2023 ■ Vol. 19 ■ no. 2

Handling planned and unplanned missing data in a longitudinal study

Mathieu Caron-Diotte [a], Mathieu Pelletier-Dumas [a], Éric Lacourse [b], Anna Dorfman [c], Dietlind Stolle [d], Jean-Marc Lina [e] & Roxane de la Sablonnière [a] ✉

[a] Department of Psychology, Université de Montréal, Montréal, Canada
[b] Department of Sociology, Université de Montréal, Montréal, Canada
[c] Department of Psychology, Bar-Ilan University, Ramat Gan, Israel
[d] Department of Political Science, McGill University, Montréal, Canada
[e] Department of Electrical Engineering, École de technologie supérieure, Montréal, Canada

Abstract ■ While analyzing data, researchers are often faced with missing values. This is especially common in longitudinal studies in which participants might skip assessments. Unwanted missing data can introduce bias in the results and should thus be handled appropriately. However, researchers can sometimes want to include missing values in their data collection design to reduce its length and cost, a method called "planned missingness." This paper review the recommended practices for handling both planned and unplanned missing data, with a focus on longitudinal studies. The current guidelines suggest to either use Full Information Maximum Likelihood or Multiple Imputation. Those techniques are illustrated with R code in the context of a longitudinal study with a representative Canadian sample on the psychological impacts of the COVID-19 pandemic.

Keywords ■ missing data, unplanned missingness, planned missingness, full information maximum likelihood, multiple imputation..

✉ mathieu.caron-diotte@umontreal.ca

10.20982/tqmp.19.2.p123

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers
■ No reviewers

Introduction

Almost every study can be affected by missing data. Participants can forget or refuse to respond to some items. Technical problems can also prevent or corrupt the recording of data. The missing data problem is especially prevalent in longitudinal studies in which, in addition to items being left unanswered, participants can decide to skip a survey assignment or even drop out. In some cases, however, missing data might be planned. Indeed, to reduce data collection costs and participant fatigue, researchers may decide that some measures will not be taken by every participant (Little & Rhemtulla, 2013; Rhemtulla & Little, 2012). This voluntary introduction of missing values in the data collection process is referred to as "planned missingness" (Graham et al., 2006).

Whatever might be the underlying cause, missing data can introduce bias in the statistical analyses (Enders, 2010). Thus, researchers should seek to handle missing data with statistical techniques designed to mitigate the impact of missing values on parameter estimates, while still yielding valid information (Graham, 2009). However, this task can be daunting, especially as there are many choices that must be made in the process of handling missing data. The present paper is aimed to review and explain two of the recommended techniques to deal with missing data, full information maximum likelihood (FIML) and multiple imputation (MI; e.g., Enders, 2010; Graham, 2009) and to provide practical advice regarding their use. Specifically, we focus on their use in the context of longitudinal studies, in conjunction with the application of planned missingness strategies in the data collection process. We provide real examples in R (R Core Team, 2023) in the context of a Canadian representative longitudinal study on the psychological impacts of the COVID-19 pandemic (de la Sablonnière et al., 2020).

Missing Data

To determine how to handle missing data, it is best to understand the relationship that the missingness has with the other data included in the analysis. This section reviews the missing data theory, and details the "planned missingness" strategy. This section concludes with a description of two methods designed to mitigate the negative impacts of missing data: FIML and multiple imputation.

Mechanisms of Missingness

In the missing data literature, "mechanisms of missingness" are used to describe how the missingness is related to other values (Rubin, 1976). Three mechanisms have been coined to describe the missingness: missing at random (MAR), missing completely at random (MCAR) and missing not at random (MNAR) (Rubin, 1976). These mechanisms describe the relation between the probability of missingness and the observed and unobserved variables (Enders, 2010).

Missing completely at random (MCAR) describes situations in which the probability of missing data has no association with observed or unobserved data (Rubin, 1976). This is the "best" case, as there are no differences between those who present missing data and those who don't. Thus, analysis on MCAR data is unbiased. An example of MCAR data could be that respondents simply forget to complete an assessment, without any relation to any other variable.

Missing at random (MAR) describes situations in which the missingness is related with data that is observed in the dataset, but not on unobserved data (Rubin, 1976; Schafer & Graham, 2002). In other words, data is MAR when the probability of it being missing depends upon the value of another variable. This is the case when an individual drop out of a longitudinal study about COVID-19 or refuse to indicate their adherence to sanitary measures because they hold conspiratorial beliefs (for which data would be available).

Missing not at random (MNAR) describes cases in which the missing data depends upon unobserved data (Rubin, 1976). In this case, the probability of missingness depends upon the value of the missing variable itself. For instance, there could be a positive or negative association between the variable and its probability of missingness. Because the missingness depends upon the value of what is missing, MNAR is difficult to handle (Enders, 2010). An example of MNAR missingness would be when high income participants are less likely to disclose their earnings or when participants who hold conspiratorial beliefs refuse to answer a scale measuring adherence to conspiratorial theories.

The way missing data is distributed among the variables investigated is referred to as "patterns of missingness" (Enders, 2010; Graham, 2009). Those patterns describe the missingness over a set of variables by regrouping together observations with similar missingness distributions. The inspection of the patterns can reveal if groups of items are frequently left unanswered (i.e., follow a monotonic pattern) (e.g., Enders, 2010; van Buuren, 2018). More importantly, they can be used to assess the plausibility of the MCAR mechanism over a set of variables by investigating if their means vary across the patterns of missingness (R. J. A. Little, 1988).

Missing data due to MCAR, MAR, and MNAR have deleterious consequences on data analysis. As less observations can be used in the analysis, the statistical power is reduced. Moreover, analyses performed on MAR and MNAR data yield biased estimates, as the probability of missingness depends on the value taken by observed or unobserved variables. Consequently, there is a need to handle any unexpected missingness with statistical techniques to regain statistical power and obtain more accurate estimates. However, the proprieties of MCAR data can be used to simplify the data collection process.

Making Missing Data Work for Us: Planned Missingness

Even if missing data come with complications, there are some reasons researchers might want to introduce missingness themselves into their data. Indeed, planning to introduce missing values completely at random in the data collection allows to reduce the length of the procedure and its global cost without compromising on quality, which represents a benefit (Enders, 2010; T. D. Little, 2013; Rhemtulla & Little, 2012; Rioux et al., 2020). Indeed, as the probability of missingness is random and unrelated to any observed or unobserved variable, no bias is introduced (Rioux et al., 2020). This voluntary introduction of missing values by researchers is called "planned missingness". There are three main planned missingness designs which can be used for longitudinal research: the multiform design, the wave missing design, and the two-method measurement design (T. D. Little, 2013; Rhemtulla & Little, 2012).

In the case of the multiform design, multiple versions of the survey are elaborated, with different groups of items included in each version. For instance, each participant may be asked to provide answers to two thirds of the items from a scale. In this type of design, scales that are not of central importance for the research question are separated into sets. Items are divided into a core set presented to all participants (X), and, for a three-form design, into three other sets (A, B, and C), presented only to a subset of respondents. Items central to the research question or which can explain away some of the missingness (Graham et al., 2006) are included in the set X. Other items are assigned to sets A, B, or C. When data collection is underway, participants are randomly assigned to one of the three versions of the questionnaire. Each version of the questionnaire contains block

X and the items from two blocks - AB, AC, or BC (T. D. Little, 2013; Rhemtulla & Little, 2012). In a longitudinal study, in each wave, participants can be assigned to a version of the questionnaire at random and independently of their assignment in the previous waves of data collection. This design can easily be adapted to more versions, by breaking the items into more sets.

The wave missing design builds on patterns of missingness over the respondent's participation to waves of data collection to introduce random missing values, that is a whole survey into the sample. In this design, some respondents are scheduled to not be contacted for some of the measurement occasions (Little & Rhemtulla, 2013). The number of missing surveys can vary, and even follow a monotonic pattern, with some respondents only being contacted for two waves (Enders, 2010; Rioux et al., 2020). The wave missing design can however be less efficient than the multi-form design, suffering from less precise estimates (e.g., see Rioux et al., 2020; Wood et al., 2019).

With the two-method measurement design, two instruments are used to measure the same construct. One of those instruments is seen as more precise and costlier, while the other instrument less precise, but cheaper. It is of paramount importance that the two instruments should still be highly correlated (Enders, 2010; Graham et al., 2006). While the cheaper instrument is responded to by the whole sample, the costlier instrument is answered only by a subset of participants. The exact number of respondents completing both measures should be determined by the expected correlation between their outcomes. Because the two measures are highly correlated, it would then be possible to estimate adequately the missing values of the costlier instrument, at a fraction of the cost.

In short, in planned missing data designs, some measurements or assessments are selected at random to be missing for some participants. Because such data are MCAR (i.e., the probability of missingness is unrelated to other variables in the study), no bias is introduced. To ensure the quality of the data collected and analysed under planned missingness designs, it is best to select with care the items composing the surveys. It is necessary to include in the core set (X) items which are strongly correlated to those of the other sets (Graham et al., 2006). The sample size is also of consideration; according to Rhemtulla and Little (2012), planned missingness should be used with samples of at least 375 participants, to estimate reliably the covariances across the items. For designs with only one time-point, the necessary sample size might be lower (Rioux et al., 2020). Under those conditions, as with unplanned missing data, some statistical power can be regained and the parameters adequately estimated through the use of two of the most up to date statistical techniques in missing data: FIML and multiple imputation.

Methods to Handle Missing Data

Two of the state-of-the-art methods that can be easily used to handle missingness are FIML and multiple imputation (Enders, 2010; Graham, 2009; Rioux et al., 2020). Both of those methods are able to tackle planned and unplanned missingness (Rioux et al., 2020) and can outperform more traditional ways to handle missing data (e.g., simple imputation, listwise or pairwise deletion) when performed correctly (Enders, 2010; Hughes et al., 2019). When using FIML or multiple imputation to analyse data containing missing values, three main analysis phases can be distinguished (e.g., Rhemtulla & Little, 2012): 1) a planning phase, 2) an examination phase, and 3) an analysis phase. In the following, each phase is explained and outlined. The mechanisms behind FIML and multiple imputation are detailed with a discussion of the analysis phase.

Planning Phase

In the planning phase, choices are made about how to perform the analysis. Variables to include in the model are selected. When using FIML or MI, it is best to not only include variables related to the research question, but also to include auxiliary variables. Precisely, it is strongly suggested to include variables that share a strong association with the variables of interest or can predict the probability of missingness (Allison, 2012; Enders, 2010; Graham, 2009). The inclusion of such variables help to provide unbiased estimates (Graham, 2009). It is important to note that the techniques for handling missing data are not causal but predictive and postdictive. As such, in a longitudinal study, variables can be imputed using data from previous or subsequent waves of data collection (Honaker et al., 2011). For a longitudinal study, it is also important to include the variable indicating time (e.g., the wave number) to model its effect.

Still, researchers should not include all the variables in the model. Indeed, with too many variables, there are higher chances that the imputation fails because of multicollinearity or to other computational problems (e.g., Graham, 2009; van Buuren, 2018). Some authors suggest including a maximum of 100 variables (Graham, 2009), while others recommend to use only 15 to 25 variables (van Buuren, 2018). To keep the number of variables to a minimum, it is possible to compute the overall values for some or all scales (Graham, 2009). FIML and multiple imputation can be performed on items, scales, or a mix of both (Enders, 2010; Graham, 2009).

Examination Phase

After having selected the variables, it is time to examine and report any unplanned missingness (Enders, 2010). Rhemtulla and Little (2012) suggest to perform a single imputation on data that is missing due to planned missingness, so that only unplanned missingness remains. Researchers can then characterize this unplanned missingness by assessing the patterns of missingness and using Little's MCAR test (Enders, 2010; R. J. A. Little, 1988). These analyses can provide indications on whether the missing data mechanism is MCAR or not (i.e., there is no underlying factor that explains why some data is unobserved). Note, however, that FIML and MI should still be considered for dealing with missing data, no matter what is the underlying missingness mechanism (e.g., see Enders, 2010). The proportion of missingness for each variable of interest can then be reported (Enders, 2010).

Full Information Maximum Likelihood

FIML is a method that is used to estimate the parameters of interest (e.g., regression coefficients) as if there was no missing values. When FIML is used, an iterative algorithm tries to find the set of parameter values for which the likelihood of having produced the observed data is maximized (Allison, 2012; Enders, 2010). This can be done with or without auxiliary variables. To include auxiliary variables in Structural Equation Models (SEM), covariances are added between the auxiliary variables (entered as manifest variables) and the error terms of other manifest variables (Graham, 2003). Auxiliary variables can also be added to multilevel models by coding their values as if it was an additional measurement of the dependent variable (Allison, 2012). FIML can be used directly in the statistical analysis model with many statistical packages such as Amos, R, Mplus, and SAS.

FIML handles missingness well on any outcome variables (Graham, 2003; van Buuren, 2018). However, multivariate normality needs to be assumed; FIML can handle some deviation from normality, but if this assumption is too heavily violated, it can yield unreliable estimates (e.g., van Buuren, 2018). In some cases, it is advised to use a robust FIML estimator to handle deviations from normality. Moreover, it is important to note that cases with missingness on the predictors are discarded when using FIML, which can reduce the statistical power. Researchers can however include auxiliary variables to prevent deletion of cases. Furthermore, in some situations, FIML can be problematic as some fit indices become unavailable with missing data (e.g., in SPSS Amos, SRMR, modification indices, standard errors and bootstrapping). A final caveat is that, while the inclusion of auxiliary variables can help to provide accurate estimates, it can also worsen a model's fit and complexify its convergence and identification. For those situations, researchers may decide to use multiple imputation instead of FIML.

Multiple Imputation

MI is a process in which multiple datasets are produced, in which missing values are replaced by new values predicted from other available data, with some random variation added (Enders, 2010). Multiple slightly different copies of the complete dataset are produced to account for the uncertainty inherent to data (Enders, 2010; van Buuren, 2018). For many softwares, the default number of imputed datasets is set to five, but there are no statistical advantages to have a small number of datasets. The recommended number of imputed datasets is at least 50 (Enders, 2010). As a general rule, higher number of imputations are recommended. The same analysis is then performed on each complete dataset and the parameters of interest are then pooled using Rubin's rules (Rubin, 1987). Analyses performed on multiply imputed datasets can give unbiased estimates, even in the presence of a large proportion of missing data. Some simulation studies found that unbiased parameter estimates can be recovered with even up to 90% of missing values when multiple imputation is used (Madley-Dowd et al., 2019).

Multiple imputation needs to be performed before the analyses. Specifying the MI model (i.e., selecting variables to include, specifying predictors, if needed) can be relatively tedious and the calculations needed for the imputation process can be quite heavy computationally and time consuming. Similarly, the analysis performed after the MI procedure can be lengthy, as it is repeated for each imputed dataset and the parameter estimates pooled. Automatic pooling of the estimates may not be available for some analyses and some statistical software. For instance, in SPSS, factorial ANOVA and ANCOVA procedures do not support the automation process to generate pooled results.

Handling Missing Data: Concrete Examples in R

We now present some concrete advice on how to use FIML and MI with R code (R Core Team, 2023). Indeed, some R packages include useful procedures to easily handle missing data, even if for some applications, some data manipulation is needed. In what follows, code snippets allowing to examine the missingness and to perform analyses with FIML and multiple imputation on a longitudinal dataset are presented and explained.

As a motivating example, we aimed to assess the impacts of the COVID-19 pandemic on Canadian's quality of sleep. The research question is to determine if there were periods in which Canadians were not getting high quality

Listing 1 ■ R code to create a figure of missing patterns.

```
sleep <- read.csv("data/sleep.csv") # load data

library(VIM) # load the VIM library, containing the aggr() function
md_patterns <- aggr(sleep[,-1], # we remove the first column (id)
                    col = c("lightblue","white"),
                    numbers = TRUE, sortVars = TRUE,
                    labels = names(sleep[,-1]), cex.axis = .7,
                    gap = 3, ylab = c("Missing data", "Pattern"))
```

sleep. We also sought to see if the respondents' emotional state could be associated with changes in their quality of sleep. The data used for this tutorial is taken from a representative panel of Canadians followed during the COVID-19 pandemic ($N = 3617$). More information on the study can be found elsewhere (de la Sablonnière et al., 2020). For the needs of the present tutorial, a random subsample of 375 respondents, and their responses to the waves 1 to 4, was selected. The data file is freely available online.[1]

Planning Phase

For our longitudinal survey on the psychological impacts of the COVID-19 pandemic (de la Sablonnière et al., 2020), we opted to use a multiform design (Graham, 2009; T. D. Little, 2013; Rhemtulla & Little, 2012) for each assessment. This type of design was selected as it would allow us to reduce the number of items presented to each respondent while keeping at a maximum (notwithstanding attrition) the number of respondents on each wave (Rioux et al., 2020; Wood et al., 2019). We created three versions of each survey, in which we divided the items into four sets. The core set, presented to every respondent, included the sociodemographic items and those of critical importance to the project's research question. The remaining items were separated into each set. We made sure to include some items from every scale into the core set; the sufficiently strong correlations between these items would help in the estimation process. Items related to sleep and the emotional state were part of the multiform questionnaire.

To explore sleep quality during the pandemic and the impact of emotions, the analysis plan includes self-reports of sleep quality (the dependent variable), and three items assessing whether participants felt lonely, nervous, and angry (the independent variables). To improve the estimation process by FIML and multiple imputation, we included auxiliary variables. The number of minutes taken to fall asleep the night before and the number of minutes slept, the valence of the last dream, and other emotions were selected as they are likely to be strongly related to the dependent and independent variables. Sociodemographic variables (gender, province of residence, and age) were also considered.

Examining Missing Data

Having selected the variables to include, we then examined the missingness in the sample. It is useful to produce the proportion of missingness for each respondent as well as for each variable. The R code found in Listing 1 can produce a figure displaying the patterns of missing data with their frequencies as well as the proportion of missing values for each variable.

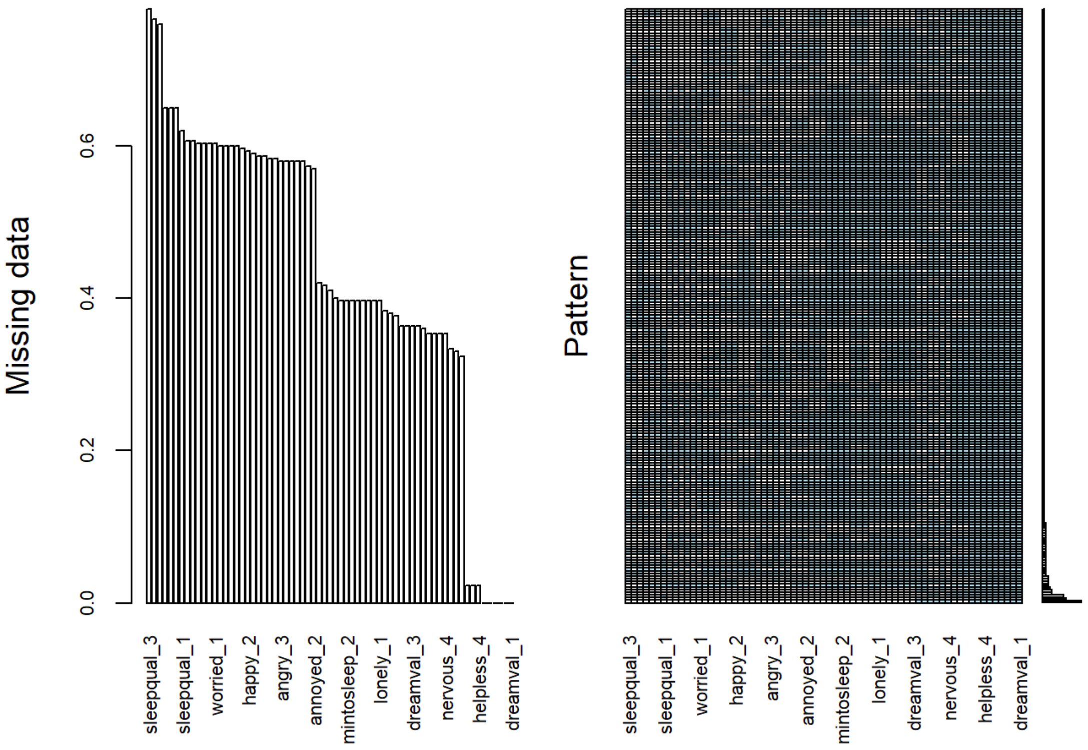
Figure 1 presents the missing patterns plot for the sample. This figure was created by plotting the patterns of missingness for the items in the database. On the left side of the figure, we find the proportion of missing values for each items (i.e., the percentage of respondents who did not answer the item). On the right side, there is a graphical representation of each missing pattern (answered items are in blue, missing items in white). On the right side of this matrix is another bar chart, indicating the frequency of each missing pattern.

Finally, we can assess the plausibility that the data are MCAR. The `mcar_test` function from the naniar package can be used to perform Little's MCAR test. The `mcar_test` function takes a data frame as parameter (see Listing 2). The data frame should not contain categorical variables (but ordinal data can be treated as continuous).

The result of the test is an indication of the plausibility of the MCAR mechanism (R. J. A. Little, 1988). If the test is significant, it is an indication that the means of the variables vary across the patterns of missingness and thus, that the data might not be MCAR. However, FIML and multiple imputation stay the recommended practice (Enders, 2010; Rioux et al., 2020), even if the MCAR assumption is not supported. Furthermore, the data could still be MAR if some variables in the dataset are related to missingness, and thus FIML and multiple imputation could provide fairly unbiased results (Enders, 2010). Little's MCAR test

[1] https://doi.org/10.5683/SP3/P8OUOT

Figure 1 ■ Example of missing data patterns for the random subsample of 375 participants. The bar chart on the left presents the proportion of non-completed surveys. The chart on the right shows the missingness patterns, blue squares denote answered items, white squares missing values. On the right of the pattern chart, bars represent the frequency of each pattern.



Listing 2 ■ R code to perform Little's MCAR test.

```
library(naniar) # load the naniar package, for the mcar_test function
mcar_test(sleep[,-c(1:3)]) # remove id and categorical data
```

is not significant, suggesting that the data could be MCAR, $\chi^2(7274, N = 300) = 720.000, p = .983$.

Full Information Maximum Likelihood

FIML estimates the parameters (i.e., the coefficients) by using all observed data. As such, whenever analyses are performed with FIML, the model needs to be specified. Auxiliary parameters should properly be included in the model to improve the estimation of the parameters (Allison, 2012; Graham, 2003). We first demonstrate the use of FIML in R with the lavaan package (Rosseel, 2012) for two types of models: regression and growth curve modeling.[2] Then, we show how to apply FIML with multilevel regression (also called mixed-effects or hierarchical models) using the lme4 package (Bates et al., 2015).

Multiple Regression

The first model we present is a multiple regression for variables found at the first longitudinal assessment, to easily explain how to include auxiliary variables. In this example, we regress loneliness, nervousness and angriness scores (i.e., three continuous predictors) on sleep quality. In this first model, we do not use FIML nor auxiliary variables: In this example, the valence of the last dream is used as an auxiliary variable. Adding auxiliary variables to a regression model can be done by using the "saturated correlates"

[2] It is to be noted that the auxiliary(),lavaan.auxiliary(), cfa.auxiliary(),sem.auxiliary(), and growth.auxiliary() functions from the semTools package (Jorgensen et al., 2022) can be used to automatically introduce auxiliary variables into a lavaan model. However, for completeness, we present a full example with lavaan model syntax.

approach: covariances between the auxiliary variable and all the other variables are included (Allison, 2012; Graham, 2003).

In Listing 3, we specify the model as a string variable using the `lavaan` syntax (Rosseel, 2012); predicted variables are on the left-hand side, followed by a tilde (~) and the prediction equation (predictor variables are separated by a +). Covariances are denoted with double tildes. To use FIML to estimate the coefficients with missing data, we add `missing = "fiml"` to the `sem()` function. Specifying `fixed.x = FALSE` allows to estimate the means, variances, and covariances on all exogenous (independent) variables, which is required for FIML. Finally, the call to the `summary()` function prints the analysis results; `fit.measures = TRUE` retrieves the model's fit, `rsquare = TRUE` yields the R^2, and `standardize = TRUE` prints the standardized coefficients. The results of regressions using FIML can be interpreted as any other regression.

Note the covariance structure of this model. Not only the means, variances, and covariances were estimated for all exogenous variables, but covariances between the auxiliary variable (valence of the last dream) and the variables of interest are also established. This "saturated correlates" model is central to the use of auxiliary variables in SEM (Graham, 2003). These same principles apply whenever we wish to use FIML with SEM to handle missing data.

Growth Curve Modeling

Growth curve modeling is a type of analysis which allows to examine how an outcome varies over time, and how other variables influence this variation. This type of analysis is well suited for longitudinal data. In `lavaan`, growth curves are estimated by a model with two latent variables: a random intercept (i) and a random slope (s). Predictors can be included for those two latent variables (i.e., here gender and age), as well as time-varying predictors (regressed on each measurement). The predictor variables can model the influence of stable characteristics on the overall level of the dependent variable (the intercept) as well as the impact of time (the slope). Time-varying predictors allow to assess if the dependent variable varies conjointly in time with another. As with regression analysis with `lavaan`, missing values can be estimated using FIML. Means, variances, and covariances of the independent variables need to be estimated. Auxiliary variables can be specified by using the "saturated correlates" approach. The code in Listing 4 presents such a model.

Multilevel Regression

As with growth curve modelling, multilevel regression allows to model how an outcome varies over time and the variables associated with this change. Multilevel regression is particularly robust with missing data on the dependent variable; there is no real need for extra steps in this case (Bates et al., 2015). However, some adjustments are needed with missing values on the independent variables. In this case, it is necessary to use auxiliary variables (Allison, 2012).

The `lme4` package can be used to fit multilevel regressions (Bates et al., 2015). However, the data frame must be in the long format, such that each row represents one measurement occasion (i.e., one wave) for one participant. This data format allows to perform a programming trick to include auxiliary variables: the auxiliary variable is treated as another measurement of the dependent variable (Allison, 2012). To make sure the computations include the auxiliary variable, an index is created (here named `D`) to distinguish the outcome from the auxiliary. The code from Listing 5 transposes the data frame from wide to long format and then performs manipulations to introduce measurements of the auxiliary variable (the valence of the last dream) on the same column as the outcome variable (sleep quality).

We then perform a multilevel longitudinal model in which we predict sleep quality with time (`wave`) and gender (see Listing 6). The variable `D` is also included for the FIML computations. It is necessary to include interaction terms between `D` and each of the other variables (Allison, 2012). To interpret the results of this model, we examine the main effects, but not the interaction terms.

Multiple Imputation

Multiple imputation is a process that produces multiple estimates for each missing value. The observed values are used as predictors. Contrary to FIML, when all predictors, dependent and auxiliary variables, are included in the imputation model, multiple imputation can be done once for a given set of analyses. This imputation process can be done on single items or on a composite score (i.e., mean on a scale) (Graham, 2009; Gottschall et al., 2012). However, it can be useful to perform some operations (i.e., computing scores, centering values) on the data frame before running the imputation. Indeed, when an interaction effect is tested, the interaction term has to be computed beforehand, so that its effect can be fully estimated on the missing (Enders, 2010). Two commonly used packages to impute data in R are `mice` (van Buuren & Groothuis-Oudshoorn, 2011) and `Amelia` (Honaker et al., 2011).

Using the mice Package

The package `mice` (van Buuren & Groothuis-Oudshoorn, 2011) is useful for dealing with missing data, because of its many features which give the user full control over the

Listing 3 ■ R code to perform a multiple regression with FIML.

```
library(lavaan) # load the lavaan library to perform analyses with FIML
model <- "
  # regressions
    sleepqual_1 ~ lonely_1 + nervous_1 + angry_1
  #covariances
    lonely_1 ~~ dreamval_1
    nervous_1 ~~ dreamval_1
    angry_1 ~~ dreamval_1
    sleepqual_1 ~~ dreamval_1
"
fit <- sem(model,
           data = data,
           missing = "fiml",
           fixed.x = FALSE)
summary(fit, fit.measures = TRUE, rsquare = TRUE, standardized = TRUE)
```

imputation model. For instance, mice allows to include a predictor in the imputation model without imputing its values (or even remove it entirely from the imputation model). Furthermore, the micemd (Audigier & Resche-Rigon, 2019) package can be used to extend the mice package for longitudinal data.

In Listing 7, we perform an empty imputation using the mice() function. By specifying only one imputed dataset (m) and no maximum number of iterations (maxit), the function returns an almost empty mice imputed data frame object, which we store in an object called ini. The predictorMatrix part of this object describes relations between the variables in the imputation model, specifying which variables are imputed and which variables are used as predictors. Rows indicate the predicted variables while columns indicate the predictors. For any row, a 1 in a column indicates that it is predicted by the variable which name is in the column; while a 0 indicates that it is not predicted by it. For longitudinal multilevel data, we set the grouping variable (here the respondent is) to -2 and all other variables to be predicted to 2, which indicates a variable predicted by a multilevel imputation model. These operations give us a custom predictor matrix, located in ini$predictorMatrix.

The multiple imputation process can then be performed by using the mice() function and specifying ini$predictorMatrix for the custom imputation model and the method for the imputation methods for each variable, as in Listing 8. The number of imputed datasets (m) is set to 50, the maximum number of iterations (maxit) to 35, and includes a seed for the random number generator to ensure reproducibility. We store the multiply imputed datasets in a new object called imp. When the multiple imputation is finished, we assess the convergence of the process with the print() and plot() functions, as an integrated procedure to estimate the parameters of the chained equations that have been used (van Buuren, 2018; van Buuren & Groothuis-Oudshoorn, 2011). All the chains should be converging towards the same values. The with() function is called to perform the analysis (e.g., the same multilevel model) on each imputed data frame and to store the results in an object named fit. Finally, results from each model are pooled using the pool() function (Audigier & Resche-Rigon, 2019), as seen in Listing 9.

Pooling procedures for many analyses have been implemented in the mice and miceadds packages (van Buuren, 2018). These can be performed using the with() and pool() combo.

Using the Amelia Package

The package Amelia (Honaker et al., 2011) is another option to deal with missing values. The algorithm used by the Amelia package has been extensively tested and is fast (Enders, 2010). However, it does not allow to control predictors and predicted variables, but can be easily used to perform multiple imputation on longitudinal and multilevel data.

We can launch the imputation process using the amelia() function, as seen in Listing 10. We set the number of imputed datasets (m) to 50, and specify nominal (noms) variables, which names are passed as character vectors.[3] Bounds for numeric variables are passed as a n (number of columns for which we specify bounds) by 3

[3] Ordinal variables can also be specified by passing a character vector to the ords parameter. However, to increase the computation and to improve the accuracy of the prediction, it is recommended to treat ordinal scales as continuous variables unless their ordinal nature is important for the analysis (e.g., see Enders, 2010; Honaker et al., 2011).

Listing 4 ■ R code to perform a growth curve analysis with lavaan.

```
model.growth <- "
  # intercept and slope
    i =~ 1*sleepqual_1 + 1*sleepqual_2 + 1*sleepqual_3 + 1*sleepqual_4
    s =~ 0*sleepqual_1 + 1*sleepqual_2 + 2*sleepqual_3 + 3*sleepqual_4
  # regressions
    i ~ gender + age
    s ~ gender + age
  # time-varying covariate
    sleepqual_1 ~ nervous_1
    sleepqual_2 ~ nervous_2
    sleepqual_3 ~ nervous_3
    sleepqual_4 ~ nervous_4
   # covariances (incl. auxiliary variable)
    dreamval_1 ~~ sleepqual_1
    dreamval_1 ~~ sleepqual_2
    dreamval_1 ~~ sleepqual_3
    dreamval_1 ~~ sleepqual_4
    dreamval_2 ~~ sleepqual_2
    dreamval_2 ~~ sleepqual_3
    dreamval_2 ~~ sleepqual_4
    dreamval_3 ~~ sleepqual_3
    dreamval_3 ~~ sleepqual_4
    dreamval_4 ~~ sleepqual_4
    dreamval_1 ~~ nervous_1
    dreamval_1 ~~ nervous_2
    dreamval_1 ~~ nervous_3
    dreamval_1 ~~ nervous_4
    dreamval_2 ~~ nervous_2
    dreamval_2 ~~ nervous_3
    dreamval_2 ~~ nervous_4
    dreamval_3 ~~ nervous_3
    dreamval_3 ~~ nervous_4
    dreamval_4 ~~ nervous_4
"
fit.growth <- growth(model.growth,
                     data = data,
                     missing = "fiml",
                     fixed.x = FALSE)
summary(fit.growth,
        fit.measures = TRUE,
        rsquare = TRUE,
        standardized = TRUE)
```

(column number, lower bound, upper bound) matrix to the bounds parameter. Of importance when imputing longitudinal data, the variable representing time should be specified with the ts parameter, and the cross section with the cs parameter. To model the effect of time, we can specify its power of polynomial with the polytime parameter (0: constant, 1: linear, 2: quadratic, 3: cubic). By default, all columns contained in the data frame passed to the function are used as predictors and are imputed. We can specify which variables will not be included in the imputation process by passing them as a character vector to the idvars parameter. We also give the random number generator (set.seed()) a seed to ensure reproductibility. To speed up the process, the computations can be done in parallel on multiple cores (see the parallel, ncpus, and cl parameters). The imputed datasets are then stored in an object called imp.amelia.

To perform the analyses over the imputed datasets, we rely on the merTools package (Knowles & Frederick, 2020). To do so, the model's formula and the list containing the imputed datasets (here, imp.amelia$imputations) are specified, as in Listing 11. The pooled estimated for this analysis can then be obtained automatically with the summary() function.

Conclusion

This paper was aimed to provide researchers and students with an understanding of planned and unplanned missingness as well as to equip them with methods to deal with these issues. FIML and multiple imputation are two vali-

Listing 5 ■ R code to restructure the data frame and to include an auxiliary variable (D).

```
sleep_long <- reshape(sleep,
                      idvar = "id",
                      varying = 5:68,
                      direction = "long",
                      timevar = "wave",
                      sep = "_")
row.names(sleep_long) <- 1:nrow(sleep_long) # fix the row numbers
sleep_long$wave <- sleep_long$wave - 1 # set first wave to 0 and so on
# Coding of the auxiliary variable (dreamval)
aux <- sleep_long[,c("id","wave","gender","dreamval")]
aux$D <- 1 # indicates that these values are the auxiliary variable
names(aux)[4] <- "sleepqual" # rename the auxiliary variable so that they are
   treated as the outcome variable
sleep_long.aux <- sleep_long[,c("id","wave","gender","sleepqual")]
sleep_long.aux$D <- 0 # indicates that these values are the real outcome
sleep_long.aux <- rbind(sleep_long.aux, aux) # merge both data framed by column
# the column containing the outcome variable now contains the auxiliary variable (
   denoted as D = 0)
```

Listing 6 ■ R code to perform a multilevel regression with an auxiliary variable.

```
library(lme4) # lme4 package to perform multilevel models
library(lmerTest) # to obtain p-values with lmer()
fit_lme.aux <- lmer(sleepqual ~ D + gender + wave + gender:wave + D:gender + D:wave
    + D:gender:wave + (1 | id), data = sleep_long.aux)
summary(fit_lme.aux)
```

dated state-of-the-art methods to handle planned and unplanned missing data (Enders, 2010; Rioux et al., 2020). Both methods allow obtaining unbiased estimates and retaining statistical power. Researchers using the data to address specific research questions must carefully choose the appropriate method, depending on the research question, the data, and the planned statistical analyses. We hope the suggestions in this report are helpful to guide researchers and students in handling the missingness in their lives.

Authors' note

This work is part of a project financed by the Canadian Institutes of Health Research (Grant No. 170633).

References

Allison, P. D. (2012). Handling missing data by maximum likelihood. *Statistics and Data Analysis (Paper 312-2012)*. https://statisticalhorizons.com/wp-content/uploads/MissingDataByML.pdf

Audigier, V., & Resche-Rigon, M. (2019). *Micemd: Multiple imputation by chained equations with multilevel data* (Version 1.6.0). https://CRAN.R-project.org/package=micemd

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01.

de la Sablonnière, R., Dorfman, A. R., Lina, J.-M., Pelletier-Dumas, M., Stolle, D., Taylor, D. M., Benoît, Z., Boulanger, A., Caron-Diotte, M., Mérineau, S., & Nadeau, A. (2020). *COVID-19 Canada: The end of the world as we know it? Technical report: Presenting the COVID-19 Canada Survey*. Université de Montréal. Montréal, Canada. https://csdc-cecd.wixsite.com/covid19csi/resultats

Enders, C. K. (2010). *Applied missing data analysis*. Guilford.

Gottschall, A. C., West, S. G., & Enders, C. K. (2012). A comparison of item-level and scale-level multiple imputation for questionnaire batteries. *Multivariate Behavioral Research*, *47*(1), 1–25. doi: 10.1080/00273171.2012.640589.

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based Structural Equation Mod-

Listing 7 ■ R code to perform an empty imputation and set the imputation model.

```
library(mice)
library(micemd)
library(miceadds)

id.gr <- 1 # index of the column indicating the grouping variable

# Set up the predictor matrix
ini <- mice(sleep_long, m = 1, maxit = 0)
ini$predictorMatrix[-id.gr, id.gr] <- -2 # set id as the grouping variable
ini$predictorMatrix["age",] <- 0 # age is not predicted
ini$predictorMatrix[ini$predictorMatrix == 1] <- 2 # variables to be predicted by
    multilevel imputation

method <- find.defaultMethod(sleep_long, id.gr) # get suggestions of imputation
    models for the data to be imputed
```

Listing 8 ■ R code to perform an imputation on longitudinal data with mice and micemd.

```
imp <- mice(sleep_long, m = 50, maxit = 35,
            predictorMatrix = ini$predictorMatrix,
            method = method, seed = 42)
print(imp)
plot(imp)
```

els. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(1), 80–100. doi: 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. doi: 10.1146/annurev.psych.58.110405.085530.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*(4), 323–343. doi: 10.1037/1082-989X.11.4.323.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, *45*(7). doi: 10.18637/jss.v045.i07.

Hughes, R. A., Heron, J., Sterne, J. A. C., & Tilling, K. (2019). Accounting for missing data in statistical analyses: Multiple imputation is not always the answer. *International Journal of Epidemiology*, *48*(4), 1294–1304. doi: 10.1093/ije/dyz032.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2022). *semTools: Useful tools for structural equation modeling* (Version 0.5-6). https://CRAN.R-project.org/package=semTools

Knowles, J. E., & Frederick, C. (2020, June 23). *merTools: Tools for analyzing mixed effect regression models* (Version 0.5.2). https://cran.r-project.org/package=merTools

Little, R. J. A. (1988). A test of Missing Completely at Random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*(404), 1198–1202. doi: 10.1080/01621459.1988.10478722.

Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford.

Little, T. D., & Rhemtulla, M. (2013). Planned missing data designs for developmental researchers. *Child Development Perspectives*, *7*(4), 199–204. doi: 10.1111/cdep.12043.

Madley-Dowd, P., Hughes, R., Tilling, K., & Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology*, *110*, 63–73. doi: 10.1016/j.jclinepi.2019.02.016.

R Core Team. (2023, April 21). *R: A language and environment for statistical computing* (Version 4.3.0). Vienna, Austria. https://www.R-project.org/

Rhemtulla, M., & Little, T. D. (2012). Planned missing data designs for research in cognitive development. *Journal of Cognition and Development*, *13*(4), 425–438. doi: 10.1080/15248372.2012.717340.

Listing 9 ■ R code to perform a longitudinal multilevel model on data multiply imputed with mice and micemd.

```
fit_lmer.mice <- with(imp.mice, lmer(sleepqual ~ gender + wave + nervous + nervous:
   gender:wave + (1 + wave | id)))
summary(pool(fit_lmer.mice))
```

Listing 10 ■ R code to perform multiple imputation with Amelia.

```
library(Amelia)

# set bounds by creating a n by 3 matrix
bds <- rbind(
  matrix(c(4, 18, 86), # bounds for age (4th column, min = 18, max = 86)
         ncol = 3, byrow = TRUE),
  cbind(matrix(6:19, ncol = 1), # column number of items
        matrix(c(1, 10), # answer scales limits
               nrow = length(6:19),
               ncol = 2, byrow = TRUE)))

set.seed(42) # set seed for reproductibility
imp.amelia <- amelia(sleep_long,
                     m = 50, # number of imputed datasets
                     noms = c("gender","prov"), # nominal variables
                     bounds = bds, # bounds for numeric variables
                     cs = "id", # participant identifier for the cross-sections
                     ts = "wave", # column name identifying the time point
                     polytime = 3, # integer indicating the power of polynomial for
    the effect of time
                     parallel = 'snow', # "snow" for windows, "multicore" for UNIX
   systems (Mac or Linux)
                     ncpus = 4, # number of cores
                     cl = parallel::makePSOCKcluster(4),
                     p2s = 2)
```

Rioux, C., Lewin, A., Odejimi, O. A., & Little, T. D. (2020). Reflection on modern methods: Planned missing data designs for epidemiological research. *International Journal of Epidemiology*, *49*(5), 1702–1711. doi: 10.1093/ije/dyaa042.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. doi: 10.18637/jss.v048.i02.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. doi: 10.1093/biomet/63.3.581.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley. doi: 10.1002/9780470316696.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. doi: 10.1037/1082-989X.7.2.147.

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). CRC. https://stefvanbuuren.name/fimd/

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3). doi: 10.18637/jss.v045.i03.

Wood, J., Matthews, G. J., Pellowski, J., & Harel, O. (2019). Comparing different planned missingness designs in longitudinal studies. *Sankhya B*, *81*(2), 226–250. doi: 10.1007/s13571-018-0170-5.

Open practices

The *Open Data* badge was earned because the data of the experiment(s) are available on doi.org/10.5683/SP3/P8OUOT

The *Open Material* badge was earned because supplementary material(s) are available on the journal's web site.

Listing 11 ■ R code to perform a multilevel regression on multiply imputed data with merTools.

```
library(merTools)
fit_lmer.amelia <- lmerModList(sleepqual ~ gender + wave + nervous + nervous:gender
    :wave + (1 + wave | id), data = imp.amelia$imputations)
summary(fit_lmer.amelia)
```

Citation

Caron-Diotte, M., Pelletier-Dumas, M., Lacourse, É., Dorfman, A., Stolle, D., Lina, J.-M., & de la Sablonnière, R. (2023). Handling planned and unplanned missing data in a longitudinal study. *The Quantitative Methods for Psychology*, *19*(2), 123–135. doi: 10.20982/tqmp.19.2.p123.

Received: 17/10/2022 ~ Accepted: 19/03/2023