

Rebutting Existing Misconceptions About Multiple Imputation as a Method for Handling Missing Data

Joost R. van Ginkel, Marielle Linting, Ralph C. A. Rippe & Anja van der Voort

To cite this article: Joost R. van Ginkel, Marielle Linting, Ralph C. A. Rippe & Anja van der Voort (2019): Rebutting Existing Misconceptions About Multiple Imputation as a Method for Handling Missing Data, Journal of Personality Assessment, DOI: [10.1080/00223891.2018.1530680](https://doi.org/10.1080/00223891.2018.1530680)

To link to this article: <https://doi.org/10.1080/00223891.2018.1530680>



© 2019 The Author(s). Published with license by Taylor & Francis Group, LLC.



Published online: 18 Jan 2019.



Submit your article to this journal [↗](#)



Article views: 1094



View Crossmark data [↗](#)

Rebutting Existing Misconceptions About Multiple Imputation as a Method for Handling Missing Data

Joost R. van Ginkel¹, Marielle Linting², Ralph C. A. Rippe², and Anja van der Voort²

¹Department of Psychology, Methodology and Statistics, Leiden University, Leiden, The Netherlands; ²Center for Child and Family Studies, Leiden University, Leiden, The Netherlands

ABSTRACT

Missing data is a problem that occurs frequently in many scientific areas. The most sophisticated method for dealing with this problem is multiple imputation. Contrary to other methods, like listwise deletion, this method does not throw away information, and partly repairs the problem of systematic dropout. Although from a theoretical point of view multiple imputation is considered to be the optimal method, many applied researchers are reluctant to use it because of persistent misconceptions about this method. Instead of providing an(other) overview of missing data methods, or extensively explaining how multiple imputation works, this article aims specifically at rebutting these misconceptions, and provides applied researchers with practical arguments supporting them in the use of multiple imputation.

ARTICLE HISTORY

Received 10 January 2018
Revised 10 September 2018

Goal and intended audience

This article addresses applied researchers within the field of social and behavioral sciences (but possibly other fields as well) facing the problem of missing data, who have heard of *multiple imputation* as a method to deal with missing data, but have concerns about actually using this method. These concerns may be based on misconceptions implying that in their specific situation multiple imputation should either not be used at all, or only with much caution. This article collects several of those misconceptions, and provides grounded rebuttal—through theory and practical argumentation—to ultimately support researchers in their deliberations regarding their statistical analyses when faced with missing data.



Missing data

Missing data are a common problem in psychological research and many other scientific areas (see studies by Van Ginkel, Sijtsma, Van der Ark, & Vermunt, 2010, in the field of psychology; Eekhout, de Boer, De Vet, & Heymans, 2012, in the field of epidemiology; and Rombach, Rivero-Arias, Gray, Jenkinson, & Burke, 2016, in the field of quality of life research). Once confronted with missing data the question is how they can be handled. The easiest way is to use *listwise deletion*, which excludes all respondents with missing data from the statistical analyses. In several statistical software packages, such as SPSS 25.0 (SPSS, Inc., 2017), the most widely applied analyses use listwise deletion by default.

Although easy to apply, listwise deletion has two important disadvantages. The first problem is wastefulness: It discards valuable information, which consequently leads to a loss of power. The second disadvantage is more serious: Results of statistical analyses may be biased. Whether or not results will be biased after listwise deletion is dependent on the underlying mechanism that caused the missing data, also known as the *missingness mechanism*. Three different missingness mechanisms can be distinguished (e.g., Little & Rubin, 2002; Van Buuren, 2012), which are discussed next.

Missingness mechanisms

The three different missingness mechanisms as defined by Little and Rubin (2002, p. 10) and Rubin (1976) are *missing completely at random* (MCAR), *missing at random* (MAR), and *not missing at random* (NMAR). MCAR means that the probability of a missing value neither depends on any observed data, nor unobserved data. An example of MCAR is a respondent accidentally skipping a question. Under MCAR, cases with missing data are a simple random subsample from all cases in the data. Consequently, when a simple random subsample is removed from the total sample, the resulting leftover sample is still as representative of the population as the original sample was. In short, under MCAR, listwise deletion reduces the sample size and by the latter, power, but does not give any biased results.

CONTACT Joost R. van Ginkel  jginkel@fsw.leidenuniv.nl  Department of Psychology, Methodology and Statistics, Faculty of Social and Behavioral Sciences, Leiden University, PO Box 9555, 2300 RB Leiden, The Netherlands.

This article was accepted under the editorship of Steven K. Huprich.

© 2019 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

Under MAR, missing data depend on observed data but not on unobserved data. For example, suppose that females skip a question about income more frequently than males, and gender is observed for all respondents. The missing data on income are thus MAR. If listwise deletion is applied under MAR, the leftover sample may not be representative of the total population anymore, consequently leading to biased results. In the preceding example, females are under-represented in the leftover sample. As a result, statistical inferences from analyses that include gender or variables that are related to gender may be biased.

Finally, under NMAR, the probability of missingness depends on data that are not observed. If, for example, respondents with higher incomes are more inclined to skip a question about income compared to those with low incomes, the missing data on income are NMAR. In general, under NMAR, listwise deletion has the same problem as it has under MAR, because there are systematic differences between the deleted cases and the cases that are left (respondents with systematically higher incomes in the preceding example). However, under MAR the causes of this systematic dropout can be traced, whereas under NMAR they cannot. Only in very specific and rare cases of NMAR (Van Buuren, 2012, p. 48; see also Vach, 1994; White & Carlin, 2010) does listwise deletion give unbiased results of statistical analyses. For a more formal explanation of missingness mechanisms, see Van Buuren (2012) and Little and Rubin (2002).

It is important to realize that the term *missing at random* does not mean that the missing data are a simple random subsample of all the data points. That scenario is MCAR. Under MAR, missing data may be more frequent in some subgroups in the data than in others, but information defining the subgroups is observed for all respondents (gender in the example). However, because of its name, MAR may easily be misinterpreted as what would technically be MCAR (see, e.g., Baraldi & Enders, 2010, p. 7; Schafer & Graham, 2002, p. 152).

Testing which of the three missingness mechanisms apply

Several statistical tests can be used to test the MCAR assumption. First, suppose that we have a group of respondents with observed values on a variable X and another group with missing values on this variable X . Using a t test or a chi-square test it can be tested whether these two groups differ significantly in means or frequency distribution on another variable Y . Such t and chi-square tests can be performed for all variable pairs. Additionally, Little (1988) provided an overall test that can test the assumption of MCAR for the whole data set at once.

When one of these tests is significant, the MCAR assumption may have been violated. However, significance of these tests does not say anything about whether the missingness is MAR or NMAR. Moreover, when none of the tests are significant, it does not automatically mean that the missingness is MCAR, either. Although unlikely, theoretically it is possible that the missing values are completely randomly scattered across the data, but that they would still

have systematically lower or higher values than the observed data if they had been observed. Because under NMAR the missing data depend on unobserved information, there is, by definition, no way of knowing whether they are NMAR, unless information about the population is available to the researcher. The only thing these statistical tests tell us is that when significant, the null hypothesis of the missing data being randomly scattered across the data has been rejected. Assuming, for the sake of argument, that no Type I error has been made, this rejection implies either MAR or NMAR. When nonsignificant, the null hypothesis of the missing data being randomly scattered has not been rejected. Assuming that no Type II error has been made here, this excludes missingness mechanism MAR but not necessarily NMAR, as NMAR is unverifiable without additional information about the population.

Because listwise deletion is very wasteful and could lead to bias in statistical analyses under MAR and NMAR, alternative methods to deal with missing data are necessary. In the next section, a number of alternatives are discussed.

Alternatives to listwise deletion

Pairwise deletion

One way to resolve the wastefulness of listwise deletion is to use pairwise deletion. For analyses that use a covariance matrix or correlation matrix as input for the computations (e.g., a principal component analysis [PCA] or a regression analysis), pairwise deletion computes each covariance or correlation from the cases with observed values on both variables for which the specific covariance or correlation is computed. Consequently, each covariance or correlation can be computed for different cases and different numbers of cases. The same can be done for other statistics, such as variable means, where for each variable the mean can be computed across a different number of cases.

Although pairwise deletion is less wasteful than listwise deletion, its applicability is limited to analyses that can work around the missing values by using a variable-by-variable basis. To illustrate, the commonly used statistical package SPSS 25.0 only has a pairwise deletion option for correlation procedures (e.g., correlations, partial correlations), linear regression, descriptive statistics procedures, PCA and its related techniques (e.g., principal axis factoring, maximum likelihood), and K-means clustering. Additionally, like listwise deletion, pairwise deletion may give bias in statistical analyses when the missing data are not MCAR. Furthermore, as each covariance or correlation could be based on a different number of cases, it is not clear what sample size to use when calculating standard errors. Finally, the fact that different cases are used for each covariance or correlation may also cause computational problems, such as negative variances or correlations outside the range of $[-1, 1]$ (Van Buuren, 2012, pp. 9–10). In short, pairwise deletion only resolves the problem of wastefulness that listwise deletion has, but comes with several additional problems, so we would generally not recommend it.

Imputation

Another way to remedy the wastefulness of listwise deletion is to fill in plausible values for the missing data. This is also known as *imputation*. Examples of imputation are variable mean imputation (filling in the variable mean for each missing value) or regression imputation (filling in the predicted values from a regression model with other variables as predictors). Although these methods solve the wastefulness from listwise deletion and computational problems from pairwise deletion, they create another problem: Filling in variable means will bias the variances and covariances downward, and predicted values from a regression model will bias variances downward and covariances upward. Consequently, subsequent statistical analyses will be biased as well.

Rather than filling in predicted values of a regression model or the variable mean for a missing value, one could also impute a predicted value from a regression model plus a random error term, drawn from a normal distribution with a variance that equals the error variance of the specific regression model. This is also known as *stochastic regression imputation* (e.g., Little & Schenker, 1995, p. 60; Van Buuren, 2012, p. 13). Stochastic regression imputation resolves the problem of biased variances and covariances. However, in the subsequent statistical analyses the imputed values are treated as if they are real values. Consequently, more certainty about the imputed values is assumed in the analyses than there actually is, which will bias both p values and widths of confidence intervals downward.

In short, simple imputation methods resolve the problem of wastefulness but they introduce additional bias in statistical analyses, regardless of the missingness mechanism. Thus, despite its wastefulness, listwise (and pairwise) deletion will still give the most guarantee of obtaining unbiased results in statistical analyses of all the methods discussed so far.

Multiple imputation

A method that resolves all of the previously mentioned problems (wastefulness, computational problems, biased [co]variances, and biased p values and confidence intervals), is *multiple imputation* (Rubin 1987). Multiple imputation works in three steps. In the first step, several plausible complete versions of the incomplete data sets are created. This is done by drawing several values for each missing data point, using a statistical model that accurately describes the data, plus a random error component. In the second step, the different complete versions of the incomplete data set are analyzed using standard statistical procedures. This will consequently result in multiple (slightly) different outcomes of the statistical analyses. In the final step, these results are combined into an overall statistical analysis in which the uncertainty about the missing data is incorporated in the standard errors and significance tests.

Considering its advantages compared to listwise deletion, pairwise deletion, and (single) imputation methods, one

would expect that since its invention (Rubin, 1987) researchers would have started using multiple imputation frequently. However, Rombach et al. (2016), Eekhout et al. (2012), and Van Ginkel et al. (2010) found that listwise deletion was by far the most frequently used method for dealing with missing data in the studies that report the presence of missing data. Multiple imputation was either rarely used (Eekhout et al., 2012; Rombach et al., 2016) or not used at all (Van Ginkel et al., 2010). Although the most recent of these studies (Rombach et al., 2016) reported an increase in multiple imputation compared to the earlier studies, the size of this increase does not suggest that multiple imputation will be the most frequently used method any time soon.

These findings raise the question of why multiple imputation is used so rarely. Although to our knowledge there is no research that investigates this question, we can think of three possible explanations. The first two explanations are unfamiliarity and user-unfriendliness (see, e.g., Baraldi & Enders, 2010, who addressed the issue of user-unfriendliness). For decades, multiple imputation was not implemented in SPSS, the most frequently used statistical software package among social scientists. Consequently, people were dependent on less well-known or less user-friendly software packages like NORM (Schafer, 1998), S-plus for Windows (2001), or the procedure PROC MI in SAS (Yuan, 2000), which may have posed too high a threshold for many researchers to start applying multiple imputation.

However, with the release of SPSS version 17.0 (PASW at the time; SPSS, Inc., 2009) multiple imputation became available for SPSS users as well. It is therefore our prediction that in time the unfamiliarity will diminish or even disappear. On the other hand, the procedure in SPSS lacks some important options that are available in less user-friendly software, such as the `mice` procedure (Van Buuren & Groothuis-Oudshoorn, 2011) in R (R Development Core Team, 2013), the `mi` procedure in Stata 14.0 (StataCorp, 2015), and the PROC MI procedure (Yuan, 2011) in SAS 9.4 (SAS Institute, 2013). Thus, to solve more complex missing-data problems, users will still need more complex software.

The third possible explanation for why multiple imputation is still rarely used nowadays is a heterogeneous set of misunderstandings of the procedure, which may cause researchers to distrust it. As statistical advisors to applied researchers we have heard many misconceptions about multiple imputation expressed by both colleagues and reviewers throughout the years. This article focuses on the most common of these misconceptions, because this general distrust in multiple imputation will not simply disappear by making the procedure more user-friendly and more widely available. For people to start accepting multiple imputation as a reliable method for handling missing data, it has to be explained why there is no reason to distrust it.

In the next section the method of multiple imputation is explained in more detail. In the sections that follow the most frequently heard misconceptions about multiple imputation held by applied researchers are rebutted. In the

conclusion, it is argued that from a theoretical point of view, multiple imputation is practically always to be preferred over listwise deletion. As an aside it should be noted that apart from listwise deletion, pairwise deletion, single imputation, and multiple imputation, various other methods for handling missing data exist. These methods are briefly commented on in the discussion.

Multiple imputation explained

In general there are two main approaches for generating multiply imputed data sets, namely the *joint modeling* approach (Schafer, 1997; Van Buuren, 2012, pp. 105–108) and the *fully conditional specification* approach (Van Buuren, 2012, pp. 116–118; Van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006). Within the fully conditional specification approach one can further make a distinction between the regression approach and predictive mean matching. Because the joint modeling approach is not used frequently nowadays, the focus is on the two varieties of fully conditional specification.

Regression

The regression approach imputes missing data as random draws from a conditional distribution based on a linear regression model when variables are continuous, or a (multinomial) logistic regression model when variables are categorical. The algorithm works as follows:

1. Starting values based on the variables' marginal distributions are filled in for the missing data on each variable (for details, see Van Buuren et al., 2006).
2. For a variable X with missing values the parameters of the (logistic) regression model for predicting its missing values are calculated.
3. The current estimates of the missing data on variable X are replaced with new random values using the (logistic) regression model with the model parameters calculated in Step 2.
4. Steps 2 and 3 are carried out for all variables with missing data.
5. Step 4 is repeated until properties of the imputed values such as means and standard deviations, stabilize.
6. Finally, to obtain M multiply imputed data sets, Step 5 is repeated M times.

More technical explanations of fully conditional specification are described in Van Buuren (2012), Van Buuren et al. (2006), and Van Buuren, Boshuizen, and Knook (1999).

Predictive mean matching

Like the regression approach, *predictive mean matching* (PMM; Rubin, 1986; Van Buuren, 2012, pp. 68–74; Van Buuren, Boshuizen, and Knook, 1999; Van Buuren et al., 2006) uses a linear regression model to impute continuous missing data, and a (multinomial) logistic regression model

for categorical variables. However, for continuous variables the imputed values in Step 3 are not random draws from the conditional distribution based on the regression model. Instead, the regression model is used to find cases with observed values on the outcome variable with predicted values that closely resemble the predicted values of the respondents with missing values. For each person with a missing value on a particular variable, a matching respondent is (randomly) selected and the observed value of this matching respondent is used for imputation. For categorical variables, PMM works in the same way as the regression approach.

One advantage of this approach over the regression approach is that it imputes values that are actually observed for other respondents, so that values outside the range of the data cannot occur. A second advantage is that it is more robust to violations of normality than the regression approach (Marshall, Altman, & Holder, 2010; Marshall, Altman, Royston, & Holder, 2010). Both the regression approach and PMM are implemented in SPSS 25.0 (SPSS Inc., 2017), the `mice` package in R (Van Buuren & Groothuis-Oudshoorn, 2011), Stata 14.0 (StataCorp, 2015), and in SAS 12.1 in the procedure PROC MI (Yuan, 2011).

Misconceptions about multiple imputation

Van Buuren (2012) argued that multiple imputation is a better alternative than both listwise deletion and single imputation, and gave an overview of many simulation studies that confirm this. However, misconceptions about the method may stand in the way of its frequent application. Next, we describe and rebut the most common of these misconceptions.

1. Multiple imputation should only be used when the missingness is MAR.
2. Multiple imputation should only be used when too few cases are left after listwise deletion.
3. If results from statistical analyses obtained from multiple imputation differ from those of listwise deletion, the results of multiple imputation must be wrong.
4. Outcome variables must not be imputed.
5. Predictor variables must not be imputed.
6. Multiple imputation must not be used because you will end up with several different outcomes of your statistical analysis.

It should be noted that this selection of misconceptions has been solely based on our own experience as statistical advisors. When applied researchers have misconceptions about multiple imputation, they will simply refrain from its use, without further mentioning any rationale for their decision. Explicit references to misconceptions are therefore not expected to occur in the literature, as supported by finding zero hits in an extensive literature search. We have, however, been able to find literature from authors with a statistical background that expresses similar experiences with applied researchers as ours (Schafer & Graham, 2002).

Furthermore, we have found references (e.g., Von Hippel, 2008) that do not express a misconception but that could easily be misinterpreted by researchers, causing misconceptions to form or persist. This literature is discussed in more detail in the rebuttals of the misconceptions, which follow next.

Multiple imputation should only be used when the missingness is missing at random

The misconception

It is our experience that researchers often refrain from using multiple imputation because statistical tests indicated that the MAR assumption had been violated. Similarly, we have had experiences with reviewers criticizing the use of multiple imputation because it had not been checked whether the assumption of MAR was met.

Rebutting the misconception

Under which missingness mechanisms does multiple imputation work? MCAR. Like listwise deletion, multiple imputation will give unbiased results under MCAR. Multiple imputation under MCAR imputes the data such that properties like means and covariances of the observed cases are simply extrapolated to cases with missing data. Consequently, the cases with multiply imputed data have similar properties as the cases with complete data. An advantage of multiple imputation over listwise deletion in the case of MCAR, however, is that multiple imputation does not throw away information, whereas listwise deletion does. Although the problem of throwing away information could be reduced by using pairwise deletion rather than listwise deletion, using pairwise deletion comes with other problems, as mentioned earlier.

MAR. Another advantage of multiple imputation over listwise deletion, apart from not throwing away useful information, is that it is capable of correcting the bias that listwise deletion suffers from when the missing data are MAR. Consider the example where females tend to leave a question measuring someone's income open more often than males. If in the imputation model gender is used as a (dummy) predictor for the imputation of income, then the dependence of the missingness on gender is incorporated in the imputed values for income. In short, when used properly, multiple imputation will give valid results of statistical analyses under both MCAR and MAR.

NMAR. Because multiple imputation only uses observed information from other variables to predict the missing data on a specific variable, it will only give unbiased results as long as the missingness can be explained completely from the observed variables. When missingness depends on information that is not observed, as is the case in NMAR, unbiased results are no longer guaranteed.

However, under NMAR, listwise deletion will yield biased results as well (Baraldi & Enders, 2010, pp. 8, 21), except in the cases described by White and Carlin (2010) and by Vach (1994), where listwise deletion may outperform

multiple imputation. Because these exceptions are very specific cases that may occur only incidentally in practice, we do not further address these exceptions in the remainder of the discussion.

Given that listwise deletion gives biased results under NMAR as well, a violation of MAR in itself is no reason to prefer listwise deletion over multiple imputation. Moreover, Schafer (1997, pp. 26–27) argued and showed that even under NMAR, multiple imputation gives less biased results than listwise deletion. There are two reasons for this. First, contrary to listwise deletion, multiple imputation still picks up dependencies of the missing data on observed information, so that at least dependencies of the missingness on observed variables are accounted for. Second, as far as missing data depend on unobserved information, these dependencies can partly be accounted for by relations with observed variables that may also be related to the unobserved variables on which the missingness depends. This implies that the more variables are included in the imputation model, the less residual dependence of the missingness on unobserved variables or information remains.¹ In short, although multiple imputation is not guaranteed to produce unbiased results under NMAR, the MAR assumption becomes more plausible as more observed variables are included in the imputation model (Schafer, 1997, p. 28).²

Preferring multiple imputation over listwise deletion is not dependent on the acting missingness mechanism.

When researchers or reviewers say that an assumption of multiple imputation is that the missing data are MAR, they are only partly right. It is not true that multiple imputation works under MAR only. MAR is the least restrictive assumption under which multiple imputation works. Because it works under MAR, it also works under the more restrictive assumption MCAR.

Furthermore, researchers and reviewers are wrong when they say that the use of multiple imputation is unjustified when statistical tests testing the MCAR assumption are significant. As pointed out earlier, these procedures test the null hypothesis that the missing data are randomly scattered across the data set, which could indicate MCAR. When significant, the missing data are thus probably not MCAR but they could still be MAR. As MCAR is a sufficient but not necessary condition for multiple imputation, a significant *t* test, chi-square test, or MCAR test does not automatically invalidate multiple imputation. What it does invalidate is listwise deletion, as MCAR is a necessary assumption of listwise deletion. In other words, an MCAR test should be used

¹Note that, as in any regression model, the number of predictors in multiple imputation models should not be too large compared to the sample size to avoid overfitting.

²It should be noted that methods for modeling NMAR have been proposed as well (e.g., Fay, 1986; Galimard, Chevret, Protopopescu, & Resche-Rigon, 2016; Heckman, 1976; Moustaki & Knott, 2000). Schafer (1997, p. 28) however, noted that such models require more parameters than can be estimated from the data alone, so to make them identifiable, restrictions must be imposed on the parameters. Furthermore, the current implementation in software packages of such NMAR models is limited.

for determining whether it is safe to use listwise deletion, not for determining whether multiple imputation is justified.

In short, neither the outcome of the statistical tests for testing MCAR, nor the actual underlying missingness mechanism are relevant for deciding whether or not to use multiple imputation. The reason many researchers use listwise deletion instead of multiple imputation when some of the tests for testing MCAR are significant is probably that they are not aware that MCAR and MAR are two different concepts. When they read or hear that MAR is a necessary assumption for multiple imputation, they think that a significant MCAR test invalidates the use of multiple imputation. Next, they use listwise deletion because they may think that doing nothing at all about the missing data handling is still better than doing something incorrect.

To conclude, regardless of the missingness mechanism, multiple imputation is always to be preferred over listwise deletion. Under MCAR it is preferred because it results in more statistical power, under MAR it is preferred because besides more power it will give unbiased results whereas listwise deletion may not, and under NMAR it is also the preferred method because it will give less biased results than listwise deletion. See also Baraldi and Enders (2010, p. 8).

Multiple imputation should only be used when too few cases are left after listwise deletion

The misconception

In our experience, applied researchers often turn to multiple imputation only after attempted analyses using listwise deletion failed because there were too few cases left for any useful statistical analysis. This suggests that these researchers think that multiple imputation should only be used if listwise deletion is infeasible.

Rebutting the misconception

Missing data are not only a problem of power reduction. Under both MAR and NMAR, the dropout that results from listwise deletion will be systematic. Consequently, results of statistical analyses could be biased if incomplete cases are dropped from the statistical analysis. Multiple imputation on the other hand, will completely eliminate this bias under MAR, and partly eliminate it under NMAR. Again, the conclusion is that multiple imputation is to be preferred over listwise deletion, even when after listwise deletion enough cases are left for statistical analysis.

If results from statistical analyses obtained from multiple imputation differ from those of listwise deletion, the results of multiple imputation must be wrong

The misconception

This misconception especially applies when the results obtained from listwise deletion are in accordance with the expectations of the researcher and the results obtained from multiple imputation are not. In our experience, applied

researchers are inclined to unjustly distrust the results from multiple imputation when they are not in accordance with their own expectations or with the results from listwise deletion, and rely on listwise deletion instead.

Rebutting the misconception

We start with a very obvious point: As multiple imputation has been developed to solve problems involved in listwise deletion, the conclusions obtained from these methods cannot always be the same. When conclusions obtained from multiple imputation differ from those obtained from listwise deletion, this is not necessarily caused by things that go wrong in multiple imputation, and even when it is, this is still no reason to automatically turn to listwise deletion instead. As already pointed out, results from listwise deletion suffer from a loss of power, but more important, they may be biased under MAR and NMAR. When carried out correctly, multiple imputation results in more power than listwise deletion, it completely corrects for bias under MAR, and partly corrects for bias under NMAR. This increase in power and correction for bias could explain possible differences in results between multiple imputation and listwise deletion.

However, results obtained from multiple imputation and listwise deletion could also differ as a result of incorrectly applying multiple imputation. Van Buuren (2012, pp. 250–251) provided a number of dos and don'ts in multiple imputation. When these guidelines are not followed, the procedure could easily impute nonsensical values and consequently, results of statistical analyses cannot be trusted.

The question is what should be done when the results obtained from listwise deletion and from multiple imputation differ. First, it is important to check the imputed values for anomalies. For example, check whether there are imputed values far beyond the minimum and maximum observed values of the variables, and check whether the imputed values follow substantially different patterns in scatter plots or histograms than the observed values do. If such anomalies occur, the imputation model should be adjusted (rather than putting multiple imputation aside as a whole). For example, for a specific variable with missing data, use more or fewer variables in the imputation model, include interaction or nonlinear (e.g., quadratic) terms, or, if you have not already done so, use PMM rather than the regression approach. It should be noted that this adjustment of the imputation model can be a complicated process, and help from a statistician may be needed. Furthermore, it should be noted that when following Van Buuren's guidelines the risk of diverging results due to anomalies in the imputed values is largely reduced, as these guidelines include checking the imputed values prior to carrying out any statistical analysis.

Second, when there are no anomalies in the imputed values, the next step is to check the MCAR assumption in the original sample without imputed data. If Little's MCAR test and some of the t or chi-square tests for testing MCAR are significant, the MCAR assumption may have been violated. This violation of MCAR could explain the differences in

results between listwise deletion and multiple imputation, and the results obtained from listwise deletion should be mistrusted, not the results obtained from multiple imputation.

Outcome variables must not be imputed

The misconception

Some of the applied researchers that we have advised to use multiple imputation have agreed on using this procedure, but were reluctant to impute the outcome variable or variables of their intended statistical analyses. When discussing their reasons, they would reply that if an outcome variable were imputed using predictors that were used as predictors in the subsequent analysis as well, the imputed values would only confirm the model that they wanted to use for their analysis.

Rebutting the misconception

If a linear regression model is used to analyze the data, and the relations of the predictors with the outcome variable are all linear, there is nothing to be incorrectly confirmed, because the imputation model (which in its standard form is based on linear regression), the model used for analysis, and the model that generated the data are the same. Consequently, this misconception only applies when the researcher has opted to use a linear regression model on data in which (some of) the predictors have a nonlinear relationship with the outcome variable. In rebutting this misconception, we thus focus on a bivariate situation where predictor X is nonlinearly related to outcome variable Y .

In [Figure 1](#), a number of graphs are shown of simulated data with a nonlinear relationship between X and Y . [Figure 1A](#) shows a scatter plot of simulated bivariate data according to a nonlinear regression model where outcome variable Y is quadratically related to X . [Figure 1B](#) shows the same simulated data but with 40% of the values of Y removed according to MCAR (corresponding cases are not shown in the plot).

Now suppose we incorrectly assume that X and Y are linearly related and that therefore we want to use a linear regression model of X on Y for both multiple imputation and the analysis. This situation is shown in [Figure 1C](#). In [Figure 1C](#) you see that the observed data (black dots) behave according to a quadratic relationship, but that the imputed values (white dots) show a linear relationship. Will this confirm the statistical model of interest (i.e., a linear regression of X on Y) any more than when the outcome variable is not imputed? The answer is no. When incorrectly assuming a linear model in both the imputation model and in the analysis, you will reach biased conclusions about the relation between X and Y , not because of multiple imputation, but because a linear relation between X and Y is assumed. When a linear relation between X and Y is assumed in both the imputation process and in the statistical analysis, the multiply imputed values will not bias the regression coefficient of X on Y and its standard error any more than not imputing would do. The imputed values are in accordance with the incorrectly assumed linear regression coefficient

that one would get when Y is not imputed, so they will give a similar (biased) regression coefficient and a similar (biased) standard error. The imputed values may not be in accordance with the nonlinear patterns in the data, but in a subsequent linear regression analysis of X on Y they will behave neutrally.

One may argue that still the incorrect regression model will be confirmed more quickly than when the outcome variable is not imputed, because of increased power. In other words, an incorrect model is estimated with more certainty. Technically speaking, someone who puts forward this argument is right. However, the same reasoning could be used for preferring a small sample size over a large sample size in general (regardless of possible presence of missing data). In practice, data never behave exactly according to the statistical model that you use for analysis and yet large sample sizes are preferred over small sample sizes all the time. In the end researchers always prefer more certainty of a not entirely correct analysis model over less certainty. The actual problem here is that there is a discrepancy between the model that is assumed for the data and the way the data actually behave in the population. Whether you assume this incorrect model from the beginning of the process (the multiple-imputation phase), at the end (the statistical analysis phase), or whether you assume it when there are no missing data at all is irrelevant.

Additionally, one should keep in mind that the given situation is rather an example of bad research practice. In practice, a researcher should first check the scatter plot of X and Y before carrying out any multiple imputation or linear regression analysis. When researchers are confronted with a plot that looks like [Figure 1B](#), they should either include a nonlinear term of X in the imputation model (the `mice` package in R has options for this) or use PMM. As mentioned before, PMM uses observed data from other cases for imputation, and keeps nonlinear relations more intact than multiple imputation using regression. [Figure 1D](#) shows what the imputed values (white dots) look like when the data are imputed using PMM. The plot shows that the imputed values follow the same nonlinear pattern as the observed data do (black dots).

Next, when the nonlinear relations among variables have been taken into account in the imputation procedure, either by means of using the same nonlinear term in the imputation model, or by means of using PMM, the model used for analysis should include a nonlinear (e.g., a quadratic) term as well. In doing so there is no reason to fear that by imputing outcome variable Y , the incorrect model will be accepted more quickly than when there are no missing data.

Some researchers may still think that multiple imputation of the outcome variable will confirm the model of interest because they may think that the relationship between X and Y for the cases with missing data on Y is different than for the cases with observed values on Y . See, for example, [Figure 1E](#). This is the same graph [Figure 1A](#), but here the black dots are the cases with observed values on both X and Y , and the white dots are the dots that would be removed in [Figure 1A](#) if the cases with the 40% highest values on X had

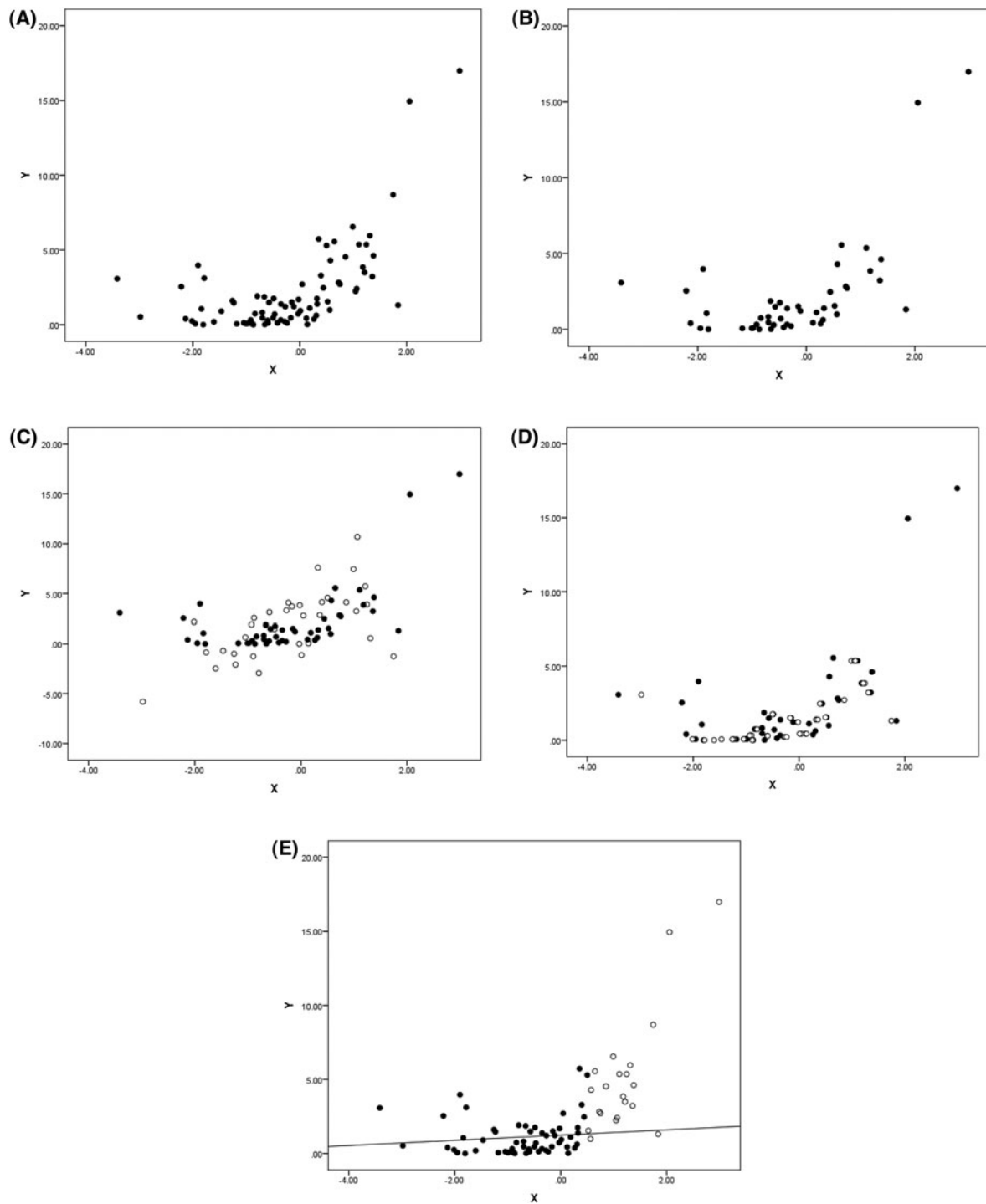


Figure 1. Five graphs of simulated bivariate data with a nonlinear relationship. (A) Neither variables have missing data. (B) 40% of the values on Y have been randomly removed. (C) White dots are imputed values using linear regression. (D) White dots are imputed values using predictive mean matching. (E) White dots are the cases with the 40% highest values on X, and the regression line is the line that is obtained when the 40% highest values are missing and not imputed.

missing data on Y. If a missingness pattern like this occurs, the researcher will see a plot like [Figure 1E](#), but without the white dots. From the remaining (black) dots the researcher may get the impression that there is a weak linear relation between X and Y. The regression line for this weak relationship is displayed in [Figure 1E](#) as well.

When a missingness pattern like this occurs in highly nonlinear data, a researcher may assume an incorrect statistical model from the scatter plot and may both impute and analyze the data using a linear regression model of X on Y.

Because there is only a weak and seemingly linear relation between X and Y for the cases with no missing data on Y, the imputation model will extrapolate the regression line and the imputed values will indeed confirm the (incorrect) statistical model of interest.

However, besides this hypothetical example being very unlikely (how likely is it that all respondents above a certain value of X have missing values on Y), this is actually an example of NMAR. Van Buuren (2012, p. 34) argued that an implication of MAR is that the multivariate distribution

of the data (X and Y in the example) given all observed data is the same for respondents without missing data and for respondents with missing data. For the missingness pattern in Figure 1E this is clearly not the case because the bivariate distribution of X and Y for the black dots is different than for the white dots. The black dots behave like there is a weak linear relation and the white dots behave like there is a strong linear relation (and if you take them together the relationship is actually quadratic).

As already argued, under NMAR neither multiple imputation nor listwise deletion (which is what technically happens when in this example the outcome variable is not imputed) are guaranteed to give unbiased results. In the specific example, both methods will incorrectly estimate a regression line that looks similar to the line in Figure 1E, only multiple imputation will extrapolate the regression line to the 40% highest cases, and impute values that lie around this regression line. In other words, in this situation all methods for handling missing data will break down, and multiple imputation of the outcome variable is not any worse than no imputation.

Of course theoretically, this problem could be resolved by using a quadratic term in the imputation model. This modification of the imputation model will cause the MAR assumption to be met for the specific example. Also, see, Schafer and Graham (2002, p. 153), who simulated MAR missingness using the same mechanism (all cases with an X above a specific threshold have missing data on Y). The difference between their example and the current one is that in their example, the relationship was linear and could already be derived from the observed part of the data, whereas in this example no clear quadratic relationship is visible from the observed part of the data alone.

Finally, there is the argument that when using multiple imputation, the power increases so that an incorrect model is estimated with more certainty. However, that argument does not hold here either, for the same reason as it did not hold for the situation of Figure 1C. In the situation of Figure 1E, the leftover sample after listwise deletion is a biased sample (only cases with the 60% lowest values on X are sampled). However, in practice (regardless of missing data), it is impossible to draw a completely unbiased sample from a population, and yet in practice, a large (biased) sample is preferred by researchers over a small sample all the time. Again, the actual problem is a discrepancy between the incorrectly assumed model for the data (here, caused by a biased sample) and the way the data actually behave in the population.

To conclude, using multiple imputation does not confirm an incorrectly assumed linear model any more than analyzing a data set without missing values. Neither does it confirm a linear relationship that only applies to the observed part of the data any more than a biased sample without missing data does. What is important is that, regardless of whether there are missing data, data are inspected in advance before blindly estimating a linear regression model on highly nonlinear data. As previously stated, when this data inspection reveals that there are nonlinear relations in the data, it is important that this nonlinearity is accounted

for in both the analysis (by including nonlinear terms) and the imputation process (by including the same nonlinear terms as in the analysis, or by means of PMM).

On a final note, Von Hippel (2008) argued that outcome variables are to be imputed because cases on missing Y may contain useful information for imputing X in other cases. However, he went on arguing that in the subsequent analysis it is better not to use the cases with imputed values on Y because it results in slightly less efficient estimates of the model. Although Von Hippel made a valid argument, it may still be justified to use the cases with imputed values on Y because first, it does not give biased results, and second, all analyses on the same multiply imputed data sets remain comparable with respect to sample size. Whatever decision one makes on this, it is important to realize that Von Hippel explicitly said that imputing the outcome variable is justified. However, in our experience, his advice not to use the imputed values on Y in the analyses is occasionally misinterpreted by researchers as not to impute the outcome variable.

Predictor variables must not be imputed

The misconception

Besides the misconception that outcome variables must not be imputed, a misconception held by other researchers is that predictor variables should not be imputed. In combination with the previous misconception, this misconception is quite remarkable because when both ideas are true, this would imply that multiple imputation should not be used at all, prior to doing an analysis with both an outcome and predictor variables.

The reason that applied researchers often give for not wanting to impute predictor variables is that conceptually it makes no sense to predict missing data on a variable that is a predictor itself. For example, suppose that in a data set someone's age is missing and that the missing value on age is predicted from someone's income. It is logically impossible that someone's age is (partly) influenced by someone's income.

Rebutting the misconception

What researchers holding this misconception do not realize about multiple imputation is that the model used for multiple imputation is not meant as a conceptually meaningful model. Multiple imputation is only used to accurately describe the relations and structures found in the data, and impute data with similar properties. As long as a variable correlates with another variable with missing data, it is a potential candidate for a predictor in the imputation model. It does not matter that we are not interested in the prediction of one variable from the other in the subsequent analysis, or that this prediction is even nonsensical. All an imputation model does is (a) determine that in general a high age coincides with a high income, (b) when age is missing for someone with a high income, infer that this

Imputation	Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
			B	Std. Error	Beta		
0	1	(Constant)	.965	.091		10.564	.000
		X	.910	.074	.880	12.298	.000
1	1	(Constant)	.949	.066		14.409	.000
		X	.960	.056	.890	17.198	.000
2	1	(Constant)	1.080	.072		14.907	.000
		X	.868	.061	.848	14.133	.000
3	1	(Constant)	.950	.071		13.303	.000
		X	.972	.060	.876	16.071	.000
4	1	(Constant)	.994	.064		15.509	.000
		X	.896	.054	.882	16.489	.000
5	1	(Constant)	.929	.063		14.793	.000
		X	.925	.053	.891	17.366	.000
Pooled	1	(Constant)	.980	.095		10.371	.000
		X	.924	.074		12.418	.000

Figure 2. SPSS output of a regression analysis with a single predictor (X) and an outcome variable (Y) on a multiply imputed data set. The pooled results are shown in bold.

person's age must probably be high as well, and (c) consequently impute a high value for age.

Multiple imputation must not be used because you will end up with several different outcomes of your statistical analysis

The misconception

Many applied researchers are reluctant to use multiple imputation because they have the idea that they will end up with several results of the same analysis, rather than one. Because they do not know what to do with that many results, they are inclined to pick one of these results. However, because they do not know which result to pick, they ultimately decide to resort to something simpler instead, such as listwise deletion or single imputation.

Rebutting the misconception

As explained before, in the process of multiple imputation statistical analyses are not only carried out on each of the multiply imputed data sets; the results of these analyses are pooled into one overall analysis as well. In multiple imputation you are not supposed to pick one of the results; it is the pooled analysis that you interpret as the final results. Figure 2 displays the SPSS output of a linear regression analysis to a simulated data set with a predictor X and an outcome variable Y , after multiple imputation. In this example, the number of times the data have been imputed is five. Normally this number should be larger (Graham, Olchowski, & Gilreath, 2007), but to keep the output small it was set to five here.

As can be seen, the SPSS output displays seven results of the same regression analysis, indicated by either a number ranging from 0 to 5 (Rows 1 to 6), or by "Pooled" (last row). The first results, indicated by 0, are the results that are obtained when data are not imputed, results indicated by the numbers 1 to 5 are the results obtained from the five imputed data sets, and the results indicated by "Pooled" are the pooled results that should be used for interpretation. In

the same way, SPSS automatically pools results of many other statistical analyses, such as correlations and their significance tests, two-sample t tests, and regression coefficients and their significance tests in logistic, ordinal, and multinomial logistic regression, and multilevel models (the Mixed Models procedure in SPSS).

Unfortunately, not all statistics are pooled by SPSS. Examples are standardized regression coefficients (as can actually be seen in Figure 2, where a pooled beta coefficient for X is lacking), (pseudo) R^2 in (logistic) regression, F tests in (multivariate) analysis of variance ([M]ANOVA), overall F tests testing the significance of R^2 in regression, Likelihood-Ratio (LR) tests testing the significance of pseudo R^2 in logistic regression, and component loadings in PCA. For some of these statistics, pooling procedures have been proposed, for example, F tests in ANOVA (Rubin, 1987; Van Ginkel & Kroonenberg, 2014a), F tests for R^2 in regression (Rubin, 1987; Van Ginkel, in press), LR tests in logistic regression (Rubin, 1987), and component loadings in PCA (Van Ginkel & Kiers, 2011; Van Ginkel & Kroonenberg, 2014b). However, these are not implemented in SPSS.

Fortunately, the pooling of some of these statistics can be done using other software packages, or additional tools in SPSS. For example, F tests in ANOVA and regression can be pooled using the MIANALYZE procedure in SAS 9.4 (2013), the XT MIXED procedure in Stata 14.0 (StataCorp, 2015), and an SPSS macro by Van Ginkel (2016). LR tests in logistic regression can be pooled using the `mice` procedure in R (Van Buuren & Groothuis-Oudshoorn, 2011). For component loadings in PCA, the 'shapes' package in R (Dryden & Mardia, 2016) may be used.

However, these procedures are less user-friendly than the automatic pooling by SPSS. Additionally, the paper promoting the SPSS macro for pooling the results of ANOVA (Van Ginkel & Kroonenberg, 2014a) has been misquoted by many authors, suggesting that combination rules for ANOVA are either not available or too complicated. Van Ginkel and Kroonenberg stated that the pooling of ANOVA results is not available in SPSS and that therefore an SPSS macro

must be used. They admitted that the use of this macro is quite involved, especially for repeated measures ANOVA, and that therefore, implementation in future releases of SPSS is desirable (p. 89). However, many authors citing this paper state that according to Van Ginkel and Kroonenberg (2014a) combination rules for ANOVA are not available in SPSS, ignoring the fact that the very paper being referenced provides a solution in the form of an SPSS macro (see Van Ginkel & Kroonenberg, 2015, for an example). Other authors focus on the statement that the procedure is quite involved, using that as a reason not to use multiple imputation at all. Such quotations perpetuate the misconception that when doing multiple imputation you will end up with several results of your statistical analysis.

Finally, for some analyses, such as MANOVA and standardized regression coefficients, combination rules have not been developed at all, at least to our knowledge.

What should we do when combination rules are not available?

Currently, as the multiple-imputation framework is not complete, continued effort is required on developing and implementing combination techniques. The question is what to do when the MCAR assumption has been violated such that discarding incomplete cases may result in serious bias, or the percentage of missing data is so large that discarding incomplete cases results in serious loss of power (or both), and pooling methods are not readily available for the preferred analysis method. First, look for a statistical software package that has more options for pooling results than the one you are currently using. If necessary, support may be provided by a statistician.

Second, one could consider using an ad hoc method for pooling the specific statistic. For example, one could simply pool standardized coefficients in regression by averaging the values across multiply imputed data sets, or giving a range of this statistic across imputed data sets. Although such ad hoc methods may have no theoretical justification, the question is to what extent this is harmful. Standardized regression coefficients are mainly used as measures of effect size in regression analysis. Even without a theoretical justification, these ad hoc solutions will still give you a rough but reasonable indication of which variables make a large contribution to the prediction of the outcome variable, and which variables do not. What is essential is that when you are forced to report an ad hoc solution for pooling a statistic, you are transparent in that this procedure is used because of a lack of a better alternative.

Conclusions

In this article a number of misconceptions about multiple imputation that we have frequently heard from applied researchers were discussed and rebutted. It was argued that from a theoretical point of view, multiple imputation is always to be preferred over listwise and pairwise deletion, and that reasons of researchers to prefer listwise deletion are based on misunderstandings about multiple imputation.

The remaining question is whether there are any reasons left for not using multiple imputation after all. Yes, but those reasons are practical ones. For example, suppose that there are very few missing values and the statistical analysis of interest is one for which pooled results are lacking. In that case, the benefits of multiple imputation may not outweigh the costs. Furthermore, some statistical analyses already have a built-in method for handling missing data. Examples are item response theory (Birnbaum, 1968; Masters, 1982; Rasch, 1960; Samejima, 1969), latent class analysis (Goodman, 1974; Lazarsfeld, 1950a, 1950b), or structural equation modeling (Jöreskog, 1969, 1977). All of these statistical techniques rely on a method called *full information maximum likelihood* (FIML). This method estimates the statistical model of interest based on the observed data, without deleting respondents with missing data. A built-in method for dealing with missing data in PCA is *missing data passive* (MDP; Takane & Oshima-Takane, 2003). This method is available in SPSS 25.0, in the procedure CATPCA (Meulman, Heiser, & SPSS, 2015). Although a disadvantage of both FIML and MDP is that they can usually only handle MAR mechanisms in which the missing data depend on variables that are included in the targeted statistical model (and ignore variables that have no part in the model), they are still good alternatives to multiple imputation when one of the previously mentioned statistical analyses is the analysis of interest. However, when FIML and MDP are not possible and you experience trouble with the multiple-imputation process, do not resort to listwise deletion, but contact a statistician first.

Finally, only the most frequently heard misconceptions about multiple imputation have been rebutted in this article. We do not know what other possible misconceptions may be held by applied researchers that we have not heard about. However, by means of this article we hope to have made clear that in most cases, multiple imputation is to be preferred over listwise deletion and single-imputation methods, as also argued by Van Buuren (2012, p. 48) and by Schafer and Graham (2002).

References

- Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48*(1), 5–37. doi:10.1016/j.jsp.2009.10.001
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores* (chapters 17–20). Reading, MA: Addison Wesley.
- Dryden, I. L., & Mardia, K. V. (2016). *Statistical shape analysis: with applications in R* (2nd ed.). New York: Wiley.
- Eekhout, I., de Boer, M. R., Twisk, J. W., de Vet, H. C., & Heymans, M. W. (2012). Missing data: A systematic review of how they are reported and handled. *Epidemiology, 23*(5), 729–732. doi:10.1097/EDE.0b013e3182576cdb
- Fay, R. E. (1986). Causal models for patterns of nonresponse. *Journal of the American Statistical Association, 81*(394), 354–365. doi:10.1080/01621459.1986.1047827
- Galimard, J. E., Chevret, S., Protopopescu, C., & Resche-Rigon, M. (2016). A multiple imputation approach for MNAR mechanisms

- compatible with Heckman's model. *Statistics in Medicine*, 35(17), 2907–2920. doi:10.1002/sim.6902
- Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2), 215–231.
- Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, 8(3), 206–213. doi:10.1007/s11121-007-0070-9
- Jöreskog, K. G. (1977). Structural equation models in the social sciences: specification estimation and testing. In P.R. Krishnaiah (Ed.), *Applications of statistics* (pp. 265–287). Amsterdam: North Holland.
- Lazarsfeld, P. F. (1950a). The interpretation and mathematical foundation of latent structure analysis. S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, and J. A. Clausen (Eds.), *Measurement and prediction* (pp. 413–472). Princeton, NJ: Princeton University Press.
- Lazarsfeld, P. F. (1950b). The logical and mathematical foundation of latent structure analysis. S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, and J. A. Clausen (Eds.), *Measurement and prediction* (pp. 361–412). Princeton, NJ: Princeton University Press.
- 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:10.2307/2290157.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York, NY: Wiley.
- Little, R. J. A., & Schenker, N. (1995). Missing data. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 39–75). New York, NY: Plenum Press.
- Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174.
- Meulman, J. J., & Heiser, W. J. & SPSS (2015). *IBM SPSS Categories 23.0*. Chicago, IL: SPSS.
- Moustaki, I., & Knott, M. (2000). Weighting for item non-response in attitude scales by using latent variable models with covariates. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(3), 445–459. doi:10.1111/1467-985X.00177
- R Development Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: The R Foundation for Statistical Computing.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.
- Rombach, I., Rivero-Arias, O., Gray, A. M., Jenkinson, C., & Burke, O. (2016). The current practice of handling and reporting missing outcome data in eight widely used PROMs in RCT publications: A review of the current literature. *Quality of Life Research*, 25(7), 1613–1623. doi:10.1007/s11136-015-1206-1
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592. doi:10.1093/biomet/63.3.581
- Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics*, 4(1), 87–94. doi:10.1080/07350015.1986.10509497
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, 34(S1), 1.
- SAS Institute Inc. (2013). *SAS® 9.4 [Computer software]*. Cary, NC: SAS Institute Inc.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London, UK: Chapman & Hall: London.
- Schafer, J. L. (1998). NORM: Version 2.02 for Windows 95/98/NT. Retrieved January 3, 2007, from <http://www.stat.psu.edu/~jls/misoftwa.html>.
- 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
- S-Plus 6 for Windows [Computer software]. (2001). Seattle, WA: Insightful Corporation.
- SPSS, Inc. (2009). *PASW 17.0 for Windows [Computer software]*. Chicago, IL: Author.
- SPSS, Inc. (2017). *IBM SPSS 25.0 for Windows [Computer software]*. Chicago, IL: Author.
- StataCorp. (2015). *Stata Statistical Software: Release 14 [Computer software]*. College Station, TX: StataCorp LP.
- Takane, Y., & Oshima-Takane, Y. (2003). Relationships between two methods for dealing with missing data in principal component analysis. *Behaviormetrika*, 30(2), 145–154.
- Vach, W. (1994). *Logistic regression with missing values in the covariates*. Berlin, Germany: Springer-Verlag.
- Van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: Chapman & Hall/CRC Press.
- Van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18(6), 681–694.
- Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12), 1049–1064. doi:10.1080/10629360600810434
- Van Buuren, S., & Groothuis-Oudshoorn, C. G. M. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45, 1–67. doi:10.18637/jss.v045.i03
- Van Ginkel, J. R. (2016). MI-MUL2.pdf [Software manual]. Retrieved October 4, 2018 from <https://www.universiteitleiden.nl/en/staffmembers/joost-van-ginkel>
- Van Ginkel, J. R. (in press). *F-tests and estimates for R² for multiple regression in multiply imputed datasets: a cautionary note on earlier findings, and alternative solutions*. *Multivariate Behavioral Research*. Advance online publication. doi:10.1080/00273171.2018.1540967.
- Van Ginkel, J. R., & Kiers, H. A. L. (2011). Constructing bootstrap confidence intervals for principal component loadings in the presence of missing data: A multiple-imputation approach. *British Journal of Mathematical and Statistical Psychology*, 64(3), 498–515. doi:10.1111/j.2044-8317.2010.02006.x
- Van Ginkel, J. R., & Kroonenberg, P. M. (2014a). Analysis of variance of multiply imputed data. *Multivariate Behavioral Research*, 49(1), 78–91. doi:10.1080/00273171.2013.855890
- Van Ginkel, J. R., & Kroonenberg, P. M. (2014b). Using generalized procrustes analysis for multiple imputation in Principal component analysis. *Journal of Classification*, 31(2), 242–261. doi:10.1007/s00357-014-9154-y
- Van Ginkel, J. R., & Kroonenberg, P. M. (2015). Comment on article entitled: “PLAY project home consultation intervention program for young children with autism spectrum disorders: A randomized controlled trial”. *Journal of Behavioral and Developmental Psychiatry*, 36, 225. doi:10.1097/DBP.0000000000000137
- Van Ginkel, J. R., Sijtsma, K., Van der Ark, L. A., & Vermunt, J. K. (2010). Incidence of missing item scores in personality measurement, and simple item-score imputation. *Methodology*, 6(1), 17–30. doi:10.1027/1614-2241/a000003
- Von Hippel, P. T. (2008). Regression with missing Ys: An improved strategy for analyzing multiply imputed data. *Sociological Methodology*, 37, 83–117. doi:10.1111/j.1467-9531.2007.00180.x
- White, I. R., & Carlin, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*, 29(28), 2920–2931. doi:10.1002/sim.3944
- Yuan, Y. C. (2000). Multiple imputation for missing data: Concepts and new development. *Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference (Paper, No. 267)*. Cary, NC: SAS Institute. Retrieved January 3, 2007, from <http://www.ats.ucla.edu/stat/sas/library/multipleimputation.pdf>
- Yuan, Y. C. (2011). Multiple imputation using SAS Software. *Journal of Statistical Software*, 45, 1–25. doi:10.18637/jss.v045.i06