Are extensive simulation studies always necessary? A critical overview

Ginkel, J.R. van; Rippe, R.C.A.

Statistical Methods & Applications
https://doi.org/10.1007/s10260-026-00877-6

ORIGINAL PAPER

Are extensive simulation studies always necessary? A critical overview

Joost R. van Ginkel[1] · **Ralph C. A. Rippe**[2]

Received: 9 January 2025 / Accepted: 23 July 2026

Abstract
Simulation studies are often carried out to study theoretical properties of a statistical procedure that cannot be derived analytically. Oftentimes, simulation studies can be very extensive, including various factors with various levels. This makes the results of such studies often difficult to report, and difficult for readers to interpret the general picture in the overabundance of results. Additionally, in many cases such extensive studies may not even be necessary. This work aims to explore and discuss some general guidelines on when (not) to include a simulation factor. Depending on the research question, oftentimes varying only one or two factors may suffice for answering the research question. In the current paper two different types of simulation studies are defined, namely proof of principle studies and robustness studies. Next, for a number of quality measures it is argued which factors are and are not useful to vary, both in proof of principle studies and robustness studies. Finally, the results of a literature study are presented confirming the need for guidelines on what factors to vary in a simulation study. A general recommendation is to not vary too many factors, especially for proof of principle studies.

Keywords Factors in simulation studies · Type-I error rate · Power · Bias · Violation of assumptions

1 Introduction

Statisticians and methodologists are often interested in performance of a statistical procedure. Imagine the following practical scenarios:

✉ Joost R. van Ginkel
jginkel@fsw.leidenuniv.nl

[1] Department of Psychology, Methodology and Statistics, Leiden University, PO Box 9500, Leiden 2300 RB, The Netherlands

[2] Department of Child and Family Studies, Leiden University, Leiden, The Netherlands

Published online: 12 August 2026

- A new statistical technique has been developed, of which the statistical properties are largely unknown, but can at least to some extent be anticipated under specific assumptions. A statistician wants to know how heavily the outcomes of the technique are affected by violation of these assumptions. Before setting up a full-scale evaluative simulation design, the statistician evaluates performance using two extreme situations, one where all assumptions are met, and one where (some of) the assumptions are violated. The results could ultimately lead to a follow-up study in which a wider variety of situations are investigated.
- A certain statistical technique has known theoretical properties under certain assumptions, based on analytical proof or on earlier simulation studies. However, its behavior is unknown under (1) varying severity of assumption violations, (2) combinations of assumption violations, or (3) otherwise potentially unfavorable but realistic scenarios. What is required is a thorough evaluation of the different conditions under which this technique performs very well, reasonably well, or not at all, resulting in a set of recommendations for using this technique.

The first example is like a viability study for a statistician to set up an upscaled follow-up study. The second example targets a wider audience, including statisticians for guidance in their consultations as well as substantive experts in the decision-making stage before running their own analyses.

The different goals of the above scenarios have implications for how extensive the specific simulation study needs to be. Scenario 1 requires only a small study with a limited number of conditions, while scenario 2 may require a larger number of conditions. However, the question is how much larger this number of conditions needs to be compared to scenario 1.

As there are an infinite number of ways in which empirical data can deviate from a situation where all assumptions of a statistical technique are met, it could be argued that under scenario 2 simulation studies should (always) include many different factors with many different levels to capture as many realistic situations as possible. However, this is often undesirable from a practical point of view, for example due to the amount of programming effort, computation time, or challenges of reporting such a large amount of results. In addition, such extensive designs may not be necessary for proving the point of the study. For some outcome measures it may be reasoned without simulations that a specific factor will not have a substantial effect on their outcomes. Thus, even under scenario 2, it could be argued that the study should not include too many conditions.

Despite the above given arguments, the authors have seen a general tendency in the literature for simulation studies to be more extensive than they practically need to be for answering the research questions. Additionally, past experiences -in which reviewers requested additional scenarios for simulation of which relevance to the underlying question was not evident - gave the impression that there may not always be full awareness of simulation factor redundancy. The current paper has been inspired by these two notions. We aim to explore and discuss some general guidelines on when (not) to include a simulation factor.

In the discussion that follows we define *proof of principle studies* as studies that fall under scenario 1, and *robustness studies* that fall under scenario 2. Furthermore,

we will refer to outcome measures to evaluate the performance of a statistical procedure, as *quality measures*. For a number of widely used factors in simulation studies, it is evaluated whether it is necessary to vary them for each type of study, and why. The factors and quality measures discussed in this paper serve as principled examples but are by no means exhaustive. To set the stage, in a small simulation literature study we discuss how frequently unnecessary factors are varied. Conclusive suggestions are provided on how extensive a simulation study should generally be, and when to vary a specific factor in a simulation study.

2 Types of simulation studies

2.1 Proof of principle studies

We define a proof of principle study as a simulation study that intends to establish the statistical properties of statistical techniques of which it is difficult to do so analytically. Suppose that for a newly developed advanced statistical technique, it has turned out to be difficult to derive its statistical properties analytically. Even though many advanced statistical techniques are based on well-established statistical or mathematical principles, it can be difficult to derive that they will lead to unbiased results when combined into a new technique. An example of such a technique is multiple imputation using Fully Conditional Specification (FCS; e.g. Van Buuren 2012). This technique imputes missing data where for each variable a separate statistical model is used to predict the missing data on this variable, conditional on other variables, using a Gibbs Sampler. Even though this technique is based on the well-established Gibbs sampling principle, it cannot be analytically derived that in FCS the imputed data are random draws from the joint distribution of the missing data, given the observed data, except in two specific conditions. These conditions are (1) a situation where all variables have a multivariate normal distribution and the conditional imputation models are all normal regression models, and (2) a situation where all variables are binary and the conditional imputation models are logistic regression models with all two-way interactions included (Van Buuren 2012, p. 116; Arnold et al. 1999, p. 186). Consequently, if the situation deviates from these specific scenarios, it cannot be derived that statistical inferences drawn from multiply imputed datasets using FCS are valid.

Analytical derivations may also be difficult for techniques based on asymptotic results or on approximations. For example, Giesbrecht and Burns (1985) proposed approximate *t*-tests testing contrasts for significance in multilevel models, based on large-sample theory. The purpose of a proof of principle study is then to serve as a substitute for analytical derivations or as an empirical demonstration of the theoretical properties. In proof of principle studies data are usually simulated under the statistical model that the statistical procedure (implicitly) assumes. When an advanced technique is based on a number of procedures that all assume multivariate normality, or when an approximation of a statistic has been derived under multivariate normality, then the data are simulated under a multivariate normal distribution, and no other distributions. Thus, one would commonly need to vary only a small number of fac-

tors; those of which it is expected that they will affect the quality of the approximation, based on theory.

In summary, when we want to gain more insight into the statistical properties of such techniques, or when we need an empirical demonstration of how good an approximation is, then a proof of principle study may be useful. Proof of principle studies can be seen as a validation study of a statistical technique, similarly to how new measurement instruments in social sciences are validated by establishing their psychometric properties.

2.2 Robustness studies

In practice, the assumptions under which the statistical properties of a statistical procedure are known are never fully met. When we would like to get a numerical impression of how robust these techniques are when these conditions have not been met, data need to be simulated under conditions that violate these assumptions. In such situations we feel a study in which we need to vary multiple factors is more appropriate. We define this type of study as a robustness study, although this is not a well-established term in the literature.

A potential problem in a robustness study is that there are many different ways of violating underlying assumptions. For example, error can be skewed to the left or right in different amounts, be uniform or bimodal with varying separation and so on. We suggest ensuring first and foremost that realistic scenarios are included, next to factors that are expected to have a large influence on the performance of the statistical procedure, or which factors vary strongly within a substantive population of interest.

One or several scenarios where no assumptions are violated may also be included in a robustness study. These scenarios can serve as a benchmark to see how the specific statistical techniques perform regarding the quality measures of interest compared to an ideal situation. Although insightful, a benchmark may not always be necessary. For example, if the quality measures are bias and type-I error rate, and we know from analytical results that when no assumptions are violated the technique produces unbiased estimates and type-I error rates, we need not demonstrate this with a benchmark scenario. However, when the quality measure in the study is the bias of some statistic that is intrinsically biased (for example, R^2 in regression, or standardized regression coefficients), or when other quality measures than bias or type-I error rates are studied (for example, RMSE, or the Standard Error), then including a benchmark may be necessary.

Robustness studies can best be compared with a sensitivity analysis in the sense that it is explored to what extent a finding in an ideal situation still holds up in various scenarios that deviate from the ideal situation. Also, in a way robustness studies are exploratory in nature in the sense that they have a wide focus and it is explored whether there could be scenarios, not necessarily based on sound theory but mostly conjecture, in which they do not hold up.

3 Possible factors to vary in a simulation study

Next, for a number of factors it will be discussed whether they should be varied (1) for different quality measures, and (2) for the two earlier defined types of simulation studies, and why (not). The quality measures considered are *bias*, *rejection rate* of a statistical test, *stability* of a parameter, and measures affected by *overfitting*, all when deemed relevant for that factor.

3.1 Sample size

Bias. For a number of estimators such as sample regression coefficients and sample means it is well-known that they are unbiased under simple random sampling. Since there are no assumptions for these estimators regarding the sample size, it is usually not very useful to vary sample size when bias of such estimators is the quality measure.

However, many estimators are intrinsically biased and become only *asymptotically* unbiased with increasing sample size. Examples are R^2 and standardized coefficients in multiple regression (e.g., Yuan and Chan 2011). For such estimators, small-sample bias may pose a practical threat. Regularized estimators, such as regression coefficients in ridge regression have a specific type of intrinsic bias, as it is intentionally introduced using shrinkage. Here, sample size has a central role as it determines the degree of shrinkage, which in turn determines the amount of bias. Finally, the effects of model misspecification are amplified with increasing sample size. In this case, sample size needs to be varied only in conjunction with other factors.

When an intrinsically biased estimator is studied, the usefulness of varying the sample size depends largely on the goal of the study. In a proof of principle study, sample size should be varied when the principle to prove is how bias of an intrinsically biased estimator diminishes with increasing sample size. In case of a robustness study however, sample size is always a useful candidate factor to vary when the quality measure is an intrinsically biased estimator. Whether it should indeed be included is partly up to the researcher, while considering the number of factors already included in the study, how influential the researcher expects sample size to be compared to the influence of the other factors, and how the researcher expects sample size to interact with the other factors.

On a side note, bias is often studied together with rejection rate and stability of the parameter. Then sample size may be useful to include as a factor, because it may influence the other quality measures as well.

Rejection rate. For many parametric statistical tests, the exact sampling distributions are known, given a number of assumptions. Sample size usually does not influence the type-I error rate of the statistical test as the sampling distribution is already corrected for sample size by means of the degrees of freedom of the test. For such statistical tests varying the sample size is not very useful when the type-I error rate is the quality measure.

However, for other statistics the sampling distributions only hold asymptotically (for example, a Wald test in logistic regression, likelihood ratio tests, and Chi-squared tests for cross tables) and consequently, approximation of the actual sampling distri-

bution oftentimes depends on sample size. Thus, in a proof of principle study where the principle to prove is that the approximation becomes better with increasing sample size, sample size may be included as a factor. In a robustness study, however, the usefulness of varying sample size when type-I errors are studied, is again up to the researcher to some extent (how many factors does the study already have, etc.).

On the other hand, when the data are simulated according to the alternative hypothesis, then rejection rate is power. As power *is* influenced by sample size, sample size is to be varied in a proof of principle study when the principle is that power increases with increasing sample size. In a robustness study sample size could also be a useful candidate factor when power is the quality measure, although again this is for a part up to the researcher.

Stability of a parameter. In a proof of principle study, the principle to prove might also be related to the standard error and hence the width of a confidence interval of a parameter, which are directly related to sample size. In a robustness study, sample size may also be a candidate factor when the stability of a parameter is the outcome variable.

Miscellaneous quality measures. The above-mentioned quality measures are not exhaustive. What is decisive for inclusion is whether you expect a quality measure to be affected by sample size or not. If you do expect sample size to be influential, the question then becomes whether the study is a proof of principle study or a robustness study. In the former case, sample size is only to be varied when it is part of the principle to prove. In the latter case, sample size is a candidate factor to include, of which its actual inclusion is partly up to the researcher to decide. Figure 1 shows a flow chart that may be used for deciding whether to include sample size as a factor in a simulation study.

3.2 Number of variables

Quality measures influenced by overfitting. For several mainstream analyses guidelines exist regarding the maximum number of variables. For example, in multiple regression it is advised to have at least 15 cases per predictor (Pituch and Stevens 2016, p. 72), and in exploratory factor analysis various rules state that the sample size should be determined as a function of the number of variables to analyze (Pituch and Stevens 2016, p. 347). These guidelines suggest that for fixed sample size there is a maximum number of variables to include (assuming that their measurement level is irrelevant).

Setting a maximum number of variables in classical prediction models is mainly necessary because of rank issues and problems of overfitting. As the issue of overfitting is quite generally known it might be a rather trivial principle to prove in a proof of principle study. However, some well-known mitigation approaches exist to accommodate for the risk of overfitting, such as random prediction and classification forests (implicit regularization; Breiman et al. 1984) or LASSO regression (explicit regularization; Tibshirani 1996). A proof of principle study could be useful to show the principle that these regularization methods are better at handling overfitting than classical prediction models are.

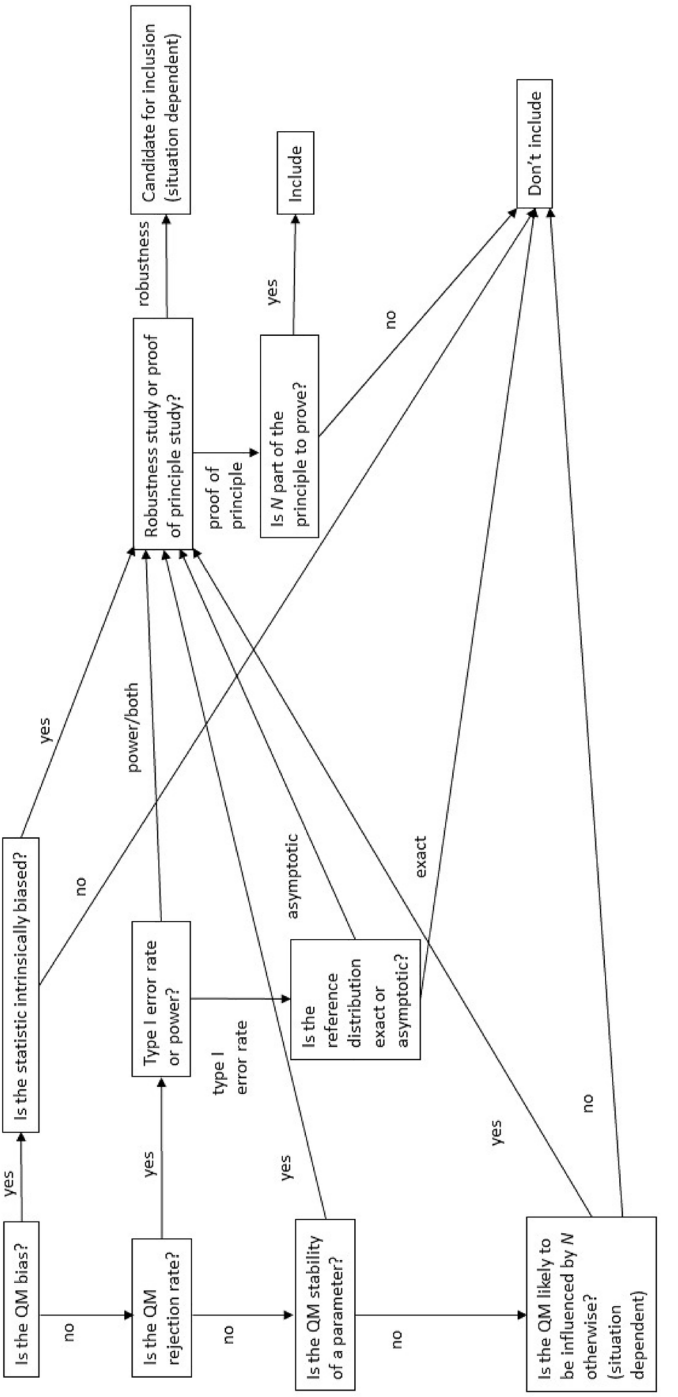

Fig. 1 Flowchart for deciding whether or not to include sample size as a factor in a simulation study

Furthermore, machine learning techniques exist that are aimed at finding the best prediction model without an a priori model specification. Here it is more difficult to simulate the effect of overfitting as the final model is unknown. The effect of overfitting may become less affected by the simulation model, but more by the machine learning technique (Du et al. 2025). Here, a proof of principle study could also be useful for showing the principle that machine learning techniques are better at the prediction of a specific criterion than classical prediction models are. Here, the focus is more on overall prediction accuracy than on the bias of specific parameters of one single model (Aliferis and Simon 2024).

In a robustness study, demonstrating that overfitting can occur is interesting when one of the research questions is up to how many variables can be added to the model before the procedure starts to break down. When that is not a research question of interest though, keep the number of variables constant at a sufficiently small number to avoid overfitting.

Figure 2 displays a flowchart for what decision to make about inclusion of the number of variables as a factor in different scenarios.

3.3 Multivariate distribution of the variables

Assumptions about the multivariate distribution. Many statistical prediction techniques neither make assumptions about the distributions of the predictors, nor about the distributions of their outcome variables. Many techniques do make assumptions about the *error* distribution of the underlying statistical model, but these are discussed separately below. Consequently, there is oftentimes not really a principle to prove about the distributions of the variables, so that varying the distributions in a proof of principle study might only be useful in specific situations.

One example of an explicit assumption about the multivariate distribution is confidence intervals of standardized regression coefficients, which have been derived under the assumption of multivariate normality (Yuan and Chan 2011). As a second example, in principal component analysis no explicit assumptions about the shape of the multivariate distribution of the variables are made (since the technique is based on correlations among variables), but the implicit assumption is that the variables are all numerically scaled and linearly related. In a proof of principle study the goal is to show the importance of these distributional assumptions. Thus, it may be desirable to simulate data under both the multivariate distribution that is accordance with the assumptions, and under one that is not.

In a robustness study, to mimic realistic situations, it could make sense to simulate data according to a mechanism that creates realistic datasets, whenever the quality measure implicitly or explicitly makes assumptions about the multivariate distribution of the data. There are a number of ways to do this. Firstly, one could simulate data according to some other distribution than the multivariate normal distribution. An example of this is Van Ginkel (2020) who simulated datasets with predictors drawn from both a multivariate normal and lognormal distribution with the same mean vector and covariance matrix. Comparing these to multivariate normal data with the same mean vector and covariance matrix, one could get an impression of how much the results are influenced by the fact that they are not multivariate nor-

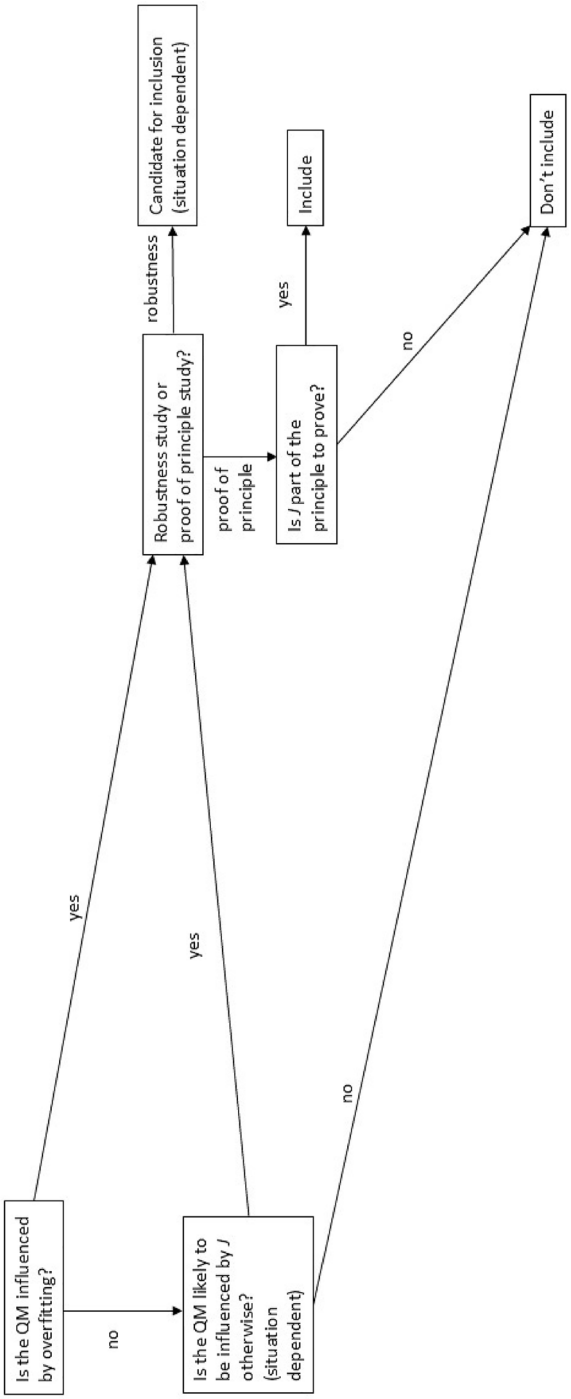

Fig. 2 Flowchart for deciding whether or not to include the number of variables as a factor in a simulation study

mally distributed. Additionally, conditional robustness of the different marginal distributions of the components can be evaluated *while* retaining the assumed Gaussian structure under Copula theory (Durante and Sempi 2016). Such an approach could make the robustness interpretation specific for each component, with exactly the same reference within that model.

Another way to create datasets with variables having a more realistic distribution is to sample from an existing large dataset as was done in, for example, Schafer et al. (1996), and Graham and Schafer (1999). When a dataset is too small to serve as a population, one could artificially enlarge the reference dataset by taking the original dataset, repeatedly adding random error (from some specified relevant distribution) to the original values to create new cases. As an example, consider Van Ginkel et al. (2014) who created random error in datasets consisting of Likert scale items by adding a number to the original scores of either −2, -1, 0, 1, and 2 with probabilities of 0.01, 0.04, 0.90, 0.04, and 0.01, respectively. If the resulting item score fell outside of the range of the data, the resulting score was set to either the lowest or highest possible score. In Fig. 3 a flowchart is displayed of when to include the multivariate distribution of the variables as a factor in a simulation study.

3.4 Error distribution of the outcome variable(s)

Rejection rate. Many statistical tests assume that the error of the model is (multivariate) normally distributed. Examples are multiple regression, (multivariate) analysis of variance, and multilevel analysis. In the case of proof of principle, varying the error distribution of the model (normal, skewed, etc.) is only useful when it is part of the principle to prove. The principle could be that statistical tests are robust to violations of the assumption of normality.

In case of a robustness study, a consideration is that in practice the errors of the outcome variable(s) are never exactly (multivariate) normally distributed, so by varying the error distribution one could capture a wide variety of realistic situations. One could consider simulating strong deviations from normality in combination with various other factors, to push the limits of this robustness, for example, under small sample sizes (the central limit theorem states that the normal approximation of the sampling distribution becomes better with increasing sample size). A flowchart for deciding on whether to include the error distribution as a factor in a simulation study, is given in Fig. 4.

3.5 Missingness mechanism

In the field of missing-data handling, one important factor that may introduce bias in the statistical analysis after handling the missing data, is the *missingness mechanism*. In general, three different types of missing data may be defined, namely missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR; Rubin 1976; Little and Rubin 1987; Van Buuren 2012). The formal definitions of all three missingness mechanisms are expressed in terms of two matrices: (1) the $N \times J$ data matrix $\mathbf{X}$ with an observed part $\mathbf{X}_{obs}$ and a missing part $\mathbf{X}_{mis}$, so that $\mathbf{X} = (\mathbf{X}_{obs}, \mathbf{X}_{mis})$, and (2) the $N \times J$ response-indicator matrix $\mathbf{R}$

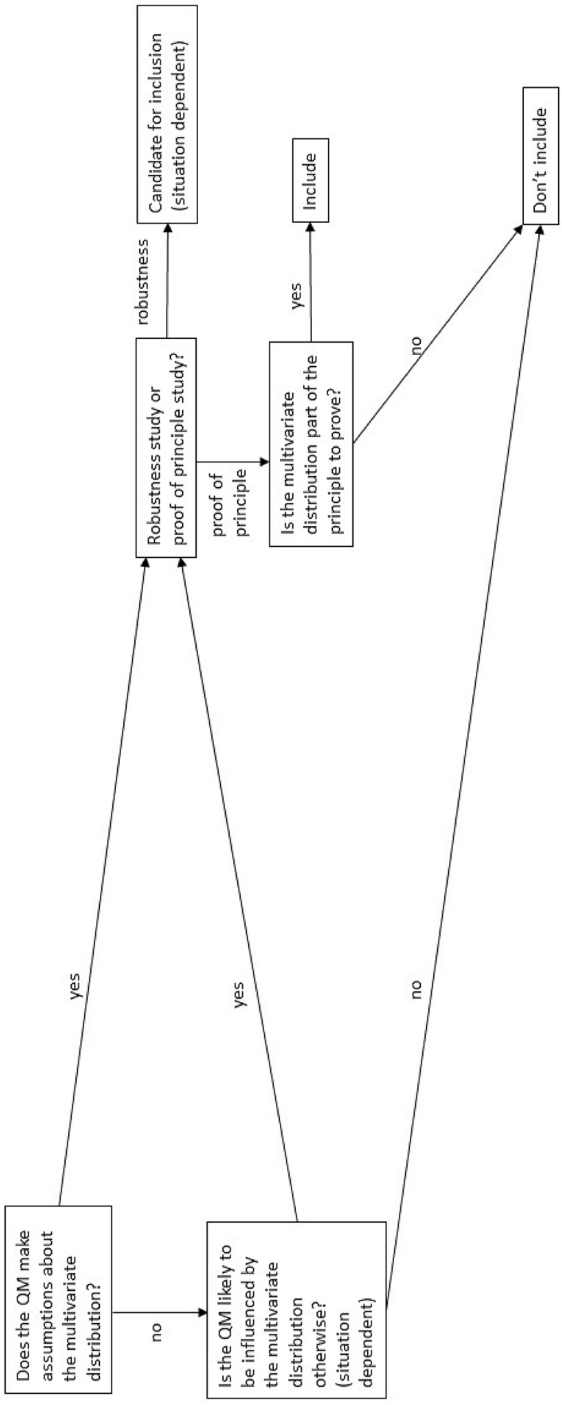

Fig. 3 Flowchart for deciding whether or not to include the multivariate distribution of the variables as a factor in a simulation study

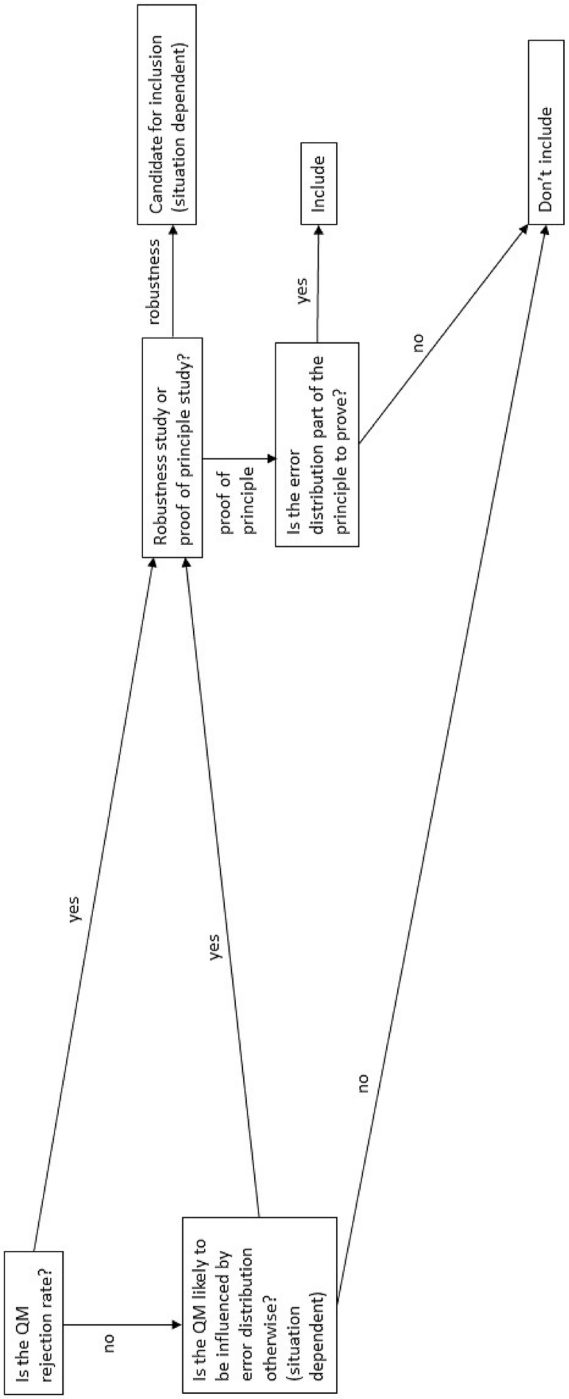

Fig. 4 Flowchart for deciding whether or not to include error distribution as a factor in a simulation study

with entries $r_{ij} = 1$ if $x_{ij} \in \mathbf{X}_{obs}$, and $r_{ij} = 0$ if $x_{ij} \in \mathbf{X}_{mis}$ ($i = 1, .., N$, and $j = 1, .., J$).

Firstly, MCAR is defined as:

$$P(\mathbf{R}|\mathbf{X}_{obs}, \mathbf{X}_{mis}) = P(\mathbf{R}).$$

Conceptually, MCAR means that the missing data neither depend on observed information nor unobserved information.

Secondly, MAR is defined as:

$$P(\mathbf{R}|\mathbf{X}_{obs}, \mathbf{X}_{mis}) = P(\mathbf{R}|\mathbf{X}_{obs}).$$

Conceptually, this means that the missing data depend on observed information within the dataset. For example, it may be the case that age is observed for all respondents and that as age increases, respondents have more missing data on the other variables.

Finally, under MNAR the conditional distribution $P(\mathbf{R}|\mathbf{X}_{obs}, \mathbf{X}_{mis})$ cannot be simplified. Conceptually, this means that the missing data depend on unobserved information. This could be a variable that has not been collected during the data collection process, or the value of the missing value itself (for example, high incomes being more frequently missing than low incomes).

It has generally been established that the traditionally most widely used method for handling missing data, namely listwise deletion (LD), only leads to unbiased results of statistical analyses under MCAR, that is, given that they are unbiased in case of complete data as well (e.g., Schafer 1997, pp. 23–24). More advanced methods to deal with missing data such as full information maximum likelihood (FIML) and multiple imputation (MI; Rubin 1987; Van Buuren 2012) produce unbiased results under both MCAR and MAR (see, e.g., Van Ginkel, Linting, Rippe, and Van der Voort 2020, p. 301). Under MNAR FIML and MI produce *less* biased results than LD does (Schafer 1997, pp. 26–27). For this reason, MCAR and MAR are also denoted *ignorable missingness mechanisms*, and methods based on FIML and MI are denoted *general ignorable procedures* (Schafer 1997, p. 23).

Under MNAR neither LD nor general ignorable procedures are guaranteed to give unbiased results. For this reason, MNAR is also denoted *nonignorable missingness*. Approaches to make MI appropriate for handling MNAR (see, for example, Galimard et al. 2016; Moustaki and Knott 2000; Fay 1986; Heckman 1976) usually require more parameters than can be estimated from the data alone so that restrictions must be imposed on them (Schafer 1997, p. 28).

Finally, it should be noted that other (less advanced) missing-data handling methods exist such a variable mean imputation, (single) regression imputation (Little and Schenker 1995, p. 60), or computing a total score based on the available item scores on a questionnaire (Schafer and Graham 2002, pp. 157–158). Like LD, these methods do not qualify as general ignorable procedures, let alone procedures that can handle MNAR, and they were never intended to be. Unlike LD, these methods will give biased results under *all* missingness mechanisms, including MCAR. Their original goal was to serve as a quick fix for the power loss of LD, in exchange for some additional (hopefully small) bias.

In the discussion of the possible inclusion of missingness mechanism as a factor in simulation studies we distinguish: (1) a scenario where the goal is to compare general ignorable procedures with non-general ignorable procedures, and (2) a scenario where this is not the goal.

3.5.1 When the goal is to compare general ignorable procedures with non-general ignorable procedures

In this scenario, firstly the goal could be to study the superiority of a general ignorable procedure over LD (e.g. Schafer 1997, pp. 24–26). It would be useful to include MCAR and MAR as missingness mechanisms, in order to show that the general ignorable procedure and LD perform equally well under MCAR, and that the general ignorable procedure performs better under MAR than LD. Possibly, MNAR could also be included to show that both procedures break down, but the general ignorable procedure still produces better results than LD does.

Secondly, the goal could be to compare a procedure that is general ignorable in one context but not in another context, with a procedure that is general ignorable in both situations. The former could be, for example, FIML while the latter could be MI. One MAR mechanism could then be included where the missingness depends on an observed variable that is included in the analysis model, and a second MAR mechanism could be included where the missingness depends on a variable outside the analysis model. In the former case, both MI and FIML are expected to give unbiased results, whereas in the latter case, only MI is expected to give unbiased results. For example, Van Ginkel et al. (2014) studied MCAR, two MAR mechanisms and MNAR in combination with a method comparable to FIML, and MI.

Lastly, the goal could be to compare a general ignorable procedure with a method that has been designed to handle specific MNAR mechanisms. If one wants to show the superiority of the latter procedure, then it is necessary to study these methods under both ignorable missingness (MAR and/or MCAR), and under nonignorable missingness (MNAR) as was done in, for example, Jolani (2012; Chap. 4).

In the comparison between general ignorable procedures and non-general ignorable procedures, the type of quality measure in the simulation study is irrelevant for deciding whether or not to include missingness mechanism as a factor. Although general ignorability is defined in terms of unbiasedness of likelihood-based estimators, these procedures may very well be compared under different missingness mechanisms on other types of quality measures as well, for example, root mean squared bias or classification error (Van Ginkel et al. 2014).

The type of study (i.e. proof of principle or robustness) is also largely irrelevant for deciding whether or not to include missingness mechanism as a factor. When the goal is to show the superiority of one of these procedures over the other under different missingness mechanisms, this may both be done in a proof of principle study to show the principle that one procedure is less sensitive to a missingness mechanism than the other, and in a robustness study where both methods are investigated under a wide variety of situations, including under different missingness mechanisms.

3.5.2 When the goal is not to compare general ignorable procedures with non-general ignorable procedures

In this scenario, the goal is to study whether the specific methods work at all, regardless of missingness mechanism. As an example, consider a study by Grund et al. (2016) in which they studied several pooled significance tests for ANOVA in multiply imputed data, under only ignorable missingness mechanisms (both MCAR and MAR were studied, though). In their study the goal was to study whether the significance tests gave satisfactory type-I error rates and power, under the assumptions under which they were derived.

In situations like this it may indeed be important to take the type of quality measure in the study and the type of study (proof of principle or robustness) into account in deciding whether or not to include missingness mechanism as a factor.

Bias. Because the three missingness mechanisms were defined as situations in which (maximum likelihood) estimators either were or were not guaranteed to be unbiased, bias of a statistic or parameter is the first quality measure that may be influenced by various missingness mechanisms. One important criterion for inclusion is whether bias of a parameter is one of the dependent variables in the study. Whether this bias is of an estimator that is intrinsically biased (such as a correlation or R^2) or not (such as an unstandardized regression coefficient) is irrelevant, since for intrinsically biased parameters one could study *how much more* biased they will become as a result of deviations from ignorable missingness.

In case of a proof of principle study, missingness mechanism should only be included when the principle is to show how much or little bias is introduced as a result of deviations from ignorability. In a robustness study, varying the missingness mechanism could indeed be useful because in practice, missingness may oftentimes deviate from ignorability. However, regardless of whether the study is a robustness study or a proof of principle study, keep in mind that when only general ignorable procedures are compared, it is not necessary to include *both* MCAR and MAR. As it is already known that general ignorable procedures give unbiased results under both missingness mechanisms, one may just pick one of the two as a baseline and compare it with MNAR. If, on the other hand, it is not clear whether the missing-data handling techniques *qualify* as general ignorable procedures, then it may be necessary to include all of the three missingness mechanisms.

Miscellaneous quality measures. When other outcome measures than bias are studied (such as the RMSE or rejection rate), it may be useful to include several missingness mechanisms since it is not always clear how the missingness mechanism will affect such measures. Van Ginkel and Kroonenberg (2017) found that for the specific the quality measure they studied (a measure related to RMSE), MNAR occasionally gave even *better* results than MAR. Although they studied a very specific situation that happened to cause this particular counterintuitive result, the more general point is that measures *other than bias* do not necessarily give the best results under MCAR and the worst results under MNAR.

In case of a proof of principle study, missingness mechanism should only be included when the principle is to show how much deviations from ignorability influ-

ence the quality measure. In a robustness study however, missingness mechanism is a good candidate factor to vary.

Following from the above, it may be useful to include MNAR in a robustness study about missing-data handling techniques that either explicitly or implicitly assume ignorable missingness, to see how robust they are to deviations from ignorable missingness. One may consider studying several types of MNAR: include an MNAR mechanism in which high values on a variable have a higher probability of being missing than low values, one in which low values have a higher probability of being missing, and one in which both low and high values have a higher probability of being missing than middle values. However, Bernaards and Sijtsma (2000) included two of the above-mentioned MNAR mechanisms in their study (one where both low and high values were more likely to be missing, and one where only high values were more likely to be missing) and showed that results did not differ strongly between MNAR mechanisms.

Figure 5 shows the flowchart for deciding whether or not to include missingness mechanism as a factor in a simulation study. The first thing that is decisive for whether or not to include missingness mechanism is not the quality measure. When comparing general ignorable procedures to non-general ignorable procedures, missingness mechanism is automatically an interesting factor to vary, regardless of the quality measure. If not, it is only then that the quality measure becomes relevant.

3.6 Percentage of missing data

Bias. It has been shown in numerous simulation studies that the general effect of increasing the percentage of missing data is that bias, uncertainty in parameter estimates, or in any other quality measure at a low percentage of missing data, becomes larger or more clearly visible at higher percentages of missing data (e.g., Van Ginkel et al. 2007; Grund et al. 2016; Van der Palm et al. 2016).

When under an ignorable missingness mechanism a missing-data handling method produces unbiased parameter estimates, it will generally also do so for higher percentages of missingness. Confidence intervals and variance of parameter estimates become larger (discussed further below). Thus, here, varying the percentage of missing data is not very useful.

The next question is whether it is useful to vary the percentage of missing data when the statistic is intrinsically biased or biased as a result of nonignorable missingness. Here, it could be useful because firstly, it is not always clear whether this bias increase is equivalent for all missing-data handling methods or all nonignorable missingness mechanisms in the study.

Secondly, when only one missing-data handling method is studied, then still for a specific percentage of missing data it is difficult to determine whether the resulting bias and/or uncertainty is large if there are no reference points of other percentages of missing data or whether this increase is, for example, linear or nonlinear. Thus, if one also wants to get an impression of *how* bias and uncertainty measures increase when the percentage of missing data increases, it may be necessary to include at least two, but preferably more than two percentages of missing data, and also report the results of the same dataset when no data are missing.

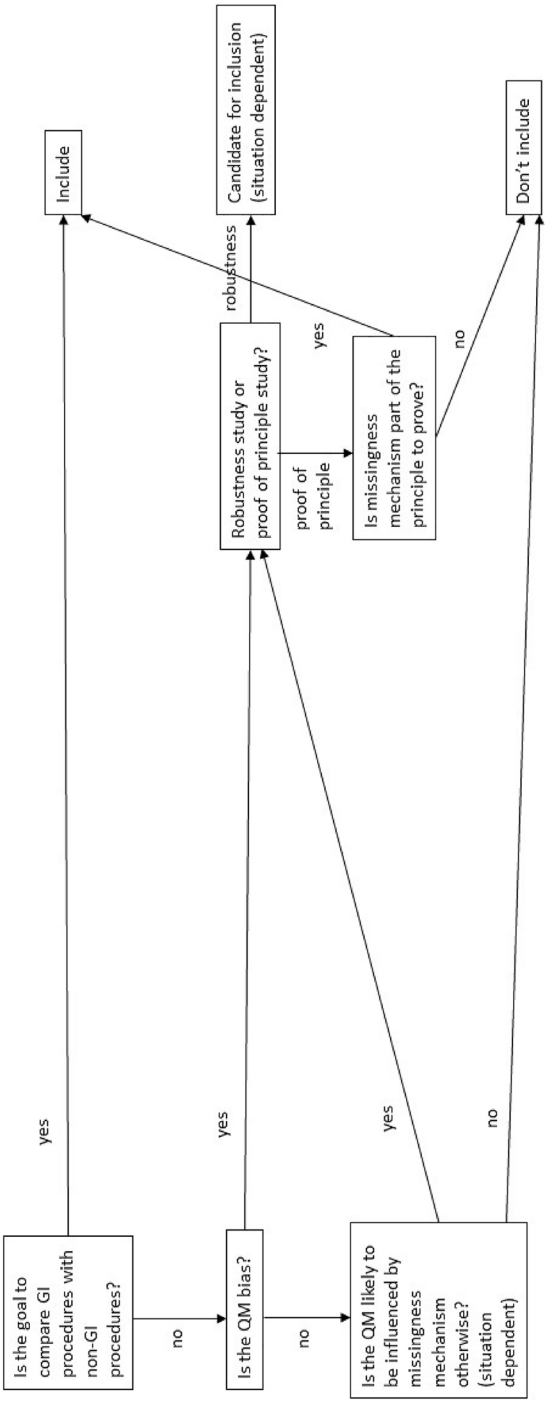

Fig. 5 Flowchart for deciding whether or not to include missingness mechanism as a factor in a simulation study

For both proof of principle and robustness studies it is useful to study what kind of an increase in bias to expect with an increasing percentage of missingness. Bias is hardly ever studied in isolation but oftentimes in combination with stability of the parameter (for example, its variance). Thus, even when an unbiased parameter is studied, then still several percentages may have to be included because of other quality measures in the study.

Rejection rate. Just as for the sample size, the first question one needs to ask in deciding whether or not to include percentages of missingness when rejection rate is the quality measure, is whether type-I errors are studied or power.

In case of type-I error rate, varying the percentage of missing data will not be useful under an exact reference distribution of the significance test. However, the reference distributions of significance tests that have been developed for multiply imputed datasets are either only approximations or have been derived under conditions that (almost) never hold. For example, Meng and Rubin (1992) developed a pooled likelihood ratio test for multiply imputed data which only holds approximately. Rubin (1987) discussed significance tests for single-parameter and multiparameter estimates in multiply imputed dataset for which several approximations of the degrees of freedom exist (Rubin 1987; Reiter 2007). Thus, when studying type-I error rates in multiply imputed data, varying the percentage of missing data is useful. In case of power, it does not even matter whether the reference distribution of a test is asymptotic or exact; as power is influenced by the percentage of missing data (the more missing data, the lower the power), percentage of missing data is useful to vary whenever power is studied. The type of simulation study is irrelevant because in practice many different percentages may occur in incomplete datasets, and it is always useful to see how well the approximation of the reference distributions still hold with increasing percentage of missing data, or how much the power drops.

Stability of a parameter estimate. For the stability of a parameter estimate (e.g. its variance or confidence width) the same arguments apply as for power.

The importance of percentage of missing data as a factor also shows in the flowchart of Fig. 6. Percentage of missing data should not be included (1) when only bias of an unbiased estimator is studied, (2) when a reference distribution for a statistical test is exact, and (3) when a quality measure is studied that is not expected to be influenced by the percentage of missing data, all of which are rare situations.

4 Literature review

In this section we will discuss a scoping literature review to support the necessity of the above comments and suggestions. To this end, we reviewed published simulation papers and related the description of the motive to the reported simulation factors. All results are summarized in Table 1 and discussed in (more) detail in the relevant sections below.

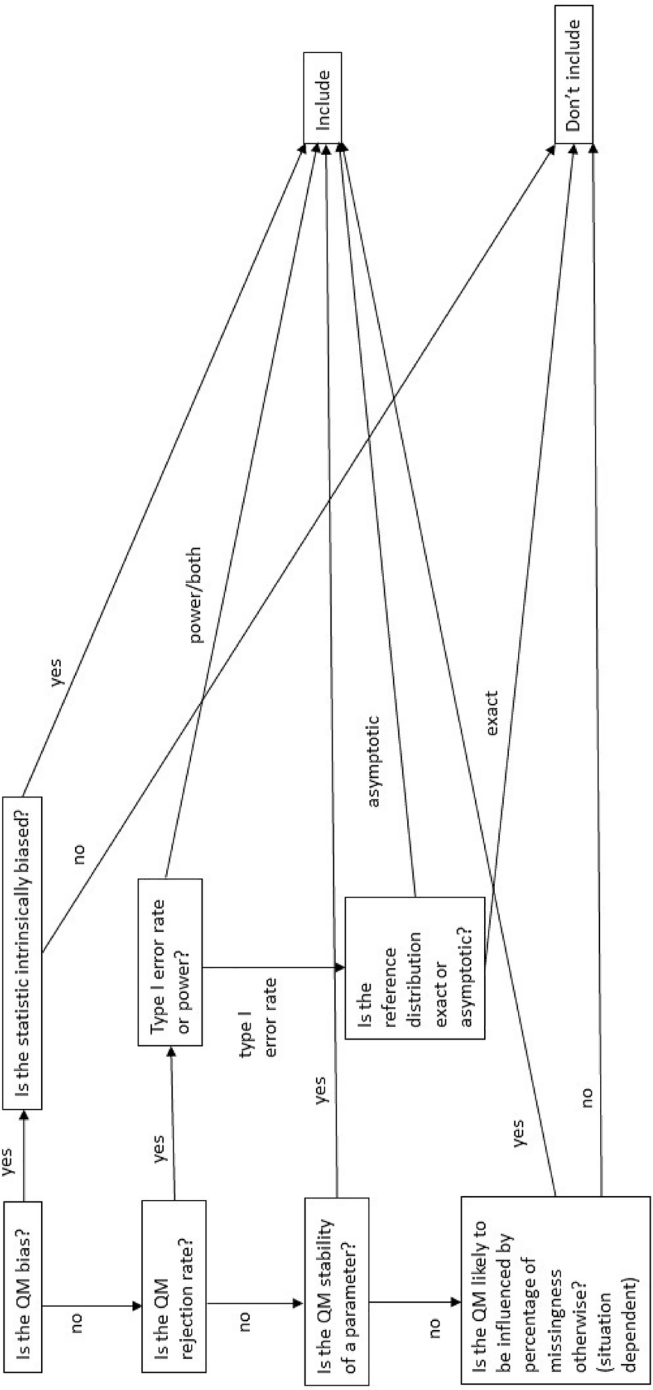

Fig. 6 Flowchart for deciding whether or not to include percentage of missingness as a factor in a simulation study

Table 1 Results literature study on factors in simulation studies

Journal*	Study type**	*n*	#Studies	#Factors	#Necessary factors	#Necessary factors/ #factors	
APM	PP	2	2	4	4	1.00	0.95
	R	3	3	15	14	0.93	
BJMSP	PP	1	1	2	2	1.00	0.96
	R	4	7	23	22	0.96	
BMA	PP	4[1]	8	16	15	0.94	0.88
	R	2[1]	2	9	7	0.78	
CSDA	PP	3	4	9	7	0.78	0.89
	R	2	3	9	9	1.00	
JC	PP	4	4	13	11	0.85	0.81
	R	1	1	3	2	0.67	
JRSSB	PP	3	4	7	6	0.86	0.71
	R	2	2	7	4	0.57	
JSCS	PP	4	7	12	11	0.92	0.82
	R	1	1	5	3	0.60	
MBR	PP	2	2	10	7	0.70	0.90
	R	3	4	21	21	1.00	
PM	PP	2[1]	2	8	6	0.75	0.78
	R	4[1]	4	15	12	0.80	
PMA	PP	3	5	10	9	0.90	0.79
	R	2	2	9	6	0.67	
Total	PP	28[2]	39	91	78	0.86	0.86
	R	24[2]	29	116	100	0.86	

* APM = Applied Psychological Measurement, BJMSP = British Journal of Mathematical and Statistical Psychology, BMA = Biometrika, CSDA = Computational Statistics and Data Analysis, JC = Journal of Classification, JRSSB = Journal of the Royal Statistical Society, series B, JSCS = Journal of Statistical Computation and Simulation, MBR = Multivariate Behavioral Research, PM = Psychological Methods, PMA = Psychometrika

** PP = Proof of Principle, R = Robustness

[1]Total does not add up to 5 because one article contains both a proof of principle study and a robustness study

[2]Total does not add up to 50 because two articles contain both a proof of principle study and a robustness study

4.1 Procedure

We identified ten journals which are known to regularly publish simulation studies. In each journal we extracted the 5 most recent articles at the time of coding (2024) in which a simulation study was explicitly described and was (part of) the main content. The 50 included papers ranged from mathematically and statistically oriented papers (e.g. on inclusion of Bayes Factors and Bayesian regularization, or on covariate selection etc.) to substantively oriented papers with a technical stance (e.g. on identifying clinical symptom sets using IRT models). We worked backward, i.e. we started from the most recent issues that were fully available online and worked backward in publication date until we found 5 articles matching the criteria. Journal issues with future releases or ongoing completion were not considered. For each article we coded for the journal, year of publication, the study type (robustness or proof of principle),

the number of simulation studies included in the paper (each in their own record), the number of simulation design factors. From these observations we coded which (and thus how many) factors were necessary and next we deduced the ratio of necessary over the total number of factors.

The examples of factors and outcome measures discussed earlier served an illustrative purpose and are by no means claimed to be exhaustive. A general and exhaustive definition of an "unnecessary" factor in a simulation study is near-impossible to provide. The current operationalization was that when a factor is considered necessary by the authors of the original paper, they had either included in the discussion of the theoretical framework, had justified its necessity for the study in advance, or both. Hence, for our literature study, a factor was thus considered "unnecessary" when it was not mentioned in the abstract or in the main introduction; when it was not described in relation to either literature or theory and was only introduced in the methods section without prior mention; or when the authors mentioned that it was included after external request. We do however recognize that on occasion, not discussing a factor in the theoretical framework or abstract may be the result of authors simply forgetting to mention its relevance to the study, limiting its rigorous interpretation as a whole. Coded, but not reported in the table are the full journal references and their DOI: these are available upon request.

4.2 Results

We found that the proportion of necessary simulation factors was relatively high (0.71–0.96), while superfluous factors were coded in all included journals. This is an interesting result, since these articles generally concern (only) the one or two most recent issues in (only) one volume. Extrapolation to other journals then suggest that in any other journal, volume or issue there will be unnecessary design factors. Thus, the prevalence of superfluous factors in absolute sense is high.

We also identified (but have not explicitly coded for) papers with no superfluous factors, but with a surplus of levels within a design factor, or where the chosen levels in the factor were not discussed nor supported. For example, in one study, five different sample sizes were studied, where two or three would have sufficed based on the main introduction or methods section.

Furthermore, some of the studies included had a relatively broad focus and included quite a high number of factors, making them robustness studies. However, the statistical procedure studied seemed to be newly developed, making it more appropriate for a first study on the procedure to be a proof of principle study rather than a robustness study. Although we cannot test this conjecture with our current data, it was our feeling that in some of these cases a reviewer might have asked to increase the number of factors, and that the authors broadened the scope of their article accordingly. Based on the above we see support for the necessity of the guidelines proposed in this manuscript, while we are inclined to interpret our review findings to be (on the) conservative (side).

5 Discussion

This work aimed to explicitly define scenarios in which certain factors in a simulation design should or should not be included. Simulation factors were discussed within their own context to which the discussion applies, and relevant exceptions were mentioned explicitly. The suggested guidelines are meant to guide the decision-making process in setting up a simulation study or when encountering (aspects of) a simulation study as a reviewer or as an applied researcher.

In pursuit of the above, two types of simulation studies were defined; those for proof of principle and those for testing robustness. They vary in aim and consequently, in required magnitude, although a general argument was made for both not to include too many factors. However, how many and which factors should be included depends largely on the quality measure and the research question. Firstly, if the quality measure is not expected to be substantially influenced by the factor, then it is not very useful to include this factor.

Secondly, if the research question is whether a new statistical technique based on well-established statistical principles performs well but of which it is difficult or impossible to prove this analytically, then the simulation study only needs to include a small number of factors. Preferably, the data should be simulated under the assumptions that underly the well-established principles.

If, on the other hand, the goal is to study how well statistical techniques perform in a wide variety of realistic situations, then the simulation study should include a larger number of factors. Varying these factors usually leads to a full factorial design with conditions, where in one (or several) condition(s) all the assumptions of the technique are met, and a host of other conditions in which they are violated to varying extents.

In general, varying many factors is mostly useful in robustness studies. That does not imply that in robustness studies one should vary as many factors as possible. Overstepping may cause the results to become more difficult to interpret and consequently, a clear summary of the results in the reporting cannot be provided.

In a literature review we investigated the practical relevance of the guidelines provided in the current paper. For each of the ten journals included, at least one out of five most recent articles studied one or more unnecessary factors, suggesting that there is a lot to gain regarding succinctness of simulation studies. By providing guidelines for which factors to vary and when, we hope to have created a framework from which only the necessary factors to include in a simulation study can be derived for future studies.

Finally, we will address a few points of discussion. Firstly, we note that the distinction between a proof of principle study and a robustness study is not always evident: any simulation study may not clearly fit into either one of these categories. For example, the principle to prove might be the central limit theorem. This theorem states that for large samples, the normal approximation of the sampling distribution of a parameter estimate fairly holds, even when the population from which is sampled is not normal. To prove the principle, one could vary both the sample size and the distribution of error of the model, which is also regularly included in robustness studies because in practice the normality-of-error assumption is often violated. In order

to avoid criticism that the study is too limited, it might be helpful to explicitly state in advance that the study is a proof of principle study, explicitly mention the principle to prove, and explain which factors are important to vary in order to prove the principle.

Secondly, it should be noted that the examples provided and contexts were not exhaustive or generic, nor were they meant to be. This raises the question how researchers can still use the framework in this paper in a more general way to design their simulation studies such that no unnecessary factors are included. The answer to this question is twofold. Firstly, although not exhaustive, the examples were based on frequently occurring situations from our own experience, so it is likely that in future simulation studies, researchers will encounter similar scenarios where they can at least partly base their design on the given flowcharts.

Secondly, when a specific scenario is not captured by any of the flowcharts, a more general guideline is to think of arguments for why a potential factor could be influential, and next, to include these arguments in the theoretical framework and the methods section of the paper. If one cannot think of a good argument, it should not be included. Reversely, if one can think of arguments not to include an intuitively appealing potential factor, then these arguments should be included in the theoretical framework or methods section as well. This general guideline of justifying the inclusion of a factor to oneself and to the reader of the study, matches the criteria that we used for defining an "unnecessary factor" in our literature study.

Given this general guideline, a researcher can ultimately make one's own flowchart for the specific scenario. As a matter of fact, this is how the flowcharts for the given examples were created in the first place.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10260-026-00877-6.

Data availability The authors confirm that all data generated or analyzed during this study are included in this published article.

Declarations

Conflict of interest None of the authors have competing interests to declare.

References

Aliferis C, Simon G (2024) Overfitting, underfitting and general model overconfidence and under-performance pitfalls and best practices in machine learning and AI. In: Simon GJ, Aliferis C (eds) Artificial intelligence and machine learning in health care and medical sciences: Best practices and pitfalls, pp 477–524. https://doi.org/10.1007/978-3-031-39355-6_10

Arnold BC, Castillo E, Sarabia JM (1999) Conditional specification of statistical methods. Springer, New York

Bernaards CA, Sijtsma K (2000) Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. Multivar Behav Res 35:321–364. https://doi.org/10.1207/S15327906MBR3503_03

Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. Wadsworth Publishing, New York

Du KL, Zhang R, Jiang B, Zeng J, Lu J (2025) Understanding machine learning principles: Learning, inference, generalization, and computational learning theory. Mathematics 13:451. https://doi.org/10.3390/math13030451

Durante F, Sempi C (2016) Principles of copula theory, vol 474. CRC, Boca Raton, FL

Fay RE (1986) Causal models for patterns of nonresponse. J Am Stat Assoc 81:354–365. https://doi.org/10.1080/01621459.1986.10478279

Galimard JE, Chevret S, Protopopescu C, Resche-Rigon M (2016) A multiple imputation approach for MNAR mechanisms compatible with Heckman's model. Stat Med 35:2907–2920. https://doi.org/10.1002/sim.6902

Giesbrecht FG, Burns JC (1985) Two-stage analysis based on a mixed model: Large sample asymptotic theory and small-sample simulation results. Biometrics 41:477–486. https://doi.org/10.2307/2530872

Graham JW, Schafer JL (1999) On the performance of multiple imputation for multivariate data with small sample size. In: Hoyle R (ed) Statistical strategies for small sample research. Sage, Thousand Oaks, CA, pp 1–29

Grund S, Lüdtke O, Robitzsch A (2016) Pooling ANOVA results from multiply imputed datasets: A simulation study. Methodology 12:75–88. https://doi.org/10.1027/1614-2241/a000111

Heckman J (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models. Ann Econ Soc Meas 5:475–492

Jolani S (2012) Dual imputation strategies for analyzing incomplete data. Dissertation, University of Utrecht

Little RJA, Schenker N (1995) Missing data. In: Arminger G, Clogg CC, Sobel ME (eds) Handbook of statistical modeling for the social and behavioral sciences. Plenum, New York, pp 39–75

Moustaki I, Knott M (2000) Weighting for item non-response in attitude scales by using latent variable models with covariates. J Roy Stat Soc A 163:445–459. https://doi.org/10.1111/1467-985X.00177

Pituch KA, Stevens JP (2016) Applied multivariate statistics for the social sciences, 6th edn. Lawrence Erlbaum Associates, Hillsdale, NJ

Reiter JP (2007) Small-sample degrees of freedom for multi-component significance tests with Q1 Q2 Q3 multiple imputation for missing data. Biometrika 94:502–508. https://doi.org/10.1093/biomet/asm028

Rubin DB (1976) Inference and missing data. Biometrika 63:581–592. https://doi.org/10.2307/2335739

Rubin DB (1987) Multiple imputation for nonresponse in surveys. Wiley, New York

Schafer JL (1997) Analysis of incomplete multivariate data. Chapman & Hall, London

Schafer JL, Ezzati-Rice TM, Johnson W, Khare M, Little RJA, Rubin DB (1996) The NHANES III multiple imputation project. Proceedings of the Survey Research Methods Section of the American Statistical Association (pp. 28–37). https://wwwn.cdc.gov/Nchs/Data/Nhanes3/7a/doc/jsm96.pdf

Tibshirani (1996) Regression shrinkage and selection via the Lasso. J R Stat Soc B 48:267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Van Buuren S (2012) Flexible imputation of missing data. Chapman & Hall/CRC, Boca Raton

Van der Palm DW, Van der Ark LA, Vermunt JK (2016) A comparison of incomplete-data methods for categorical data. Stat Methods Med Res 25:754–774. https://doi.org/10.1177/0962280212465502

Van Ginkel JR (2020) Standardized regression coefficients and newly proposed estimators for R^2 in multiply imputed data. Psychometrika 85:185–205. https://doi.org/10.1007/s11336-020-09696-4

Van Ginkel JR, Kroonenberg PM (2017) Evaluation of multiple-imputation procedures for three-mode component models. J Stat Comput Sim 87:3059–3081. https://doi.org/10.1080/00949655.2017.1355368

Van Ginkel JR, Van der Ark LA, Sijtsma K (2007) Multiple imputation for item scores when test data are factorially complex. Brit J Math Stat Psy 60:315–337. https://doi.org/10.1348/000711006X117574

Van Ginkel JR, Kroonenberg PM, Kiers HAL (2014) Missing data in principal component analysis of questionnaire data: a comparison of methods. J Stat Comput Sim 84:2298–2315. https://doi.org/10.1007/s11336-020-09696-4

Yuan KH, Chan W (2011) Biases and standard errors of standardized regression coefficients. Psychometrika 76:670–690. https://doi.org/10.1007/S11336-011-9224-6