

**Simulation Studies for Methodological Research in Psychology:
A Standardized Template for Planning, Preregistration, and Reporting**

Björn S. Siepe^{*1}, František Bartoš^{*2}, Tim P. Morris³, Anne-Laure Boulesteix⁴, Daniel W.
Heck¹, and Samuel Pawel^{*5}

¹Department of Psychology, University of Marburg

²Department of Psychological Methods, University of Amsterdam

³MRC Clinical Trials Unit at UCL, University College London

⁴Institute for Medical Information Processing, Biometry and Epidemiology, Faculty of
Medicine, Ludwig Maximilian University of Munich

⁵Epidemiology, Biostatistics and Prevention Institute (EBPI), Center for Reproducible Science
(CRS), University of Zurich

Author Note

*Björn S. Siepe, František Bartoš, and Samuel Pawel contributed equally.

Björn S. Siepe  <https://orcid.org/0000-0002-9558-4648>

František Bartoš  <https://orcid.org/0000-0002-0018-5573>

Tim P. Morris  <https://orcid.org/0000-0001-5850-3610>

Anne-Laure Boulesteix  <https://orcid.org/0000-0002-2729-0947>

Daniel W. Heck  <https://orcid.org/0000-0002-6302-9252>

Samuel Pawel  <https://orcid.org/0000-0003-2779-320X>

The authors made the following contributions: BSS: Conceptualization, Methodology, Formal Analysis, Software, Investigation, Data Curation, Visualization, Writing – original draft, Writing – review & editing; FB: Conceptualization, Methodology, Formal Analysis, Software, Investigation, Data Curation, Visualization, Writing – original draft, Writing – review & editing; TPM: Writing – review & editing; ALB: Writing – review & editing; DWH: Writing – review & editing; SP: Conceptualization, Methodology, Formal Analysis, Software, Investigation, Data Curation, Visualization, Writing – original draft, Writing – review & editing.

Correspondence concerning this article should be addressed to Email:

bjoern.siepe@uni-marburg.de

Abstract

Simulation studies are widely used for evaluating the performance of statistical methods in psychology. However, the quality of simulation studies can vary widely in terms of their design, execution, and reporting. In order to assess the quality of typical simulation studies in psychology, we reviewed 321 articles published in *Psychological Methods*, *Behavioral Research Methods*, and *Multivariate Behavioral Research* in 2021 and 2022, among which $100/321 = 31.2\%$ report a simulation study. We find that many articles do not provide complete and transparent information about key aspects of the study, such as justifications for the number of simulation repetitions, Monte Carlo uncertainty estimates, or code and data to reproduce the simulation studies. To address this problem, we provide a summary of the ADEMP (Aims, Data-generating mechanism, Estimands and other targets, Methods, Performance measures) design and reporting framework from [Morris, White, and Crowther \(2019\)](#) adapted to simulation studies in psychology. Based on this framework, we provide ADEMP-PreReg, a step-by-step template for researchers to use when designing, potentially preregistering, and reporting their simulation studies. We give formulae for estimating common performance measures, their Monte Carlo standard errors, and for calculating the number of simulation repetitions to achieve a desired Monte Carlo standard error. Finally, we give a detailed tutorial on how to apply the ADEMP framework in practice using an example simulation study on the evaluation of methods for the analysis of pre–post measurement experiments.

Keywords: experimental design, Monte Carlo experiments, meta-research, preregistration, reporting

Simulation studies are experiments and should be treated as such by authors and editors.

[Hauck and Anderson \(1984, p. 215\)](#)

Introduction

Simulation studies are an experimental method for evaluating the properties of statistical methods. They allow researchers to study the statistical properties of methods under complex conditions which would be difficult or impossible to study theoretically, for instance, with formal analyses or mathematical proofs. The idea is to simulate data with known characteristics, analyze these data using the methods under investigation, and then (ideally) compare the results with the known truth. By repeating this procedure under various conditions, the performance and robustness of a method can be assessed and compared to that of other methods. Simulation studies thus represent the “controlled experiment” in the toolbox of methodologists, whereas benchmarking of methods on a real data set would be analogous to a case study. Both are important, but simulation studies in particular allow us to understand when a method works well and when it does not, and ultimately to make recommendations on when to use a particular method in practice. We note that simulation can also be used for other purposes, such as experimental design (e.g., sample size planning or power analysis for complex statistical analyses where no closed-form solutions exist as in [Heck & Erdfelder, 2019](#); [Lakens & Caldwell, 2021](#)) or numerical integration (e.g., Markov chain Monte Carlo methods for computing posterior distributions in Bayesian statistics), but this use of simulation is typically not called “simulation study” in methodological research and is not the focus of the present paper.

As with any experiment, the quality of evidence from a simulation study depends on how the study is designed, conducted, analysed, and reported. However, unlike many other types of experiments, simulation studies offer much greater flexibility, as it is usually easy and financially inexpensive to change the design of the study and generate new results. This can be seen both as a strength but also as a reason for caution, since there are considerably more *researcher degrees of freedom* than in other types of experiments ([Simmons, Nelson, & Simonsohn, 2011](#)). For instance, researchers often have a high degree of flexibility in selecting certain methods and data-generating mechanisms, and in deciding which results are reported.

Issues with the conduct and reporting of simulation studies were described almost half a century ago (Hoaglin & Andrews, 1975). However, the attention afforded to researcher degrees of freedom in psychology and other empirical sciences has recently led to more critical reflection on the state of methodological research (Boulesteix, Binder, Abrahamowicz, & Sauerbrei, 2018; Boulesteix, Hoffmann, Charlton, & Seibold, 2020; Boulesteix, Hoffmann, et al., 2020; Friedrich & Friede, 2023; Heinze, Boulesteix, Kammer, Morris, & White, 2023; Luijken et al., 2023; Pawel, Kook, & Reeve, 2023; Strobl & Leisch, 2022).

Some may argue that simulation studies are often conducted at a more exploratory stage of research and therefore do not require as much rigor and transparency (including measures such as sample size planning, preparation and preregistration of a study protocol, or code and data sharing) as other types of studies. However, many simulation studies are not conducted and reported as exploratory, but rather with the explicit goal of deriving recommendations for the use of methods. It is important to realize that such simulation studies often have a large impact. For example, the simulation study by Hu and Bentler (1999) on cut-off criteria for structural equation models has been cited over 100,000 times, presumably justifying thousands of choices in structural equation modeling. It would be detrimental if the results of such a study were flawed or reported suboptimally. Another example is the simulation study that recommended the “1 variable per 10 events” heuristic as a sample size criterion for logistic regression (Peduzzi, Concato, Kemper, Holford, & Feinstein, 1996). This heuristic has been cited over 8,000 times and was widely adopted as a minimum sample size criterion, but the influential simulation study advocating it was later found to be non-replicable (van Smeden et al., 2016).

In non-methodological research, it has been repeatedly emphasized that research results should be accompanied by measures of statistical uncertainty, such as p -values, standard errors, or confidence intervals (Cumming, Fidler, Kalinowski, & Lai, 2012; van der Bles et al., 2019). Clear guidelines are now available in most fields, for example, the APA guidelines require that “*when point estimates [...] are provided, always include an associated measure of variability*” (American Psychological Association, 2020, p. 88). It is perhaps surprising, therefore, that methodological researchers rarely report measures of uncertainty

associated with the results of simulation studies, even though these researchers tend to be more familiar with statistical reasoning than non-methodological researchers (see the literature reviews by [Burton, Altman, Royston, & Holder, 2006](#); [Harwell, Kohli, & Peralta-Torres, 2018](#); [Hauck & Anderson, 1984](#); [Koehler, Brown, & Haneuse, 2009](#); [Morris et al., 2019](#)).

To help navigate the complexities of conducting simulation studies, various guidelines, recommendations and tutorials have been published over the years—both in statistics (e.g., [Boulesteix, Groenwold, et al., 2020](#); [Burton et al., 2006](#); [Chipman & Bingham, 2022](#); [Hoaglin & Andrews, 1975](#); [Kelter, 2023](#); [Koehler et al., 2009](#); [Lange, 2022](#); [Morris et al., 2019](#); [White, Pham, Quartagno, & Morris, 2023](#)) and in psychology (e.g., [Boomsma, 2013](#); [Carsey & Harden, 2014](#); [Chalmers & Adkins, 2020](#); [Feinberg & Rubright, 2016](#); [McNeish, Lane, & Curran, 2018](#); [Paxton, Curran, Bollen, Kirby, & Chen, 2001](#); [Psychometric Society, 1979](#); [Skrondal, 2000](#)). A recent tutorial from the statistical literature is provided by [Morris et al. \(2019\)](#), which we recommend as a first read on the state-of-the-art methodology of simulation studies.

A general and accessible introduction to simulation studies that builds on recent guidelines from the statistics literature is currently lacking in psychology. We therefore provide an introduction to the ADEMP (Aims, Data-generating mechanisms, Estimands and other targets, Methods, Performance measure) design and reporting structure of [Morris et al. \(2019\)](#), aimed at researchers in psychology. Based on ADEMP's structure, we provide a literature review of simulation studies published in the journals *Psychological Methods (PM)*, *Behavioural Research Methods (BRM)*, and *Multivariate Behavioral Research (MBR)*, which represent three prominent journals for methodological research in psychology. To help researchers conduct rigorous simulation studies, we provide the ADEMP preregistration (ADEMP-PreReg) template (<https://github.com/bsiepe/ADEMP-PreReg>) that methodological researchers can use to preregister their simulation studies, clarifying important aspects of their study in advance and helping them to avoid common pitfalls. For simulation studies that are not preregistered, the template can still be used as a blueprint for structured planning and reporting of their study. We also give formulae for the most commonly used performance measures, their Monte Carlo standard errors, and for the associated sample size

calculations. Finally, we illustrate ADEMP and the template with an example simulation study on a typical application from psychological research—a comparison of methods for the analysis of pre–post measurements.

The ADEMP structure

[Morris et al. \(2019\)](#) introduced ADEMP as a structured approach to planning and reporting of simulation studies. Table 1 provides an overview of the approach. Here, we summarize the main points of ADEMP in the context of methodological research in psychology, but refer to Section 3 in [Morris et al. \(2019\)](#) for a more comprehensive treatment of the topic. In addition, we also summarize recommendations for computational aspects and reporting.

Table 1

Summary of the ADEMP Planning and Reporting Structure for Simulation Studies.

Step	Explanation	Example
Aims	What is the aim of the study?	To evaluate the hypothesis testing and estimation characteristics of different methods for analyzing pre–post measurements in terms of efficiency and robustness
Data-generating mechanism	How are data sets generated?	Pre–post measurements are simulated from a bivariate normal distribution for two groups, with varying treatment effects and pre–post correlations
Estimands and other targets	What are the estimands and/or other targets of the study?	The null hypothesis of no effect between groups is the primary target, the treatment effect is the secondary estimand of interest
Methods	Which methods are evaluated?	ANCOVA, change-score analysis, and post-score analysis
Performance measures	Which performance measures are used?	Type I error rate, power, and bias

Aims

The aim of a simulation study refers to the goal of the methodological research project and shapes subsequent choices. Aims are typically related to evaluating the properties of a method (or multiple methods) with respect to a particular statistical task. In psychological simulation studies, common statistical tasks and exemplary aims (taken from the literature review) can include:

- *Estimation*, e.g., assessing the effect of different parametrizations of covariance structures in mixed-effect models when estimating an intervention effect ([McNeish & Bauer, 2022](#)).
- *Hypothesis testing*, e.g., comparing different tests of publication bias ([Rodgers & Pustejovsky, 2021](#)).
- *Model selection*, e.g., comparing different fit indices for selecting the best structural equation model ([Shi, DiStefano, Maydeu-Olivares, & Lee, 2022](#)).
- *Design*, e.g., comparing different methods for determining sample size in mixed-effect modeling ([Murayama, Usami, & Sakaki, 2022](#)).
- *Prediction*, e.g., comparing different algorithms for predicting participants' problem-solving strategies ([Moss, Wong, Durriseau, & Bradshaw, 2022](#)).
- *Other aims*, e.g., assessing tools for quantifying complexity ([Moulder, Daniel, Teachman, & Boker, 2022](#)), clustering data sets into equivalent parts ([Papenberg & Klau, 2021](#)), or comparing implementations of principal component analysis rotations across software packages ([Grieder & Steiner, 2021](#)).

These statistical tasks are often closely related, for example, hypothesis testing and model selection may be seen as the same task from a Bayesian perspective; the duality of p -values and confidence intervals enables both to be used for estimation and hypothesis testing from a frequentist perspective; model selection may be used for the purpose of description, prediction or estimation.

Data-generating mechanism

The data-generating mechanism (often also called data-generating process) corresponds to the process of simulating data sets for assessing the performance of the compared methods in accordance with the aims of the simulation study. In general, the data sets can be simulated from a known *parametric model* or by *resampling* an existing real data set. Sometimes part of the simulation uses real data and part simulates data, which is the basis of “plasmode simulation” (Franklin, Schneeweiss, Polinski, & Rassen, 2014).

When simulating from a parametric model, researchers need to specify the data-generating mechanism. For example, the data sets can be generated from a normal distribution with varying values of the underlying true parameters. These can be determined either based on theory and previous research (e.g., depression scores in clinically depressed or healthy populations), estimated from an existing data set, specified according to conventional thresholds (e.g., small, medium, and large effect sizes), or set to arbitrary values to test performance across a wide variety of conditions. A common feature is, however, that the true values of the data-generating mechanism are known by the researcher and can be used to evaluate the performance of the compared methods. Especially when data generation is not based on a real-data model, the choice of data-generating parameters should be explained and justified to enable an understanding of the choice for readers, as well as for researchers wishing to perform similar simulation studies. In some cases this may be “extreme” to see when and how methods break, or not, as the case may be.

If multiple factors are varied, there are different possible ways to combine them: *fully factorial* (all possible combinations), *partially factorial* (considering some combinations but not all possible ones), *one-at-a-time* (varying one factor while holding the other/s constant), or *scattershot* (creating a set of distinct conditions). The fully factorial approach is typically preferred because it allows us to disentangle the individual effects of the factors and their interactions, but it may not always be feasible computationally or because some combinations of factors make no sense. For example, in a simulation study involving missing data, we may wish to vary the proportion of missing data and the missing data mechanism. When the proportion is zero, the mechanism is not applicable. Complex simulation designs can also

make the reporting and interpretation of results more difficult. To reduce the complexity of the design, a partially factorial design may then be chosen, for which [Morris et al. \(2019\)](#) and [Skrondal \(2000\)](#) give recommendations.

Figure 1

Example of Different Ways to Combine Factors in the Design of a Simulation Study.

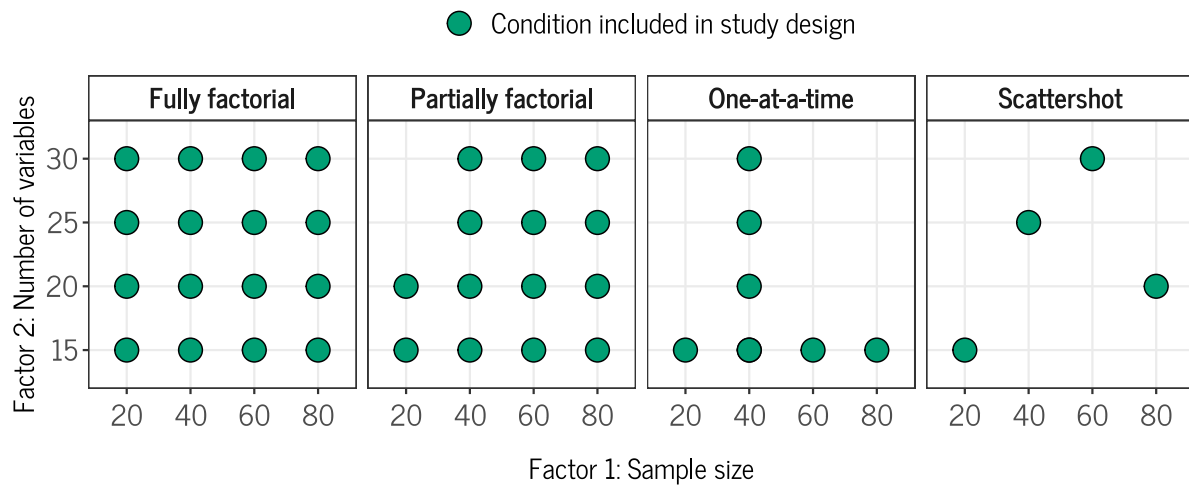


Figure 1 gives an example of how two factors, sample size and the number of variables, could be combined for a simulation study comparing different regression methods: The fully factorial approach would include all possible combinations (left panel). However, this may not be possible because, for example, the regression methods under study may not be able to handle situations where the number of variables is greater than the sample size. In this case, these conditions may be excluded and a partially factorial design adopted (middle left panel). With the one-at-a-time approach, one may fix the sample size to a value of 40 and then vary the number of variables across all levels, and vice versa, fix the number of variables to 15 and vary the sample size across all levels (middle right panel). Finally, with the scattershot approach, one may create distinct conditions of sample size and number of variables, for example, inspired by actual data sets that feature these combinations (right panel).

When resampling an existing data set, researchers rely on a (usually large) existing data set to sample smaller data sets for the simulation. Alternatively, one may sample equally large data sets with replacement from the existing data set. The data-generating mechanism is thus implicitly determined by the data set while researchers only need to specify the

resampling mechanism.

Estimands and other targets

Estimands and other targets jointly refer to the practical aims of the compared methods, Table 2 provides an overview of common targets of simulation studies. For example, if a simulation study aims to compare different methods for estimating the effect of an intervention *versus* the absence of that intervention, the estimand of interest is a contrast of these groups rather than, say, a group mean. An estimand is a target quantity of a statistical analysis (see [ICH, 2019](#); [Keene, Lynggaard, Englert, Lanius, & Wright, 2023](#); [Lundberg, Johnson, & Stewart, 2021](#), for accessible introductions to estimands). In simulation studies, an estimand is typically, but not always, a parameter of the underlying data-generating model. When it is not, care is needed to define and compute the true or ideal value of an estimand. If the simulation study aims to compare different methods for hypothesis testing, the true hypothesis is the target of interest. Again, care is needed to distinguish between different ways of translating substantive hypotheses into statistical hypotheses (e.g., whether a null hypothesis of no effect is specified as a sharp null hypothesis or an average null hypothesis). Similarly, the targets appropriate for the statistical tasks of other simulation studies include the true model (when the statistical task is model selection), the design characteristics (when the statistical task is design), or new data (when the statistical task is prediction).

Methods

The term “methods” corresponds to the different procedures evaluated in the simulation study for estimating parameters, testing hypotheses, predicting new data, etc. There is some ambiguity about what exactly constitutes a method in simulation studies. Often, different methods correspond to different statistical procedures, e.g., different tests for assessing publication bias such as the Egger regression or selection-model tests. Methods can also encompass different specifications or settings of a single statistical procedure, e.g., different parameterizations of covariance structures in mixed-effect models, different metrics of a statistical procedure, e.g., fit indices, or different software implementations of the same procedure, or the performance of different variance estimators for a given point estimator, e.g. model-based, robust, and bootstrap based standard errors. Less often, a simulation study

evaluates only a single method, for instance, to verify that the method can recover the targeted estimands in the first place or to study robustness when its assumptions are violated. Of course, it may also be the case that no competing method is available.

If only a single method is evaluated and there is no clear benchmark for what constitutes “good” performance (e.g., the successful control of the type I error rate at 5% in the case of a statistical test), the lack of a comparator method can introduce ambiguity into a simulation study. If multiple methods are compared, caution is required along various steps of planning and conducting a simulation study to ensure that comparisons are both neutral and meaningful. As in research on new drugs or treatments, methodological researchers have warned against over-optimism in the evaluation of new statistical methods for some time (e.g., [Boulesteix, 2015](#)). Such over-optimism can occur, among other reasons, when researchers are not neutral with respect to the evaluation of a method but rather choose data-generating mechanisms that favor a certain method ([Nießl, Hoffmann, Ullmann, & Boulesteix, 2023](#)). Neutral comparisons aimed at identifying data characteristics that determine the performance of the investigated methods (and their contextualized relative advantages) are often more informative than searching for specific conditions under which a method appears to be the “best” ([Strobl & Leisch, 2022](#)).

Comparative simulation studies can benefit from approaches that decrease over-optimism and allegiance bias used in other scientific fields such as experimental psychology or clinical trials. These include blinding the data analysts to the method ([Pawel et al., 2023](#)) or using separate research teams for data simulation and analysis ([Kreutz et al., 2020](#)). Further, “adversarial collaboration”, the collaboration of researchers with different theoretical or methodological views ([Cowan et al., 2020](#); for an example see [Binder, Sauerbrei, & Royston, 2012](#)), could be introduced to simulation studies to achieve useful comparisons between different methods. Researchers can also build on previous research by combining the conditions and methods of previous simulation studies into a single, large simulation study, extending previous simulation designs when necessary, to assess the robustness of their results to different experimental settings that have already been investigated in isolation by others (see [Bartoš, Maier, Wagenmakers, Doucouliagos, &](#)

[Stanley, 2023](#); [Hong & Reed, 2021](#), for an example).

Performance measures

Performance measures are the summary statistics used for quantifying how well methods can achieve their task for a given data-generating mechanism. For instance, a performance measure may quantify how well a method can estimate an estimand. As such, the estimated performance corresponds to the “inferences” of a simulation study that allow researchers to draw conclusions about the methods. The selection of appropriate performance measures depends on the aims of the simulation study, but also the estimands and other targets. For example, bias, (root) mean square error, and confidence interval coverage can be used to evaluate methods for estimating intervention effects, while power and type I error rate might be used to evaluate methods for testing hypotheses about publication bias. Table 2 shows typical performance measures for different simulation study aims.

The same statistical method may be applied for different statistical tasks and in different contexts, such as estimation and prediction, for which different performance measures can be used. Typically, multiple performance measures for a method should be interpreted together ([Morris et al., 2019](#)). In general, providing a rationale for the choice of performance measure as well as defining it clearly (ideally, with a formula-based representation) avoids ambiguity. This is especially important when less familiar performance measures are used, and when performance is estimated conditional on some sample statistic (e.g., bias of a study given that it converged in a given simulated data set).

An important aspect that affects the estimation and interpretation of performance measures is the convergence of methods. By “convergence”, we mean broadly that a method successfully produces the outcomes of interest (e.g., an estimate, a prediction, a p -value, a sample size, etc.) required for estimation of performance. Although convergence may not always be the main interest of a simulation study, it should be reported whether or not non-convergence occurred, and if so, under what conditions and for which methods. Non-convergent repetitions are the “missing values” of a simulation study, and they impact the interpretation of other measures, since these can only be estimated when a method converged. This is complicated because one method may converge more often than others, and so the

comparison conditional-on convergence is delicate. The way non-convergent repetitions are handled can have a major impact on results and conclusions, especially if repetitions are not missing at random. For example, if a method fails to converge under the most challenging conditions (e.g., small sample sizes), excluding such repetitions only for the problematic method while keeping the results for all other methods that did converge may bias performance measure estimates in favor of the excluded method.

Table 2

Different Types of Statistical Tasks, their Target(s), and Typical Performance Measures.

Statistical Task	Target(s)	Typical performance measures
Estimation	Estimand(s) (True effects/parameters)	Bias, variance, mean square error, coverage, CI width
Hypothesis testing	True hypothesis(es)	Type I error rate, power
Model selection	True model(s)	Model-recovery rate, sensitivity, specificity
Prediction	New data	Prediction error, calibration, discrimination
Clustering	True cluster(s)	Fidelity to the true cluster structure
Design	Design characteristics	Expected sample size, minimum/maximum sample size, power, precision (for a fixed sample size)

Note. Table adapted from [Morris et al. \(2019, Table 3\)](#).

Table 3 provides definitions of the performance measures that were most commonly used in our literature review of simulation studies in psychology (which will be presented in the next section). We refer to Table 6 in [Morris et al. \(2019\)](#) for a definition and MCSEs of various other performance measures, such as bias-eliminated coverage or average model standard error, that are less frequently used in psychology. Performance measures estimated from simulated data are subject to sampling variability, similarly to any other quantity estimated from a finite set of data ([Koehler et al., 2009](#)). Much like the sample size in other empirical studies, the number of simulation repetitions n_{sim} determines the precision of these estimates. Table 3 therefore provides formulae for approximate Monte Carlo standard errors (MCSEs) associated with the estimated performance measures. All MCSEs are based on the assumptions of independent simulations and approximate normality of the estimated

Table 3

Definitions of Common Performance Measures, their Estimates, Monte Carlo Standard Errors (MCSE), and Number of Simulation Repetitions n_{sim} to Achieve a Desired MCSE_{}.*

Performance measure	Definition	Estimate	MCSE	n_{sim}
Bias	$E(\hat{\theta}) - \theta$	$(\sum_{i=1}^{n_{\text{sim}}} \hat{\theta}_i / n_{\text{sim}}) - \theta$	$\sqrt{S_{\hat{\theta}}^2 / n_{\text{sim}}}$	$S_{\hat{\theta}}^2 / \text{MCSE}_*^2$
Relative bias	$\{E(\hat{\theta}) - \theta\} / \theta$	$\{(\sum_{i=1}^{n_{\text{sim}}} \hat{\theta}_i / n_{\text{sim}}) - \theta\} / \theta$	$\sqrt{S_{\hat{\theta}}^2 / (\theta^2 n_{\text{sim}})}$	$S_{\hat{\theta}}^2 / (\text{MCSE}_*^2 \theta^2)$
Mean square error (MSE)	$E\{(\hat{\theta} - \theta)^2\}$	$\sum_{i=1}^{n_{\text{sim}}} (\hat{\theta}_i - \theta)^2 / n_{\text{sim}}$	$\sqrt{S_{(\hat{\theta}-\theta)^2}^2 / n_{\text{sim}}}$	$S_{(\hat{\theta}-\theta)^2}^2 / \text{MCSE}_*^2$
Root mean square error (RMSE)	$\sqrt{E\{(\hat{\theta} - \theta)^2\}}$	$\sqrt{\sum_{i=1}^{n_{\text{sim}}} (\hat{\theta}_i - \theta)^2 / n_{\text{sim}}}$	$\sqrt{S_{(\hat{\theta}-\theta)^2}^2 / (4n_{\text{sim}} \widehat{\text{MSE}})}$	$S_{(\hat{\theta}-\theta)^2}^2 / (4\widehat{\text{MSE}} \text{MCSE}_*^2)$
Empirical variance	$\text{Var}(\hat{\theta})$	$S_{\hat{\theta}}^2$	$S_{\hat{\theta}}^2 \sqrt{2 / (n_{\text{sim}} - 1)}$	$1 + 2(S_{\hat{\theta}}^2)^2 / \text{MCSE}_*^2$
Empirical standard error	$\sqrt{\text{Var}(\hat{\theta})}$	$\sqrt{S_{\hat{\theta}}^2}$	$\sqrt{S_{\hat{\theta}}^2 / \{2(n_{\text{sim}} - 1)\}}$	$1 + S_{\hat{\theta}}^2 / (2\text{MCSE}_*^2)$
Coverage	$\Pr(\text{CI includes } \theta)$	$\sum_{i=1}^{n_{\text{sim}}} \mathbb{1}(\text{CI}_i \text{ includes } \theta) / n_{\text{sim}}$	$\sqrt{\widehat{\text{Cov}}(1 - \widehat{\text{Cov}}) / n_{\text{sim}}}$	$\widehat{\text{Cov}}(1 - \widehat{\text{Cov}}) / \text{MCSE}_*^2$
Power (or type I error rate)	$\Pr(\text{Test rejects } H_0)$	$\sum_{i=1}^{n_{\text{sim}}} \mathbb{1}(\text{Test}_i \text{ rejects } H_0) / n_{\text{sim}}$	$\sqrt{\widehat{\text{Pow}}(1 - \widehat{\text{Pow}}) / n_{\text{sim}}}$	$\widehat{\text{Pow}}(1 - \widehat{\text{Pow}}) / \text{MCSE}_*^2$
Mean CI width	$E(\text{CI}_{\text{upper}} - \text{CI}_{\text{lower}})$	$\sum_{i=1}^{n_{\text{sim}}} (\text{CI}_{i,\text{upper}} - \text{CI}_{i,\text{lower}}) / n_{\text{sim}}$	$\sqrt{S_W^2 / n_{\text{sim}}}$	S_W^2 / MCSE_*^2
Mean of generic statistic G	$E(G)$	$\sum_{i=1}^{n_{\text{sim}}} G_i / n_{\text{sim}}$	$\sqrt{S_G^2 / n_{\text{sim}}}$	S_G^2 / MCSE_*^2

Note. Table adapted from Table 6 in [Morris et al. \(2019\)](#)

$E(X)$ and $\text{Var}(X)$ are the expected value and variance of a random variable X , respectively. Summation is denoted by $\sum_{i=1}^n x_i = x_1 + x_2 + \cdots + x_{n-1} + x_n$.

$\hat{\theta}$ is an estimator of the estimand θ , and $\hat{\theta}_i$ is the estimate obtained from simulation i

$\mathbb{1}(\text{CI}_i \text{ includes } \theta)$ and $\mathbb{1}(\text{Test}_i \text{ rejects } H_0)$ are 1 if the respective event occurred in simulation i and 0 otherwise

$\widehat{\text{MSE}}$, $\widehat{\text{Cov}}$, and $\widehat{\text{Pow}}$ denote the estimated MSE, coverage, and power, respectively. MCSE_* denotes the desired MCSE when calculating the number of repetitions n_{sim} .

The sample variance of the estimates is $S_{\hat{\theta}}^2 = \sum_{i=1}^{n_{\text{sim}}} \{\hat{\theta}_i - (\sum_{i=1}^{n_{\text{sim}}} \hat{\theta}_i / n_{\text{sim}})\}^2 / (n_{\text{sim}} - 1)$

The sample variance of the square errors is $S_{(\hat{\theta}-\theta)^2}^2 = \sum_{i=1}^{n_{\text{sim}}} [(\hat{\theta}_i - \theta)^2 - \{\sum_{i=1}^{n_{\text{sim}}} (\hat{\theta}_i - \theta)^2 / n_{\text{sim}}\}]^2 / (n_{\text{sim}} - 1)$

The sample variance of the CI widths is $S_W^2 = \sum_{i=1}^{n_{\text{sim}}} [(\text{CI}_{i,\text{upper}} - \text{CI}_{i,\text{lower}}) - \{\sum_{i=1}^{n_{\text{sim}}} (\text{CI}_{i,\text{upper}} - \text{CI}_{i,\text{lower}}) / n_{\text{sim}}\}]^2 / (n_{\text{sim}} - 1)$

The sample variance of a generic statistic G is $S_G^2 = \sum_{i=1}^{n_{\text{sim}}} \{G_i - (\sum_{i=1}^{n_{\text{sim}}} G_i / n_{\text{sim}})\}^2 / (n_{\text{sim}} - 1)$ with G_i the statistic obtained from simulation i . For example, G may be a measure of predictive performance.

performance measures. More accurate jackknife-based MCSEs are available through various R packages such as `rsimsum` (Gasparini, 2018) and `simhelpers` (Joshi & Pustejovsky, 2022). The `SimDesign` R package (Chalmers & Adkins, 2020) can compute confidence intervals for performance measures *via* bootstrapping.

MCSEs (or other measures of uncertainty) should be provided alongside the estimates of performance to indicate the associated uncertainty. Failing to calculate and report Monte Carlo uncertainty can lead to erroneous interpretations of results and unsupported claims about the performance of different methods (see, e.g., the illustration by Koehler et al., 2009). In situations where MCSEs are tiny relative to the estimated performance and may distract, one could, for example, provide the maximum MCSE across all conditions to give the reader reassurance about the worst case.

When planning a simulation study, researchers should choose a number of simulation repetitions that ensures a desired precision for estimating the chosen performance measures. The last column of Table 3 gives simple formulae for this purpose. Many of these depend on quantities that are not known but have to be estimated from the simulated data. For example, the MCSE of the estimated coverage depends on the coverage itself. In this case, one can either assume a certain value for which the desired MCSE should be achieved (e.g., 95%), take a “worst-case value” (this would be 50% for coverage), or estimate it from a small pilot study (e.g., taking the estimated coverage closest to 50% across all conditions and methods obtained in the pilot study). The latter approach may be especially advisable for performance measures where there is no conventional benchmark, such as 95% is for coverage.

Reporting

As with any experiment, transparent reporting of study design, execution, and results is essential to put the outcomes from a simulation study into context. The ADEMP structure is a useful template for researchers to follow when reporting the design and results of their simulation study. Furthermore, the results should be reported in a way that clearly answers the main research questions and acknowledges the uncertainty associated with the estimated performance. It is often difficult to find a balance between streamlining the results of simulation studies for the reader and exhaustively reporting all conditions in detail. However,

it is important that researchers avoid selectively reporting only certain conditions that favor their preferred method or are in line with their expectations, as this can lead to overoptimism (Pawel et al., 2023).

Figures are often helpful for interpreting large quantities of results and identifying general trends. However, for most plot types, there is a limit to how many factors can be communicated visually (see section 7.2 in Morris et al., 2019 for some recommendations, see also Rücker & Schwarzer, 2014). On the other hand, presenting results only with figures can hinder the accurate interpretation of results and also make it more difficult for researchers replicating the simulation study to verify whether they have been successful (Luijken et al., 2023). Figures should therefore ideally be combined with quantitative summaries of results, such as tables. For complex simulation designs with a large number of conditions, the communication of results can be improved using interactive tools such as R Shiny applications (Chang et al., 2023, see e.g., Carter, Schönbrodt, Gervais, & Hilgard, 2019; Gasparini, Morris, & Crowther, 2021).

Computational aspects

The computational implementation of simulation studies can often be complex. Conclusions critically depend on the soundness of the underlying code for data-generation, model fitting, and computations of performance measures, and even small mistakes can have a big impact (Schwab & Held, 2021). Morris et al. (2019); White et al. (2023) give detailed advice on how to code a simulation study, here we want to focus more on the aspects of reproducibility and code sharing. Code that is not openly available online prevents an assessment of the computational reproducibility of simulation studies. It is also an obstacle for reviewers and readers who want to understand, inspect, or replicate the implementation of a simulation study, or for researchers who seek to build on the previous literature. It is therefore recommended for researchers to make their code openly available, share all relevant information about their computational environment, and strive to use a robust computational workflow to ensure the reproducibility and replicability of their results (Luijken et al., 2023; Pawel et al., 2023).

Platforms such as the Open Science Framework and Zenodo can be used to

persistently store and share data and code, independently of specific journals and according to the FAIR principles (Wilkinson et al., 2016). The computational reproducibility of simulation studies can be further enhanced by sharing complete or intermediate results of simulation studies, such as parameter estimates of computed models. This enables independent reproduction and evaluation of the results by other researchers without the full computational effort that large simulation studies require.

Information on the computational environment and operating system is relevant to reproduce simulation studies. Different software packages or package versions can lead to different results, even when the apparently same method is used (Hodges et al., 2023). Operating systems can differ in a variety of aspects that may subtly influence the results of analyses (Glatard et al., 2015). There are several helpful tools that facilitate sharing information on the computational environment and operating system. For example, when using R, the output of the `sessionInfo()` command includes information about the operating system, R package versions, and auxiliary dependencies (e.g., the installed linear algebra programs such as BLAS/LAPACK). Furthermore, Peikert and Brandmaier (2021) and Epskamp (2019) provide accessible tips for reproducible workflows in R, which can serve as a starting guide for other statistical software as well. For instance, in advanced workflows, a snapshot of the current version of all software required to reproduce the analysis is stored (e.g., via Docker or the R package `renv`, Ushey & Wickham, 2023).

An important computational aspect of simulation studies is the use of pseudo-random numbers. It is important to initialize the random number generator with a seed and store this seed, so that the same sequence of pseudo-random numbers can be reproduced in the future (assuming other dependencies, such as operating system and software versions, remain the same). The primary purpose of the seed is to ensure computational reproducibility and to facilitate debugging. At the same time, the seed should not matter for simulation studies with a sufficient number of repetitions, because the seed should have a negligible effect on the results (estimated performance measures, patterns, and conclusions). Morris et al. (2019) recommend to set the seed *once and only once* at the beginning of the entire simulation study, and then store the state of the random number generator at various places (e.g., at the start of a

condition/iteration) to be able to reproduce the results at that place (e.g., to inspect why an error occurred). The reason for this recommendation is that if the seed is set multiple times (e.g., at the start of every condition), the generated pseudo-random numbers may become dependent because, by chance, the deterministic sequence of the generated pseudo-random numbers can overlap.¹ However, modern random number generators typically have extremely long periods before a sequence of pseudo-random numbers repeats. For example, the Mersenne Twister used in R has a period of $2^{19937} - 1$, rendering the chance of correlated simulations rather small (but not zero).² Things become more complicated when multiple cores, clusters, or computers are used for running the simulation study since the seed has to be set for each parallelized instance to ensure reproducibility. One solution is to use “streams” rather than seeds, which fixes the random number generator to the actual starting position in the deterministic sequence of generated numbers (Morris et al., 2019). Streams are available in Stata, SAS, and R³. However, when using streams, one needs to know how many pseudo-random number are required per instance, so that the streams can be set to avoid overlap. This can be challenging, especially when the methods evaluated in the simulation study also use pseudo-random numbers (e.g., Markov Chain Monte Carlo sampling or bootstrap methods).

Literature review

In this section, we use the ADEMP structure to assess the current state of simulation studies in psychology. For each ADEMP component, we summarize the findings, highlight their relevance, and suggest improvements for future simulation studies. We compare some of our results with the results of Morris et al. (2019) who reviewed 100 simulation studies

¹ The following code executed with R version 4.3.1 illustrates two correlated sequences of pseudo-random numbers: `set.seed(0); x<-runif(50); set.seed(1); y<-runif(50); plot(x[-1], y[-50])`

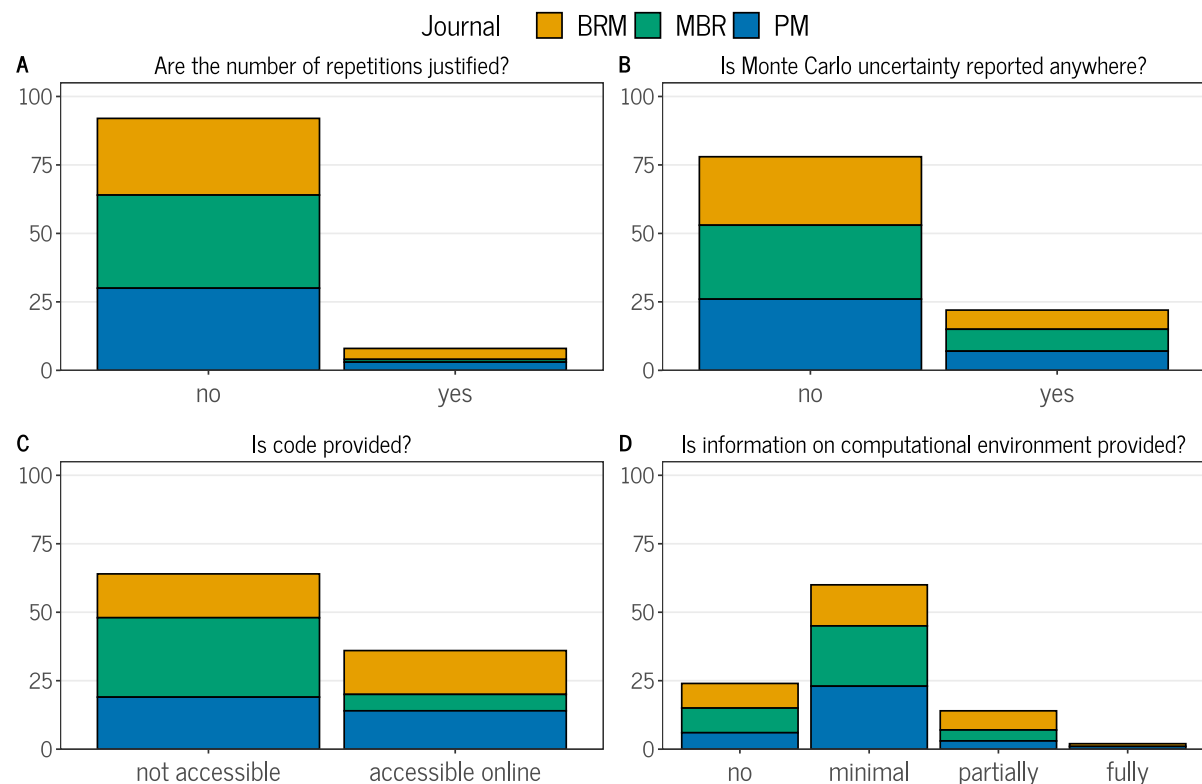
² The authors have had several intense discussions among themselves and with other researchers about whether or not it is problematic to set the seed once per condition (e.g., the default behavior of the `SimDesign` package (Chalmers & Adkins, 2020), which randomly generates starting seeds in the range 1 to 2,147,483,647 for each condition). Fact is, there is a non-zero probability that sequences of pseudo-random numbers may overlap, but the probability of this happening can be minuscule and depends on how many conditions and random numbers are used in the study. In our opinion, the dangers associated with this behavior are negligible as long as enough simulation repetitions are used.

³ Random number streams are available in R through different packages, e.g., `parallel` (R Core Team, 2023), `rstream` (Leydold, 2022), or `doRNG` (Gaujoux, 2023).

published in *Statistics in Medicine*. Visual summaries of the review are provided in Figure 2 and Figure 3. The preregistration, data, and code to reproduce the results are available at the Open Science Framework (<https://osf.io/dfgvu/>).

Figure 2

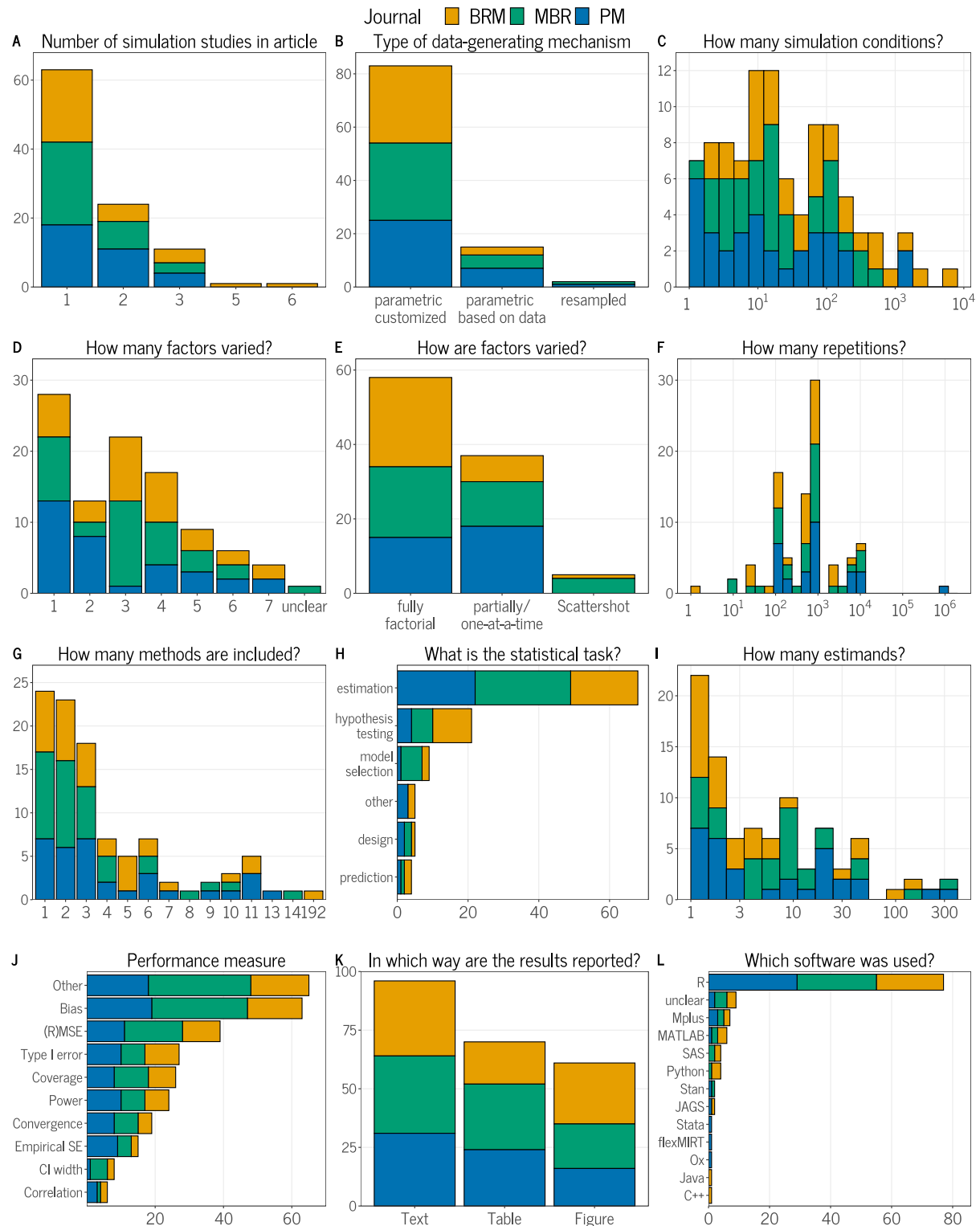
Common Issues of Simulation Studies in Psychology as Identified in the Literature Review.



Note. 100 articles were reviewed that included simulation studies and were published in *Psychological Methods*, *Behavioral Research Methods*, and *Multivariate Behavioral Research* in 2021 and 2022.

We extracted 321 articles until we reached 100 articles containing at least one simulation study. We extracted articles by going through the 2022 issues of the journals in chronological order. After assessing the number of articles containing a simulation study from each journal, we then continued chronologically in the 2021 issues, aiming for a roughly equal split of simulation studies from the three journals.⁴ The proportion of articles containing a simulation study (31.2%) was considerably lower than the 75.4% proportion reported by

⁴ Due to an oversight, we did not review the last issue of BRM in 2022 but rather continued with the first 2021 issues of the journals.

Figure 3*Descriptive Results from Literature Review of Simulation Studies in Psychology.*

Note. 100 articles were reviewed that included simulation studies and were published in *Psychological Methods*, *Behavioral Research Methods*, and *Multivariate Behavioral Research* in 2021 and 2022. In Panel J, absolute and relative bias are combined in the bias category. In Panel E, partially factorial and one-at-a-time are combined. Within-panel totals are greater than 100 in panels H, J, K, L due to the possibility of more than one category.

[Morris et al. \(2019\)](#) for the 100 simulation studies published in *Statistics in Medicine*. The lower proportion of simulation studies in our review is mainly due to articles in *BRM*, which generally published the most articles, but only 15.6% of them contained a simulation study. We extracted roughly equal numbers of articles containing a simulation study from the three journals, with 32 from *BRM* and 35 each from *PM* and *MBR*. Of these articles, 63 contained only a single simulation study, while the rest contained up to 6 simulation studies (see Panel A of Figure 3).

Three authors (BS, FB, SP) each reviewed around one-third of all simulation studies and assessed the overall confidence in their rating of each study as “low”, “medium”, or “high”. To assess inter-rater agreement, each rater also reviewed six studies that were assigned to the other raters and which had a “low” or “medium” confidence rating, thereby representing the most challenging simulation studies that were reviewed. Nevertheless, an agreement larger than 75% was achieved for the majority of questions (Median = 83.3%). The lowest agreement was with respect to whether the estimands were stated and the number of estimands (above 30%). All studies where we disagreed about the number of estimands were studies about latent variable models, where it was often unclear which parameters were of interest and how their number varied across conditions, with many studies even showing varying numbers of parameters per condition. The results of the agreement analysis are shown in Figure 5 in the Appendix.

Aims

In 94% of the reviewed articles, the aims of the study were defined in some form. We did not rate how specific or vague the aims were defined. By far most studies had estimation as one of their statistical tasks (68%), followed by hypothesis testing (21%) and model selection (9%; Panel H in Figure 3). This resembles the results of [Morris et al.](#), who also found these three tasks to be the most prominent ones with similar frequencies.

Data-generating mechanism

In our review, the clear majority of simulation studies (83%) generated data based on parametric models with parameters specified by researchers (‘parametric customized’), while 15% were directly based on parameter estimates from real data (Panel B of Figure 3). The

remaining 2% used resampling techniques. In almost all of the studies (95%), the data-generating parameters were provided, which mirrors the results from [Morris et al.](#) (91% studies).

Researchers used between 1 and 6,000 simulation conditions (Median = 16; Panel C in Figure 3). In these, they varied between 1 and 7 factors, with 1 and 3 being the most common choices (Panel D in Figure 3). Of all designs, 58% were fully factorial, meaning that all possible combinations of factor levels were investigated. Moreover, 37% of the studies were either partially factorial or varied factors one-at-a-time (including studies with a single design factor) and 5% used distinct scenarios in a scattershot design (Panel E in Figure 3). As in experimental psychology, a fully factorial design enables the study of the main and interaction effects of the varied factors. In our review, some studies made use of this fact by using analysis of variance to assess the effects of simulation factors (see also [Chipman & Bingham, 2022](#)).

The number of repetitions per simulation condition ranged between 1 and 1,000,000 (Panel F in Figure 3). The median number was 900, whereas the most frequently selected options were 1,000 repetitions followed by 500 repetitions, similar to the results from [Morris et al.](#) Only 8% of the studies provided a justification for the specific number of repetitions used, while only 3% of these actually performed a sample size calculation (Panel A in Figure 2). This is very similar to the results from [Morris et al.](#), who also found only 4% of studies presenting a justification for their choice of the number of repetitions. This lack of justification is, unfortunately, consistent with the findings from similar surveys of the methodological literature ([Harwell et al., 2018](#); [Hauck & Anderson, 1984](#); [Hoaglin & Andrews, 1975](#); [Koehler et al., 2009](#)).

Estimands and other targets

In 20% of the studies, the estimands or targets of the simulation were either not reported or unclear to us. Of those that were clear, most studies focused only on a single estimand, while the median number of estimands was 4. In at least 17% of the studies, estimated performance measures related to different estimands were later aggregated to calculate average performance, while this was unclear in 4% of studies. We noticed that especially when evaluating models with many parameters, such as latent variable models or

certain time series models, it can easily become unclear which parameters are of interest. Clear definition and reporting of estimands and (potentially aggregated) performance measures is particularly important in these situations.

Methods

While the number of methods evaluated in the simulation studies ranged from 1 to 192, more than half (65%) evaluated 3 or fewer methods, and 24% evaluated only a single method (Panel G in Figure 3).

Performance measures

Reflecting the popularity of estimation as a statistical task, bias (used in 63% of studies) and (root) mean square error ([R]MSE, in 39% of studies) were the most common performance measures in our review (Panel J in Figure 3). Convergence was reported only in 19% of the studies. In 10% of the studies, performance measures were unclear, for example, how a certain performance measure was defined mathematically. Many studies also included other performance measures not explicitly listed here. For example, the correlation between true and estimated parameters was sometimes used as a measure of performance. While there may be cases where this metric provides valuable information, interpreting it without considering the bias and variance of the estimates gives only a very limited insight into the performance of a statistical method. A positive example of clear reporting of both model specifications and performance measures is [H. Liu, Yuan, and Wen \(2022\)](#), who provided formulae for both their models and the performance measures used.

Presentation of results

Simulation results were most commonly reported in the text of an article and accompanied by tables and figures (Panel K in Figure 3). The vast majority of studies (77%) did not report the uncertainty of performance measures (Panel B in Figure 2), despite our liberal approach of including visualizations such as boxplots as indicative of Monte Carlo uncertainty. The proportion is comparable to the stricter approach of [Morris et al. \(2019\)](#) who counted 93% of their studies not reporting Monte Carlo standard errors. To cite two positive examples from our review, [J. Liu and Perera \(2022\)](#) ran a pilot simulation study to obtain the empirical standard errors for parameter estimates, which they then used to calculate the

needed sample size (i.e., number of repetitions) to keep the MCSE below a desired level.

Rodgers and Pustejovsky (2021) provided the upper bound of the MCSE of their performance measures to indicate their precision.

Computational aspects

R was the most commonly used statistical software to conduct simulation studies and was used in 77% of the studies (Panel L in Figure 3). Notably, the software used was unclear or not mentioned in 9% of the studies. In Morris et al. (2019), 38% of the studies did not mention the software used for their simulation study. In around half of the studies we reviewed, authors also indicated that they used some form of user-written commands, such as custom model code, or packages for their simulations. To fully understand these simulations, it would be crucial to share code alongside the manuscript. However, code was not available for almost two-thirds of the simulation studies (64%; Panel C in 2). This also includes cases in which code was supposed to be provided, but the repository was not available, and cases in which code was supposedly available “upon (reasonable) request”. In multiple cases, authors supposedly provided code on the journal website or on a university homepage, but the code was not available at the designated location. Our results are similar to the findings of Kucharskỳ, Houtkoop, and Visser (2020), who analyzed articles in three methodological journals (including *Psychological Methods* and *Behavior Research Methods*) and found that 56% of studies that contained coded analyses did not share their code. Of the 36% of studies in our review in which code was provided, 21% also provided a seed in their code. We did not check if this seed and the supplied code would be sufficient to reproduce the reported results.

Beyond the code and software used, we reviewed whether articles contained information on the computational environment and operating system used. We coded information on the computational environment as “fully” when packages with versions and auxiliary dependencies were provided, for example in a “sessionInfo” output from R or via a Docker container. We rated the information as “partially” or “minimal” when the main packages used were reported with or without versions, respectively. Full information on the computational environment was only reported in 2% of the studies, while 24% did not report on their computational environment at all (Panel D in Figure 2). Even more studies (93%) did

not provide any information whatsoever on their operating system. Full information (naming the operating system and its version) was provided in 4% of the studies, while 3% at least provided the operating system without stating its version. Papenberg and Klau (2021) are a positive example that included full information on their computational environment and operating system which they used, as well as code and data to reproduce the simulations.

The ADEMP-PreReg template

Our literature review highlights varying standards of reporting of simulation studies in psychology. To promote more structured and detailed reporting and to simplify preregistration of simulation studies, we developed the ADEMP-PreReg template. This template closely follows the outlined ADEMP structure and provides a list of questions, their explanation, and example answers. This prompts researchers to describe all relevant parts of their simulation studies a priori, and the template can additionally be used as a blueprint for reporting. As such, the ADEMP-PreReg template is not only suited for experienced researchers who want to plan, preregister, or report their simulation studies but also for (under)graduate students embarking on a first exploration of simulation studies. Once a simulation study is conducted, the ADEMP-PreReg template can be transformed into the method section of the simulation study with minimal effort. The template is available at GitHub (<https://github.com/bsiepe/ADEMP-PreReg>) in different versions (L^AT_EX, Microsoft Word, Google Docs, Overleaf) and can be uploaded and timestamped at OSF, AsPredicted, or Zenodo among others. The ADEMP-PreReg template is intended as a “living document” and we welcome feedback and suggestions via opening an issue or a pull request on the GitHub repository.⁵

It is worth emphasizing that preregistration of simulation studies has many parallels, but also several differences to the preregistration of other studies. Traditionally, preregistration is often used to distinguish between exploratory and confirmatory research. This distinction is more blurry for simulation studies, as it is often not clear when “data collection” starts and which part of the research process can be considered confirmatory. This parallels the

⁵ We used version 0.1.0 of the template in this example as archived at <https://zenodo.org/records/10057884>.

challenges with preregistrations of secondary analyses on observational data (den Akker et al., 2021). At the same time, preregistration in the traditional sense may be more appropriate in later stages of methodological research (Heinze et al., 2023) when researchers attempt to neutrally compare already well-established methods. Beyond that, preregistration of simulation studies can serve many other purposes. It can help to structure the planning of a study, guard against cognitive biases in the interpretation of results, promote a minimum degree of neutrality and transparency, and save work in the preparation of manuscripts. Finally, even if researchers do not want to preregister their simulation, the preregistration template can serve as a blueprint for planning and reporting.

Example simulation study on methods for the analysis of pre–post measurements

To illustrate the application of the preregistration template, we conduct a simulation study to evaluate different methods for analyzing data from pre–post measurement experiments. The filled-out ADEMP-PreReg template is available at <https://osf.io/dfgvu/>. We made a minor modification of the study design from the preregistration, increasing the number of simulation repetitions to achieve a lower MCSE than was originally planned since our preregistration did not guarantee a sufficiently small MCSE for the worst-case power. Importantly, the sole purpose of this simulation study is to illustrate both the template and the transparent reporting of results using easy-to-understand simulation conditions, not to contribute new knowledge to the literature. See Clifton and Clifton (2019); Lüdtke and Robitzsch (2023); Senn (2006); Van Breukelen (2013); Vickers (2001) for a comprehensive treatment of the topic.

Aims

The aim of the simulation study is to evaluate the hypothesis testing and estimation characteristics of different methods for estimating the treatment effect in pre–post measurement experiments. We compare three different methods (ANCOVA, change score analysis, and post score analysis) in terms of power and type I error rate related to the hypothesis test of no effect, and bias related to the treatment effect estimate. We vary the true treatment effect and the correlation of pre- and post-measurements.

Data-generating mechanism

In each simulation repetition, we generate $n = 50$ pre–post measurements in the control group ($g = \text{control}$) and $n = 50$ pre–post measurements in the experimental group ($g = \text{exp}$) from a bivariate normal distribution

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \mathbf{N} \left(\begin{bmatrix} 0 \\ \mu_{g,2} \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right), \quad (1)$$

where the first argument of the normal distribution in (1) is the mean vector and the second argument the covariance matrix. The numerical subscript 1 indicates measurement time ‘pre’ and 2 indicates ‘post’. The parameter $\mu_{g,2}$ denotes the post-treatment mean. It is fixed to zero in the control group ($\mu_{\text{control},2} = 0$), whereas it is varied across simulation conditions in the experimental group. The parameter ρ denotes the pre–post correlation and is also varied across simulation conditions.

We use the following values for the manipulated parameters of the data-generating mechanism:

- $\mu_{\text{exp},2} \in \{0, 0.2, 0.5\}$
- $\rho \in \{0, 0.5, 0.7\}$

We vary the conditions in a fully factorial manner which results in 3 (post-treatment mean in experimental group) \times 3 (pre–post measurement correlation) = 9 simulation conditions. We select the specific values as they correspond to the conventions for no, small, and medium standardized mean difference effect sizes in psychology (Cohen, 2013). The pre–post measurement correlations correspond to no, one quarter, and approximately one half of shared variance that, based on our experience, are both realistic and also allow us to observe differences between the examined methods.

Estimands and other targets

Our primary target is the null hypothesis of no difference between the outcomes of the control and treatment groups. Our secondary estimand is the treatment effect size defined as the expected difference between the control and the experimental group measurements at

time-point two

$$E(Y_2 \mid g = \text{exp}) - E(Y_2 \mid g = \text{control}),$$

for which the true value is given by the parameter $\mu_{\text{exp},2}$ for the considered data-generating mechanisms.

Methods

We compare the following methods:

1. **ANCOVA** (ANalysis of COVariance): A regression of the post-treatment measurement using the pre-treatment measurement and the treatment indicator as covariates, which is specified in R as `lm(post ~ pre + treatment)`
2. **Change score analysis**: A regression of the difference between post-treatment and pre-treatment measurement using the treatment indicator as covariate, which is specified in R as `lm(post ~ offset(pre) + treatment)`
3. **Post score analysis**: A regression of the post-treatment measurement using the treatment indicator as covariate, which is specified in R as `lm(post ~ treatment)`

Both change score and post score analysis can be seen as special cases of ANCOVA. Change score analysis fixes the `pre` coefficient to 1 (using the `offset()` function), and post score analysis omits the `pre` variable from the model (effectively fixing its coefficient to 0).

Performance measures

Our primary performance measures are the type I error rate (in conditions where the true effect is zero) and the power (in conditions where the true effect is non-zero) to reject the null hypothesis of no difference between the control and treatment condition. The null hypothesis is rejected if the two-sided Wald test p -value for the null hypothesis of no effect is less than or equal to the conventional threshold of 0.05. The rejection rate (the type I error rate or the power, depending on the data generating mechanism) is estimated by

$$\widehat{\text{RRate}} = \frac{\sum_{i=1}^{n_{\text{sim}}} \mathbb{1}(p_i \leq 0.05)}{n_{\text{sim}}}$$

where $\mathbb{1}(p_i \leq 0.05)$ is the indicator of whether the p -value in simulation i is equal to or less than 0.05. We use the following formula to compute the MCSE of the estimated rejection rate

$$\widehat{\text{MCSE}}_{\widehat{\text{RRate}}} = \sqrt{\frac{\widehat{\text{RRate}}(1 - \widehat{\text{RRate}})}{n_{\text{sim}}}}.$$

Our secondary performance measure is the bias of the treatment effect estimate. It is estimated by

$$\widehat{\text{Bias}} = \frac{\sum_{i=1}^{n_{\text{sim}}} \hat{\theta}_i}{n_{\text{sim}}} - \theta$$

where θ is the true treatment effect and $\hat{\theta}_i$ is the effect estimate from simulation i . We compute the MCSE of the estimated bias with

$$\widehat{\text{MCSE}}_{\widehat{\text{Bias}}} = \frac{S_{\hat{\theta}}}{\sqrt{n_{\text{sim}}}}$$

where $S_{\hat{\theta}} = \sqrt{\frac{1}{n_{\text{sim}}-1} \sum_{i=1}^{n_{\text{sim}}} \{\hat{\theta}_i - (\frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \hat{\theta}_i)\}^2}$ is the sample standard deviation of the effect estimates.

Based on these performance measures, we perform 10,000 repetitions per condition. This number is determined by using the formulae from Table 3 in [Siepe et al. \(2023\)](#) aiming for 0.005 MCSE of type I error rate and power under the worst case performance (50% rejection rate: $0.50 \times (1 - 0.50) / 0.005^2 = 10,000$). Our simulation protocol also illustrates how to determine the number of repetitions for bias based on a small pilot simulation study to estimate the unknown effect estimate variance.

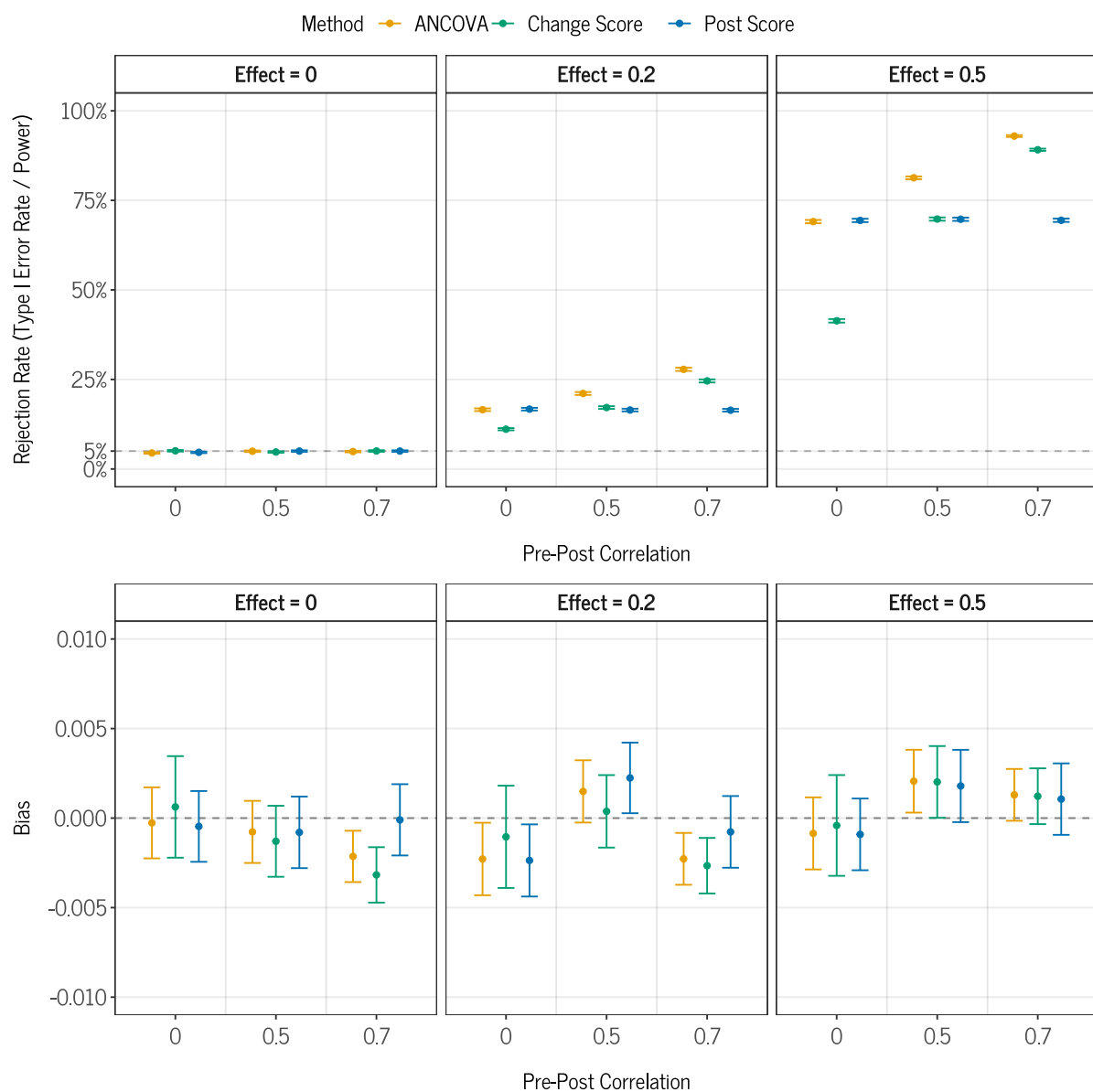
Computational aspects

The simulation study is performed using R version 4.3.1 ([R Core Team, 2023](#)) and the following R packages: the `mvtnorm` package (Version 1.2-3, [Genz & Bretz, 2009](#)) to generate data, the `lm()` function included in the `stats` package (Version 4.3.1, [R Core Team, 2023](#)) to fit the different models, the `SimDesign` package (Version 2.12, [Chalmers & Adkins, 2020](#)) to set up and run the simulation study, and the `ggplot2` package (Version 3.4.3, [Wickham, 2016](#)) to create visualizations. A `sessionInfo` output with more

information on the computational environment, and code and data to reproduce the study and its analysis are available at the Open Science Framework (<https://osf.io/dfgvu/>).

Figure 4

Estimated rejection rate (Power / Type I Error depending on DGM) and Bias of ANCOVA, Change Score Analysis and Post Score Analysis.



Note. Error bars correspond to ± 1 Monte Carlo standard error. The y-axis in the bias plot is scaled only from -0.01 to 0.01 , meaning that the bias can be considered negligible.

Results

Figure 4 shows the results of the simulation study visually, Table 5 in the Appendix shows the same results numerically. No missing/non-convergent values were observed. We see from the Effect = 0 panel/rows that all methods maintain a Type I error rate close to 5% irrespective of the correlation between the pre–post measurements. For non-zero effects, when the pre–post measurement correlation is zero, ANCOVA and post-score analysis exhibit similar levels of power and both surpass change-score analysis. However, when the pre–post measure correlation increases to 0.7, change-score analysis shows higher power than post-score analysis, yet ANCOVA shows higher power than both other methods.

The lower panels in Figure 4 show the estimated bias of the methods. We see that all methods had essentially equivalent bias across all simulation conditions. Furthermore, the bias of all methods in all conditions was close to zero and can be considered negligible relative to the MCSE.

In sum, under the investigated scenarios, all methods produced unbiased effect estimates while ANCOVA consistently showed the highest power among the three methods. In almost all conditions, the Type I error rate was within one MCSE of the nominal rate of 5%, and all differences to the nominal rate were smaller than 0.5%. For this simple setting and the methods under study, there is a substantial amount of statistical theory that explains and predicts our results (see, e.g., [Senn, 2008](#)), which is not often the case for simulation studies. Our findings are also in line with previous simulation studies ([Vickers, 2001](#)).

Discussion and recommendations

Our review of 100 articles published in prominent journals for methodological research in psychology shows considerable room for improvement in the design and reporting of simulation studies. The precision and uncertainty of estimated results are often neglected, as evidenced by the lack of justification for the number of simulation repetitions and the limited reporting of Monte Carlo uncertainty in most studies. The unavailability of openly accessible code and detailed information on the computational environment in most studies is a major barrier to critical evaluation, reproducibility, and incremental progress in methodological research. At the same time, we have also highlighted several positive

examples from the literature that stood out for their transparent reporting or clear sample size justification. These positive examples serve as role models for how simulation studies ought to be conducted and reported. Based on the insights from our literature review, we provide recommendations for researchers to improve the quality of simulation studies in the following (see Table 4, for an overview).

Table 4

Recommendations for Methodological Research Using Simulation Studies.

Recommendation
1. Provide a rationale for all relevant choices in design and analysis (e.g., justifications for data-generating mechanism conditions and analysis methods)
2. Use a standardized structure for planning and reporting of simulation studies (e.g., ADEMP)
3. Report Monte Carlo uncertainty (e.g., Monte Carlo standard errors)
4. Choose the number of simulation repetitions to achieve desired precision
5. Write (and possibly preregister) study protocol to guide simulation design and to disclose the state of knowledge, prior expectations, and evaluation criteria before seeing the results (e.g., using the ADEMP-PreReg template)
6. Avoid selective reporting of results that lead to desired outcomes
7. Acknowledge the limited generalizability of a single simulation study
8. Report software versions and environment (e.g., using <code>sessionInfo()</code> in R)
9. Upload code, data, results, and other supplements to a FAIR research data repository (e.g., OSF or Zenodo)
10. Journals/Editors/Reviewers: Promote higher reporting standards and open code/data

The quality of simulation studies can benefit from standardized design and reporting, such as with the ADEMP structure (Morris et al., 2019) that we have reviewed in this article. A standardized structure ensures that researchers think about important issues when designing their study and that all important information is reported. However, comprehensive reporting of what was done is not enough; researchers should also provide a rationale for the choices

made in the design and analysis of their simulation study (e.g., justifications for data-generating mechanisms and analysis methods), as this is essential for readers and reviewers to assess the quality of evidence provided by the study.

As with any empirical study, it is important to acknowledge the uncertainty of the results within each simulation study, for example, by reporting Monte Carlo standard errors for estimated performance measures. Any observed pattern should only be interpreted if the associated Monte Carlo uncertainty (e.g., MCSE) is sufficiently small. In order for researchers to draw meaningful conclusions, the uncertainty should already be taken into account in the design of the study, for example, by choosing the number of repetitions such that a sufficiently small Monte Carlo standard error is achieved. The formulae in Table 3 can be used for this purpose. The choice of the desired MCSE, and hence the number of repetitions required, is embedded in the trade-off between the generalizability and the precision of a simulation study. Researchers aiming for high precision in their performance measure estimates will usually be able to study fewer conditions, restricting the scope of their investigation and limiting the external validity and generalizability of their results. Therefore, setting the number of simulation repetitions too high can also waste computational resources that could be better spent investigating additional settings. Choosing the number of simulation repetitions to achieve desired precision, as explained in our article, can help researchers to make informed choices in this trade-off. However, even when simulation studies are carefully designed in advance, their scope is often narrow compared to all possible realistic settings. Researchers should avoid discrepancies between the scope of their simulation and the generality with which their results are reported.

Preregistration of a study protocol helps to make a transparent distinction between knowledge, decisions, and evaluation criteria that were present before or after the results were observed. At the same time, preregistration does not mean that the researcher's hands are tied and that modifications to the study cannot be made, but rather that they should be transparently disclosed through amendments to the protocol. Fortunately, the issue of "double-dipping" on the same data to formulate and test hypotheses is less of a problem in simulation studies as new data can typically be generated cheaply. Rather, the purpose of preregistered protocols is

to guide with planning of rigorous simulation studies and to provide other researchers with a transparent picture of the research process, and potentially receive peer feedback independent of the results. This concerns especially the selection of methods, data-generating mechanisms, conditions, and performance measures, which are often highly flexible in simulation studies. Researchers can also obtain feedback on their protocols from other researchers, especially if the protocol is publicly available (see, e.g., [Kipruto & Sauerbrei, 2022](#)). Moreover, fixing the criteria for the evaluation of the results a priori protects researchers from cognitive biases in the interpretation of results, such as hindsight bias, confirmation bias, or allegiance bias, that can blur their interpretation of simulation study results. Our ADEMP-PreReg template (<https://github.com/bsiepe/ADEMP-PreReg>) can be used for preparing a (possibly preregistered) simulation study protocol or as a blueprint for the structured reporting of a simulation study.

To foster computational reproducibility and enable other researchers to build on a simulation study, we recommend to share code, data, and other supplementary material. We recommend to upload files to a research data repository that accords with the FAIR principles ([Wilkinson et al., 2016](#)), such as OSF or Zenodo, as we encountered various dead links in our review, even from journals. Zenodo, in particular, offers great integration with GitHub which facilitates developing simulation code on GitHub (using the git version control system) and then archiving time-stamped versions or snapshots of the repository in a FAIR way on Zenodo with one click. Moreover, we recommend to report software versions and the computational environment used to run the study in detail. For example, for R users (the vast majority of researchers based on our review), we recommend to at least report the output of `sessionInfo()` in the supplementary material or code repository as a low-effort step for reporting necessary software versions and the computational environment.

Besides researchers conducting simulation studies themselves, other academic stakeholders can help raise the standards of methodological research. For example, during the peer-review process, reviewers and editors can encourage proper design and reporting of simulation studies, for instance, by guiding authors to justify the number of replications or to report Monte Carlo standard errors. Similarly, journals can promote higher standards for

simulation studies by requiring authors to share code and/or data for articles that include simulation studies. This seems appropriate since conclusions from simulation studies heavily depend on the validity of their underlying code, and since there are usually no ethical concerns with publishing code and simulated data (with the exception of studies with data generating mechanisms based on resampling, where sharing the resampled data could be problematic). Mandatory code and data sharing, along with reproducibility checks and reproducibility badges, have already been adopted by several journals, for example, *Meta-Psychology* or *Biometrical Journal* which both have dedicated reproducibility teams that (partially) rerun simulation studies of submitted articles (Hofner, Schmid, & Edler, 2015; Lindsay, 2023). In a similar vein, journals could have specific calls for the replication and/or generalization of influential simulation studies (Lohmann, Astivia, Morris, & Groenwold, 2022; Luijken et al., 2023).

Conclusions

Simulation studies are a remarkably powerful tool. They allow methodologists to study the behavior of methods in virtually any situation they can imagine, often less limited by ethical, resource, or time constraints than most other types of research. At the same time, large-scale simulation studies can involve a considerable amount of costly human labor and energy-intensive computational resources. To reduce the waste of valuable resources, simulation studies should thus be planned and executed with care. Even more important, with the potentially long-lasting impact of simulation studies on scientific practice comes great responsibility to recognize the inherent uncertainty and the limited generalizability of their results. After all, simulation studies are experiments, and their success depends on the same factors as any other type of experiment—careful experimental design, consideration of sampling uncertainty, neutral evaluation of results, transparent reporting, and sharing of data and code so that other researchers can build on them.

Over the past two decades, psychology has proven to be a remarkably adaptive discipline, as exemplified by the widespread adoption of preregistration as well as open data, code, and materials (Munafò et al., 2017). We believe, therefore, that the time is ripe for a similar shift in methodological research toward more rigor and transparency in simulation

studies.

References

- American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). American Psychological Association.
- Bartoš, F., Maier, M., Wagenmakers, E.-J., Doucouliagos, H., & Stanley, T. (2023). Robust Bayesian meta-analysis: Model-averaging across complementary publication bias adjustment methods. *Research Synthesis Methods*, 14(1), 99–116. doi: 10.1002/jrsm.1594
- Binder, H., Sauerbrei, W., & Royston, P. (2012). Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. *Statistics in Medicine*, 32(13), 2262–2277. doi: 10.1002/sim.5639
- Boomsma, A. (2013). Reporting Monte Carlo studies in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 20(3), 518–540. doi: 10.1080/10705511.2013.797839
- Boulesteix, A.-L. (2015). Ten simple rules for reducing overoptimistic reporting in methodological computational research. *PLoS Computational Biology*, 11(4), e1004191. doi: 10.1371/journal.pcbi.1004191
- Boulesteix, A.-L., Binder, H., Abrahamowicz, M., & Sauerbrei, W. (2018). On the necessity and design of studies comparing statistical methods. *Biometrical Journal*, 60(1), 216–218. doi: 10.1002/bimj.201700129
- Boulesteix, A.-L., Groenwold, R. H., Abrahamowicz, M., Binder, H., Briel, M., Hornung, R., ... Sauerbrei, W. (2020). Introduction to statistical simulations in health research. *BMJ Open*, 10(12), e039921. doi: 10.1136/bmjopen-2020-039921
- Boulesteix, A.-L., Hoffmann, S., Charlton, A., & Seibold, H. (2020). A replication crisis in methodological research? *Significance*, 17(5), 18–21. doi: 10.1111/1740-9713.01444
- Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24), 4279–4292. doi: 10.1002/sim.2673
- Carsey, T., & Harden, J. (2014). *Monte Carlo simulation and resampling methods for social*

- science*. SAGE Publications. doi: 10.4135/9781483319605
- Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, 2(2), 115–144. doi: 10.1177/2515245919847196
- Chalmers, R. P., & Adkins, M. C. (2020). Writing effective and reliable Monte Carlo simulations with the SimDesign package. *The Quantitative Methods for Psychology*, 16(4), 248–280. doi: 10.20982/tqmp.16.4.p248
- Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., . . . Borges, B. (2023). shiny: Web application framework for R [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=shiny> (R package version 1.7.5)
- Chipman, H., & Bingham, D. (2022). Let's practice what we preach: Planning and interpreting simulation studies with design and analysis of experiments. *Canadian Journal of Statistics*, 50(4), 1228–1249. doi: 10.1002/cjs.11719
- Clifton, L., & Clifton, D. A. (2019). The correlation between baseline score and post-intervention score, and its implications for statistical analysis. *Trials*, 20(1), 1–6. doi: 10.1186/s13063-018-3108-3
- Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.
- Cowan, N., Belletier, C., Doherty, J. M., Jaroslawska, A. J., Rhodes, S., Forsberg, A., . . . Logie, R. H. (2020). How do scientific views change? Notes from an extended adversarial collaboration. *Perspectives on Psychological Science*, 15(4), 1011–1025. doi: 10.1177/1745691620906415
- Cumming, G., Fidler, F., Kalinowski, P., & Lai, J. (2012). The statistical recommendations of the American Psychological Association publication manual: Effect sizes, confidence intervals, and meta-analysis. *Australian Journal of Psychology*, 64(3), 138–146. doi: 10.1111/j.1742-9536.2011.00037.x
- den Akker, O. R. V., Weston, S., Campbell, L., Chopik, B., Damian, R., Davis-Kean, P., . . . Bakker, M. (2021). Preregistration of secondary data analysis: A template and tutorial. *Meta-Psychology*, 5. doi: 10.15626/mp.2020.2625
- Epskamp, S. (2019). Reproducibility and replicability in a fast-paced methodological world.

- Advances in Methods and Practices in Psychological Science*, 2(2), 145–155. doi: 10.1177/25152459198474
- Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice*, 35(2), 36–49. doi: 10.1111/emip.12111
- Franklin, J. M., Schneeweiss, S., Polinski, J. M., & Rassen, J. A. (2014). Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Computational Statistics & Data Analysis*, 72, 219–226. doi: 10.1016/j.csda.2013.10.018
- Friedrich, S., & Friede, T. (2023). On the role of benchmarking data sets and simulations in method comparison studies. *Biometrical Journal*, 2200212. doi: 10.1002/bimj.202200212
- Gasparini, A. (2018). rsimsum: Summarise results from Monte Carlo simulation studies. *Journal of Open Source Software*, 3(26), 739. doi: 10.21105/joss.00739
- Gasparini, A., Morris, T. P., & Crowther, M. J. (2021). INTEREST: Interactive tool for exploring REsults from simulation sTudies. *Journal of Data Science, Statistics, and Visualisation*, 1(4). doi: 10.52933/jdssv.v1i4.9
- Gaujoux, R. (2023). dornrg: Generic reproducible parallel backend for 'foreach' loops [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=doRNG> (R package version 1.8.6)
- Genz, A., & Bretz, F. (2009). *Computation of multivariate normal and t probabilities*. Heidelberg: Springer-Verlag.
- Glatard, T., Lewis, L. B., Ferreira da Silva, R., Adalat, R., Beck, N., Lepage, C., ... others (2015). Reproducibility of neuroimaging analyses across operating systems. *Frontiers in Neuroinformatics*, 9, 12. doi: 10.3389/fninf.2015.00012
- Grieder, S., & Steiner, M. D. (2021, June). Algorithmic jingle jungle: A comparison of implementations of principal axis factoring and promax rotation in R and SPSS. *Behavior Research Methods*, 54(1), 54–74. doi: 10.3758/s13428-021-01581-x
- Harwell, M., Kohli, N., & Peralta-Torres, Y. (2018). A survey of reporting practices of

- computer simulation studies in statistical research. *The American Statistician*, 72(4), 321–327. doi: 10.1080/00031305.2017.1342692
- Hauck, W. W., & Anderson, S. (1984). A survey regarding the reporting of simulation studies. *The American Statistician*, 38(3), 214–216. doi: 10.1080/00031305.1984.10483206
- Heck, D. W., & Erdfelder, E. (2019). Maximizing the expected information gain of cognitive modeling via design optimization. *Computational Brain & Behavior*, 2, 202–209. doi: 10.1007/s42113-019-00035-0
- Heinze, G., Boulesteix, A.-L., Kammer, M., Morris, T. P., & White, I. R. (2023). Phases of methodological research in biostatistics—Building the evidence base for new methods. *Biometrical Journal*, 2200222. doi: 10.1002/bimj.202200222
- Hoaglin, D. C., & Andrews, D. F. (1975). The reporting of computation-based results in statistics. *The American Statistician*, 29(3), 122–126. doi: 10.1080/00031305.1975.10477393
- Hodges, C. B., Stone, B. M., Johnson, P. K., Carter III, J. H., Sawyers, C. K., Roby, P. R., & Lindsey, H. M. (2023). Researcher degrees of freedom in statistical software contribute to unreliable results: A comparison of nonparametric analyses conducted in SPSS, SAS, Stata, and R. *Behavior Research Methods*, 55(6), 2813–2837. doi: 10.3758/s13428-022-01932-2
- Hofner, B., Schmid, M., & Edler, L. (2015). Reproducible research in statistics: A review and guidelines for the biometrical journal. *Biometrical Journal*, 58(2), 416–427. doi: 10.1002/bimj.201500156
- Hong, S., & Reed, W. R. (2021). Using Monte Carlo experiments to select meta-analytic estimators. *Research Synthesis Methods*, 12(2), 192–215. doi: 10.1002/jrsm.1467
- Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, 6(1), 1–55. doi: 10.1080/10705519909540118
- ICH. (2019). *Addendum on estimands and sensitivity analyses in clinical trials to the guideline on statistical principles for clinical trials, ICH E9(R1)*.
<https://database.ich.org/sites/default/files/>

[E9-R1_Step4_Guideline_2019_1203.pdf](#).

- Joshi, M., & Pustejovsky, J. (2022). *simhelpers*: Helper functions for simulation studies [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=simhelpers> (R package version 0.1.2)
- Keene, O. N., Lynggaard, H., Englert, S., Lanius, V., & Wright, D. (2023). Why estimands are needed to define treatment effects in clinical trials. *BMC Medicine*, 21(1). doi: 10.1186/s12916-023-02969-6
- Kelter, R. (2023). The Bayesian simulation study (BASIS) framework for simulation studies in statistical and methodological research. *Biometrical Journal*, 2200095. doi: 10.1002/bimj.202200095
- Kipruto, E., & Sauerbrei, W. (2022). Comparison of variable selection procedures and investigation of the role of shrinkage in linear regression-protocol of a simulation study in low-dimensional data. *PLOS ONE*, 17(10), e0271240. doi: 10.1371/journal.pone.0271240
- Koehler, E., Brown, E., & Haneuse, S. J.-P. (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *The American Statistician*, 63(2), 155–162. doi: 10.1198/tast.2009.0030
- Kreutz, C., Can, N. S., Bruening, R. S., Meyberg, R., Mérai, Z., Fernandez-Pozo, N., & Rensing, S. A. (2020). A blind and independent benchmark study for detecting differentially methylated regions in plants. *Bioinformatics*, 36(11), 3314–3321. doi: 10.1093/bioinformatics/btaa191
- Kucharský, Š., Houtkoop, B. L., & Visser, I. (2020). Code sharing in psychological methods and statistics: An overview and associations with conventional and alternative research metrics. *OSF Preprints*. doi: 10.31219/osf.io/daews
- Lakens, D., & Caldwell, A. R. (2021). Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science*, 4, 2515245920951503. doi: 10.1177/2515245920951503
- Lange, F. J. D. (2022). *Confirmatory studies in methodological statistical research: Concept*

- and illustration* (Master's thesis, Ludwig-Maximilians-Universität, Munich, Germany).
doi: 10.5282/ubm/epub.94460
- Leydold, J. (2022). rstream: Streams of random numbers [Computer software manual].
Retrieved from <https://CRAN.R-project.org/package=rstream> (R
package version 1.3.7)
- Lindsay, D. S. (2023). A plea to psychology professional societies that publish journals:
Assess computational reproducibility. *Meta-Psychology*, 7. doi:
10.15626/mp.2023.4020
- Liu, H., Yuan, K.-H., & Wen, Z. (2022). Two-level moderated mediation models with
single-level data and new measures of effect sizes. *Behavior Research Methods*, 54,
574—596. doi: 10.3758/s13428-021-01578-6
- Liu, J., & Perera, R. A. (2022). Estimating knots and their association in parallel bilinear
spline growth curve models in the framework of individual measurement occasions.
Psychological Methods, 27(5), 703. doi: 10.1037/met0000309
- Lohmann, A., Astivia, O. L. O., Morris, T. P., & Groenwold, R. H. H. (2022). It's time! Ten
reasons to start replicating simulation studies. *Frontiers in Epidemiology*, 2. doi:
10.3389/fepid.2022.973470
- Lüdtke, O., & Robitzsch, A. (2023). ANCOVA versus change score for the analysis of
two-wave data. *The Journal of Experimental Education*, Advance Online Publication.
doi: 10.1080/00220973.2023.2246187
- Luijken, K., Lohmann, A., Alter, U., Gonzalez, J. C., Clouth, F. J., Fossum, J. L., ...
Groenwold, R. H. H. (2023). *Replicability of simulation studies for the investigation of
statistical methods: The RepliSims project*. doi: 10.48550/arXiv.2307.02052
- Lundberg, I., Johnson, R., & Stewart, B. M. (2021). What is your estimand? Defining the
target quantity connects statistical evidence to theory. *American Sociological Review*,
86(3), 532–565. doi: 10.1177/000312242110041
- McNeish, D., & Bauer, D. J. (2022). Reducing incidence of nonpositive definite covariance
matrices in mixed effect models. *Multivariate Behavioral Research*, 57(2-3), 318–340.
doi: 10.1080/00273171.2020.1830019

- McNeish, D., Lane, S., & Curran, P. (2018). Monte Carlo simulation methods. *The Reviewer's Guide to Quantitative Methods in the Social Sciences*, 269–276.
- Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074–2102. doi: 10.1002/sim.8086
- Moss, J., Wong, A. Y., Durriseau, J. A., & Bradshaw, G. L. (2022). Tracking strategy changes using machine learning classifiers. *Behavior Research Methods*, 54, 1–23. doi: 10.3758/s13428-021-01720-4
- Moulder, R. G., Daniel, K. E., Teachman, B. A., & Boker, S. M. (2022). Tangle: A metric for quantifying complexity and erratic behavior in short time series. *Psychological Methods*, 27(1), 82.
- Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N. P., . . . Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1(0021). doi: 10.1038/s41562-016-0021
- Murayama, K., Usami, S., & Sakaki, M. (2022). Summary-statistics-based power analysis: A new and practical method to determine sample size for mixed-effects modeling. *Psychological Methods*, 27(6), 1014–1038. doi: 10.1037/met0000330
- Nießl, C., Hoffmann, S., Ullmann, T., & Boulesteix, A.-L. (2023). Explaining the optimistic performance evaluation of newly proposed methods: A cross-design validation experiment. *Biometrical Journal*, 2200238. doi: 10.1002/bimj.202200238
- Papenberg, M., & Klau, G. W. (2021). Using anticlustering to partition data sets into equivalent parts. *Psychological Methods*, 26(2), 161–174. doi: 10.1037/met0000301
- Pawel, S., Kook, L., & Reeve, K. (2023). Pitfalls and potentials in simulation studies: Questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method. *Biometrical Journal*, 2200091. doi: 10.1002/bimj.202200091
- Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling: A Multidisciplinary Journal*, 8(2), 287–312. doi: 10.1207/S15328007SEM0802_7
- Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation

- study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12), 1373–1379. doi: 10.1016/s0895-4356(96)00236-3
- Peikert, A., & Brandmaier, A. M. (2021). A reproducible data analysis workflow. *Quantitative and Computational Methods in Behavioral Sciences*, 1(Article e3763). doi: 10.5964/qcmb.3763
- Psychometric Society. (1979). Publication policy regarding Monte Carlo studies. *Psychometrika*, 44, 133–134.
- R Core Team. (2023). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <https://www.R-project.org/>
- Rodgers, M. A., & Pustejovsky, J. E. (2021). Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychological Methods*, 26(2), 141–160. doi: 10.1037/met0000300
- Rücker, G., & Schwarzer, G. (2014). Presenting simulation results in a nested loop plot. *BMC Medical Research Methodology*, 14(129). doi: 10.1186/1471-2288-14-129
- Schwab, S., & Held, L. (2021). Statistical programming: Small mistakes, big impacts. *Significance*, 18(3), 6–7. doi: 10.1111/1740-9713.01522
- Senn, S. (2006). Change from baseline and analysis of covariance revisited. *Statistics in Medicine*, 25(24), 4334–4344. doi: 10.1002/sim.2682
- Senn, S. (2008). *Statistical issues in drug development* (second ed., Vol. 69). Chichester: John Wiley & Sons.
- Shi, D., DiStefano, C., Maydeu-Olivares, A., & Lee, T. (2022). Evaluating SEM model fit with small degrees of freedom. *Multivariate Behavioral Research*, 57(2-3), 179–207. doi: 10.1080/00273171.2020.1868965
- Siepe, B. S., Bartoš, F., Morris, T. P., Boulesteix, A.-L., Heck, D. W., & Pawel, S. (2023). Simulation studies for methodological research in psychology: A standardized structure for planning, preregistration, and reporting. (Preprint)
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology:

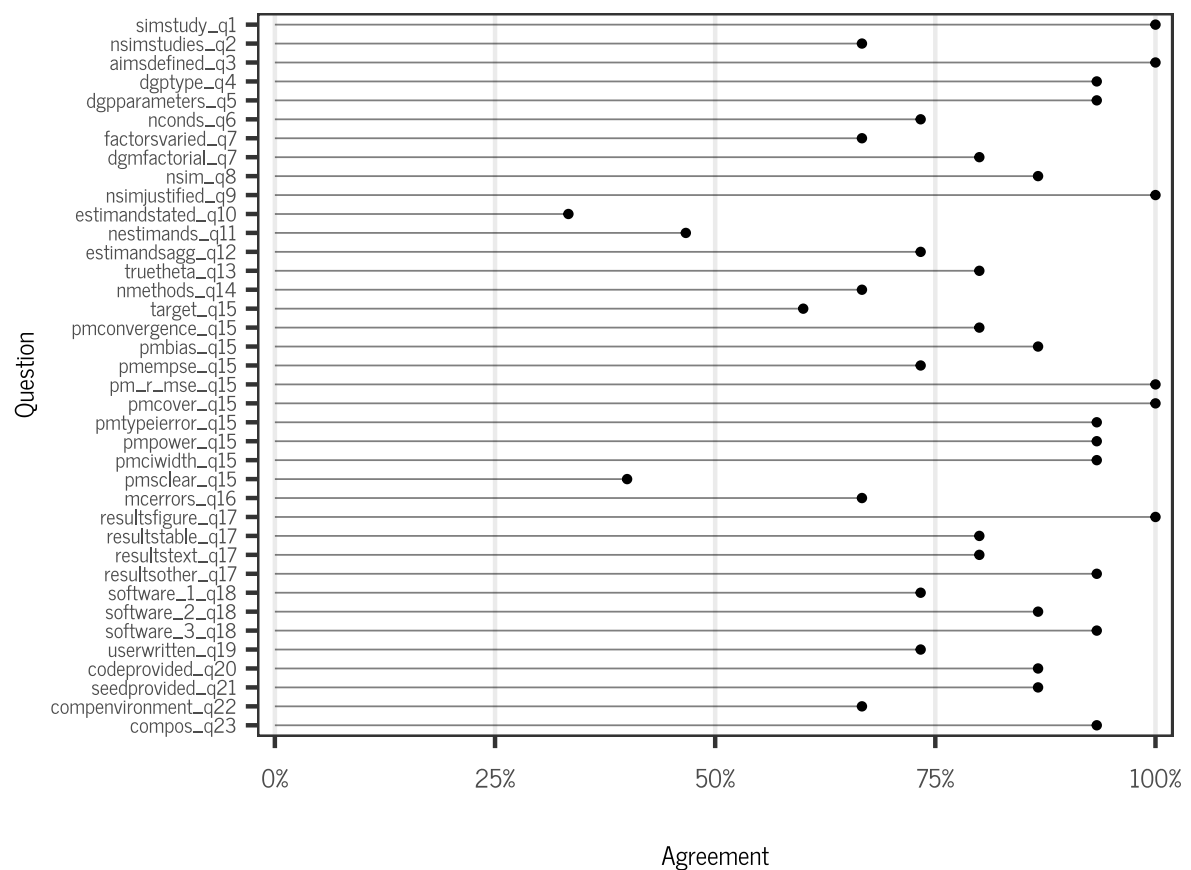
- Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. doi: 10.1177/0956797611417632
- Skrondal, A. (2000). Design and analysis of Monte Carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research*, 35(2), 137–167. doi: 10.1207/s15327906mbr3502_1
- Strobl, C., & Leisch, F. (2022). Against the “one method fits all data sets” philosophy for comparison studies in methodological research. *Biometrical Journal*. doi: 10.1002/bimj.202200104
- Ushey, K., & Wickham, H. (2023). renv: Project environments [Computer software manual]. (<https://rstudio.github.io/renv/>, <https://github.com/rstudio/renv>)
- Van Breukelen, G. J. (2013). ANCOVA versus CHANGE from baseline in nonrandomized studies: The difference. *Multivariate Behavioral Research*, 48(6), 895–922. doi: 10.1080/00273171.2013.831743
- van der Bles, A. M., Van Der Linden, S., Freeman, A. L., Mitchell, J., Galvao, A. B., Zaval, L., & Spiegelhalter, D. J. (2019). Communicating uncertainty about facts, numbers and science. *Royal Society Open Science*, 6(5), 181870. doi: 10.1098/rsos.181870
- van Smeden, M., de Groot, J. A. H., Moons, K. G. M., Collins, G. S., Altman, D. G., Eijkemans, M. J. C., & Reitsma, J. B. (2016). No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology*, 16, 163. doi: 10.1186/s12874-016-0267-3
- Vickers, A. J. (2001). The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: A simulation study. *BMC Medical Research Methodology*, 1(6). doi: 10.1186/1471-2288-1-6
- White, I. R., Pham, T. M., Quartagno, M., & Morris, T. P. (2023). How to check a simulation study. *International Journal of Epidemiology*. doi: 10.1093/ije/dyad134
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved 2023-10-17, from <https://ggplot2.tidyverse.org>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ...

Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(160018). doi: 10.1038/sdata.2016.18

Appendix

A summary of inter-rater agreement for the studies that were most difficult to code is provided in Figure 5. The wording of the questions and their answer options are available in the preregistration of our review (<https://osf.io/8cbfd>). Overall, the agreement between raters appears to be acceptable. Low agreement is especially pronounced in questions about the estimands, where the ambiguity of reporting combined with the complexity of some models often made the assessment of a specific number of estimands very difficult. As we used studies that we identified as most difficult for our assessment of agreement, we consider the proportions found here as a lower bound for the overall agreement across all studies. Also, higher rates of disagreement in some questions here again indicate the need for more clarity in the reporting of studies. In the case of disagreement, we kept the rating of the initial reviewer for the analyses in the manuscript.

Numerical results of the example simulation study on methods for the analysis of pre–post measurements are given in Table 5.

Figure 5*Agreement between Raters*

Note. Proportion of agreement between the three raters for 15 papers with a low or medium confidence rating. Two studies that were also rated for agreement were not assessed here, as the raters chose different simulation studies therein.

Table 5

Estimated Power / Type I Error Rate (MCSE) and Bias (MCSE) of ANCOVA, Change Score Analysis, and Post Score Analysis Across Simulation Conditions

Condition		Power / Type I Error Rate			Bias		
Correlation	Effect	ANCOVA	Change score	Post score	ANCOVA	Change score	Post score
0.00	0.00	0.0447 (0.0021)	0.0508 (0.0022)	0.0464 (0.0021)	−0.0003 (0.0020)	0.0006 (0.0028)	−0.0005 (0.0020)
0.00	0.20	0.1655 (0.0037)	0.1111 (0.0031)	0.1671 (0.0037)	−0.0023 (0.0020)	−0.0010 (0.0029)	−0.0024 (0.0020)
0.00	0.50	0.6907 (0.0046)	0.4137 (0.0049)	0.6940 (0.0046)	−0.0009 (0.0020)	−0.0004 (0.0028)	−0.0009 (0.0020)
0.50	0.00	0.0496 (0.0022)	0.0474 (0.0021)	0.0500 (0.0022)	−0.0008 (0.0017)	−0.0013 (0.0020)	−0.0008 (0.0020)
0.50	0.20	0.2108 (0.0041)	0.1715 (0.0038)	0.1646 (0.0037)	0.0015 (0.0017)	0.0004 (0.0020)	0.0022 (0.0020)
0.50	0.50	0.8130 (0.0039)	0.6978 (0.0046)	0.6973 (0.0046)	0.0021 (0.0018)	0.0020 (0.0020)	0.0018 (0.0020)
0.70	0.00	0.0487 (0.0022)	0.0503 (0.0022)	0.0500 (0.0022)	−0.0021 (0.0014)	−0.0032 (0.0015)	−0.0001 (0.0020)
0.70	0.20	0.2782 (0.0045)	0.2459 (0.0043)	0.1641 (0.0037)	−0.0023 (0.0014)	−0.0027 (0.0016)	−0.0008 (0.0020)
0.70	0.50	0.9296 (0.0026)	0.8913 (0.0031)	0.6944 (0.0046)	0.0013 (0.0014)	0.0012 (0.0016)	0.0011 (0.0020)