Missing values in exploratory factor analysis: A 'best of all possible worlds' approach to imputation for incomplete survey data

Bas Bosma[1] and Arjen van Witteloostuijn[1]

[1]Affiliation not available

January 14, 2021

Abstract

In the social sciences, multi-item scales and factor analyses are standard tools in survey research. In the social sciences, such tools are omnipresent, as are, unavoidably, nonresponses. The question is how to handle missing values when an exploratory factor analysis is intended. Deletion methods will result in — oftentimes substantial and damaging — reduction of power. The seemingly obvious alternative is to keep all respondents and apply imputation to missing values. However, with the true factor structure unknown, theoretically recommendable multiple imputation methods cannot simply be applied. Instead of declaring an entire method unsuitable for exploratory analysis, we propose an approach that keeps the relevant aspects of various methods and combines these by sacrificing less relevant aspects. Doing so, we keep understanding and ease of use in mind, aiming for an approach that is more rigorous and 'correct' than what is commonly used in practice, whilst still being straightforward enough to actually be used.

Missing values in exploratory factor analysis

A 'best of all possible worlds' approach to imputation for incomplete survey data

Bas Bosma[*,1] and Arjen van Witteloostuijn[1,2]

[1] VU Amsterdam, School of Business and Economics, the Netherlands

[2] University of Antwerp / Antwerp Management School, Belgium

August 2021

[*] Corresponding author: De Boelelaan 1105, 1041 HV Amsterdam, the Netherlands. Email: b.bosma@vu.nl.

Missing values in exploratory factor analysis

A ‘best of all possible worlds’ approach to imputation for incomplete survey data

Abstract. In the social sciences, multi-item scales and factor analyses are standard tools in survey research. These tools are omnipresent, as are, unavoidably, nonresponses. The question is how to handle missing values when an exploratory factor analysis is intended. Deletion methods will result in — oftentimes substantial and damaging — reduction of power. The seemingly obvious alternative is to keep all respondents and apply imputation to missing values. However, with the true factor structure unknown, theoretically recommendable multiple imputation methods cannot simply be applied. Instead of declaring an entire method unsuitable for exploratory analysis, we propose an approach that keeps the relevant aspects of various methods and combines these by sacrificing less relevant aspects. Doing so, we keep understanding and ease of use in mind, aiming for an approach that is more rigorous and ‘correct’ than what is commonly used in practice, whilst still being straightforward enough to be used.

Keywords. Imputation, missing data, survey, exploratory factor analysis.

1 Introduction

The process of theory building iteratively cycles through two major stages, which Christensen and Carlile (2009) refer to as the descriptive and the normative. In the quantitative Popperian tradition, this translates into the descriptive stage providing us with constructs, frameworks, and models through observation, categorization, and association, which eventually result in hypotheses and their subsequent testing in the normative stage. Formulating these hypotheses, researchers focus on the expected relationships between constructs and how the (differences between) attributes of frameworks correlate with the outcome variables of interest.

In this descriptive stage of theory building, applied social scientists often use surveys, (partly) consisting of Likert scale items, as measurement instrument. Subsequently studying the data thus collected for possible underlying structure is most often performed by means of exploratory factor analysis (EFA) (see, e.g., Auerswald & Moshagen, 2019; Bandalos & Boehm-Kaufman, 2009; Fabrigar et al., 1999). Other salient characteristics of the descriptive stage of theory building in applied social science tend to be small sample sizes and missing data. In what follows, we will call this specific combination of characteristics (small sample of survey data plagued with missing values on which the researcher wants to perform EFA) the *exploratory context* of applied social science research, and provide practical guidelines in dealing with its challenges.[1]

Although all four characteristics of the exploratory context are extensively researched, much less attention has been given to their interdependencies and the challenges these present

[1] In practice, the descriptive and normative stages are often combined in a single study. Indeed, in theory-testing social science studies, we observe quite frequently the combination of a small sample size, missing values, and an exploratory element in measurement, when the to-be-tested theory includes a new construct for which a validated measure is not available in the extant literature. For the sake of the argument, we simply refer to 'the exploratory context,' which captures any study jointly involving small sample sizes, missing values, and the use of EFA on survey data.

us with jointly. Without even trying to be exhaustive, there exist extensive literatures on survey data (e.g., Agresti, 2007), EFA and its variants (e.g., Jöreskog, 2007), required sample size (e.g., Price, 2017), and the treatment of missing data (e.g., Allison, 2002). Within these literatures, one can also find studies combining two characteristics – for example, recommendations on how to treat missing data specifically for survey data (e.g., Little, 1988; Rubin, 1987), or on (the determinants of) required sample size when performing EFA (MacCallum et al., 1999). However, far less common are studies addressing the small sample performance of missing data treatments. Whilst Roth et al. (1999, p. 230), more than two decades ago, already noted this lacuna in stating that "We strongly urge future research on varying sample sizes. [… as] we wonder if the quality of estimates of missing data would change for smaller *N*s," and similar calls to action have appeared in many studies since, this deficiency largely persists.

Missing data treatments are commonly compared based on converged performance in large(r) samples, leaving out the potential differential performance under the typical small sample conditions of the exploratory context.[2] Of the relatively few studies that do take the small sample conditions of missing data treatments into account, the predominant focus is on statistical concerns mostly relevant to theory testing, like bias and efficiency of estimated standard errors (e.g., Graham & Schafer, 1999; von Hippel, 2016) or the proper degrees of freedom that should be taken into account (e.g., Barnard & Rubin, 1999; Reiter, 2007). However, as McNeish (2017, p. 638) rightly points out, "EFA is rather unique in terms of analytic goals because it is one of the few prominent statistical methods where hypothesis testing is often not a central interest." The applied social scientist in the descriptive stage of

[2] Newman (2014, p. 397) more recently observed that "More research would still be useful on a wide variety of imaginable boundary conditions under which the various missing data techniques might have different degrees of relative performance (e.g., […] small sample size conditions […])."

theory building is mostly interested in an informative and interpretable underlying structure that can guide and incite further research. Using EFA, the focus will be on extracting the proper number of factors and getting a feel for the strength of factor loadings far more than any formal testing. So, although the exploratory context is ubiquitous in practice and important in theory building, little is known about how to approach the key difficulty of missing data under its practical conditions of occurrence. As a result, we argue, current practices regarding the handling of missing values in the exploratory context are clearly suboptimal.

The current article aims to address this deficiency. It will do so in a manner similar to Newman (2014), being a companion piece to the already existing high-quality reviews and recommendations of the separate characteristics of the exploratory context. It will add to current literature by specifically addressing the challenges stemming from the interdependencies of the characteristics of the exploratory context. The intention is not to be as statistically and methodologically exact as possible, but to offer a practical 'best of all possible worlds' approach that combines the strengths of various best practices into a way of working that is well within reach of applied social scientists. We suggest a compromise that takes current research practices, qualitative considerations, and statistical theory into account, and that nevertheless results in clear improvements in terms of the reliability and accuracy of research outcomes.

In Section 2, the relevant aspects of the exploratory context are introduced. The ubiquity of small sample conditions will be illustrated, and the mechanisms of missing data and their most common treatments discussed. The challenges pertinent to the exploratory context will be explained, and our 'best of all possible worlds' approach proposed. Thereafter, Section 3 will set out the design and conditions of the simulation study by which our proposal and several likely alternatives are evaluated, followed by a discussion of the results in Section 4. We discuss and conclude in the final Section 5.

2 The exploratory context

Collecting primary data using survey methods is a very common practice for applied social scientists, both in the descriptive and normative stages of theory building. The rule of thumb in terms of required sample size for psychometric purposes seems to have become that 200 suffices for most descriptive analyses, whilst reaching 500 or more respondents is strongly recommended (MacCallum et al., 1999).[3] With an average response rate[4] in organizational research of 52% (Anseel et al., 2010), collecting a sample of sufficient size can be challenging. On top of challenges related to the response rate, at least half of the respondents to surveys do not answer one or more questions (cf. King et al., 2001). How these missing data are dealt with is seen to greatly influence the resulting sample size, and thereby the accuracy and credibility of research outcomes.

Reviewing over 800 articles in political science journals, King et al. (2001) find 94% of the studies dealing with missing data to do so by listwise deletion, possibly after partially replacing missing values by 'educated guesses.' The loss of information in doing so is, on average, about a third of the information collected. And although many studies have shown the harmful consequences of suboptimal missing data treatment (see, e.g., Newman, 2014; Peugh & Enders, 2004), ignoring everything but the full respondents is still popular in current research practice.

Taking the prevalence of listwise deletion into account, it is hardly surprising that many studies that include a descriptive component, published in top-ranking outlets, are based on sample sizes smaller than recommended. Russell (2002), for example, investigated all articles appearing in the *Personality and Social Psychology Bulletin* over a three-year period, and

[3] Comrey and Lee's (1992, p. 217) guideline is more fine-grained than the rule of thumb: "50 – very poor; 100 – poor; 200 – fair; 300 – good; 500 – very good; ≥ 1000 – excellent."

[4] Defined as (#partial respondents + #full respondents) / #contacted.

found that 27% of all articles explore underlying structures in survey data, of which 39% on sample sizes of 100 or smaller and another 23% on sample sizes between 100 and 200. Fabrigar et al. (1999), in their review of contributions to the *Journal of Applied Psychology* and *Journal of Personality and Social Psychology*, comparably found that 50% of the studies exploring underlying structure in survey data had sample sizes below 200, whilst Conway and Huffcutt (2003) report very similar statistics for organizational research outlets.

As Newman (2014, p. 384) lucidly observes, "The prevalence of listwise deletion is attested by phrases of the following sort, which are regularly found in the Method sections of our top journals: 'Out of 542 surveys returned, 378 provided usable data and were included in the analysis.' The problem with this sort of statement is that it is inherently false. *All* of the respondents who provided data provided 'usable data,' but the researcher chose to throw away some of this precious information. Indeed, listwise deletion compounds the problem of sample nonresponse, by adding to it the extra problem that the researcher herself or himself creates additional missing data by discarding the partial respondents".

So, although sample size will remain a challenge in applied social science research, missing data treatment is arguably the most promising avenue of improvement.

2.1 Missing data treatments

We can clarify the missing data challenges and treatments in their different forms and shapes by introducing some notation. Let $\boldsymbol{D}$ denote the data matrix. For a typical survey, this matrix consists of a row for each respondent and a column for each of the items. In the presence of missing data, $\boldsymbol{D}$ has an observed and a missing part, denoted by $\boldsymbol{D} = \{\boldsymbol{D}^{obs}, \boldsymbol{D}^{mis}\}$. The latter

concerns values that 'exist' (in a specific metaphysical sense), but are not observed.[5] Also, let $\boldsymbol{R}$ be a matrix of identical dimensions to $\boldsymbol{D}$ containing response indicators. That is, r_{ij} is equal to 1 if d_{ij} is observed, and equal to 0 if it is missing.

The main missing data mechanisms can now be described in terms of our ability to predict the values of $\boldsymbol{R}$ (King et al., 2001). Missing data is *missing completely at random* (MCAR) if $\boldsymbol{R}$ cannot be predicted any better by making use of information in $\boldsymbol{D}$. More formally, $\boldsymbol{R}$ is independent of $\boldsymbol{D}$: $\mathbb{P}(\boldsymbol{R}|\boldsymbol{D}) = \mathbb{P}(\boldsymbol{R})$. This could, for example, happen if individuals make the decision to answer questions by flipping a coin, or if the missing data results from random errors in the software used. In practice, missing data will almost never be MCAR; rather, not answering a question is often related to some other characteristic in the survey (Graham, 2009).[6]

Missing data is *missing at random* (MAR) if $\boldsymbol{R}$ cannot be predicted any better by making use of information in $\boldsymbol{D}^{mis}$, but might be by using information in $\boldsymbol{D}^{obs}$. That is, $\mathbb{P}(\boldsymbol{R}|\boldsymbol{D}) = \mathbb{P}(\boldsymbol{R}|\boldsymbol{D}^{obs})$. The probability of missing data on a variable is, under MAR, related to other variables on which data is collected, but not to the variable with the missing data itself. If, for example, young respondents do not answer questions on their preferred leadership style, this missing data is not related to the variable with missing data (preferred leadership style),

[5] Whilst some respondents might prefer to not disclose their income, they do have an income (positive, negative, or zero). This is what is meant by 'exist'. If a survey contains specialist questions whilst also being conducted outside of the matching population of specialists, it may be the case that a missing value or a 'don't know' answer does not signal a preference to not reveal the true existing value, but signals the question being inapplicable (Kroh, 2006). Assigning all respondents an opinion on every item (for example, by replacing missing values on questions that might be inapplicable to them) causes an underrepresentation of the less informed, and thereby introduces a bias.

[6] Whether or not missing data is MCAR can be tested for (see, for example, Jamshidian & Jalal, 2010). In practice, MCAR will almost always be rejected in favor of MAR/MNAR. Given its dependence on missing data, the presence or absence of MNAR can never be demonstrated based only on observed data.

but to another variable (age) on which data is collected. Its patterns of missingness can thus be predicted by making use of $\boldsymbol{D}^{obs}$.[7]

Finally, missing data is *missing not at random* (MNAR) if $\boldsymbol{R}$ depends on both the observed and the missing part of $\boldsymbol{D}$. That is, $\mathbb{P}(\boldsymbol{R}|\boldsymbol{D})$ does not further simplify and the probability of missing data on a variable is related, even after controlling for all other variables on which data is collected, to that variable itself. A classic example of MNAR would be high-income individuals refusing to answer questions about income while simultaneously being unable to predict which individuals have high income based on all other data collected (King et al., 2001; Little & Rubin, 2014; Rubin, 1976).

The extensive statistical literature on missing data (see, e.g., Allison, 2003; Little, 1988; Little & Rubin, 2014; Raghunathan et al., 2001) almost universally recommends maximum likelihood[8] (ML) or multiple imputation (MI) as optimal treatment. The consistency of these recommendations over the past five decades and the weight of evidential support notwithstanding, an implementation gap versus current research practices of many applied social scientists persists. To further assist researchers in identifying and applying missing data treatments, and to bridge the implementation gap, excellent reviews (e.g., Enders, 2010; Newman, 2009; Schafer & Graham, 2002) and practical guidelines "offering a set of compromised standards that are midway between current research practice (e.g., in which listwise and pairwise deletion are routinely implemented) and statistical best practice" (Newman, 2014, p. 397) have been made available, but to little avail.

Although the practical guidelines are less strict than the statistical literature, "only recommending state-of-the-art missing data routines (ML and MI) be used in those instances

[7] The prediction requires no causality. If the response pattern in $\boldsymbol{R}$ can be better predicted by making use of the information in $\boldsymbol{D}^{obs}$, in any way whatsoever, the missing data is MAR.

[8] Full Information Maximum Likelihood (FIML) and Expectation Maximization (EM) are often recommended as exponents of the ML treatment of missing data.

when they are likely to make the biggest difference (e.g., when the percentage of respondents who are partial respondents >10%)" (Newman, 2014, p. 397), it is safe to say that the applied social scientist could never go wrong in applying the best practices, and is even recommended to do so in a substantial fraction of the practically relevant situations discussed (cf. Figure 1 in Newman, 2014).

2.2 Guidelines and qualitative considerations

Of the five major categories of missing data treatment, listwise deletion (LD) and single imputation (SI) are consistently rejected as valid alternatives in both the statistical literature (e.g., Rubin, 1987) as well as practical guidelines, on grounds ranging from inefficiency and bias (Newman & Sin, 2009) all the way to ethics (Rosenthal, 1994).[9] The two most important practical guidelines of Newman (2014, p. 373), "use all available data (e.g., do not use listwise deletion)" and "do not use single imputation," leave little room for interpretation and are, if possible, even more apt within the exploratory context. Simply put, if sample size is a challenge even when data would be complete, as tends to be the case in the exploratory context, any further reduction of sample size (e.g., through LD) decreases statistical power quite possibly to the point where valid conclusions can no longer be reached. Besides the reduction in power, disregarding information introduces bias for all practically relevant mechanisms of missing data (cf. Graham, 2009, p. 567).[10] This might also be the case, but for different reasons, when

[9] "In sum, because listwise deletion often leads to extreme levels of inferential error (low power) and missing data bias (over- or underestimation of effect sizes), and because it is based on theoretically and ethically indefensible rationales, it should be avoided outright" (Newman, 2014, p. 384).

[10] As quantified by King et al. (2001), even under the idealized and practically less relevant condition of missing data being MCAR, the efficiency loss of using deletion methods is, on average, one standard deviation. That is, the point estimates reported in many applied social science studies are, on average, about one standard deviation farther away from their true population value due to the common practice of using deletion methods in addressing missing values. Under more realistic conditions concerning missing data mechanisms, the inefficiency will be more severe, and results will also be biased.

applying SI. Imputing, for example, a constant mean (across an item or respondent) downwardly biases the sample estimates of variance and correlation. Again aptly summarized by Newman (2014, p. 385), "the main reason to place a moratorium on single imputation is because multiple imputation has all of the advantages of single imputation, but none of its major drawbacks."

The remaining major categories of missing data treatments are pairwise deletion (PD), ML, and MI. The first of these calculates summary statistics based on all available data for pairs of variables. Within the exploratory context, where the summary statistic will most likely be the correlation matrix, each separate cell within the correlation matrix will thus be calculated using the observed cases for both variables involved.[11] PD introduces bias under practically relevant missing data mechanisms (MAR and MNAR), results in inaccurate standard errors under all missing data mechanisms, presents a single measure of association that nevertheless is based on different subpopulations (each cell of the correlation matrix can be based on a different number of cases), and might result in correlation matrices that are not positive definite (Newman, 2009). As EFA requires positive definite correlation matrices, using PD as missing data treatment might result in non-convergence. Additionally, and from a more qualitative perspective, incorporating the missing data treatment into the process of estimating summary statistics is less transparent and intuitive. The applied social scientist is no longer explicitly partaking in the missing data treatment, but using a black-box estimation technique of which the inner workings and technicalities might very well be beyond her expertise. As this situation negatively affects the adoption of new practices (Greenwood et al., 2019), considerations like these should be taken into account if our collective aim is to close the implementation gap.

[11] As the correlation matrix is a scaled derivative of the covariation matrix, we'll make use of both throughout the text. In the analyses performed, we consistently make use of correlation matrices.

A similar qualitative argument can be made when considering variants of ML as missing data treatment. FIML tries to maximally incorporate information from all observed data by allowing the dimensions of the mean and covariance matrix to vary for individuals (cf. Enders & Bandalos, 2001). It is a direct estimation technique that similarly incorporates the missing data treatment and withdraws it from the conscious considerations of the applied social scientist.[12] In terms of converged performance in large(r) samples, FIML is expected to perform better than PD, and struggle somewhat less with non-convergence. It only introduces bias for MNAR (as all missing data treatments do), and results in accurate standard errors for all missing data mechanisms.

MI is comparable to FIML in terms of converged performance in large(r) samples. It also only introduces bias for MNAR and results in accurate standard errors for all missing data mechanisms. In terms of qualitative considerations, however, MI treats missing data at the level of the data set. It consists of three stages: imputation, analysis, and pooling. In the first stage, m completed data sets are created by predicting each d_{ij} for which $r_{ij} = 0$. The observed d_{ij}, for which $r_{ij} = 1$, are the same in each of the m completed data sets. The researcher thus ends up with $\boldsymbol{D}_{(m)} = \{\boldsymbol{D}^{obs}, \boldsymbol{D}_{(m)}^{mis}\}$, where $\boldsymbol{D}_{(m)}$ are the m completed data sets, and $\boldsymbol{D}_{(m)}^{mis}$ are the m sets of predicted values for the missing data. In the second stage, the intended analysis is performed on each of these completed data sets. Denoting the quantity of interest of the analysis performed (like a matrix of factor loadings, a regression coefficient, or some other statistic) by

[12] Whilst EM is another popular ML variant commonly recommended, we focus on FIML. Both techniques are unbiased under practically relevant missing data mechanisms, but only FIML also has accurate standard errors. We furthermore see no advantages over FIML of earlier similar techniques focused on optimal estimation of the covariance or correlation matrix, like the nonlinear iterative partial least squares (NIPALS) estimation of Wold (1966) or the iterative principal components analysis (IPCA) method, that includes a single imputation of the missing data in its operation, of Kiers (1997). Many of these methods were shown to be prone to overfitting – a problem becoming more severe in the exploratory context. Versions including regularization parameters (e.g., Ilin & Raiko, 2010; McNeish, 2015) can control the problem of overfitting to some degree, but only through an extensive tuning of regularization parameters with which the applied social scientist might very well be unfamiliar and/or uncomfortable.

Q, each completed data set $\boldsymbol{D}_m \in \boldsymbol{D}_{(m)}$ results in an estimate, $\hat{Q}_m = \hat{Q}(\boldsymbol{D}_m)$, and the estimate's sampling variance $\hat{U}_m$. In the third stage, the final inference is arrived at through pooling the m results. The MI point estimate, $\bar{Q}$, is the average of the estimated quantities of interest on each of the m completed data sets. The variance of $\bar{Q}$ takes two sources of uncertainty into account. The first, denoted by $\bar{U}$, is the average of $\hat{U}_m$ over the m completed data sets. This is the average of the sampling variances of each of the estimated quantities of interest *within* each completed data set $\boldsymbol{D}_m$. The second source is the excess variance due to the missing data being imputed instead of observed (the variance *between* completed data sets):

$$\bar{B} = \frac{1}{m-1}\sum_{m}\left(\hat{Q}_m - \bar{Q}\right)^2.$$

The total variance of $\bar{Q}$ can then be expressed as:

$$\bar{T} = \bar{U} + \left(1 + \frac{1}{m}\right)\bar{B},$$

where the variance between completed data sets is multiplied by a factor that corrects for the fact that we have constructed a finite amount of completed data sets (Rubin, 1987).

Being consciously involved in creating $\boldsymbol{D}_{(m)}$, MI closely aligns with the way of working applied social scientists are already familiar with. Furthermore, having multiple completed data sets that could have been observed, and agree on the data that is observed, quite intuitively represent both the presence of missing data and the inherent uncertainty of its treatment.

In all, taking all considerations into account, MI seems to be the preferred missing data treatment for applied social scientists in the descriptive stage of theory building, expected to be closely followed by FIML. PD is expected to perform considerably worse. This is in sharp contrast with current practice, as the most popular current research practices of applied social scientists are LD and SI, which both address missing data at the level of the data set and

consciously create a 'complete data set' on which EFA is subsequently performed. But how would MI work out in a context where EFA is used subsequently?

2.3 MI and EFA

EFA is arguably the most widely used statistical technique in studying potential underlying structure of observed variables without *a priori* justification for a theoretical model (e.g., van der Eijk & Rose, 2015). In general terms, EFA conceptualizes each of the interrelated *manifest* variables observed as composed of a function of more fundamental quantities, called the *systematic part*, and an unrelated error. The systematic part, assumed to be a linear combination of *latent* variables that we do not observe, accounts for the interrelation in the manifest variables, in the sense that partialling the systematic part out by means of regression would remove all manifest partial covariation (Anderson, 2003).

Denoting the manifest variables by a $(p \times 1)$ vector $\boldsymbol{x}$, and the latent variables, or *factors*, by a $(q \times 1)$ vector $\boldsymbol{\phi}$, EFA conceptualizes every manifest variable as:

$$x_i = \lambda_{i1}\phi_1 + \lambda_{i2}\phi_2 + \cdots + \lambda_{iq}\phi_q + u_i, \qquad 1 \leq i \leq p,$$

where u_i is independent of all factors $\phi_1, \phi_2, \ldots, \phi_q$ and all u_j for which $j \neq i$ (Auerswald & Moshagen, 2019).[13] The weights, λ_{ij}, of manifest variable i to factor j are called *factor loadings*. Figure 1 shows a path diagram for a common factor model in which two, possibly correlated, factors each account for the covariation in four manifest variables. The errors only determine item-specific variance, and are independent from both the factors and the errors of other manifest variables.

[13] To be precise, u_i is composed of a *unique factor*, specific to the manifest variable, and true measurement error. However, as we cannot distinguish between these components based on the survey data collected, we jointly refer to them as 'error' and think of the u_i's as the parts of the manifest variables not explained by the systematic part.

[Insert Figure 1 about here]

Collecting all manifest variables, we end up with the following equivalent expression in matrix notation:

$$\boldsymbol{x} = \boldsymbol{\Lambda\phi} + \boldsymbol{u},$$

where the $(p \times q)$ matrix $\boldsymbol{\Lambda}$ contains the factor loadings, and all error terms are included in the $(p \times 1)$ vector $\boldsymbol{u}$. Assuming the error to be distributed with zero mean and diagonal $(p \times p)$ covariance matrix $\boldsymbol{\Psi}$, and the factors to be distributed with zero mean and $(q \times q)$ covariance matrix $\boldsymbol{\Phi}$, the $(p \times p)$ covariance matrix $\boldsymbol{\Sigma}$ of the manifest variables can be written as:[14]

$$\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}) = \boldsymbol{\Lambda\Phi\Lambda}^{\mathrm{T}} + \boldsymbol{\Psi}.$$

Intuitively, EFA thus tries to approximate the observed covariance matrix of the manifest variables, $\boldsymbol{\Sigma}$, using a small set of q factors.[15] It operates on the covariation between survey items, trying to uncover a lower dimensional structure able to reproduce the observed covariation in manifest variables with as little loss of information possible.

In applying MI as missing data treatment when the intended analysis is EFA, the initial pre-processing performed in the first stage poses no challenges. Imputation can be performed with various prediction models (see, e.g., van Buuren, 2007; Schafer, 2002), fine-tuning the missing data treatment to the situation at hand. Whilst certainly beneficial, research has also shown the exact distributional details to contribute less to the soundness of the outcome than MI's general process. Graham and Schafer (1999) and Schafer and Olsen (1998), amongst

[14] More formally, $\mathbb{E}(\boldsymbol{\phi}\boldsymbol{u}^{\mathrm{T}}) = 0, \mathbb{E}(\boldsymbol{u}) = 0, \mathbb{E}(\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}) = \boldsymbol{\Psi}, \mathbb{E}(\boldsymbol{\phi}) = 0,$ and $\mathbb{E}(\boldsymbol{\phi}\boldsymbol{\phi}^{\mathrm{T}}) = \boldsymbol{\Phi}$. The factor structure is then said to 'fit the q-dimensional common factor model' when there exists a diagonal matrix $\boldsymbol{\Psi}$ with nonnegative diagonal entries such that $\boldsymbol{\Sigma} - \boldsymbol{\Psi}$ is positive semidefinite of rank q. The matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{\Phi}$ are determined, in any solution, up to a rotation. In case the factors are furthermore *orthogonal* with unit variance, the approximation simplifies further to $\boldsymbol{\Sigma} = \boldsymbol{\Lambda\Lambda}^{\mathrm{T}} + \boldsymbol{\Psi}$. For more elaborate explanations, see Anderson (2003) or Jöreskog (2007).

[15] Auerswald and Moshagen's (2019, p. 468) statement that "EFA determines the underlying structure using a data-driven approach assuming a common factor model" is technically more accurate. Our focus is on EFA, but some of the statements concern the common factor model generally and are thus broader applicable.

others, have shown prediction models that assume the variables to be jointly multivariate normally distributed, an obvious approximation for survey data, to work almost as well as more sophisticated models specifically taking distributional assumptions for ordinal data into account. The improvements that can be realized versus the research practices of applied social scientists thus are, for the most part, related to the general use of MI as a three-staged process, and only secondarily to the sophistication of the prediction model used to create the completed data sets. Incorporating the inherent uncertainty of missing data treatment all the way to the parameter level, ideally only to be collapsed into a single point estimate at the stage of pooling, is key to (the performance of) MI.

However, for pooling to make sense, the estimated quantities of interest must be comparable. And to be comparable, the completed data sets should be analyzed using a single hypothesis. This hypothesis can be an assumed factor structure, like in confirmatory factor analysis (CFA), or a regression model, or some other structure, but it must be the *same* for all independent analyses on the m completed data sets. Only if this condition is met will the estimated parameters be comparable, and can they be pooled to obtain a single point estimate $\bar{Q}$ and its total variance $\bar{T}$.

EFA, however, has no single hypothesis. It commences without assuming any specific number of factors to be present and without having a clear-cut prior on factor structure. For each of the $\boldsymbol{D}_m$, a different estimated covariance matrix $\widehat{\boldsymbol{\Sigma}}_m$ will result, possibly influencing the estimated quantities of interest and turning the subsequent pooling into comparing apples to oranges. More specifically, even when using the same approach to identify the number of factors that can be extracted, it cannot be guaranteed to give the same outcome for all $\widehat{\boldsymbol{\Sigma}}_m$. Additionally, even if the identified number of factors would be the same, their order might not be. What we consider to be the first factor in the EFA performed on $\widehat{\boldsymbol{\Sigma}}_1$, might very well be the second factor in the EFA performed on $\widehat{\boldsymbol{\Sigma}}_2$. In pooling the factor loading matrices of these

analyses, rather than properly taking the uncertainty of imputation into account, we would be averaging parameters related to different factors into a meaningless result.

The exploratory context thus confronts applied social scientists with a situation in urgent need of missing data treatment, and clearly recommending MI to do so. Due to interdependencies between the defining characteristics of the exploratory context, though, the largest part of the potential improvement, related to implementing MI's three-staged process, seemingly cannot be realized. For the third stage (pooling) to be sensible, the second stage (analysis) needs to be performed based on a single hypothesis. Yet as both the number and order of extracted factors can vary for EFA performed on each of the completed data sets, EFA lacks a single hypothesis and seems to be incompatible with MI.

2.4 Suggested solutions

Given the strong expected benefits of MI, both quantitative and qualitative, and the popularity of EFA, there have been attempts to work around their incompatibility issues. Dray and Josse (2014), for example, suggested to circumvent the problems of not having a single hypothesis by collapsing the completed data sets into one before the second stage of MI. That is, after creating the m completed data sets, these are collapsed into a single completed data set by averaging over m.[16]

Although the missing data is treated at the level of the data set and multiple completed data sets that could have been observed are created, the resulting completed data set after averaging is not very intuitive. The averaged values of $\boldsymbol{D}^{obs}$ remain ordinal, but those of $\boldsymbol{D}^{mis}_{(m)}$ become real-valued, turning the resulting averaged data set into a mix of observed values and values that could not even have been observed. One can furthermore question how well the

[16] Each d_{ij} in the resulting data set thus is the average of the d_{ij} in the m completed data sets.

correlation matrix of this resulting averaged data set incorporates the uncertainty in covariation between items. The applied social scientist can indeed proceed with EFA as she is accustomed to, but the quality of the missing data treatment that collapses the uncertainty almost before the process of MI has started is questionable. Too much of MI might be sacrificed to save (the doing of) EFA.[17]

At the other end of the spectrum, Josse et al. (2011) and Lorenzo-Seva and van Ginkel (2016) suggested to treat the problems of not having a single hypothesis by introducing an additional step into MI's three-staged process aimed at increasing comparability. More specifically, they suggest to perform the first two stages of MI as one would in case of a single hypothesis, then post-process the outcomes of analysis using a rotation that is based on the generalized Procrustes rotation[18] to make the matrices of factor loadings "as similar to one another as possible" (Lorenzo-Seva & van Ginkel, 2016, p. 599), and subsequently collapse the uncertainty in the third stage of pooling.

There are several challenges with this suggestion. One cannot, for example, ascertain if 'as similar to one another as possible' is indeed similar enough to warrant the pooling stage of MI. Furthermore, what is known about the rotation that needs to be performed is described almost exclusively in highly technical outlets that might not be within reach of the applied social scientist. Related to this, existing implementations of the required rotation in software accessible to applied social scientist often still requires coding, and developing some intuition about the correctness of the results requires at least some understanding of the, quite

[17] Other, somewhat similar, suggestions 'solve' the challenges of MI by reverting back to SI. McNeish (2017, p. 641), for example, notes and suggests: "One difficult issue to reconcile with the multiple imputation conditions was how to deal with replications in which the different imputations yielded a different number of factors. To combat this, because standard error estimates are not often a concern with EFA, we restricted the number of imputations to 1 across the multiple imputation conditions. Theoretically, the loading estimates should retain their asymptotically unbiased properties with only a single imputation."

[18] See ten Berge (1977) and Lorenzo-Seva et al. (2002) for technical details.

demanding, underlying mathematics. More importantly, though, the actual doing of EFA has to be standardized to some extent to make this approach practically applicable. As Lorenzo-Seva and van Ginkel (2016, p. 599) already mention: "It must be noted that the decision on how many r factors to extract (one factor for each latent trait) has to be the same for the K copies of data." With precisely the identification of the number of factors seen as (one of) the most important decisions to be made in the doing of EFA, fixing this beforehand can certainly be argued to be too restrictive, being disproportionate for what it makes possible in terms of treating missing data (cf. Auerswald & Moshagen, 2019; van der Eijk & Rose, 2015; McNeish, 2017).

Many applied social scientists will acknowledge the doing of EFA to be 'part art, part science.' Oftentimes, the researcher will be, for example, comparing outcomes in terms of factors to retain using many different criteria in combination with the domain expertise of the researcher herself. These will, most likely, come to different conclusions, and choices must be made along the way. The same goes for the type of rotation to use, possibly taking out cross-loading items in iterations of achieving simple structure, checking the various goodness of fit and reliability statistics whilst making choices, and more. How this doing of EFA interacts with the rotation aimed at for maximal comparability is unknown. Each of the steps normally taken in an EFA process, will have to be conducted on all m completed data sets, and the results can only ever be examined after the additional rotation, of which the effect remain largely unknown, and the subsequent pooling. So, although the applied social scientist can now perform the three-staged process of MI, plus an extra, technically demanding step, too much of the doing of MI is sacrificed to make this possible.

With one suggested solution possibly sacrificing too much of MI to salvage the doing of EFA, and the other, we argue, sacrificing too much of EFA to salvage MI, is a combined solution possible that outperforms both?

2.5 Best of all possible worlds

Whilst the possible combination of MI and EFA arguably offers the best of both worlds for the applied social scientist, the adaptation of the former to make it applicable in combination with the latter is very challenging. The representation of imputation uncertainty in the form of multiple completed data sets is very intuitive and should, ideally, be incorporated into as many stages of the MI process as possible. The second stage is pivotal in this regard. Staying true to the spirit of MI means EFA is performed on all m completed data sets. However, this means an extra rotation needs to be performed that is both technically demanding, and has unknown effects on the typical doing of EFA and the many choices made along the way.

If, rather, one aims for the doing of EFA, with all its subtleties and interdependent choices along the way, as the applied social scientist is used to and familiar with, this means it should be performed once on an estimated covariance matrix that incorporates and expresses the uncertainty of imputation as properly as can be. Averaging the completed data sets, however, was shown to result in a counterintuitive mix of observed ordinal values and real-valued averaged imputed values that could never have been observed.

The 'best of all possible worlds' approach we suggest, therefore, is the intermediate solution between averaging completed data sets and averaging artificially adapted matrices of factor loadings.[19] That is, for each of the completed data sets $\boldsymbol{D}_m$, we estimate a covariance matrix $\widehat{\boldsymbol{\Sigma}}_m$. Next, and in accordance with MI, we average these covariance matrices to obtain a single point estimate that properly incorporates the uncertainty of imputation:

[19] This approach was already suggested by Nassiri, Lovik, Molenberghs, and Verbeke (2018). Their argument, however, is mainly focused on estimating confidence intervals for the proportion of explained variance by the first k factors, in combination with MI. The suggestion of averaging covariance matrices is offered almost as an aside and has not, as far as we are aware, been taken up by other researchers, neither in follow-up theoretical research nor in applied research.

$$\bar{\boldsymbol{\Sigma}} = \frac{1}{m}\sum_{m} \hat{\boldsymbol{\Sigma}}_m \,.$$

EFA is performed on this point estimate in the manner familiar to applied social scientists, circumventing the issues of comparability as there is no pooling of multiple quantities of interest from separate EFAs.

There also are no unknown effects of and possible interactions between common decision made in the doing of EFA, and a somewhat obscure rotation executed before pooling. Nor are there any problem regarding the ordering of the identified factors. Mature software implementations can be used in generating the completed data sets with a minimum of coding, all well within the quantitative operations applied social scientists are familiar with. The resulting $\boldsymbol{D}_{(m)}$ are all data sets containing values that could have been observed (ordinal values). The EFA, as a contextual conversation between the data collected and the (domain expertise of the) researcher, proceeds as it would when LD or SI would have been used. No observed values are disregarded, no information is thrown away, very few singularity issues emerge in the estimation of the covariance matrices due to the imputation of likely and possible values, and no bias is introduced that cannot be controlled by the number of completed data sets generated. This indeed seems to be the best of all possible worlds.

There are, of course, downsides. Properly executing all stages of MI, without artificial interventions, like one can when performing a CFA, would be better still. Furthermore, working with a single collapsed covariance matrix on which EFA is performed precludes the application of pooling for precision purposes. However, given the (aims of the) exploratory context, the applied social scientist is, at this point in time, mostly interested in descriptive purposes. Possible inferences and formal tests that require precision are more common in later stages of research, possibly after additional data collection based on the outcomes of the exploratory stage. Although these thus are, from a methodological standpoint, shortcomings,

they might be less acute within the descriptive stage of theory building. Put differently, the benefits of our suggested approach are not in any way 'for free,' but do outweigh the problems introduced, offering a way forward that is both more rigorous and more sound, and as intuitive and easy-to-use as applied social scientist are used to.

So, the question remains: how does this 'best of all possible worlds' approach perform under realistic conditions? It is this question we turn to now.

3 Method

The goal of the present study is to assess and compare the three major categories of missing data treatment (PD, ML, and MI) under realistic conditions for the exploratory context. For ML, we focus on FIML; for MI we include both the suggestion by Dray and Josse (2014) as well as our proposal. The former will be denoted as *mean values* (MV), as it collapses the completed data sets into one by directly averaging their values; the latter, our proposed 'best of all possible worlds' approach, we refer to as *mean correlations* (MC), as it collapses the correlation matrices of each of the completed data sets into one by averaging. With extensive simulation, we will compare the performance of these missing data treatments, focusing on the key aspects of the typical exploratory context oftentimes observed in applied social science.

Specifically, controlling the experimental conditions, we create many data sets of which the statistical characteristics are known. For each simulated data set without missing data, variants with given percentages of missing data are constructed. Subsequently, the four missing data treatments under investigation are applied, and EFA is performed. Outcome measures relevant to the exploratory context are examined and the various missing data treatments are critically reviewed.

In what follows, we introduce the experimental conditions in the conceptual order of the simulation process.[20]

3.1 Data generation

Data sets were generated from a two-factor model.[21] In accordance with the literature reviews mentioned in the Introduction, the number of respondents was set to 100, 160, and 240, covering approximately the lower 50% of all comparable descriptive empirical studies. The number of items per factor was set to 4, 6, and 8, in accordance with the typical range of the majority of scales (cf. Fabrigar et al., 1999; Jackson et al., 2009). For each of the resulting nine unique combinations, 200 data sets were generated. To ensure variation in terms of statistical characteristics like spread of central tendencies and skew for the 200 data sets generated for a given number of respondents and items per factor, the sampling frequency for both respondents and items per factor was set to 5.[22]

Assuming the population distributions defined over the factors to be, respectively, a normal distribution with mean 50 and standard deviation 20 (denoted by $\mathcal{N}(50, 20)$), and a bimodal Gaussian mixture that draws with equal probability from $\mathcal{N}(25, 10)$ and $\mathcal{N}(75, 10)$, each of the generated respondents is characterized by a draw from each of the factor population

[20] An accompanying script file is provided as supplementary material that can be used for replication or, by adapting parameter settings, simulating different specifications. The description matches the sections of the script and can be read as a vignette. Furthermore, an additional script file is provided that illustrates the use of MC and can be adapted to fit the analysis needs of the applied social scientist.

[21] The most important settings were varied beyond the ranges mentioned. Sensitivity of the results to these additional specifications is discussed below.

[22] The 1,800 data sets generated using these settings collectively contain 300,000 respondents and 21,600 items. A sampling frequency of 5 means 60,000 respondents and 4,320 items are generated. On average, every generated respondent and item is included in 1 out of every 5 generated data sets.

distributions (cf. van der Eijk & Rose, 2015).[23] Respondent 1 could thus, for example, be characterized by the tuple $(68, 33)$, representing the positions on factors 1 and 2, respectively.

Each item is characterized by a tuple of $n - 1$ boundaries, with n being the number of Likert categories. With n set to 5, item 1 could thus, for example, be characterized by the tuple $(21, 38, 51, 59)$ and item 2 by $(32, 45, 61, 88)$. Together, a respondent and an item define a response. That is, if item 1 would be sampled from the items determined by factor 1, the response of respondent 1 would be 5.[24] Similarly, if item 2 would be sampled from the items determined by factor 2, the response of respondent 1 would be 2. Taken together, a $(60{,}000 \times 4{,}320)$ matrix of five-point ordinal 'true scores' (in terms of classical test theory) results, where each row characterizes a respondent and each column an item (half determined by factor 1, and half by factor 2).

These 'true scores' are then translated into observed data by adding errors in a manner preserving respondents' positions on the factors, taking the smaller probability of larger errors into account, and maintaining both the discreteness and the 'lumpy' character of ordinal scores (van der Eijk & Rose, 2015). More specifically, for each cell in the matrix of 'true scores,' an independent random variable $z \sim \mathcal{N}(0, 1)$ is drawn. The 'true score' is then corrected upwards or downwards according to the sign of z, taking the following cut-offs: no adjustment for $|z| \leq 1.2$, adjustment of one for $1.2 < |z| \leq 2.2$, adjustment of two for $2.2 < |z| \leq 3$, and an

[23] In actual applied social science research, the population distribution of factors is, of course, unknown and unknowable. All the researcher has access to are the observed responses to survey items (including measurement error, missing values, and other deviations).

[24] Respondents characterized by positions on factor 1 smaller than 21 would answer 1, with positions of 21 or larger but smaller than 38 would answer 2, *et cetera*. Given that respondent 1 is characterized by a position of 68 on factor 1, her answer to item 1 would be 5 (68 > 59).

adjustment of three for $3 < |z|$ (all truncated to remain within the five-point response categories).[25]

Finally, each of the 1,800 data sets without missing data is generated by randomly sampling rows and columns. A data set with 100 respondents and 4 items per factor would thus sample 100 rows (out of the 60,000), 4 columns out of the 2,160 determined by factor 1, and another 4 out of the 2,160 determined by factor 2. We refer to the resulting 1,800 data sets without missing data as the *non-missing data* (NMD), and assign a unique identifier to all.

3.2 Missing data

For each missing data percentage p (10%, 20%, and 40%), another 1,800 data sets with MAR missing data are created by censoring.[26] More specifically, for each data set in the NMD, half the columns are randomly assigned to be *controls* and matched to a column in which MAR missing values will be created. For each of these pairs of columns, the rows in the control column containing the $2p$ smallest values are identified, and missing values are created in these rows of the matched column. A (100×8) data set in which 10% MAR missing data is created thus has 20% missing values in four of its columns based on the values of the four other columns.

[25] In this manner, slightly less than one out of four 'true scores' will be adjusted. Furthermore, the resulting response data is realistic in that it resembles response distributions observed in actual surveys and demonstrates the often-ignored fact that the distribution of observed responses in no way indicates the population distribution of underlying factors. See Figures 2 and 3 in van der Eijk and Rose (2015) for a more detailed analysis and discussion.

[26] See Santos et al. (2019) and Rockel (2020) for more details. We focus on the more practically relevant MAR over the stylized MCAR. Whilst we recognize that nearly all missing data occurring in practice is somewhere in-between MAR and MNAR, our aim to suggest potential improvement of missing data treatments justifies the focus on MAR as MNAR requires the collection of additional later that is typical of later stages in the theory building process.

3.3 Imputation

To each of the 5,400 data sets with missing data, the four treatments (PD, FIML, MV, and MC) are applied and correlation matrices are estimated (21,600 in total).[27] For PD and FIML, the treatment is directly incorporated into the estimation.[28] For MV and MC, the missing data is treated at the level of the data set. More specifically, for each of the 10,800 data sets of MV and MC, eight completed data sets $\boldsymbol{D}_m$ are created using a chained implementation of random forests (Breiman, 2001; van Buuren & Groothuis-Oudshoorn, 2011; Stekhoven & Bühlmann, 2012). An extra predictive mean matching step (using 10 candidates) is included to ensure ordinal predictions that are more attuned in terms of distributional properties. Each variable with missing data is imputed by predicted values from 100 random forests grown using all other observed data as covariables.[29]

Recapitulating, for each of the 1,800 data sets with unique identifier in the NMD, three variants are created with different percentages of missing data. To each identifier we can thus link four data sets that differ for $\boldsymbol{D}_{mis}$, but exactly agree on the observed d_{ij} they happen to share. To each of these latter 5,400 data sets, four missing data treatments are applied that result in an estimated correlation matrix. The 7,200 data sets generated thus result in 23,400

[27] As ordinal scores can only be compared in terms of their ordering, covariation and correlation should be based on Spearman's rank-order correlation, which can be understood as a nonparametric version of the more familiar Pearson product-moment correlation. Spearman's rank-order correlation determines the strength and direction of monotonic relationships (not just linear). It can be used for variables measured on interval or ratio scales that violate (one of) the important, but often not explicitly checked, underlying assumptions of Pearson's product-moment correlation. Ordinal data violates the assumed measurement on a continuous scale, meaning that association can only be measured by either Spearman's correlation or the polychoric correlation coefficient (Pearson's alternative). Although believed to be very different measures, they are not. The former can even be shown to be a deterministic transformation of the empirical version of the latter. As the polychoric correlation coefficient *requires* the factors to be continuous, Spearman's rank-order correlation often is the slightly more appropriate measure of association.

[28] We try to solve non-convergence caused by singularity issues through the application of smoothing. More specifically, possible negative eigenvalues are made slightly positive whilst adjusting the other eigenvalues to compensate for the change. Reported convergence is after potential smoothing.

[29] See Mayer (2021) for specific implementation details.

correlation matrices; 1,800 for the NMD, and another 1,800 for each of 12 combinations of percentage missing data and missing data treatment applied.

3.4 Outcome measures

Many studies consider the identification of the number of factors to be the most important issue in EFA (e.g., Auerswald & Moshagen, 2019; van der Eijk & Rose, 2015; Larsen & Warne, 2010; Zwick & Velicer, 1986). Both under- and overestimating the number of factors has severe detrimental effects on research outcomes. The former, for example, causes significant errors in all factor loadings (Wood et al., 1996), whereas the latter can result in non-parsimonious models that include constructs with little to none explanatory value, loadings that split on multiple factors after rotation, and even negative variance estimates (Fava & Velicer, 1996).

There are numerous methods to identify the number of factors. Arguably the most popular in current research practices of applied social scientists are the Kaiser-Guttman criterion (Guttman, 1954), the scree test (Cattell, 1966), and parallel analysis (PA; Horn, 1965), of which the latter has received the most support and is generally considered ‘best practice’(see, e.g., Hayton et al., 2004; Humphreys & Montanelli Jr., 1975; Schmitt, 2011). Given that we apply the same methods of factor identification across all experimental conditions, and that we are interested in the *relative* performance of various missing data treatments, the specific method chosen will not be a decisive factor for this study.

For all 23,400 correlation matrices, we identify the number of factors using PA as well as a model selection perspective based on sequential model tests (SMT). In all instances, we use unweighted least squares (ULS) in estimation. For PA, we determine the eigenvalues of each of the correlation matrices after replacing the diagonal with squared multiple correlations, and compare these with the eigenvalues obtained from 1,000 random samples of the same

dimension consisting of uncorrelated variables. For the largest eigenvalue, we take the 95th percentile of the 1,000 largest eigenvalues from the random samples. For all other eigenvalues, we compare against the mean (Crawford et al., 2010). We refer to this method as PA_SMC.

For SMT, we sequentially estimate factor structures with increasing number of factors. The number of factors identified is the smallest number for which the lower bound of the 90% confidence interval of the root mean square error of approximation (RMSEA) drops below 0.05 (Preacher et al., 2013). The value of 0.05 as cut-off for close fit is based on Browne and Cudeck (1992), and comparable to the 0.06 of Hu and Bentler's (1999) large-scale simulations. We refer to this method as SMT_RMSEA.

The second outcome measure considered is the median factor loading bias. This measure is based on a comparison of each of the 12 combinations defined by the percentage missing data and the missing data treatment applied, to the NMD. To meaningfully interpret this outcome measure within the exploratory context, several steps need to be taken. Firstly, an EFA is estimated for each of the 1,800 correlation matrices of the NMD, each with the number of factors previously identified by PA_SMC as well as SMT_RMSEA. For all resulting estimated factor loading matrices $\widehat{\mathbf{\Lambda}}$, the factor loading largest in absolute value for each of the rows of $\widehat{\mathbf{\Lambda}}$ (i.e., each of the manifest variables) is determined and stored, including its original sign.[30] We do so to limit the impact of the fact that the order of factors is not necessarily the same. Combined with a relatively 'pure' specification (observed variables determined by one factor, independent factors, average factor loading of around 0.75, no enforced cross-loadings or minor factors), we can be confident to capture the dominant loadings for all manifest

[30] In reference to **Error! Reference source not found.**, determine $\max_j |\lambda_{ij}|\,, \forall i$ and store the thus identified loadings without transformation (i.e., as λ_{ij}).

variables, especially for the NMD. Denote the resulting vectors of dominant loadings for the NMD by $\boldsymbol{l}_c$.[31]

In the second step, we perform the same analysis on each of the 1,800 correlation matrices for the 12 combinations of missing data percentage and treatment applied. Denote the resulting vectors by $\boldsymbol{l}_m$. Finally, in step 3, we calculate, for each of the 12 combinations, the absolute difference between that combination's $\boldsymbol{l}_m$ and $\boldsymbol{l}_c$. Of the resulting 1,800 vectors of absolute factor loading differences per combination, we take the median to end up with 1,800 values of the median factor loading bias for each of the 12 combinations.

3.5 Sensitivity analysis

We extensively explored the sensitivity of the results discussed below to changes in the experimental conditions. For one, data was also generated from one-, three-, and four-factor models, also using, in various permutations and in addition to the population distributions over factors already mentioned, a uniform $U(0, 100)$ and a beta $\beta(2, 5)$ distribution scaled to the interval $(0, 100)$ (van der Eijk & Rose, 2015). In terms of percentages missing data, both smaller percentages (e.g., 2%, 5%, and 8%) and percentages in-between the main values of 10%, 20%, and 40% were examined (e.g., 12%, 15%, and 18%).

Furthermore, numbers of items per factor larger than eight were tested to mimic the phase in scale development before the elimination of indicators (Auerswald & Moshagen, 2019). Also, both smaller (e.g., 40, 60, and 80) and larger (e.g., 300, 400, and 500) numbers of respondents were examined. The number of replications was varied between 50 and 250, the

[31] Given the number of items per factor and the fact that data is generated by a two-factor model, the vectors $\boldsymbol{l}_c$ are of dimensions 8, 12, and 16 (600 of each).

number of completed data sets[32] $\boldsymbol{D}_m$ between 5 and 20, the number of candidates considered in the PMM step of imputing predicted values between 5 and 10, and the number of trees grown between 50 and 250. Simulations were also run with seven rather than five-point ordinal scores, and with smaller sets of respondents and items to be sampled from in constructing the data sets.

Although the various sensitivity settings tested all had their effects, they were, generally speaking, minor. None of them necessitated a reconsideration of our main conclusions, as presented below. Where relevant, the outcomes of the sensitivity analyses will be discussed in what follows. All detailed sensitivity analyses are available upon request.

4 Results

4.1 Factor identification

For the experimental conditions introduced in the previous section, Table 1 shows an overview of the main results in terms of factor identification. The top row, based on the 1,800 correlation matrices of the NMD, reveals that both PA_SMC and SMT_RMSEA converged in all instances, and identified the correct number of factors in, respectively, 73% and 58% of these converged instances. For all outcomes reached that did not identify the correct number of factors, the average deviation was, respectively, 0.32 and 0.53 factors. Notwithstanding the challenging conditions of the exploratory context, both methods perform reasonably well, with PA_SMC even correctly identifying a two-factor model for 1,314 of the correlation matrices, being, on average, only 0.32 factor off for the remaining 486.

[Insert Table 1 about here]

[32] For many practically relevant situations, a small number of 3–10 completed data sets suffices when using chained equations (van Buuren, 2007).

PD and FIML can be seen to struggle, even in terms of convergence, for missing data percentages of 10% and higher. On top of that, both treatments far less often identify the correct number of factors in this already smaller set of converged outcomes. And, when identifying an incorrect number of factors, both are significantly further off the mark. This can also be seen, quite dramatically, in the boxplots of Figure 2.

[Insert Figure 2 about here]

In contrast, by treating missing data at the level of the data set through imputation, the convergence of both MV and MC remains high.[33] Adding to that, moreover, the fraction of correctly identified number of factors within these converged outcomes also remains significantly higher.

Combined, these results clearly show the outperformance, in the exploratory context, of both MV and MC when it comes to the important aspect of identifying the number of factors. Compared to one another, MC is the better missing data treatment. Even for high percentages of missing data, MC not only results in high convergence, but correctly identifies the number of factors for a fraction of converged outcomes close to the results for the NMD. And when it does not correctly identify the number of factors, its absolute deviation, on average, also remains close to the results for the NMD.

Although we are mainly interested in the *relative* performance of missing data treatments within the exploratory context, the results of SMT_RMSEA after MC has been applied are worth noticing. Instead of a performance slightly worse than the benchmark and dropping off for higher percentages of missing data, MC's performance is better than the

[33] There might, of course, still be instances of, for example, an entire column of a completed data set having the same value and thus zero variance. For MC, we decided not to work around these issues. For this example, it would mean that one of the eight correlation matrices cannot be estimated. Instead of disregarding the one problematic completed data set and using the other seven (or generating additional instances), as one would probably do in practice, we allow this issue to result in a non-convergence for that entire correlation matrix.

benchmark and increasing when conditions worsen. This effect disappears for higher ratios of respondents to items per factor, but is consistently present in many conditions relevant to the exploratory context.[34]

Corroborating Newman's (2014, p. 384) observation that, under conditions of relatively low percentages of missing data, "pairwise deletion might be similarly as accurate as a state-of-the-art (ML or MI) technique," Table 3 shows that the performance of PD (and FIML) is indeed markedly better for lower percentages of missing data. Importantly, however, now having investigated the specific challenges of the exploratory context, we can extend Newman's insights for the small sample conditions he mentions as topic for further research. Simply put, although low missing data percentages indeed make the performance of PD in terms of identifying the number of factors comparable to MC, the latter is, even in these conditions, the overall better choice. Under conditions more representative of the reality applied social scientists face in the descriptive stage of theory building, where double digit percentages of missing data are endemic, the advantage of MC only becomes larger.

[Insert Table 3 about here]

4.2 Factor loading bias

Table 4 shows the performance in terms of median factor loading bias. First of all, the convergence numbers reported indicate that correlation matrices resulting in convergence for the number of factors identified, nearly always also converge for EFA.

[Insert Table 4 about here]

[34] As can be seen in Table 2 of the supplementary material contained in the appendix, the effect is no longer present for data sets consisting of 500 respondents and 4 items per factor (100 replications).

If we exclude the combinations of missing data percentage and treatment that reach an outcome for only a very small fraction of correlation matrices (PD and FIML for 40% missing data), we can collect the data sets for which all remaining combinations reach an outcome. Denote this collection of data sets by $\mathcal{C}$. Compared to the outcomes reached on the NMD, and with the number of factors identified by PA_SMC, the *median* factor loading, on average and within $\mathcal{C}$, deviates 0.083 for 20% missing data and PD as missing data treatment. With an average factor loading magnitude of 0.75, half of the factor loadings thus approximately lie *outside* of the interval [0.67, 0,83]. For MV (0.041) and MC (0.035) the comparable factor loading bias is less than half that of PD. FIML (0.057) performs better than PD, but significantly worse than MV and, especially, MC.[35]

For lower percentages of missing data, as Table 6 reveals, the performance of PD, this time in terms of factor loading bias, again improves and surpasses that of FIML. And whilst it could be argued that convergence and factor loading bias, for the condition of 2% missing data, qualify PD for the 'similarly as accurate' category, MC again remains the better choice. Combined with the qualitative advantages, and the knowledge that PD's performance will deteriorate for slightly higher percentages of missing data, MC remains 'best of all possible worlds.'

[Insert Table 6 about here]

Percentages of missing data common to the exploratory context, but combined with higher numbers of respondents (300, 400, and 500) see the performance of PD dropping below that of FIML. As Table 7 reveals, the comparative advantage for MC increases for larger number of respondents.

[Insert Table 7 about here]

[35] The same conclusion is reached if the outcome measure is based on SMT_RMSEA as method for factor identification. See Table 5 of the supplementary material contained in the appendix.

Whereas Table 4 (based on 100, 160, and 240 respondents) indicated a factor loading bias for MV that is slightly lower than that of MC for the condition of 40% missing data, Table 7 (based on 300, 400, and 500 respondents) shows MC performing universally best.[36]

5 Discussion and conclusion

Applied social scientists, including organizational researchers, frequently engage in theory building. In doing so, many are confronted with situations typified by four salient characteristics: survey data (partly consisting of Likert items), small sample sizes, missing data, and an initial research aim involving the identification of possible underlying structure by means of EFA. Although this exploratory context is ubiquitous, little is known about how to best deal with the challenges the interdependencies of these salient characteristics jointly present us with.

With missing data treatment arguably the most promising avenue of improvement, we must have a clear understanding of the small sample performance of missing data treatments, as well as their compatibility with, and effect upon, EFA. This clear need, voiced by many over the years, notwithstanding, we have seen very few studies investigating the practicalities of the exploratory context.

In addressing this deficiency, the present study investigated the performance of the major categories of missing data treatments under conditions relevant to the exploratory

[36] Conditions must become quite extreme for MC to not come out on top. For example, analyzing the performance based on 40, 60, and 80 respondents, well below the already problematic sample sizes analyzed for the exploratory context, we see MC struggling for the condition of 40% missing data. As can be seen in Table 8 and Table 9 of the supplementary material contained in the appendix, the convergence, number of factors identified, and factor loading bias, especially for SMT_RMSEA, performs worse than MV. The drop in performance between the conditions having less than 40% missing data, for which MC still performs best, and the condition of 40% missing data seems to indicate we've reached the limit of our harsh treatment of MC in terms of completed data sets not translating into proper correlation matrices.

context. In addition to the most promising missing data treatments already in use, we suggest a ‘best of all possible worlds’ approach that combines the strengths of various best practices into a way of working that is well within reach of applied social scientists. More specifically, our proposed approach uses as much of the expected strong benefits of the most recommended missing data treatment, MI, without compromising the doing of EFA as applied social scientists are accustomed to. It represents both the presence of missing data and the uncertainty of their treatment in an intuitive manner by creating multiple completed data sets. Instead of averaging these, or ignoring all but one of them, the ‘best of all possible worlds’ approach estimates correlation matrices for each of them, and collapses these into a single correlation matrix through averaging. Maintaining the benefits of an intuitive representation, incorporating the uncertainty into as many steps of the MI process as possible, and preventing intransparent operations to force factor loading matrices into a similarity that cannot be properly evaluated, the proposed approach closely aligns with applied social scientists’ familiar way of working and, as such, offers the best possibility of closing the implementation gap between current suboptimal research practices and best practices recommended by the statistical literature.

As our quantitative results indicate, the ‘best of all possible worlds’ approach outperforms the most viable alternatives. Both in terms of identifying the number of factors required to adequately describe the covariation in the observed data, as well as in terms of the bias in factor loadings. Furthermore, it does so consistently within conditions relevant to the exploratory context, but also proves to be remarkably robust outside of these conditions.

It goes without saying that additional research on a wide range of choices and assumptions made is necessary. A better understanding of the boundary conditions of performance across a range of conditions outside of the exploratory context, and the unique strengths under particular conditions of various methods, like those identifying the number of factors, and their interdependencies, being chief among them.

All considering, though, we deem the ‘best of all possible worlds’ approach to have the greatest potential to change a common and, unfortunately, highly suboptimal practice within applied social science for the better. Adopting it will, we believe, pose no great difficulties for applied social scientist, whilst substantially increasing the quality of their work. The reliability, credibility, and accuracy of research outcomes will improve through the avoidance of information loss, without the need to significantly change the way of working or having to use technically demanding and intransparent methods.

References

Agresti, A. (2007). *An Introduction to Categorical Data Analysis* (2 ed.). John Wiley & Sons.

Allison, P. D. (2002). *Missing Data*. Sage.

Allison, P. D. (2003). Missing Data Techniques for Structural Equation Modeling. *Journal of Abnormal Psychology, 112*(4), 545-557. https://doi.org/10.1037/0021-843X.112.4.545

Anderson, T. (2003). *An Introduction to Multivariate Statistical Analysis* (3 ed.). John Wiley & Sons.

Anseel, F., Lievens, F., Schollaert, E., & Choragwicka, B. (2010). Response Rates in Organizational Science, 1995-2008: A Meta-Analytic Review and Guidelines for Survey Researchers. *Journal of Business and Psychology, 25*(3), 335-349. https://doi.org/10.1007/s10869-010-9157-6

Auerswald, M., & Moshagen, M. (2019). How to Determine the Number of Factors to Retain in Exploratory Factor Analysis: A Comparison of Extraction Methods Under Realistic Conditions. *Psychological Methods, 24*(4), 468-491. https://doi.org/10.1037/met0000200

Bandalos, D. L., & Boehm-Kaufman, M. R. (2009). Four Common Misconceptions in Exploratory Factor Analysis. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and Methodological Myths and Urban Legends: Doctrine, Verity and Fable in the Organizational and Social Sciences* (pp. 61-87). Routledge.

Barnard, J., & Rubin, D. B. (1999). Small-Sample Degrees of Freedom with Multiple Imputation. *Biometrika, 86*(4), 948-955. https://doi.org/10.1093/biomet/86.4.948

ten Berge, J. M. (1977). Orthogonal Procrustes Rotation for Two or More Matrices. *Psychometrica, 42*(2), 267-276. https://doi.org/10.1007/BF02294053

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Browne, M. W., & Cudeck, R. (1992). Alternative Ways of Assessing Model Fit *Sociological Methods & Research, 21*(2), 230-258. https://doi.org/10.1177/0049124192021002005

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

Cattell, R. B. (1966). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research, 1*(2), 245-276. https://doi.org/10.1207/s15327906mbr0102_10

Christensen, C. M., & Carlile, P. M. (2009). Course Research: Using the Case Method to Build and Teach Management Theory. *Academy of Management Learning & Education, 8*(2), 240-251. https://doi.org/10.5465/amle.2009.41788846

Conway, J. M., & Huffcutt, A. I. (2003). A Review and Evaluation of Exploratory Factor Analysis Practices in Organizational Research. *Organizational Research Methods, 6*(2), 147-168. https://doi.org/10.1177/1094428103251541

Crawford, A. V., Green, S. B., Levy, R., Lo, W.-J., Scott, L., Svetina, D., & Thompson, M. S. (2010). Evaluation of Parallel Analysis Methods for Determining the Number of Factors. *Educational and Psychological Measurement, 70*(6), 885-901. https://doi.org/10.1177/0013164410379332

Dray, S., & Josse, J. (2014). Principal Component Analysis With Missing Values: A Comparative Survey of Methods. *Plant Ecology, 216*(5), 657-667. https://doi.org/10.1007/s11258-014-0406-z

van der Eijk, C., & Rose, J. (2015). Risky Business: Factor Analysis of Survey Data — Assessing the Probabilty of Incorrect Dimensionalisation. *PLoS ONE, 10*(3), 1-31. https://doi.org/10.1371/journal.pone.0118900

Enders, C. K. (2010). *Applied Missing Data Analysis*. Guilford Press.

Enders, C. K., & Bandalos, D. L. (2001). The Relative Performance of Full Information Maximum Likelihood Esstimation for Missing Data in Structural Equation Models. *Structural Equation Modeling, 8*(3), 430-457. https://doi.org/10.1207/S15328007SEM0803_5

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the Use of Exploratory Factor Analysis in Psychological Research. *Psychological Methods, 4*(3), 272-299. https://doi.org/10.1037/1082-989X.4.3.272

Fava, J. L., & Velicer, W. F. (1996). The Effects of Underextraction in Factor and Component Analyses. *Educational and Psychological Measurement, 56*(6), 907-929. https://doi.org/10.1177/0013164496056006001

Graham, J. W. (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology, 60*, 549-576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Graham, J. W., & Schafer, J. L. (1999). On the Performance of Multiple Imputation for Multivariate Data with Small Sample Size. In R. Hoyle (Ed.), *Statistical Strategies for Small Sample Research* (pp. 1-33). Sage.

Greenwood, B. N., Agarwal, R., Agarwal, R., & Gopal, A. (2019). The Role of Individual and Organizational Expertise in the Adoption of New Practices. *Organization Science, 30*(1), 191-213. https://doi.org/10.1287/orsc.2018.1246

Guttman, L. (1954). Some Necessary Conditions for Common-Factor Analysis. *Psychometrica, 19*(2), 149-161. https://doi.org/10.1007/BF02289162

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor Retention Decisions in Exploratory Factor Analysis: a Tutorial on Parallel Analysis. *Organizational Research Methods, 7*(2), 191-205. https://doi.org/10.1177/1094428104263675

von Hippel, P. T. (2016). New Confidence Intervals and Bias Comparisons Show That Maximum Likelihood Can Beat Multiple Imputation in Small Samples. *Structural Equation Modeling, 23*(3), 422-437. https://doi.org/10.1080/10705511.2015.1047931

Horn, J. L. (1965). A Rationale and Test for the Number of Factors in Factor Analysis. *Psychometrica, 30*(2), 179-185. https://doi.org/10.1007/BF02289447

Hu, L. T., & Bentler, P. M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. *Structural Equation Modeling, 6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Humphreys, L. G., & Montanelli Jr., R. G. (1975). An Investigation of the Parallel Analysis Criterion for Determining the Number of Common Factors. *Multivariate Behavioral Research, 10*(2), 193-205. https://doi.org/10.1207/s15327906mbr1002_5

Ilin, A., & Raiko, T. (2010). Practical Approaches to Principal Component Analysis in the Presence of Missing Values. *Journal of Machine Learning Research, 11*(66), 1957-2000.

Jackson, D. L., Gillaspy Jr., J. A., & Purc-Stephenson, R. (2009). Reporting Practices in Confirmatory Factor Analysis: An Overview and Some Recommendations. *Psychological Methods, 14*(1), 6-23. https://doi.org/10.1037/a0014694

Jamshidian, M., & Jalal, S. (2010). Tests of Homoscedasticity, Normality, and Missing Completely at Random for Incomplete Multivariate Data. *Psychometrica, 75*(4), 649-674. https://doi.org/10.1007/S11336-010-9175-3

Jöreskog, K. G. (2007). Factor Analysis and Its Extensions. In R. Cudeck & R. C. MacCallum (Eds.), *Factor Analysis at 100: Historical Developments and Future Directions*. Lawrence Erlbaum Associates.

Josse, J., Pagès, J., & Husson, F. (2011). Multiple Imputation in Principal Component Analysis. *Advances in Data Analysis and Classification, 5*(3), 231-246. https://doi.org/10.1007/s11634-011-0086-7

Kiers, H. A. (1997). Weighted Least Squares Fitting Using Ordinary Least Squares Algoithms. *Psychometrica, 62*(2), 251-266. https://doi.org/10.1007/BF02295279

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *American Political Science Review, 95*(1), 49-69. https://doi.org/10.1017/S0003055401000235

Kroh, M. (2006). Taking 'Don't Knows' as Valid Responses: A Multiple Complete Random Imputation of Missing Data. *Quality & Quantity, 40*(2), 225-244. https://doi.org/10.1007/s11135-005-5360-3

Larsen, R., & Warne, R. T. (2010). Estimating Confidence Intervals for Eigenvalues in Exploratory Factor Analysis. *Behavior Research Methods, 42*(3), 871-876. https://doi.org/10.3758/BRM.42.3.871

Little, J. R. (1988). Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics, 6*(3), 287-296. https://doi.org/10.1080/07350015.1988.10509663

Little, J. R., & Rubin, D. B. (2014). *Statistical Analysis with Missing Data* (2 ed.). John Wiley & Sons. https://doi.org/10.1002/9781119482260

Lorenzo-Seva, U., & van Ginkel, J. R. (2016). Multiple Imputation of Missing Values in Exploratory Factor Analysis of Multidimensional Scales: Estimating Latent Trait Scores. *Annals of Psychology, 32*(2), 596-608. https://doi.org/10.6018/analesps.32.2.215161

Lorenzo-Seva, U., Kiers, H. A., & ten Berge, J. M. (2002). Techniques for Oblique Factor Rotation of Two or More Loading Matrices to a Mixture of Simple Structure and Optimal Agreement. *British Journal of Mathematical & Statistical Psychology, 55*(2), 337-360. https://doi.org/10.1348/000711002760554624

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample Size in Factor Analysis. *Psychological Methods, 4*(1), 84-99. https://doi.org/10.1037/1082-989X.4.1.84

Mayer, M. (2021). *missRanger: Fast Imputation of Missing Values.* In (Version 2.1.3) https://CRAN.R-project.org/package=missRanger

McNeish, D. (2015). Using Lasso for Predictor Selection and to Assuage Overfitting: A Method Long Overlooked in Behavioral Sciences. *Multivariate Behavioral Research, 50*(5), 471-484. https://doi.org/10.1080/00273171.2015.1036965

McNeish, D. (2017). Exploratory Factor Analysis With Small Samples and Missing Data. *Journal of Personality Assessment, 99*(6), 637-652. https://doi.org/10.1080/00223891.2016.1252382

Nassiri, V., Lovik, A., Molenberghs, G., & Verbeke, G. (2018). On Using Multiple Imputation for Exploratory Factor Analysis of Incomplete Data. *Behavior Research Methods, 50*(2), 501-517. https://doi.org/10.3758/s13428-017-1000-9

Newman, D. A. (2009). Missing Data Techniques and Low Response Rates: The Role of Systematic Nonresponse Parameters. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and Methodological Myths and Urban Legends: Doctrine, Verity, and Fable in the Organizational and Social Sciences*. Routledge.

Newman, D. A. (2014). Missing Data: Five Practical Guidelines. *Organizational Research Methods, 17*(4), 372-411. https://doi.org/10.1177/1094428114548590

Newman, D. A., & Sin, H.-P. (2009). How Do Missing Data Bias Estimates of Within-Group Agreement? Sensitivity of *SD wg*, *CV wg*, *r wg(J)*, *r wg(J)**, and ICC to Systematic

Nonresponse. *Organizational Research Methods, 12*(1), 113-147. https://doi.org/10.1177/1094428106298969

Peugh, J. L., & Enders, C. K. (2004). Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement. *Review of Educational Research, 74*(4), 525-556. https://doi.org/10.3102/00346543074004525

Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the Optimal Number of Factors in Exploratory Factor Analysis: A Model Selection Perspective. *Multivariate Behavioral Research, 48*(1), 28-56. https://doi.org/10.1080/00273171.2012.710386

Price, L. R. (2017). *Psychometric Methods: Theory into Practice*. Guilford Press.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., & Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology, 27*(1), 85-95.

Reiter, J. P. (2007). Small-Sample Degrees of Freedom for Multi-Component Significance Tests with Multiple Imputation for Missing Data. *Biometrika, 94*(2), 502-508. https://doi.org/10.1093/biomet/asm028

Rosenthal, R. (1994). Science and Ethics in Conducting, Analyzing, and Reporting Psychological Research. *Psychological Science, 5*(3), 127-134.

Roth, P. L., Switzer III, F. S., & Switzer, D. M. (1999). Missing Data in Multiple Item Scales: A Monte Carlo Analysis of Missing Data Techniques. *Organizational Research Methods, 2*(3), 211-232. https://doi.org/10.1177/109442819923001

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*. John Wiley & Sons. https://doi.org/10.1002/9780470316696

Russell, D. W. (2002). In Search of Underlying Dimensions: The Use (and Abuse) of Factor Analysis in *Personality and Social Psychology Bulletin*. *Personality and Social Psychology Bulletin, 28*(12), 1629-1646. https://doi.org/10.1177/014616702237645

Schafer, J. L. (2002). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.

Schafer, J. L., & Graham, J. W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods, 7*(2), 147-177. https://doi.org/10.1037/1082-989X.7.2.147

Schafer, J. L., & Olsen, M. K. (1998). Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. *Multivariate Behavioral Research, 33*(4), 545-571. https://doi.org/10.1207/s15327906mbr3304_5

Schmitt, T. A. (2011). Current Methodological Considerations in Exploratory and Confirmatory Factor Analysis. *Journal of Psychoeducational Assessment, 29*(4), 304-321. https://doi.org/10.1177/0734282911406653

Stekhoven, D. J., & Bühlmann, P. (2012). missForest — Non-Parametric Missing Value Imputation for Mixed-Type Data. *Bioinformatics, 28*(1), 112-118. https://doi.org/10.1093/bioinformatics/btr597

Wold, H. (1966). Estimation of Principal Components and Related Models by Iterative Least Squares. In P. R. Krishnaiah (Ed.), *Multivariate Analysis* (pp. 391-420). Academic Press.

Wood, J. M., Tataryn, D. J., & Gorsuch, R. L. (1996). Effects of Under- and Overextraction on Principal Axis Factor Analysis With Varimax Rotation. *Psychological Methods, 1*(4), 354-365. https://doi.org/10.1037/1082-989X.1.4.354

Zwick, W. R., & Velicer, W. F. (1986). Comparison of Five Rules for Determining the Number of Components to Retain. *Psychological Bulletin, 99*(3), 432-442. https://doi.org/10.1037/0033-2909.99.3.432

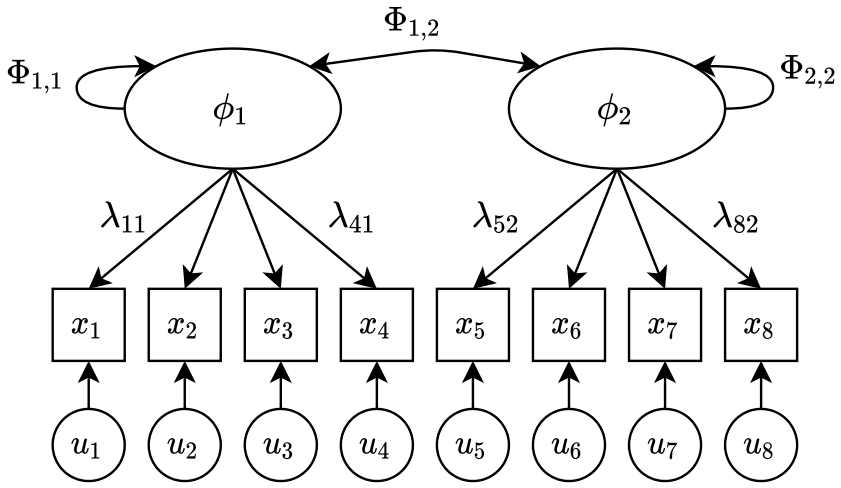

Figure 1: Path diagram for a common factor model with two, possibly correlated, factors (ϕ_1, ϕ_2) determining the covariation in eight observed variables ($x_1, \ldots, x_8$). Error terms ($u_1, \ldots, u_8$) are independent from both the factors and the error terms of other observed variables. Arrows from the factors to the observed variables represent the factor loadings λ_{ij}.

	PA_SMC			SMT_RMSEA		
	Converged	Correct	mean AD	Converged	Correct	mean AD
Complete	1.00	0.73	0.32	1.00	0.58	0.53
PD						
10	0.91	0.32	1.11	0.87	0.13	2.48
20	0.32	0.04	2.07	0.32	0.02	2.72
40	0.00	0.00	2.33	0.00	0.00	1.50
FIML						
10	0.97	0.42	0.86	0.97	0.19	1.69
20	0.70	0.19	1.37	0.69	0.07	2.43
40	0.02	0.00	2.77	0.01	0.00	1.69
MV						
10	1.00	0.66	0.41	1.00	0.46	0.72
20	1.00	0.57	0.56	1.00	0.33	0.99
40	0.98	0.27	1.22	0.98	0.24	1.35
MC						
10	1.00	0.73	0.31	1.00	0.64	0.42
20	1.00	0.70	0.36	1.00	0.72	0.32
40	0.99	0.64	0.44	0.99	0.84	0.16

Table 1: Overview of identified factors for Parallel Analysis using Squared Multiple Correlations (PA_SMC), and Sequential Model Tests based on the lower bound of the 90% confidence interval of the Root Mean Square Error of Approximation (SMT_RMSEA). The leftmost column has a row for the correlation matrices based on NMD ('complete'), and rows for all combinations of percentages missing data (10%, 20%, and 40%) and missing data treatment (PD, FIML, MV, and MC). The numbers in each of the rows are based on 1,800 correlation matrices (23,400 in total). *Converged* is the fraction for which outcomes are reached. *Correct* is the fraction of converged outcomes identifying the correct number of factors. *Mean AD* is the average absolute deviation, in number of factors, of converged outcomes identifying an incorrect number of factors.

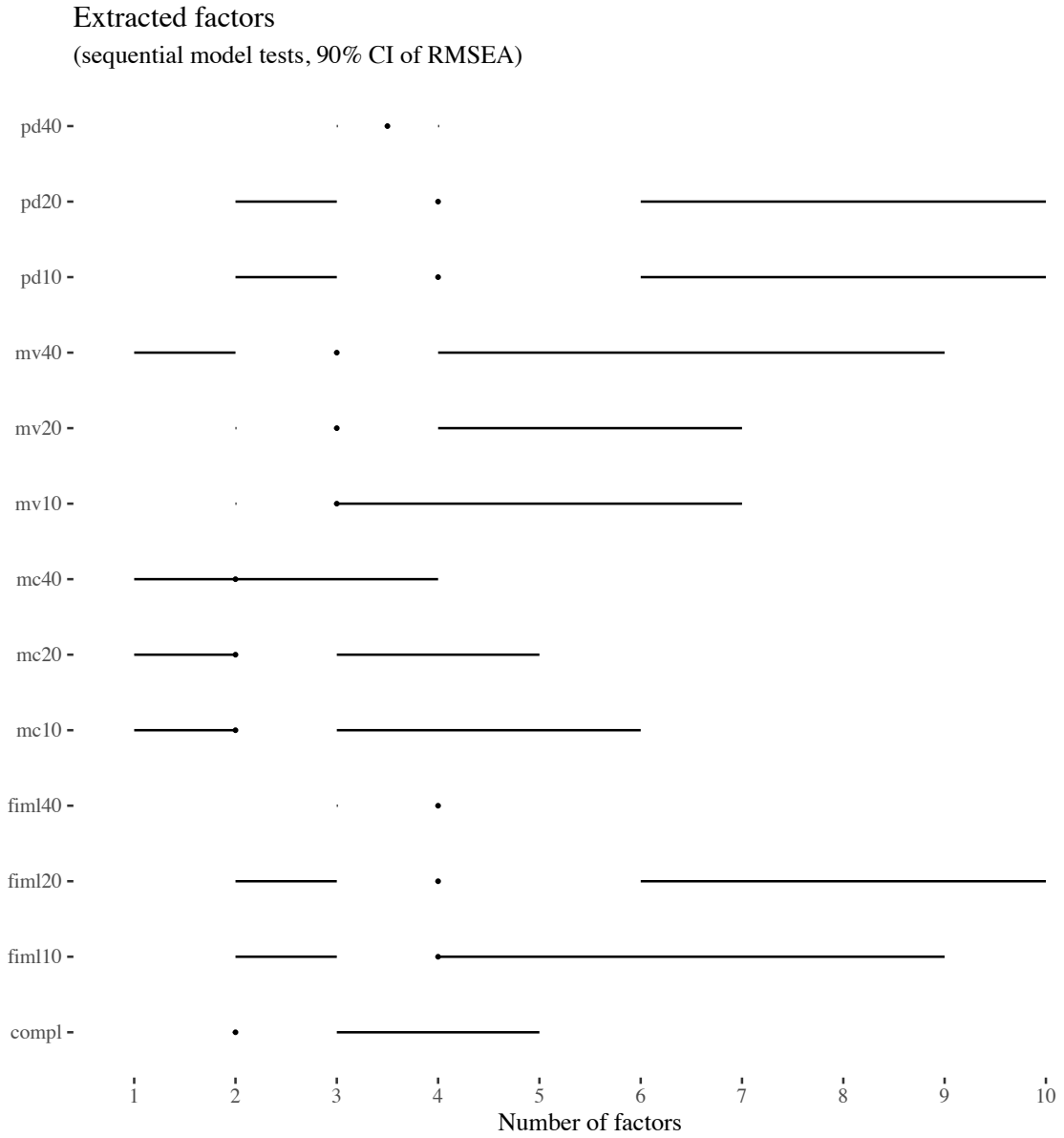

Figure 2: Distributional characteristics of the number of factors identified using SMT_RMSEA. Each line is based on the converged outcomes for 1,800 correlation matrices. The line to the left-hand side of the gap indicates the range of values below the 25th percentile (lower extremes). The gap itself represents the IQR, the range from the 25th to the 75th percentile. The dot represents the median, and the line to the right-hand side indicates the range of values above the 75th percentile (higher extremes). For the data sets without missing data, the 25th percentile and the median both equal two. There were no instances for which one factor was identified. The 75th percentile equals three, and the highest extreme equals five. PD and FIML have median values of four and identify up to ten factors. Convergence for 40% missing data is very low for both treatments. MV has median values of three and deteriorating results for increasing percentages of missing data. MC correctly has median values of two and actually *improves* for increasing percentages of missing data.

	PA_SMC			SMT_RMSEA		
	Converged	Correct	mean AD	Converged	Correct	mean AD
Complete	1.00	0.73	0.34	1.00	0.73	0.31
PD						
10	1.00	0.50	0.72	0.98	0.45	0.77
20	1.00	0.17	1.50	0.78	0.12	1.19
40	0.09	0.00	3.11	0.03	0.00	1.67
FIML						
10	1.00	0.55	0.64	1.00	0.49	0.71
20	1.00	0.41	0.77	0.98	0.33	0.92
40	0.75	0.00	2.45	0.31	0.00	1.94
MV						
10	1.00	0.64	0.48	1.00	0.62	0.48
20	1.00	0.49	0.65	0.98	0.40	0.77
40	1.00	0.13	1.55	0.89	0.07	1.34
MC						
10	1.00	0.68	0.44	1.00	0.71	0.34
20	1.00	0.63	0.48	1.00	0.71	0.38
40	1.00	0.37	0.87	1.00	0.64	0.39

Table 2: Overview of identified factors for PA_SMC and SMT_RMSEA on data sets consisting of 500 respondents and 4 items per factor. The numbers in each of the rows are based on 100 correlation matrices (1,300 in total). *Converged* is the fraction for which outcomes are reached. *Correct* is the fraction of converged outcomes identifying the correct number of factors. *Mean AD* is the average absolute deviation, in number of factors, of converged outcomes identifying an incorrect number of factors.

	PA_SMC			SMT_RMSEA		
	Converged	Correct	mean AD	Converged	Correct	mean AD
Complete	1.00	0.66	0.42	1.00	0.59	0.50
PD						
2	1.00	0.59	0.54	1.00	0.46	0.79
5	0.99	0.45	0.82	0.99	0.31	1.38
8	0.94	0.29	1.22	0.94	0.20	2.02
FIML						
2	1.00	0.48	0.76	1.00	0.36	1.02
5	1.00	0.43	0.86	1.00	0.30	1.24
8	0.99	0.38	1.00	0.99	0.24	1.48
MV						
2	1.00	0.64	0.45	1.00	0.57	0.54
5	1.00	0.59	0.51	1.00	0.55	0.60
8	1.00	0.57	0.59	1.00	0.49	0.67
MC						
2	1.00	0.67	0.41	1.00	0.61	0.46
5	1.00	0.66	0.43	1.00	0.62	0.45
8	1.00	0.66	0.41	1.00	0.64	0.42

Table 3: Overview of identified factors for PA_SMC and SMT_RMSEA. Percentages of missing data equal to 2%, 5%, and 8%. The numbers in each of the rows are based on 900 correlation matrices (11,700 in total). *Converged* is the fraction for which outcomes are reached. *Correct* is the fraction of converged outcomes identifying the correct number of factors. *Mean AD* is the average absolute deviation, in number of factors, of converged outcomes identifying an incorrect number of factors.

	Converged	Mean	IQR	Mean[a]	IQR[a]
PD					
10	0.894	0.046	0.043	0.043	0.040
20	0.266	0.086	0.076	0.083	0.078
40	0.002	0.105	0.023	—	—
FIML					
10	0.957	0.045	0.034	0.041	0.032
20	0.678	0.066	0.060	0.057	0.043
40	0.007	0.101	0.071	—	—
MV					
10	0.997	0.022	0.017	0.022	0.017
20	0.991	0.037	0.029	0.041	0.032
40	0.951	0.082	0.061	0.089	0.068
MC					
10	0.999	0.019	0.013	0.021	0.016
20	0.998	0.033	0.026	0.035	0.025
40	0.980	0.091	0.079	0.092	0.080

[a] Based on the 382 data sets all approaches have results for. Excluded due to convergence lower than 0.25 are: fiml40, pd40

Table 4: Overview of factor loading bias based on the number of factors identified by PA_SMC. Each of the rows represent a combination of percentages missing data (10%, 20%, and 40%) and missing data treatment (PD, FIML, MV, and MC). The numbers in each of the rows are based on 1,800 correlation matrices (23,400 in total). *Converged* is the fraction for which outcomes are reached. *Mean* is the average of the median factor loading biases versus NMD for the converged outcomes. *IQR* is the interquartile range (75^{th} percentile minus 25^{th} percentile) of these same median factor loading biases. The two rightmost columns show the *Mean* and *IQR* calculated on those data sets for which all combinations of percentage missing data and missing data treatment reached an outcome (excluding combination for which convergence was below 25%).

	Converged	Mean	IQR	Mean[a]	IQR[a]
PD					
10	0.867	0.086	0.091	0.064	0.063
20	0.323	0.116	0.098	0.115	0.096
40	0.002	0.095	0.015	—	—
FIML					
10	0.967	0.066	0.060	0.055	0.048
20	0.692	0.096	0.087	0.082	0.075
40	0.007	0.099	0.034	—	—
MV					
10	0.998	0.028	0.022	0.025	0.020
20	0.997	0.048	0.038	0.047	0.038
40	0.984	0.098	0.067	0.105	0.075
MC					
10	0.999	0.023	0.017	0.021	0.017
20	0.999	0.037	0.030	0.036	0.027
40	0.993	0.106	0.087	0.114	0.100

[a] Based on the 497 data sets all approaches have results for. Excluded due to convergence lower than 0.25 are: fiml40, pd40

Table 5: Overview of factor loading bias based on the number of factors identified by SMT_RMSEA. Each of the rows represent a combination of percentages missing data (10%, 20%, and 40%) and missing data treatment (PD, FIML, MV, and MC). The numbers in each of the rows are based on 1,800 correlation matrices (23,400 in total). *Converged* is the fraction for which outcomes are reached. *Mean* is the average of the median factor loading biases versus NMD for the converged outcomes. *IQR* is the interquartile range (75^{th} percentile minus 25^{th} percentile) of these same median factor loading biases. The two rightmost columns show the *Mean* and *IQR* calculated on those data sets for which all combinations of percentage missing data and missing data treatment reached an outcome (excluding combination for which convergence was below 25%).

	Converged	Mean	IQR	Mean[a]	IQR[a]
PD					
2	0.997	0.015	0.010	0.015	0.010
5	0.987	0.029	0.023	0.028	0.022
8	0.922	0.043	0.039	0.042	0.038
FIML					
2	0.997	0.038	0.030	0.036	0.028
5	0.993	0.041	0.030	0.039	0.027
8	0.980	0.045	0.035	0.044	0.033
MV					
2	0.998	0.010	0.006	0.010	0.006
5	0.996	0.018	0.013	0.017	0.012
8	0.994	0.026	0.020	0.025	0.019
MC					
2	0.998	0.009	0.005	0.008	0.005
5	0.998	0.015	0.010	0.014	0.009
8	0.997	0.019	0.013	0.018	0.013

[a] Based on the 807 data sets all approaches have results for.

Table 6: Overview of factor loading bias based on the number of factors identified by PA_SMC. Each of the rows represent a combination of percentages missing data (2%, 5%, and 8%) and missing data treatment (PD, FIML, MV, and MC). The numbers in each of the rows are based on 900 correlation matrices (11,700 in total). *Converged* is the fraction for which outcomes are reached. *Mean* is the average of the median factor loading biases versus NMD for the converged outcomes. *IQR* is the interquartile range (75^{th} percentile minus 25^{th} percentile) of these same median factor loading biases. The two rightmost columns show the *Mean* and *IQR* calculated on those data sets for which all combinations of percentage missing data and missing data treatment reached an outcome.

	Converged	Mean	IQR	Mean[a]	IQR[a]
PD					
10	0.984	0.071	0.075	0.065	0.071
20	0.688	0.121	0.096	0.123	0.095
40	0.011	0.151	0.100	—	—
FIML					
10	0.997	0.065	0.064	0.060	0.063
20	0.960	0.090	0.073	0.085	0.066
40	0.083	0.113	0.068	—	—
MV					
10	0.999	0.031	0.027	0.030	0.026
20	0.996	0.055	0.048	0.054	0.047
40	0.967	0.110	0.073	0.111	0.070
MC					
10	0.999	0.024	0.019	0.024	0.018
20	0.997	0.042	0.040	0.041	0.038
40	0.999	0.083	0.068	0.083	0.070

[a] Based on the 574 data sets all approaches have results for.

Table 7: Overview of factor loading bias based on the number of factors identified by SMT_RMSEA. Each of the rows represent a combination of percentages missing data (10%, 20%, and 40%) and missing data treatment (PD, FIML, MV, and MC). The number of respondents is set to 300, 400, and 500. The numbers in each of the rows are based on 900 correlation matrices (11,700 in total). *Converged* is the fraction for which outcomes are reached. *Mean* is the average of the median factor loading biases versus NMD for the converged outcomes. *IQR* is the interquartile range (75^{th} percentile minus 25^{th} percentile) of these same median factor loading biases. The two rightmost columns show the *Mean* and *IQR* calculated on those data sets for which all combinations of percentage missing data and missing data treatment reached an outcome (excluding combination for which convergence was below 25%).

	PA_SMC			SMT_RMSEA		
	Converged	Correct	mean AD	Converged	Correct	mean AD
Complete	1.00	0.88	0.13	1.00	0.65	0.50
PD						
10	0.62	0.27	0.91	0.46	0.10	2.16
20	0.12	0.03	1.51	0.11	0.01	1.70
40	0.00	0.00	1.00	0.00	0.00	—
FIML						
10	0.69	0.39	0.65	0.67	0.19	1.73
20	0.23	0.08	0.99	0.23	0.04	1.62
40	0.00	0.00	—	0.00	0.00	2.00
MV						
10	1.00	0.84	0.19	1.00	0.56	0.67
20	0.99	0.75	0.31	1.00	0.44	0.88
40	0.87	0.34	1.02	0.88	0.52	0.49
MC						
10	1.00	0.90	0.12	1.00	0.81	0.21
20	1.00	0.88	0.13	1.00	0.89	0.11
40	0.87	0.68	0.22	0.81	0.30	0.63

Table 8: Overview of identified factors for PA_SMC and SMT_RMSEA. Percentages of missing data equal to 10%, 20%, and 40%. Number of respondents equal to 40, 60, and 80. The numbers in each of the rows are based on 900 correlation matrices (11,700 in total). *Converged* is the fraction for which outcomes are reached. *Correct* is the fraction of converged outcomes identifying the correct number of factors. *Mean AD* is the average absolute deviation, in number of factors, of converged outcomes identifying an incorrect number of factors.

	Converged	Mean	IQR	Mean[a]	IQR[a]
PD					
10	0.439	0.088	0.087	0.080	0.080
20	0.104	0.112	0.088	—	—
40	0.000	—	—	—	—
FIML					
10	0.641	0.078	0.074	0.064	0.055
20	0.217	0.102	0.093	—	—
40	0.001	0.333	0.000	—	—
MV					
10	0.976	0.035	0.026	0.029	0.019
20	0.970	0.059	0.044	0.049	0.032
40	0.708	0.149	0.125	0.109	0.080
MC					
10	0.980	0.030	0.023	0.024	0.017
20	0.946	0.051	0.041	0.037	0.029
40	0.293	0.190	0.116	0.190	0.120

[a] Based on the 116 data sets all approaches have results for. Excluded due to convergence lower than 0.25: fiml20, fiml40, pd20, pd40.

Table 9: Overview of factor loading bias based on the number of factors identified by SMT_RMSEA. Each of the rows represent a combination of percentages missing data (10%, 20%, and 40%) and missing data treatment (PD, FIML, MV, and MC). The number of respondents is set to 40, 60, and 80. The numbers in each of the rows are based on 900 correlation matrices (11,700 in total). *Converged* is the fraction for which outcomes are reached. *Mean* is the average of the median factor loading biases versus NMD for the converged outcomes. *IQR* is the interquartile range (75^{th} percentile minus 25^{th} percentile) of these same median factor loading biases. The two rightmost columns show the *Mean* and *IQR* calculated on those data sets for which all combinations of percentage missing data and missing data treatment reached an outcome (excluding combination for which convergence was below 25%).