UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Chaput-Langlois, S., Stickley, Z. L., Little, T. D., & Rioux, C. (2024). Multiple Imputation When Variables Exceed Observations: An Overview of Challenges and Solutions. *Collabra: Psychology*, *10*(1). https://doi.org/10.1525/collabra.92993

Methodology and Research Practice

# Multiple Imputation When Variables Exceed Observations: An Overview of Challenges and Solutions

Sophie Chaput-Langlois[1][a], Zachary L. Stickley[2], Todd D. Little[3], Charlie Rioux[4]

[1] School of Psychoeducation, University of Montreal, Montreal, QC, Canada, [2] Yhat Entreprises, Lubbock, Texas, United States, [3] Department of Educational Psychology and Leadership, College of Education, Texas Tech University, Lubbock TX, United States, [4] Department of Psychology, University of Oklahoma, Norman, OK , United States

Keywords: missingness, inclusive imputation, broad imputation, MICE, joint modeling, auxiliary variable

https://doi.org/10.1525/collabra.92993

Collabra: Psychology

Vol. 10, Issue 1, 2024

Missing data are a prevalent problem in psychological research that can reduce statistical power and bias parameter estimates. These problems can be mostly resolved with multiple imputation, a modern missing data treatment that is increasingly used. Imputation, however, requires the number of variables to be smaller than the number of observations (i.e., non-missing values), and this number is often exceeded due to, e.g., large assessments, high missing data rates, the inclusion of variables predictive of missing values, and the inclusion of non-linear transformations. Even when the ratio of variables to observations meets the minimum requirement, convergence failure can occur in large, complex models. Specialized techniques have been developed to overcome the challenges related to having too many variables in an imputation model, but they are still relatively unknown by researchers in psychology. Accordingly, this paper presents an overview of four imputation techniques that can be used to reduce the number of predictors in an imputation model: item aggregation with scales and parcels, passive imputation, principal component analysis (PcAux) and two-fold fully conditional specification. The purpose, advantages, limitations, and applications of each method are discussed, along with recommendations and illustrative examples, with the aims of (1) understanding different imputation methods and (2) identifying methods that could be useful for one's imputation problem.

Missing data are a pervasive problem in psychological research that, if not treated correctly, can lead to loss of power and biased results. While deleting cases with missing data is still often done (Lang & Little, 2018; Rioux, Lewin, et al., 2020), modern missing data techniques, like multiple imputation and full information maximum likelihood, can be used to recover the missing information and reduce bias. Multiple imputation is increasingly used (Rioux & Little, 2021), but its use can be held back by its challenging nature, especially when faced with a large imputation, which in this article refers to an imputation where *the number of variables is larger than the number of observations*. Although this situation may seem uncommon, it can arise quickly when including all relevant variables in the imputation, which includes at a minimum all variables and non-linear transformations (e.g., interaction terms) that are in the planned analytical model and the variables that predict the likelihood of missingness on the analyzed variables (van Buuren, 2018). The likelihood of a large imputation problem is higher when the sample is small, the missing data rate is high, or the study is longitudinal, which are all common circumstances in psychological research. Specialized techniques have been developed to overcome the challenges related to large imputations, but they are still relatively unknown. Accordingly, the present paper aims to introduce researchers to these specialized techniques and guide them to the proper literature for their imputation needs. To do so, after reviewing missing data mechanisms and the basics of multiple imputation, we present an overview of four techniques that can be used to help in imputing large databases: item aggregation, passive imputation, principal component analysis, and two-fold fully conditional specification. Table 1 provides a summary of the main terms covered in this article for reference.

## Missing Data Mechanisms

Three missing data mechanisms can explain the presence of missing data in any study: *Missing completely at random* (MCAR), *missing at random* (MAR) and *missing not at*

a Correspondence concerning this article should be addressed to: Sophie Chaput-Langlois, School of Psychoeducation, University of Montreal, 90 Vincent d'Indy, Montréal QC, H2V 2S9, Canada. Email: sophie.chaput-langlois@umontreal.ca

**Table 1. Acronyms and Definitions for Missing Data Terminology**

| Term | Acronym | Definition |
|---|---|---|
| Analytical model | | Analysis model that aims to answer a research question. |
| Auxiliary variable | | Variable that is not related to the research question, but that is correlated with the missingness found in the variable(s) of interest. |
| Broad imputation | | Imputation done for an entire dataset that can then be used for all research questions and analyses using that dataset. |
| Fully conditional specification | FCS | Multiple imputation technique where each variable is imputed separately with its own regression model. |
| Imputation model | | Analytical model planned specifically to impute missing data values. |
| Impute-then-transform | | Strategy where an imputation is done using the observed variables, before computing variable transformations in the imputed datasets. |
| Inclusive imputation | | Imputation strategy where all possible auxiliary variables are included. |
| Joint modeling | | Multiple imputation technique where a single multivariate model is used to fill in missing data in all incomplete variables. |
| Just another variable | JAV | Synonym for transform-then-impute. |
| Missing at random | MAR | Missingness mechanism where the probability of missingness on a specific variable is related to other variables. |
| Missing completely at random | MCAR | Missingness mechanism referring to a completely random process; missing data truly happens by chance. |
| Missing not at random | MNAR | Missingness mechanism where missing data on a specific variable are related to the values of the variable itself. |
| Multiple imputation by chained equation | MICE | Synonym for fully conditional specification (FCS). |
| Narrow imputation | | Imputation done specifically for one research question or analysis. |
| Restrictive imputation | | Imputation strategy where few or no auxiliary variables are included. |
| Transform-then-impute | | Strategy where variable transformations are done in the original dataset and included in the imputation model. |

*random* (MNAR; Rubin, 1976; Seaman et al., 2013). MCAR refers to missingness that is due to a completely random process; it truly happens by chance. For example, it could be because of a blackout during a laboratory task or because a participant moved between two waves of data collection and has become unreachable. Missingness is not related in any way to the variable(s) with missing data or to other variables. It is rarely the most prevalent pattern in a study, except when random missingness is implemented by the researcher through a planned missing data design (Rioux, Lewin, et al., 2020). MCAR is the least problematic of the missing data mechanisms. As the missing data are truly random, they do not introduce bias in the results and while doing analyses on complete cases only can enlarge standard errors (SEs) and reduce statistical power, SEs and power are easily recovered when using multiple imputation (Enders, 2010).

When data are MAR, the probability of missingness on a specific variable is related to other variables. In other words, a systematic relationship exists between missingness on a variable and values on other variables (Enders, 2010; Seaman et al., 2013). For example, this mechanism could happen if, in a survey, minors had higher probabilities of skipping a question on drug use frequency because they are worried about disclosing illegal information. Missing data would then be related to age, an observed variable. MAR also occurs frequently in longitudinal studies, where dropout is associated with factors such as age, education, socioeconomic status, or ethnicity (Robinson et al., 2016). If only complete cases were used in analyses, it would result in a loss of power and in biased parameter estimates, including effect sizes, SEs, variances, and p-values (van Buuren, 2018). As with MCAR, however, MAR data are recoverable when multiple imputation is used. Indeed, when multiple imputation is used with MAR data, power is recovered and the missing data do not introduce bias in the results, as long as auxiliary variables, which are variables that are not related to the research question, but that are correlated with the missingness found in the variables of interest (i.e., predict the MAR mechanism), are included (Enders, 2010).

When data are MNAR, missing data on a specific variable are related to the values of the variable itself (Enders, 2010). For example, this would be the case if frequent drug users had higher probabilities of skipping a question on drug use frequency. There would be more missing data on drug use frequency for frequent drug users than for infrequent drug users, but there would be no way to test this as the data providing this information are missing. MNAR can severely bias estimates, and even with multiple imputation, these biases are unrecoverable or only partly recoverable (Enders, 2010; van Buuren, 2018). Importantly, if a variable predicts MAR missingness but is not measured, it will be equivalent to MNAR in the analyses. For example,

if age predicted missing data on the drug use frequency question, but age was not measured, there is no way for the researcher to know or to control for it in the analysis, which will lead to biased results. Accordingly, a proactive approach in the study design phase is to carefully plan and include measures of potential predictors of missingness to help prevent bias (Rioux, Lewin, et al., 2020).

## Multiple Imputation

Multiple imputation is a missing data treatment that can recover power and provide unbiased parameter estimates under MAR and MCAR (Rubin, 1987). Increasingly used in research (Rioux & Little, 2021), multiple imputation involves making a copy of the original dataset while replacing each missing cell with a plausible estimation of its value. This procedure creates $m > 1$ complete datasets, where $m$ corresponds to the number of imputations, which are larger than one. During the imputation process, each missing value is estimated using the observed scores of the variables in the original dataset – as well as previously imputed values in some cases (see fully conditional specifications, next paragraph). For each of the $m$ datasets, a different but plausible estimation of the missing value is estimated using that information. Once the $m$ datasets are imputed, the next step consists in conducting the statistical analysis in each dataset separately. The estimates and standard errors (SEs) obtained from each analysis are then aggregated using Rubin's rule (Rubin, 1987).

These multiple estimations of possible values for each missing data point have many advantages over single imputation or other techniques where only one dataset with complete data is generated. Indeed, it accounts for the uncertainty of estimating the missing data by integrating the variability within and between each imputed data set. As such, the SEs obtained take this uncertainty into account and are more trustworthy than those stemming from a single imputation (Rioux & Little, 2021; Rubin, 1987).

Two main approaches may be used for multiple imputation: joint modelling (a.k.a., multivariate normal imputation or MVNI; Schafer, 1997) and fully conditional specification (FCS, a.k.a., multiple imputation by chained equation or MICE; van Buuren, 2007). In joint modelling, a single multivariate model is used to fill in missing data in all incomplete variables. All missing observations are imputed in a single computational step. Thus, the imputation model (i.e., the statistical model used to impute variables with missingness) is the same for all variables (Graham, 2012; Schafer, 1997). FCS uses a series of conditional models where each variable is imputed separately with its own regression model. FCS has the advantage of being more flexible because the imputation model can be tailored to each individual variable and can maintain skip patterns, bracketed responses, and bounds, but has the disadvantage of being more computationally intensive than joint modelling (van Buuren, 2007, 2018).

### *How Many Imputed Datasets are Needed?*

As mentioned above, more than one imputed dataset is needed to take into account the uncertainty stemming from the missing information and to compute more trustworthy estimates and *SE*s (Rioux & Little, 2021; Rubin, 1987). How many imputed datasets are necessary, however, is an enduring debate. Initial recommendations suggested 3 to 10 imputed datasets (Rubin, 1987; Schafer, 1999), but further research has shown that more is better and technological advances have made a large number of imputations feasible. However, a high number of imputed datasets may still take a long time to estimate (i.e., weeks), especially for a complex model and for researchers who do not have access to high-performance computing systems. Furthermore, because every imputation estimation needs to converge, adding more imputed datasets to estimate increases the chances of non-convergence or other bugs, which is more likely with large imputation models.

Modern recommendations vary, with some recommending 5–20 imputed datasets when the percentage of missing values is moderate and 20–100 to be more precise or when the percentage of missing values is high (van Buuren, 2018). More specific recommendations suggest using information from the original dataset (before imputation) to decide. For example, von Hippel (2009) suggested using as many imputed datasets as the percentage of incomplete cases in the sample. This strategy can easily bring the number of imputed datasets to 100, however, as even one missing scale item would result in an incomplete case. As a compromise, van Buuren (2018) suggested considering the average missing data rate instead of the fraction of incomplete cases. Other recommendations use the fraction of missing information (FMI) to estimate the number of imputed datasets needed (von Hippel, 2020; White et al., 2011). However, FMI is a measure of uncertainty about the imputed values that can only be calculated after the multiple imputation is done, complicating its use (Q. Pan & Wei, 2016). Furthermore, when missing data are MCAR, the FMI is usually equal or smaller to the fraction of incomplete cases, bringing us back to the recommendation by von Hippel (2009). Missing data that are MAR may increase the FMI beyond this later fraction, especially if the variables predicting the missingness are not included in the imputation model, encouraging the use of an inclusive imputation strategy.

In summary, 100 imputed datasets is a conservative number that covers most recent recommendations. If this is impossible due to convergence failure, computational limitations, or time demands, a smaller number based on the percentage of incomplete cases or average missing data rate can be used. Because imputation models are usually built iteratively, it is recommended to build the model using a small number of imputations, such as five, and set the number higher only for the last round of imputation when the model is finalized (van Buuren, 2018).

### *Level of Inclusivity in the Imputation Model*

Regardless of the method chosen, an important step is to choose the level of inclusivity of the imputation model. Inclusivity is based on the number of included auxiliary variables. At a minimum, to adequately maintain the relationships between variables, an imputation model must include all variables in the analytical model. A few auxiliary variables predicting missingness should also be included since multiple imputation yields unbiased parameters under the MAR mechanism only if relevant auxiliary variables are included. Inclusivity can go from *restrictive*, where few or no auxiliary variables are included, to *inclusive*, where all possible auxiliary variables are included (Collins et al., 2001). Inclusive strategies are often recommended based on their advantage of utilizing a large number of variables that can have some predictive power on missingness, permitting a more precise estimation of the missing values (Collins et al., 2001; van Buuren, 2018).

Researchers can also decide to do a *narrow imputation*, where the imputation is done specifically for one research question or analysis, or a *broad imputation* where the imputation is done for an entire dataset that can then be used for all research questions and analyses using that dataset (van Buuren, 2018). While a narrow imputation usually involves fewer variables than a broad imputation, it can still use an inclusive strategy. For both inclusive auxiliary variable approaches and broad imputations, datasets may include hundreds or thousands of variables.

### *Using Transformed Variables in the Imputation Model*

In psychological research, researchers may need to use many transformed variables in their imputation model, which are computed using linear or non-linear transformations of observed variable(s) from the original dataset. Linear transformation often includes scales made from an average or a sum of multiple questionnaire items while non-linear transformations often include interactions between two or more variables (e.g., x*y), quadratic or cubic terms of one variable (e.g., $x^2$ or $x^3$), or other combinations of variables (e.g., x/y). They can be needed as part of the analytic model used to test the research question (e.g., if physical activity levels predict BMI, formed from the measured variables height and weight) or contribute to the prediction of missingness in MAR mechanisms.

For linear transformations, a technique known as *impute-then-transform* can be used, in which the imputation model includes the untransformed, observed variables. Once the multiple imputation is done, the transformed variables can be computed using operations like sums or averages on the imputed datasets (von Hippel, 2009). Item-level imputation is sometimes favored as it has a statistical power advantage over scale-level imputation (Gottschall et al., 2012; Rioux, Stickley, et al., 2020), but transformation before imputation can also be used and can be particularly useful for reducing the size of imputation models (see *Item aggregation* below).

The impute-then-transform technique does not introduce bias for linear transformations because the associations between the original variables are preserved during these transformations. However, these associations are not maintained in non-linear transformations. As such, these transformed variables *must* be computed and added to the imputation model before the multiple imputation is done, a technique known as *transform-then-impute*, or "just another variable" (von Hippel, 2009). This method involves computing transformed variables first and including them in the imputation model. Observed variables used to compute the transformations are usually kept in the imputation model as they are usually included in the analytic model (e.g., testing main and interaction effects), but they could be removed if they won't be used in the analytical model (von Hippel, 2009). While this method keeps the relations between the transformed and original variables, it can lead to impossible values in the imputed transformed variables such as negative quadratic terms (Austin et al., 2021). Multiple studies have tested this method and shown that estimates are unbiased under some circumstances, including when the missing data mechanism is MCAR and under certain specifications of MAR (Lüdtke et al., 2020; Vink & van Buuren, 2013; Zhang & Wang, 2017). Because certain situations can lead to bias, some authors have proposed procedures and methods to reduce this bias (e.g., Vink & van Buuren, 2013; Zhang & Wang, 2017), including passive imputation which will be presented below (van Buuren, 2018; van Buuren & Groothuis-Oudshoorn, 2011). However, especially when the imputation model is complex, all methods may introduce bias and the recommendation is to conduct sensitivity analyses after the imputation is completed to check if the method performed as intended (Austin et al., 2021; Zhang & Wang, 2017).

Non-linear transformations increase the complexity of the imputation model since both the original variables and the transformed variables often need to be included. While for narrow imputations, only the non-linear transformations to be included in the analytic model will be imputed, broad imputations may require the inclusion of a large number of transformations to be included for a multitude of subsequent analyses. If the imputation aims to meet the needs of secondary data analysis projects that are not known at the time of the imputation, all possible transformations may have to be included in the imputation. This inclusiveness can yield an exceedingly large number of variables. In addition, if the original variables used to compute the transformed variables also have missing data, the imputation of the transformed variables themselves can add to the complexity of the imputation model.

### *Why Are Too Many Variables a Problem?*

As the number of included variables increases, the risks of failure or lack of convergence also increases (for an explanation of convergence, see Little et al., 2016). This is one of the main considerations and a source of problems when imputing using a broad or inclusive strategy. As a minimum condition for the imputation model to converge, each variable has to have at least as many observations

(i.e., non-missing values) as the number of predictor variables in the imputation model (Little et al., 2016). Small samples, large assessments, high missing data rates, inclusive and broad imputation strategies, and the inclusion of transformed variables can all lead to having more variables than observations, making this ratio quickly exceeded in many studies. A longitudinal cohort followed throughout the lifespan will quickly have more variables than observations, but a study that seems "small" can also encounter these problems. For example, a cross-sectional study of 100 participants with a 75-item questionnaire, the inclusion of 5 interaction terms, and 20% missing data rates on some items would have more variables than observations on some variables.

Accordingly, for the purpose of this paper, *large imputations* refer to any imputation where the number of variables is larger than the number of observations, making it inadmissible from the get-go. Numerous techniques have been developed to try and overcome the challenges related to the number of variables included in these large imputations, but they are still relatively unknown. As such, the remainder of this paper gives an overview of these techniques to familiarize researchers with large imputations and guide those looking to conduct one to the appropriate resources. Note that while the ratio of variables per observations is a minimum requirement that should be met before even trying to run an imputation model, other factors, such as variable types, variable correlations, model complexity, missing data rates, and missing data patterns, can contribute to lack of convergence and exceedingly long imputation run times (Graham, 2012; Little et al., 2016). As the following techniques will decrease model complexity, they may also help in those instances.

## Imputation Techniques to Reduce the Number of Variables

### *Item Aggregation*

In many psychological studies, the high number of variables can be explained by the use of multi-item questionnaires. As questionnaire scales are usually the result of a linear transformation of their items (sum or average), combining these items in some way before imputing can help reduce the number of variables for both joint modelling and FCS (Enders, 2010; Graham, 2012), without introducing bias in the imputation. Two main methods of aggregating items exist: scales and parcels. Questionnaires usually include one or more scales that are computed using the sum or average of their respective items. These scales can be computed before imputing the data, thus considerably reducing the size of the dataset. Parcels are aggregates of two or more items from a larger multi-item scale, which are used to reduce a multi-item scale to a few indicators to estimate latent factors in structural equation modelling. Thus, researchers planning to use this analytical method can compute parcels before imputing the data to reduce the size of their dataset by following recommendations to construct valid parcels, see Little et al. (2022). Parcelling as a data reduction technique for multiple imputation can also be useful for some projects that are not planning to use latent variables. Indeed, it is recommended that scales only be computed for participants with complete data on that scale since computing a scale by averaging non-missing items when some items are missing can introduce bias (Matsunaga, 2008). Thus, parcels could be preferred when only a few items on the scale are complete and the others need to be imputed.

Using aggregated items instead of individual items in multiple imputation has been validated in a few studies (Eekhout et al., 2018; Gottschall et al., 2012; Plumpton et al., 2016). In the study by Gottschall et al. (2012), no differences in bias were found in item-level vs. scale-level imputation, but scale-level imputation was associated with lower statistical power compared to item-level imputation, which could be troublesome for small samples. A simulation study by Rombach et al. (2018) used instruments with general scale and specific sub-scale scores to compare item-level to subscale- and to scale-level imputation ($n$ = 100–1,030, missing data rates = 5–40%). Item-level imputation resulted in more precise score estimations, especially when the rate of item-nonresponse was high. In small samples (<200), however, item-level imputation had convergence issues, and both subscale- and scale-level imputation still yielded acceptable estimations. Thus, the authors recommended the use of subscale-level imputation when item-level was not feasible and, if further item aggregation was necessary, to then use full-instrument scales (Rombach et al., 2018). Finally, a recent simulation by Vera & Enders (2021) looked at scale vs. item imputation in a longitudinal study, more specifically examining whether imputation results were influenced by the number of items per scale (five to 15), magnitude of inter-item correlations (.30 to .50), rate of missing item data across waves (5–40%), and sample size ($n$ = 250–1,000). When all items of a scale were missing for a participant, scale-level imputation was superior to item-level, especially when the number of items was large or the missing data rate was high. No study has examined imputation with parcels, although it could be expected to perform similarly to subscale-level imputation (Rioux, Stickley, et al., 2020).

Item aggregation prior to imputation has been used in recent empirical studies, although little information on the missing data and imputation process is reported (Lobatto et al., 2020; Steffgen et al., 2020). As mentioned before, the main limitation of this technique is that scales or parcels can be aggregated only for participants with complete data on that scale or parcel to avoid bias. If one wants to compute scales or parcels before imputation but is using items with missing data, then passive imputation, presented next, may be preferred.

### *Passive Imputation*

Passive imputation can be used to handle multi-item scales as well as any type of variable transformations during the imputation process. This technique has two main applications. The first is to reduce the number of variables in the imputation model, for example by computing a multi-item scale, especially when all the items have some missing data.

The second is to include composite variables made from non-linear transformations of observed variables (e.g., interactions, quadratic terms) as their inclusion before imputation is needed to preserve their relations with other variables of the analytical model, as explained above. When the observed variables to be used in the transformed variables have missing data, passive imputation is advantageous over the transform-then-impute technique described above. Indeed, passive imputation can compute transformations using variables that have missing values while maintaining the relationship between items/observed variables and their scale/transformed variable (van Buuren, 2018; van Buuren & Groothuis-Oudshoorn, 2011).

Because of its specification, this method needs to be implemented using FCS. In a classic imputation model, a variable with missing data is regressed on all other variables included in the model to predict its missing values. With passive imputation, a transformed variable is regressed on all its own items, but only on aggregated and transformed scores for other variables. For example, a scale item with missing data would be regressed on all items from the same scale, but only on the aggregated scores from the other scales and transformed variables included in the imputation model. The item is not regressed on its own scale score to avoid multicollinearity. For an interaction term, the interaction variable would be regressed on all items included in the interaction, but on aggregated and transformed scores for all other variables. With this technique, missing transformed variables are not directly imputed. Rather, they are computed using the imputed observed variables and updated after each imputation. Their updated score is used to better predict item-level missing data in subsequent iterations (van Buuren, 2018; van Buuren & Groothuis-Oudshoorn, 2011).

Passive imputation has been implemented in Stata (StataCorp, 2023), S-PLUS (Venables & Ripley, 2002), IVEware (Raghunathan et al., 2022) and in the R package MICE (van Buuren, 2018). Some authors have also proposed a three-step method that resembles passive imputation but can be done with any FCS imputation software; scales are manually computed before the imputation – ideally with their own imputation if they have items with missing data (Enders, 2010, p. 271) – and, when imputing, items with missing data are regressed on items from the same scale and aggregated scores from the other scales. Only item scores are kept after imputing. This manual method differs from passive imputation because the scale scores are not updated after each imputation.

Eekhout et al. (2018) performed a simulation study ($n$ = 150–250, 30% of subjects with missing data) and found that passive imputation performed better than item-level imputation in small samples, with fewer convergence failures. Concerning the imputation of a multi-category nominal variable computed from binary observed variables, Y. Pan et al. (2020) also found passive imputation to be superior to regular FCS imputation regardless of whether the binary observed variables used in the composite were included as auxiliary variables or not and across various model specifications ($n$ = 1,000–5,000, missing data rates = 5–60%). Similarly, passive imputation has been found to perform well in the estimation of linear and non-linear data transformations, using binary, categorical and continuous observed variables and various model specifications ($n$ = 1,000–5,000, missing data rates = 5–60%, MAR and MCAR) in complex study designs like longitudinal ($\geqslant$ 3 waves) and multilevel models ($k$ = 40–500, $n$ per cluster = 5–30, two or three levels; Grund et al., 2018; Mitani et al., 2015; Wijesuriya et al., 2021). Several studies conducted passive imputation using real-world data under various conditions ($n$ = 120–12,115, missing data rates = 1–30%, usually with a MAR mechanism, longitudinal and multilevel models), and found similar encouraging results (Eekhout et al., 2018; Hayati Rezvan et al., 2022; Mitani et al., 2015; Y. Pan et al., 2020; Sep et al., 2020; Wijesuriya et al., 2021).

Item aggregation and passive imputation can thus be used to reduce the number of items when scale scores and other transformed variables are used. However, if variables cannot be reduced in these ways (e.g., there are many single item variables) or the dataset is not reduced enough, principal component analysis, presented next, can be used with all types of variables.

### *Principal Component Analysis*

Principal component analysis (PCA; Hotelling, 1933; Jolliffe & Cadima, 2016) is a dimensionality reduction method. It identifies common latent structures among the variables of an original dataset to form components, which are new and uncorrelated variables that capture the variance of these common structures. As such, large datasets are reduced to smaller ones that still contain most of the original information (Jolliffe & Cadima, 2016). PCA can form up to $n$ components where $n$ is equal to the number of variables in the original dataset, but a smaller number of components is usually retained for data reduction purposes. The first principal component derived from the procedure always captures the most variance, with each further component capturing a decreasing amount. This allows researchers to retain a minimal number of components while capturing a maximum amount of variance (Schreiber, 2021). Accordingly, using principal components as predictors in the imputation model, instead of individual variables, can drastically reduce the number of predictors while preserving as much information as possible and lowering the risks of convergence failures due to multicollinearity or overfitting (Howard et al., 2015).

PCA for imputation is done in multiple steps. First, since PCA can only be conducted using complete data, a single imputation is done for each observed variable. While any single imputation method could be used, the imputation model is usually specified under the FCS approach to avoid problems related to too many predictors. While in narrow imputation, this specification could be done manually for each variable, other techniques or criteria can be used for a broad imputation. For example, only variables that are correlated to a certain degree with a specific variable could be included as predictors in its single imputation model. At this step, variable transformations (e.g., polynomials, interactions) are computed using the original observed vari-

ables using transform-then-impute. After this single imputation, principal components are extracted from the complete data. The number of components retained is usually determined by the researcher based on the percentage of variance explained wanted (either in total across components or as a minimum percentage of variance explained per component). The principal components retained are then used as the predictor variables in the multiple imputation procedure instead of the original auxiliary variables (Howard et al., 2015). Single imputation is usually not recommended to treat missing data as it underestimates SEs, increasing Type I errors, and does not consider the uncertainty of estimating missing values (Rioux & Little, 2021). When using PCA for imputation, however, the data generated with the single imputation in the first step of the procedure are not used for statistical inferences, but only to compute the principal component variables, where SEs are not relevant (Howard et al., 2015).

Each step can be done manually and separately by the researcher in any statistical software that supports both PCA and imputation with FCS. However, the R (R Core Team, 2021) package PcAux (Lang et al., 2018) is usually recommended since it does all the steps almost automatically, from help with data preparation to the final multiple imputation, based on syntax specified by the researcher. Principal component variables can also be extracted and used in other multiple imputation procedures or as auxiliary variables in analyses using full-information maximum-likelihood (FIML; Howard et al., 2015). These component score predictors also benefit from being orthogonal, eliminating any collinearity issues among the predictors.

A growing number of studies have examined the validity of this technique, with both simulated and real-world data. First, Howard et al. (2015) compared the PCA imputation technique to the inclusive approach using original variables in a simulation study ($n$ = 50–1,000, missing data rates = 10–80%). They compared the use of one linear principal component (made from eight auxiliary variables), one non-linear principal component (made from the same eight auxiliary variables, the square of those variables, and 28 two-way interactions), and eight auxiliary variables without PCA (inclusive approach). Linear and non-linear PCA were as effective as the inclusive approach in removing the bias linked to missing data. Furthermore, the authors compared the PCA (one, seven, and fourteen components) and inclusive (49 observed variables) approaches with a real-data example, using a sample of 2,299 participants with 49% missing data. Results showed that the PCA approach was more efficient than the inclusive one and yielded the smallest SEs. In this analysis, using seven components (40% variance explained) was better (i.e., smaller SEs) than using only one, but using 14 components instead of seven did not further improve SEs (Howard et al., 2015).

Next, a study by Hayati Rezvan et al. (2022) compared multiple missing data techniques with real-world data ($n$ = 1,487, missing data rates = 20%, MAR, 30 covariates, 17 items from two scales, four auxiliary variables, and one binary outcome). They specifically examined the imputation of missing items in multiple-item scales using the PCA (2 principal components, 56% variance explained), inclusive (17 scale items) and passive imputation (2 scales) approaches. The PCA approach performed as well as inclusive or passive imputation but had less convergence problems. All three techniques also performed slightly better than simply using scale scores as auxiliary variables, and Hayati Rezvan et al. (2022) hypothesized that this difference would increase as the number of items included and the rates of missing data increased as well. The effectiveness of different PCA approaches when hundreds of variables are included in the imputation model was verified recently by a team of researchers via a series of simulations and real-world studies (Costantini, Lang, Reeskens, et al., 2023; Costantini, Lang, Sijtsma, et al., 2023). Their studies included up to 500 original variables, with up to 56 extracted principal components. The PCA approach was overall effective for this number of variables, with small bias and good statistical efficiency. Interestingly, Costantini, Lang, Reeskens, & Sijtsma (2023) found that when the potential auxiliary variables were highly correlated and components were chosen based on a total variance explained of 50%, using only one component met the rule, but it predominantly contained noise variance. This suggests that using the variance *per component* rather than the total variance across components may be preferable for determining the number of components. Both studies (Costantini, Lang, Reeskens, et al., 2023; Costantini, Lang, Sijtsma, et al., 2023) suggested that an efficient and reliable method to select components would be the Kaiser criterion (Guttman, 1954; Kaiser, 1960), which is to drop all components with eigenvalues under 1.0 – meaning all components explaining less variance than single variables.

Regarding categorical data, Kim et al. (2021) examined how the PCA approach performed by comparing it to the inclusive imputation approach in a simulation study with one categorical outcome, one continuous predictor, and eight continuous auxiliary variables. They found that using one principal component led to lower bias in the imputation of the categorical outcome than the inclusive approach without PCA, especially when the sample size was small, the distribution of the categorical variable was asymmetrical, or the correlations between auxiliary variables were high (Kim et al., 2021). It is important, however, to declare categorical data as such in imputation syntax to prevent bias in estimation (Lang et al., 2018; Manisera et al., 2010). In recent years, the PCA technique was also used to multiply impute data in a number of studies. While not all studies reported details on their implementation of the procedure, those who did report using anywhere from 11 to 65 principal components to summarize 40–75% of the variance from the original variables (Gedal Douglass et al., 2020; Shogren et al., 2020; Shubert et al., 2020).

Some limitations and questions concerning PCA still remain. Among the validation studies to date, components have been formed only using observed/measured variables and two-way interactions (Hayati Rezvan et al., 2022; Howard et al., 2015; Kim et al., 2021). Furthermore, the reduction of only up to 500 variables has been examined so far, but broad and/or inclusive approaches could far ex-

ceed that number. Further validation studies are needed to better understand how PCA works with non-linear variables (Costantini, Lang, Sijtsma, et al., 2023) and to test more complex simulations such as the inclusion of triple interactions and polynomials in the principal components and the use of principal components to reduce hundreds or thousands of variables. This would strengthen confidence in this technique as well as help decision-making for contexts often encountered in psychological research. Further guidance on how many principal components to retain is also needed, although the best evidence so far is to use the Kaiser criterion.

Item aggregation, passive imputation, and PCA can greatly reduce the number of predictors in an imputation model and can be used for any study design. In the case of a longitudinal study, however, one can also reduce data for imputation by restricting the number of data collection waves to include as predictors, which is done using the two-fold fully conditional specification.

### *Two-Fold Fully Conditional Specification*

Although the other three techniques presented can be used for longitudinal studies, the two-fold FCS technique was specifically developed for longitudinal data (Nevalainen et al., 2009). Including all data collection waves in one imputation model, especially in large long-term longitudinal studies, can easily lead to having too many variables as well as multicollinearity problems. Accordingly, two-fold FCS proposes that the imputation model for each variable includes the other variables from the same time point (*t*) along with variables from the waves just before *(t-1*) and just after (*t+1*), as these time points are the most likely to hold important auxiliary information (Welch, Bartlett, et al., 2014). As with other techniques, in a broad imputation, this may include all variables at each time point, while in a narrower imputation it may include the variables in the analytical model along with the auxiliary variables (in a longitudinal dataset, auxiliary variables would be expected to include, at minimum, the repeated-measure variables from t-1 and t+1). The time window can also be changed to include, for example, two times before and after (Welch, Bartlett, et al., 2014). The algorithm allows the imputation of all variables and data collection waves of a dataset, while also limiting the number of predictors for each variable to the adjacent time points (Nevalainen et al., 2009). It can also be useful to include time-independent variables (e.g., birth weight) in the prediction models of all variables (Welch, Bartlett, et al., 2014). Despite limiting the number of time points included in the imputation model, the number of variables may still be too high. As a solution, Nevalainen et al. (2009) recommend including as auxiliary predictors only the variables that explain a certain amount of variance in the variable to be imputed instead of all variables measured in *t-1*, *t*, and *t+1*, with their suggestion being to fix the minimum variance explained as low as possible for the imputation model to converge, starting as low as $R^2 > 0.001$ and increasing as needed.

In the two-fold FCS algorithm, time-independent variables are imputed first so they can be used as predictors in the imputation of each time point. Afterward, each wave is imputed in chronological order using time-independent variables, same-wave variables and adjacent wave variables as predictors (Nevalainen et al., 2009). Two-fold FCS can be implemented in IVEware (Raghunathan et al., 2022), SAS (SAS Institute Inc., 2023) and Stata (StataCorp, 2023).

Two-fold FCS has been shown to be a valid imputation method in three simulation studies ($n$ = 561-5000, missing data rates = 50-70%, up to ten data collection waves). Results showed that two-fold FCS introduced almost no bias compared to results with no missing data, although biases were a bit higher but still acceptable when missing data rates were higher (Nevalainen et al., 2009; Welch, Petersen, et al., 2014). A study by De Silva et al. (2017) used five waves of data and found that standard imputation (with all data collection waves included) was slightly more precise and less biased than two-fold FCS, but the latter became more precise when using two adjacent times before and after *t* instead of one. In their study, Welch, Petersen, et al. (2014) used ten waves of data to compare two-fold FCS to a standard multiple imputation method but ran in too many collinearity issues to use all waves of data in the prediction model of the standard imputation model, highlighting the usefulness of two-fold FCS for long-term longitudinal studies. Multiple recent empirical studies have used two-fold for their real-world data imputation. The datasets used varied from 10 to 45 time points to impute samples of 277 to over 150,000 participants and used up to four adjacent time points (before and after) as predictors (Blackburn et al., 2018; Ferreira et al., 2018; Huybregts et al., 2019), but studies did not provide rates of missing data, or the number of variables included in the imputation model.

By only keeping the two adjacent times as predictors, two-fold FCS has less risk of having multicollinearity or overfitting problems. It may nevertheless be desirable to integrate more waves as predictors since one study found that this decreases bias (De Silva et al., 2017). Further methodological research comparing the inclusion of different numbers of waves would help better guide researchers. Still, the ideal number of adjacent waves included most likely depends on the study design, population studied, and variables measured. As such, a good theoretical understanding of the variables studied and exploratory analyses will help to construct a proper imputation model by identifying waves that independently predict variables in time *t*, beyond their association with adjacent time points (Welch, Bartlett, et al., 2014).

## Recommended Imputation Techniques

Each technique presented in this paper aims to lower the number of variables in an imputation model to aid a successful imputation. They all have different requirements, strengths and limitations and can be used with different types of data. To help the reader choose which technique to investigate further and use in their data imputation, the following section provides a summary of recommendations

**Table 2. Software and Syntax Guidance for Each Imputation Technique**

| Technique | MI Framework | Software | Syntax guidance and examples |
|---|---|---|---|
| **Multiple imputation** | Joint modeling | Stata | p.207-233 in StataCorp. (2023a) |
| | | R package Amelia II | Honaker et al. (2011) |
| | FCS | R Package MICE | van Buuren (2018)<br>Berglund (2015) |
| | | Stata | p.140-168 in StataCorp. (2023a)<br>Royston & White (2011)<br>Berglund (2015) |
| | | SAS | SAS Institute Inc. (2023)<br>Berglund (2015) |
| | | SPSS | IBM Corp. (2021) |
| | | IVEware | Raghunathan et al. (2022)<br>Survey Research Center, University of Michigan (2023)<br>Berglund (2015) |
| **Item aggregation** | Joint modeling<br>FCS | All software | Basic syntax for creating variable using the sum or average of other specific variables (pre-imputation step) |
| **Passive imputation** | FCS | R package MICE | Chapter 6, section 6.4 in van Buuren (2018)<br>Vink & van Buuren (n.d.) |
| | | Stata | p. 289-292 of StataCorp. (2023a) |
| | | IVEware | Chapter 2 in Raghunathan et al. (2022) |
| **Principal component analysis** | FCS | R package PcAux | Lang et al. (2017)<br>Lang et al. (2018). |
| **Two-Fold Fully Conditional Specification** | FCS | Stata (recommended) | Welch, 2022<br>Welch, Bartlett & Petersen (2014) |
| | | IVEware add-in with SAS | Raghunathan et al. (2022) |

and requirements for each technique and Table 2 provides information on how to implement each technique.

### *Multiple Imputation (basic)*

This technique is recommended when the number of observations does not exceed the number of columns in the data, there are little if any interaction terms, and a relatively low percentage of missing observations. In this case, the chance of having convergence issues due to the number of variables exceeding the number of observations is low and using a traditional multiple imputation method is a viable choice that is easier to implement.

### *Item Aggregation*

This technique is recommended when the data contain a relatively large number of scale items that can be combined into either scale scores or parcels, and the analytical model to be conducted with the imputed data is going to be done on the scale or parcel level instead of on the item level (e.g., would not be used if the analysis planned is a factor analysis at the item-level).

### *Passive Imputation*

This technique is recommended when the data contain multiple scales with items that need to be imputed at the item level, as well as items that are non-linear transformations of other items which will be included in the analytical model or may be informative to include in the imputation model.

### *Principal Component Analysis*

This technique is recommended when reducing the number of items by creating scale scores or parcels is not feasible (e.g., item-level analysis planned) and/or will not reduce the dataset enough. While the number of scale scores or parcels is predetermined, the number of principal components can be chosen to be under the variable-to-observation threshold for the imputation model, often allowing a larger data reduction than the other techniques while allowing for the imputation of items and transformed variables.

### *Two-Fold Fully Conditional Specification*

This technique can be particularly efficient at reducing large-scale longitudinal data with many data collection waves. It is only recommended for studies that have at least four waves of longitudinal data since the method imputes three waves of data at a time (i.e., a focal time point, and the two adjacent time points). If the number of columns in three waves of data still exceeds the number of rows,

**Table 3. Summary of Observations and Columns for the Fictional Vignettes Across Imputation Methods**

| | | Columns | | | | | |
|---|---|---|---|---|---|---|---|
| | | Full dataset | Item aggregation | | Passive imputation | PCA | 2F-FCS |
| | Observations | | Scale | Parcel | | | |
| **Pre-post study** | 78-102 | 140 | 32* | 56* | 21-40* | ≤65* | N/A |
| **Long-term longitudinal** | 200-800 | 1100 | 200～ | 400～ | 199-208～ | ≤65* | 329～ |

*Note.* Observation ranges depend on missing data rates on each variable. Column ranges, when applicable, depend on the predictors used for each type of variable. Parcels are done at 3 parcels/scale.

*Method meets column-to-observation ratio alone (without adaptation or combining with other methods) across all time points and variables.

～Method meets column-to-observation ratio alone (without adaptation or combining with other methods) for some time points and variables.

then researchers can adjust the minimum amount of variance explained by a predictor or implement one of the other three methods as they are also suitable for longitudinal data.

## Examples

To illustrate how the previous techniques might be applied to research, we will be using two fictional vignettes with studies of different complexities. These examples are explained based on a wide data format (i.e., each participant is a single row and each variable, including repeated measures, is a column). The first example is of a **pre-post test study** with several scales used to evaluate student attitude and performance over the course of an after-school intervention. This example study has six 10-item scales being used to evaluate students at both time points, resulting in 120 columns of scale data, along with 20 single items representing demographic information and academic performance that were measured once, resulting in a total data frame that is 140 columns. Unfortunately, the program was only able to evaluate students in one school in the district, limiting the sample size to 120 students. This pre-post example resulted in 15–35% missing data due to students not completing the survey, or not being present at one of the two time points of assessment. This means that there are between 78 (35% missing) and 102 (15% missing) valid observations for each measured variable.

The second example is of a **long-term longitudinal study** evaluating how prejudicial attitudes towards marginalized communities change over time as participants develop through middle school, high school, and post-high school education or employment, encompassing 10 time points in total. This study is using ten 10-item scales, resulting in 100 scale items at each time point, which are assessed along with 10 single items measuring demographic information at each time point, for a total of 1100 columns of data. This study has a sample of 800 participants at the first time point, but attrition reduces the sample to 500 by the end of high school, and post-high school follow-ups have a sample of 200. This attrition results in increasingly more missing data as the study progresses.

Below, the result of all four techniques for both examples is explained, with the results for the observation-to-column ration summarized in Table 3.

### *Item Aggregation*

Applying item aggregation to our first illustration case which described a pre-post test study, these researchers would take the six scales they included in their survey and create six scale scores for each participant, reducing the number of scale items to 6 per time point, and reducing the overall number of items across both time points to 32 columns. If the researchers instead wanted to create three parcelled indicators for each scale, the number of items would be reduced to 18 per time point and a final column count of 56 (36 parcels and 20 demographic questions). In this case, the researchers should ensure that the same parcelling scheme is used at both time points. The reduction to 32 or 56 columns would meet the required column-to-observation ratio.

For our second illustration case of a 10-time-point longitudinal study, the same general techniques would be applied. Here, using scale scores instead of the individual items would result in 200 columns in the final data set (10 scale scores and 10 demographic questions at each of the 10 time points). If parcels were used instead of scale scores, this would result in 30 parcelled indicators and 10 demographic questions per time point, totalling 400 columns in the final data set. Here, item aggregation would not be sufficient for post-high school time points, where there are 200 participants. Thus, another method would need to be used with or instead of item aggregation.

### *Passive Imputation*

To apply this technique to either of our illustration cases, researchers would first create scale scores for each of the scales in the data. Next, researchers would inform the imputation program that the scale scores are composed of the items that went into each scale. Following this, the researchers would generate a predictor matrix that identified the predictors of each scale item as the scale scores of the other scales in the data set. This predictor matrix would tell the imputation program to only use the scale scores from the other scales along with the other within-scale items to predict plausible values for each item in the data, reducing the number of predictors to a more manageable column to row ratio.

In the first example, each scale item would have 40 predictors (9 within-scale items, 11 scales and 20 single vari-

ables of demographic/academic indicators). The demographic and academic variables, which are single-item scores, would have 21 predictors (12 scales and the 19 other single variables). Using passive imputation, the imputation model of the first example now respects the predictor-to-observation ratio. For each variable, there are less predictors (40 or 21) than the minimum number of valid observations (78, as there is a sample of 120 with a maximum of 35% of missing data).

In the second example, each scale item would have 208 predictors (9 within-scale items, 99 scales, and 100 demographic variables) instead of the original 1099. The demographic variables would have 199 predictors (100 scales and 99 other demographic variables). The imputation model of variables measured before the end of high school would meet the predictor-to-observation ratio (208 or 199 predictors for 500-800 participants). However, imputation of the post-high school time points may run into convergence issues due to attrition as there are only 200 participants left. Researcher would have to adjust the prediction matrix to reduce the number of predictors, by leaving out the less correlated one for example, or choose another imputation method to be used with or instead of passive imputation.

### *Principal Component Analysis*

To apply this technique, researchers using the PcAux package in R (Lang et al., 2018) would either specify a finite number of principal components to extract (e.g., 8) or the percentage of explained variance wanted across components (e.g., 70%). As mentioned before, there are no firm guidelines on how many principal components to retain as predictors, but existing studies have used between one and 65 to explain up to 75% of the variance in the original dataset. Furthermore, studies reviewed above also found that using a large number of principal components may not be necessary to increase the predictive power in the imputation, as the proportion of variance explained is lower in each additional component, and that including too many components with low prediction power can introduce bias. In our two examples, even if researchers were to opt to retain as much variance as possible while following specifications that have been previously conducted, they would specify 65 principal components. Even this larger number of components would greatly reduce the number of predictors and meet the predictor-to-observation ratio requirement for both examples. Furthermore, it would ensure that there are no multicollinearity problems between variables, which could especially happen in the second example as the same items are measured year after year.

### *Two-Fold Fully Conditional Specification*

This method would not be available to our first example, as it only has two time points, the pre- and post-test. For researchers in our second illustration depicting a 10-year longitudinal study, at a minimum, they would design a predictor matrix that included variables in the focal time point, the previous time point (t-1), and the following time point (t+1). In this way, each variable's set of predictors only has three time points of predictors, which would result in 329 predictors. While this would reduce the ratio of predictors to observations, the number of predictors would still exceed observations in the later time points when the sample dropped to 200 participants. Like before, the number of predictors would need to be reduced even further. It could be done by increasing the correlation threshold as explained above or by using multiple techniques, for example by computing parcels with non-missing scale items and using those parcels in the two-fold conditional specification.

## Conclusion

In conclusion, the techniques presented in this paper can help reduce datasets when conducting multiple imputation, helping convergence and overfitting problems when the number of variables is too high for the sample size. The more flexible FCS (MICE) imputation approach can be advantageous when facing large imputations; all techniques presented in this paper can be implemented with this approach. However, some reduction is also possible in joint modelling imputation, for example by using item aggregation. It may also be possible to further reduce datasets by combining some of the methods presented in this paper, for example by implementing two-fold FCS with principal components or scale scores as auxiliary variables, although this has yet to be examined by methodological research. The methods presented are helpful for an inclusive imputation strategy with more auxiliary variables, thus allowing better estimation of the missing values (Collins et al., 2001), and for conducting broad imputations that allow researchers to impute a dataset only once rather than each time an analysis is conducted. Still, a narrower strategy restricted to a small number of variables can also be valid if one wants to conduct a simpler imputation model (Graham, 2012). To avoid bias, however, such a model should always include auxiliary variables that are most predictive of the MAR mechanism (Enders, 2010) and not just the variables in the analysis model.

## Contributions

Contributed to conception and design: SCL, TDL, CR

Contributed to acquisition of data (literature review): SCL, CR

Contributed to design of the Example section: ZLS, CR

Contributed to analysis and interpretation of reviewed papers: SCL, CR

Drafted and/or revised the article: SCL, ZLS, TDL, CR

Approved the submitted version for publication: SCL, ZLS, TDL, CR

## Funding Information

This work was supported in part by the Canadian Institutes for Health Research and the Fonds de Recherche du Québec - Santé through fellowships to CR.

## Competing Interests

The authors declare that there were no conflicts of interest with respect to the authorship or the publication of this article. TDL owns and receives remuneration from Yhat Enterprises, which runs educational workshops such as Stats Camp (statscamp.org).

Submitted: November 02, 2022 PST, Accepted: January 22, 2024 PST

## References


Austin, P. C., White, I. R., Lee, D. S., & van Buuren, S. (2021). Missing data in clinical research: A tutorial on multiple imputation. *Canadian Journal of Cardiology*, *37*(9), 1322–1331. https://doi.org/10.1016/j.cjca.2020.11.010

Berglund, P. A. (2015). *Multiple imputation using fully conditional specification method: A comparison of SAS, Stata, IVEware, and R*. SAS Support. https://support.sas.com/resources/papers/proceedings15/2081-2015.pdf

Blackburn, R., Osborn, D., Walters, K., Nazareth, I., & Petersen, I. (2018). Statin prescribing for prevention of cardiovascular disease amongst people with severe mental illness: Cohort study in UK primary care. *Schizophrenia Research*, *192*, 219–225. https://doi.org/10.1016/j.schres.2017.05.028

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351. https://doi.org/10.1037/1082-989x.6.4.330

Costantini, E., Lang, K. M., Reeskens, T., & Sijtsma, K. (2023). High-dimensional imputation for the social sciences: A comparison of state-of-the-art methods. *Sociological Methods & Research*, 1–52. https://doi.org/10.1177/00491241231200194

Costantini, E., Lang, K. M., Sijtsma, K., & Reeskens, T. (2023). Solving the many-variables problem in MICE with principal component regression. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02117-1

De Silva, A. P., Moreno-Betancur, M., De Livera, A. M., Lee, K. J., & Simpson, J. A. (2017). A comparison of multiple imputation methods for handling missing values in longitudinal data in the presence of a time-varying covariate with a non-linear association with time: a simulation study. *BMC Medical Research Methodology*, *17*(1). https://doi.org/10.1186/s12874-017-0372-y

Eekhout, I., de Vet, H. C., de Boer, M. R., Twisk, J. W., & Heymans, M. W. (2018). Passive imputation and parcel summaries are both valid to handle missing items in studies with many multi-item scales. *Statistical Methods in Medical Research*, *27*(4), 1128–1140. https://doi.org/10.1177/0962280216654511

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

Ferreira, I., Hovind, P., Schalkwijk, C. G., Parving, H.-H., Stehouwer, C. D. A., & Rossing, P. (2018). Biomarkers of inflammation and endothelial dysfunction as predictors of pulse pressure and incident hypertension in type 1 diabetes: A 20 year life-course study in an inception cohort. *Diabetologia*, *61*(1), 231–241. https://doi.org/10.1007/s00125-017-4470-5

Gedal Douglass, A., Roche, K. M., Lavin, K., Ghazarian, S. R., & Perry, D. F. (2020). Longitudinal parenting pathways linking Early Head Start and kindergarten readiness. *Early Child Development and Care*, *191*(16), 2570–2589. https://doi.org/10.1080/03004430.2020.1725498

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. https://doi.org/10.1080/00273171.2012.640589

Graham, J. W. (2012). *Missing data: Analysis and design*. Springer Science & Business Media.

Grund, S., Lüdtke, O., & Robitzsch, A. (2018). Multiple imputation of missing data for multilevel models. *Organizational Research Methods*, *21*(1), 111–149. https://doi.org/10.1177/1094428117703686

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika*, *19*(2), 149–161. https://doi.org/10.1007/bf02289162

Hayati Rezvan, P., Comulada, W. S., Fernández, M. I., & Belin, T. R. (2022). Assessing alternative imputation strategies for infrequently missing items on multi-item scales. *Communications in Statistics: Case Studies, Data Analysis and Applications*, *8*(4), 682–713. https://doi.org/10.1080/23737484.2022.2115430

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, *45*(7), 1–47. https://doi.org/10.18637/jss.v045.i07

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417–441. https://doi.org/10.1037/h0071325

Howard, W. J., Rhemtulla, M., & Little, T. D. (2015). Using principal components as auxiliary variables in missing data estimation. *Multivariate Behavioral Research*, *50*(3), 285–299. https://doi.org/10.1080/00273171.2014.999267

Huybregts, L., Le Port, A., Becquey, E., Zongrone, A., Barba, F. M., Rawat, R., Leroy, J. L., & Ruel, M. T. (2019). Impact on child acute malnutrition of integrating small-quantity lipid-based nutrient supplements into community-level screening for acute malnutrition: A cluster-randomized controlled trial in Mali. *PLOS Medicine*, *16*(8), e1002892. https://doi.org/10.1371/journal.pmed.1002892

IBM Corp. (2021). *IBM SPSS Missing Values 28*. IBM Corp. https://www.ibm.com/docs/en/SSLVMB_28.0.0/pdf/IBM_SPSS_Missing_Values.pdf

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Kaiser, H. F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*, *20*(1), 141–151. https://doi.org/10.1177/001316446002000116

Kim, Y., Lee, J., & Little, T. D. (2021). Multiple imputation with principal components for non-normal categorical data. *Multivariate Behavioral Research*, *56*(1), 165–166. https://doi.org/10.1080/00273171.2020.1869516

Lang, K. M., Curtis, J., & Bontempo, D. E. (2017). *PcAux*. https://github.com/PcAux-Package/PcAux/blob/master/documentation/PcAux_Field_Guide.pdf

Lang, K. M., & Little, T. D. (2018). Principled missing data treatments. *Prevention Science: The Official Journal of the Society for Prevention Research*, *19*(3), 284–294. https://doi.org/10.1007/s11121-016-0644-5

Lang, K. M., Little, T. D., Chesnut, S., Gupta, V., Jung, B., Panko, P., & Waggenspack, L. (2018). *PcAux: Automatically extract auxiliary features for simple, principled missing data analysis*. R package Version 0.0.0.9013. https://github.com/Statscamp/PcAux

Little, T. D., Lang, K. M., Wu, W., & Rhemtulla, M. (2016). Missing data. In D. Cicchetti (Ed.), *Developmental Psychopathology* (pp. 1–37). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119125556.devpsy117

Little, T. D., Rioux, C., Odejimi, O. A., & Stickley, Z. L. (2022). *Parceling in Structural Equation Modeling: A Comprehensive Introduction for Developmental Scientists (Elements in Research Methods for Developmental Science)*. Cambridge University Press. https://doi.org/10.1017/9781009211659

Lobatto, D. J., Vliet Vlieland, T. P. M., van den Hout, W. B., de Vries, F., de Vries, A. F., Schutte, P. J., Verstegen, M. J. T., Pereira, A. M., Peul, W. C., Biermasz, N. R., & van Furth, W. R. (2020). Feasibility, safety, and outcomes of a stratified fast-track care trajectory in pituitary surgery. *Endocrine*, *69*(1), 175–187. https://doi.org/10.1007/s12020-020-02308-2

Lüdtke, O., Robitzsch, A., & West, S. G. (2020). Analysis of interactions and nonlinear effects with missing data: A factored regression modeling approach using maximum likelihood estimation. *Multivariate Behavioral Research*, *55*(3), 361–381. https://doi.org/10.1080/00273171.2019.1640104

Manisera, M., van der Kooij, A. J., & Dusseldorp, E. (2010). Identifying the component structure of satisfaction scales by nonlinear principal components analysis. *Quality Technology & Quantitative Management*, *7*(2), 97–115. https://doi.org/10.1080/16843703.2010.11673222

Matsunaga, M. (2008). Item parceling in structural equation modeling: A primer. *Communication Methods and Measures*, *2*(4), 260–293. https://doi.org/10.1080/19312450802458935

Mitani, A. A., Kurian, A. W., Das, A. K., & Desai, M. (2015). Navigating choices when applying multiple imputation in the presence of multi-level categorical interaction effects. *Statistical Methodology*, *27*, 82–99. https://doi.org/10.1016/j.stamet.2015.06.001

Nevalainen, J., Kenward, M. G., & Virtanen, S. M. (2009). Missing values in longitudinal dietary data: A multiple imputation approach based on a fully conditional specification. *Statistics in Medicine*, *28*(29), 3657–3669. https://doi.org/10.1002/sim.3731

Pan, Q., & Wei, R. (2016). Fraction of Missing Information (γ) at Different Missing Data Fractions in the 2012 NAMCS Physician Workflow Mail Survey. *Applied Mathematics*, *07*(10), 1057–1067. https://doi.org/10.4236/am.2016.710093

Pan, Y., He, Y., Song, R., Wang, G., & An, Q. (2020). A passive and inclusive strategy to impute missing values of a composite categorical variable with an application to determine HIV transmission categories. *Annals of Epidemiology*, *51*, 41-47.e2. https://doi.org/10.1016/j.annepidem.2020.07.012

Plumpton, C. O., Morris, T., Hughes, D. A., & White, I. R. (2016). Multiple imputation of multiple multi-item scales when a full imputation model is infeasible. *BMC Research Notes*, *9*(1). https://doi.org/10.1186/s13104-016-1853-5

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Raghunathan, T., Solenberger, P., Berglund, P., & van Hoewyk, J. (2022). *IVEware: Imputation and Variance Estimation Software*. University of Michigan. https://src.isr.umich.edu/wp-content/uploads/iveware_manual_updated_19nov2022.pdf

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. https://doi.org/10.1093/ije/dyaa042

Rioux, C., & Little, T. D. (2021). Missing data treatments in intervention studies: What was, what is, and what should be. *International Journal of Behavioral Development*, *45*(1), 51–58. https://doi.org/10.1177/0165025419880609

Rioux, C., Stickley, Z. L., Odejimi, O. A., & Little, T. D. (2020). Item Parcels as Indicators. In R. van de Shoot & M. Miočević (Eds.), *Small sample size solutions: A guide for applied researchers and practitioners* (pp. 203–214). Routledge. https://doi.org/10.4324/9780429273872-17

Robinson, L., Adair, P., Coffey, M., Harris, R., & Burnside, G. (2016). Identifying the participant characteristics that predict recruitment and retention of participants to randomised controlled trials involving children: a systematic review. *Trials*, *17*(1). https://doi.org/10.1186/s13063-016-1415-0

Rombach, I., Gray, A. M., Jenkinson, C., Murray, D. W., & Rivero-Arias, O. (2018). Multiple imputation for patient reported outcome measures in randomised controlled trials: advantages and disadvantages of imputing at the item, subscale or composite score level. *BMC Medical Research Methodology*, *18*(1). https://doi.org/10.1186/s12874-018-0542-6

Royston, P., & White, I. A. (2011). Multiple imputation by chained equation (MICE): Implementation in Stata. *Journal of Statistical Software*, *45*(4), 1–20. https://doi.org/10.18637/jss.v045.i04

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

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

SAS Institute Inc. (2023). *SAS/STAT® 15.3 User's Guide, The MI Procedure*. SAS Institute Inc. https://documentation.sas.com/api/collections/pgmsascdc/9.4_3.5/docsets/statug/content/mi.pdf?locale=en#nameddest=statug_mi_toc

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall. https://doi.org/10.1201/9781439821862

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, *8*(1), 3–15. https://doi.org/10.1177/096228029900800102

Schreiber, J. B. (2021). Issues and recommendations for exploratory factor analysis and principal component analysis. *Research in Social and Administrative Pharmacy*, *17*(5), 1004–1011. https://doi.org/10.1016/j.sapharm.2020.07.027

Seaman, S., Galati, J., Jackson, D., & Carlin, J. (2013). What is meant by "missing at random"? *Statistical Science*, *28*(2), 257–268. https://doi.org/10.1214/13-sts415

Sep, M. S. C., Joëls, M., & Geuze, E. (2020). Individual differences in the encoding of contextual details following acute stress: An explorative study. *European Journal of Neuroscience*, *55*(9–10), 2714–2738. https://doi.org/10.1111/ejn.15067

Shogren, K. A., Little, T. D., Grandfield, E., Raley, S., Wehmeyer, M. L., Lang, K. M., & Shaw, L. A. (2020). The Self-determination inventory–student report: Confirming the factor structure of a new measure. *Assessment for Effective Intervention*, *45*(2), 110–120. https://doi.org/10.1177/1534508418788168

Shubert, J., Wray-Lake, L., & McKay, B. (2020). Looking ahead and working hard: How school experiences foster adolescents' future orientation and perseverance. *Journal of Research on Adolescence*, *30*(4), 989–1007. https://doi.org/10.1111/jora.12575

StataCorp. (2023a). *Stata multiple-imputation reference manual, Release 18*. Stata Press. https://www.stata.com/manuals/mi.pdf

StataCorp. (2023b). *Stata Statistical Software: Release 18*. StataCorp LLC.

Steffgen, G., Sischka, P. E., & Fernandez de Henestrosa, M. (2020). The Quality of work index and the Quality of employment index: A multidimensional approach of job quality and its links to well-being at work. *International Journal of Environmental Research and Public Health*, *17*(21), 7771. https://doi.org/10.3390/ijerph17217771

Survey Research Center. (2023). *IVEware: Imputation and variance estimation software*. University of Michigan, Institute for Social Research. https://src.isr.umich.edu/software/iveware/

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16*(3), 219–242. https://doi.org/10.1177/0962280206074463

van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). **mice**: Multivariate Imputation by Chained Equations in*R*. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer. http://catdir.loc.gov/catdir/toc/fy042/2002022925.html

Vera, J. D., & Enders, C. K. (2021). Is item imputation always better? An investigation of wave-missing data in growth models. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(4), 506–517. https://doi.org/10.1080/10705511.2020.1850289

Vink, G., & van Buuren, S. (n.d.). *mice: Passive imputation and Post-processing*. Retrieved July 22, 2023, from https://www.gerkovink.com/miceVignettes/Passive_Post_processing/Passive_imputation_post_processing.html

Vink, G., & van Buuren, S. (2013). Multiple imputation of squared terms. *Sociological Methods & Research*, *42*(4), 598–607. https://doi.org/10.1177/0049124113502943

von Hippel, P. T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, *39*(1), 265–291. https://doi.org/10.1111/j.1467-9531.2009.01215.x

von Hippel, P. T. (2020). How many imputations do you need? A two-stage calculation using a quadratic rule. *Sociological Methods & Research*, *49*(3), 699–718. https://doi.org/10.1177/0049124117747303

Welch, C. A. (2022). *TWOFOLD: Stata module to perform multiple imputation using the two-fold fully conditional specification algorithm to impute missing values in longitudinal data*. https://EconPapers.repec.org/RePEc:boc:bocode:s457690

Welch, C. A., Bartlett, J., & Petersen, I. (2014). Application of multiple imputation using the two-fold fully conditional specification algorithm in longitudinal clinical data. *The Stata Journal*, *14*(2), 418–431. https://doi.org/10.1177/1536867x1401400213

Welch, C. A., Petersen, I., Bartlett, J. W., White, I. R., Marston, L., Morris, R. W., Nazareth, I., Walters, K., & Carpenter, J. (2014). Evaluation of two-fold fully conditional specification multiple imputation for longitudinal electronic health record data. *Statistics in Medicine*, *33*(21), 3725–3737. https://doi.org/10.1002/sim.6184

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*(4), 377–399. https://doi.org/10.1002/sim.4067

Wijesuriya, R., Moreno-Betancur, M., Carlin, J. B., De Silva, A. P., & Lee, K. J. (2021). Evaluation of approaches for accommodating interactions and non-linear terms in multiple imputation of incomplete three-level data. *Biometrical Journal*, *64*(8), 1404–1425. https://doi.org/10.1002/bimj.202000343

Zhang, Q., & Wang, L. (2017). Moderation analysis with missing data in the predictors. *Psychological Methods*, *22*(4), 649–666. https://doi.org/10.1037/met0000104

## Supplementary Materials

## Peer Review History

Download: https://collabra.scholasticahq.com/article/92993-multiple-imputation-when-variables-exceed-observations-an-overview-of-challenges-and-solutions/attachment/194667.docx?auth_token=cKEzq8JAmFuPO0qX8xGv