to illustrate the simplest form of FIML that LG offers.

Figure 7 illustrates code for a multinomial dependent variable of any level with four independent variables: A and B are complete and are multinomial of any level, C is an incomplete 2-level multinomial variable, and D is an incomplete 3-level multinomial variable. Applying Fuchs' approach to the independent variables requires the use of quasi-latent variables: a quasi-latent variable is a latent variable used to define a single manifest variable. In LG, latent variables can be specified on either side of an equation, but manifest variables can only be used on one side (independent variables on the right side; dependent variables on the left side). In our five-variable example, above, to fit a model for Y where C and D have been estimated using FIML techniques rather

Figure 6. Fuchs-Mean model syntax in Latent GOLD 5.0

```

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;
    
```

than mean imputation, independent variables *C* and *D* must be modeled in a latent framework. This is done by specifying categorical latent variables (line 11) and equations (lines 13 and 14) in Figure 7. The use of the

weight statement (*w2~wei* and *w3~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). LG by default applies Fuchs' FIML

estimation to variables on the left side of an equation. 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 estimation-maximization fashion until convergence is reached for each model specified.

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 Figures 8 and 9, only the dependent variable is modeled using Fay's approach; the response indicator for the dependent variable (*iY*) is added to the variables section on line 9. In Figure 8, the missing values on independent variables *C* and *D* are imputed via the default mean imputation. In Figure 9, these values are estimated using Fuchs' FIML approach. Notice on line 14 of Figure 8 and line 18 of Figure 9 that *iY* is

Figure 7. Fuchs-Fuchs model syntax in Latent GOLD 5.0

```

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 latent q_C nominal 2, q_D nominal 3;
12 equations
13 q_C <- (w2~wei) C;
14 q_D <- (w3~wei) D;
15
16 Y <- 1 + A + B + q_C + q_D;
17
18 w2 <- {1 0 0 1};
19 w3 <- {1 0 0 0 1 0 0 0 1};

```

Figure 8. Fay-Mean (MAR) model syntax in Latent GOLD 5.0

```

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, iY nominal;
10 independent A nominal, B nominal, C nominal, D nominal;
11 equations
12 Y <- 1 + A + B + C + D;
13
14 iY <- 1 + A + B + A * B;

```

Figure 9. Fay-Fuchs (MAR) model syntax in Latent GOLD 5.0

```

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, iY nominal;
10 independent A nominal, B nominal, C nominal, D nominal;
11 latent q_C nominal 2, q_D nominal 2;
12 equations
13 q_C <- (w2~wei) C;
14 q_D <- (w2~wei) D;
15
16 Y <- 1 + A + B + q_C + q_D;
17
18 iY <- 1 + A + B + A * B;
19
20 w2 <- {1 0 0 1};

```

dependent on the structural variables A and B only. Under the MAR assumption, the model for iY can depend on any of the complete structural variables other than Y . Therefore, several response pattern models 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 .

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.

Models with more than one variable for Fay's method are more complicated. Every variable with missingness 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 Figure 10). 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.

The equation section begins by estimating the quasi-latent independent variables using all complete independent data. The equations for the quasi-latent variables must be defined working from least amount of missing to most amount of missing. 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. Following Fay's instruction, the response indicators are specified and modeled after the model of interest. In Figure 10, 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, several MAR models could be specified here. Lines 22 to 25 connect the quasi-latent variables to the observed data. When this set of equations is estimated at the same time using estimation-maximization techniques, Fay's FIML approach is applied to both the independent and dependent variables with a MAR response mechanism.

Figure 10. Fay-Fay (MAR) model syntax in Latent GOLD 5.0

```

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, C nominal, D nominal,
10     iY nominal, iC nominal, iD nominal;
11   independent A nominal, B nominal;
12   latent q_C nominal 2, q_D nominal 2,
13     q_iY nominal 2, q_iC nominal 2;
14 equations
15   q_C <- 1 + A + B;
16   q_D <- 1 + A + B + q_C;
17
18   Y <- 1 + A + B + q_C + q_D;
19
20   q_iY <- 1 + A + B;
21   q_iC <- 1 + A + q_iY;
22   iD <- 1 + B + q_iY + q_iC + q_iY*q_iC;
23
24   iY <- (w2~wei) q_iY;
25   iC <- (w2~wei) q_iC;
26   C <- (w2~wei) q_C;
27   D <- (w2~wei) q_D;
28
29   w2 <- {1 0 0 1};

```

Modeling Assuming Data Missing Not at Random (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; see Figure 11.

Consider the case of the Fay-Mean MAR application in Figure 8. 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 sides of separate equations, a quasi-latent variable for Y must be created. By creating a latent variable for Y , we can now use the unobserved latent value of Y to predict the response probabilities and the observed values of Y . Therefore, the MNAR code for a Fay-Mean model

Figure 11. Missing not at random (MNAR) model path diagram

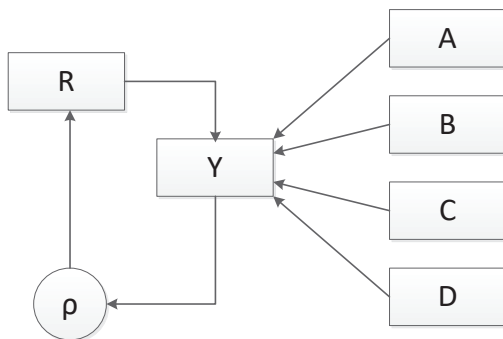


Figure 12. Fay-Mean (MNAR) model syntax in Latent GOLD 5.0

```

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, iY nominal;
10 independent A nominal, B nominal, C nominal, D nominal;
12 latent q_Y nominal 3;
13 equations
14 q_Y <- 1 + A + B + C + D;
15
16 iY <- 1 + q_Y;
17
18 Y <- (w3~wei) q_Y;
19
20 w3 <- {1 0 0 0 1 0 0 0 1};

```

looks similar to Figure 12. 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 be mixed with Fuchs' method.

Discussion

This report demonstrates how to fit FIML models that compensate for item nonresponse in categorical data analysis using Latent GOLD 5.0. Here we consider and contrast the two approaches (Fuchs' and Fay's) and discuss a few noteworthy results that emerged.

Implementing Fay's method in LG presents a few unique challenges. Applying Fay's method to only one variable in the model with LG requires the creation of a response indicator, two quasi-latent variables, and four equations. When missingness for more than one variable is being modeled, the addition of these variables and equations can result in models that sometimes do not converge. Specifically, the models invoking Fay's FIML approach on all variables may produce convergence warnings regarding boundaries and rank deficiencies. These warnings may be the result of the additional response indicators, quasi-latent variables, and equations required to apply Fay's approach in LG, which may indicate models with questionable stability. In some cases, Bayes smoothers can be used to resolve these warnings; directions for

implementing Bayes smoothers in LG can be found in the technical manual.

An important advantage of the ability to fit MNAR models is to test whether missingness for one or more variables is MNAR or MAR. The appropriate missing data mechanism can be tested

by visually comparing estimates from MAR and MNAR models; when estimates differ significantly between the two methods in either direction, then a MNAR nonresponse mechanism may be present. The disadvantages of treating item missingness as MNAR are larger variances and increased model complexity, which can lead to model instability due to weak identifiability and local minima (Bartolucci, Farmcomeni, & Pennoni, 2013; Biemer, 2011).

It is possible that not all variables in a model are subject to the same missing data mechanism. Given the flexibility of LG, the code demonstrated in this methods report may be adjusted to model a mix of FIML MAR and MNAR approaches. Mixing these techniques might produce better estimates, but the impact on complexity, variance, and burden of fitting such models must be considered.

In conclusion, these FIML techniques are best used as a set. When MAR and MNAR FIML models are

fit to categorical data, the nature of the missing data mechanism can be identified through model comparison. Given the potential difficulties of coding and fitting Fay's response indicator models in LG, we recommend using LG's default procedure that uses Fuchs' approach whenever a MAR mechanism seems appropriate based on visual comparison of the estimates and fit statistics of the considered models. However, Fay's response indicator approach is currently the only choice for fitting MNAR models in LG. For testing whether missingness is MNAR or MAR, Fay's method should be used—but for best results, use it on the MNAR variable. Fitting various models under MNAR and MAR nonresponse mechanisms is recommended to identify variables with missing data that are MNAR. This approach will also identify when MNAR models produce invalid estimates due to either model convergence errors or software coding errors.

References

- Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, *52*(278), 200–203. <https://doi.org/10.1080/01621459.1957.10501379>
- Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling* (pp. 243–277). Mahwah, NJ: Erlbaum.
- Asparouhov, T. & Muthén, B. (2010). Multiple imputation with MPlus. *MPlus Web Notes*. Retrieved from <http://www.statmodel2.com/download/Imputations7.pdf>
- Bartolucci, F., Farmcomeni, A., & Pennoni, F. (2013). *Latent Markov models for longitudinal data*. Boca Raton, FL: CRC Press.
- Biemer, P. P. (2011). *Latent class analysis of survey error*. Hoboken, NJ: Wiley.
- Edwards, S., Berzofsky, M. E., & Biemer, P. P. (2017). Effect of missing data on classification error in panel surveys. *Journal of Official Statistics*, *33*(2), 551–570. <https://doi.org/10.1515/jos-2017-0026>
- 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. <https://doi.org/10.1080/01621459.1986.10478279>
- 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. <https://doi.org/10.1080/01621459.1982.10477795>
- Graham, J. W. (2003). Adding missing-data relevant variables to FIML-based structural equation models. *Structural Equation Modeling*, *10*(1), 80–100. https://doi.org/10.1207/S15328007SEM1001_4
- Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*(1), 549–576. <https://doi.org/10.1146/annurev.psych.58.110405.085530>

- Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *The Annals of Economic and Social Measurement*, 5, 475–492.
- Little, R. J. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421), 125–134.
- Little, R. J., & Rubin, D. B. (2002). *Wiley series in probability and statistics: Statistical analysis with missing data* (2nd ed.). Somerset, NJ: Wiley. <https://doi.org/10.1002/9781119013563>
- Muthén, L., & Muthén, B. (1998–2011). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592. <https://doi.org/10.1093/biomet/63.3.581>
- Vermunt, J. K. (1997). *Log-linear models for event histories*. London: Sage.
- Vermunt, J. K., & Magidson, J. (2005). *Latent GOLD 4.0 user's guide*. Belmont, MA: Statistical Innovations.
- Vermunt, J. K., & Magidson, J. (2016). *Technical guide to Latent Gold 5.1: Basic, advanced, and syntax*. Belmont, MA: Statistical Innovations.

RTI International is an independent, nonprofit research institute dedicated to improving the human condition. We combine scientific rigor and technical expertise in social and laboratory sciences, engineering, and international development to deliver solutions to the critical needs of clients worldwide.

www.rti.org/rtipress

RTI Press publication MR-0038-1809