Universität Marburg

Simulation Studies for Methodological Research: State of the Art, Issues, and Potential Solutions

Björn Siepe[1]

June 26, 2025 – Münster Statistics & Methods Colloquium

[1]Psychological Methods Lab, Department of Psychology, Universität Marburg

Agenda

Introduction

Questionable research practices in simulation studies

Simulation studies in Psychology

Potential improvements

Discussion

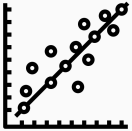
Introduction

Quantitative methodological research

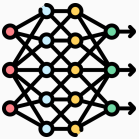
- **Diverse fields**: Statistics, psychometrics, bioinformatics, ecology, econometrics, machine learning, …

Quantitative methodological research

- **Diverse fields**: Statistics, psychometrics, bioinformatics, ecology, econometrics, machine learning, …
- Common question: **Which data analysis methods work well when?**

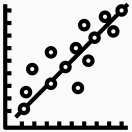


vs.

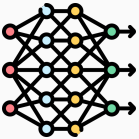


Quantitative methodological research

- **Diverse fields**: Statistics, psychometrics, bioinformatics, ecology, econometrics, machine learning, …
- Common question: **Which data analysis methods work well when?**



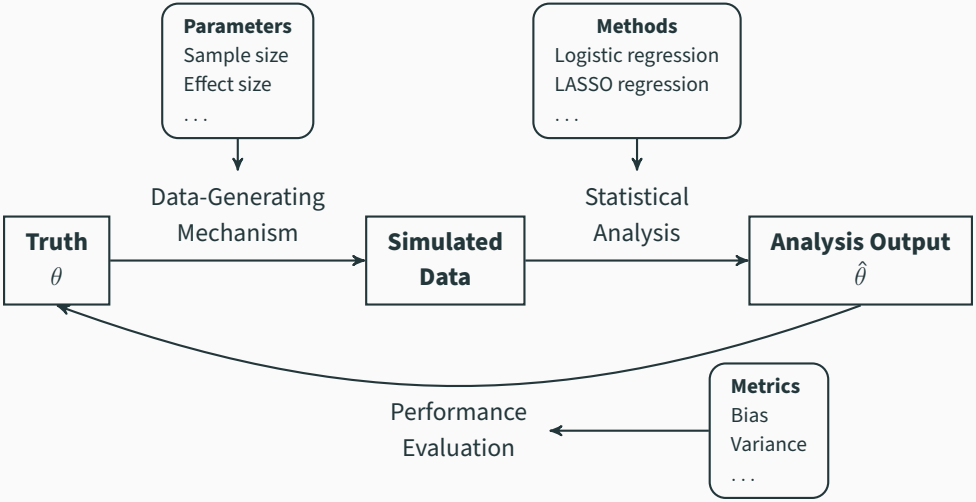
vs.



- Tools:
  - Formal analysis and **mathematical proofs** → theory
  - Application to **real data sets** → case studies
  - **Simulation studies** → controlled experiments

Icons taken from flaticon.com

Simulation studies



Simulation studies are commonly used

Journal	Article contains simulation study
Journal of the American Statistical Association	186/200 = **93%**
Statistics in Medicine	104/115 = **90%**
Psychological Methods	98/179 = **55%**
Research Synthesis Methods	94/306 = **31%**

Literature review from Pawel et al. (2024a)

Simulation studies can be influential

Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives

L **Hu**, PM **Bentler** - Structural equation modeling: a …, 1999 - Taylor & Francis

This article examines the adequacy of the “rules of thumb” conventional cutoff criteria and several new alternatives for various fit indexes used to evaluate model fit in practice. Using a 2-…

☆ Save Cite Cited by 116305 Related articles All 9 versions

A **simulation study** of the number of events per variable in logistic regression analysis

P **Peduzzi**, J Concato, E Kemper, TR Holford… - Journal of clinical …, 1996 - Elsevier

… In a **simulation study** of forward stepwise multiple linear regression, Freedman and Pee [3] demonstrated that the … In **simulation studies** of the effect of EPV on proportional … Peter **Peduzzi**. …

☆ Save Cite Cited by 8827 Related articles All 9 versions

Simulation studies impact implementation of research

Post-anaesthesia pulmonary complications after use of muscle relaxants (POPULAR): a multicentre, prospective observational study

E Kirmeier, LI Eriksson, H Lewald… - The Lancet …, 2019 - thelancet.com

Background Results from retrospective studies suggest that use of neuromuscular blocking agents during general anaesthesia might be linked to postoperative pulmonary …

☆ Save 99 Cite Cited by 303 Related articles All 37 versions

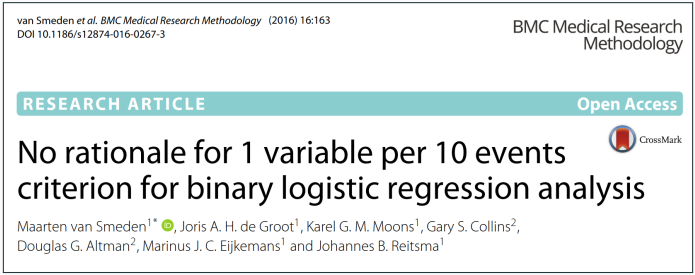
Statistical analysis

Sample size was estimated using the rule of ten.[19]

$$\text{Sample size} = \frac{10 \times \text{number of factors and cofactors}}{\text{Incidence of postoperative pulmonary complications}}$$

19 Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996; **49**: 1373–79.

There can be problems with simulation studies



van Smeden *et al. BMC Medical Research Methodology* (2016) 16:163
DOI 10.1186/s12874-016-0267-3

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

CrossMark

No rationale for 1 variable per 10 events criterion for binary logistic regression analysis

Maarten van Smeden[1*], Joris A. H. de Groot[1], Karel G. M. Moons[1], Gary S. Collins[2], Douglas G. Altman[2], Marinus J. C. Eijkemans[1] and Johannes B. Reitsma[1]

"*The current **evidence supporting [the rule of ten] is weak** [...] there is an urgent need for new research to provide guidance for supporting sample size considerations for binary logistic regression*" van Smeden et al. (2016)

Related new arXiv preprint

Handling Missingness, Failures, and Non-Convergence in Simulation Studies: A Review of Current Practices and Recommendations

Samuel Pawel [1], František Bartoš [2,*], Björn S. Siepe [3,*], Anna Lohmann [4,5,*]

Related new arXiv preprint

Handling Missingness, Failures, and Non-Convergence in Simulation Studies: A Review of Current Practices and Recommendations

Samuel Pawel [1], František Bartoš [2,*], Björn S. Siepe [3,*], Anna Lohmann [4,5,*]

- Review of 482 simulation studies published in JASA, SiM, PM, RSM:
  - **23.0%** mention missingness / failures / non-convergence

Related new arXiv preprint

Handling Missingness, Failures, and Non-Convergence in Simulation Studies: A Review of Current Practices and Recommendations

Samuel Pawel [1], František Bartoš [2,*], Björn S. Siepe [3,*], Anna Lohmann [4,5,*]

- Review of 482 simulation studies published in JASA, SiM, PM, RSM:
  - **23.0%** mention missingness / failures / non-convergence
  - **19.1%** report frequency
  - **13.9%** report handling
  - **46.7%** share code

Related new arXiv preprint

Handling Missingness, Failures, and Non-Convergence in Simulation Studies: A Review of Current Practices and Recommendations

Samuel Pawel [1], František Bartoš [2,*], Björn S. Siepe [3,*], Anna Lohmann [4,5,*]

- Review of 482 simulation studies published in JASA, SiM, PM, RSM:
  - **23.0%** mention missingness / failures / non-convergence
  - **19.1%** report frequency
  - **13.9%** report handling
  - **46.7%** share code
- Missingness classification, handling approaches, case-study

doi:10.48550/arXiv.2409.18527

Issues in simulation studies

> *"…extensive simulation studies show that the proposed method performs* ***better than existing methods*** *…"*

Issues in simulation studies

> *"…extensive simulation studies show that the proposed method performs* ***better than existing methods*** *…"*

- Over-Optimism (e.g., Ullmann et al., 2022)

Issues in simulation studies

> *"…extensive simulation studies show that the proposed method performs* ***better than existing methods*** *…"*

- Over-Optimism (e.g., Ullmann et al., 2022)
- Issues similar to other empirical research (Boulesteix et al., 2020)

Issues in simulation studies

> *"…extensive simulation studies show that the proposed method performs* ***better than existing methods*** *…"*

- Over-Optimism (e.g., Ullmann et al., 2022)
- Issues similar to other empirical research (Boulesteix et al., 2020)
- Insufficient reporting standards (e.g., Hoaglin and Andrews, 1975)

Issues in simulation studies

> *"…extensive simulation studies show that the proposed method performs* ***better than existing methods*** *…"*

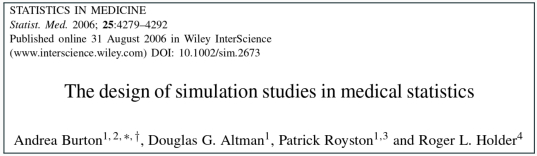
- Over-Optimism (e.g., Ullmann et al., 2022)
- Issues similar to other empirical research (Boulesteix et al., 2020)
- Insufficient reporting standards (e.g., Hoaglin and Andrews, 1975)
- Reproducibility? (e.g., Luijken et al., 2023)

Issues in simulation studies

> *"…extensive simulation studies show that the proposed method performs* ***better than existing methods*** *…"*

- Over-Optimism (e.g., Ullmann et al., 2022)
- Issues similar to other empirical research (Boulesteix et al., 2020)
- Insufficient reporting standards (e.g., Hoaglin and Andrews, 1975)
- Reproducibility? (e.g., Luijken et al., 2023)

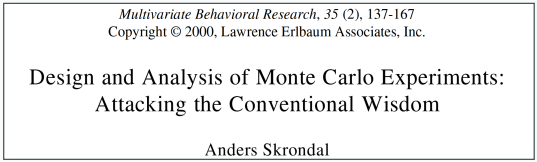
Meta-research on simulation studies



STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:4279–4292
Published online 31 August 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2673

The design of simulation studies in medical statistics

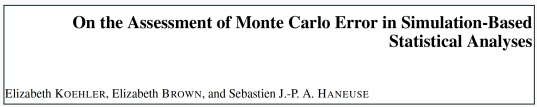
Andrea Burton[1, 2, *, †], Douglas G. Altman[1], Patrick Royston[1,3] and Roger L. Holder[4]



Multivariate Behavioral Research, *35* (2), 137-167
Copyright © 2000, Lawrence Erlbaum Associates, Inc.

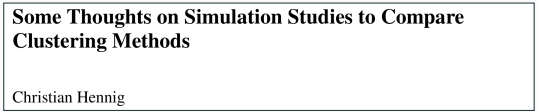
Design and Analysis of Monte Carlo Experiments: Attacking the Conventional Wisdom

Anders Skrondal



On the Assessment of Monte Carlo Error in Simulation-Based Statistical Analyses

Elizabeth KOEHLER, Elizabeth BROWN, and Sebastien J.-P. A. HANEUSE



Some Thoughts on Simulation Studies to Compare Clustering Methods

Christian Hennig

DOI: 10.1002/bimj.202200104

Biometrical Journal

DISCUSSION

Against the "one method fits all data sets" philosophy for comparison studies in methodological research

Carolin Strobl[1] | Friedrich Leisch[2]

DOI: 10.1002/bimj.202200222

Biometrical Journal

RESEARCH ARTICLE

Phases of methodological research in biostatistics—Building the evidence base for new methods

Georg Heinze[1] | Anne-Laure Boulesteix[2] | Michael Kammer[1,3] | Tim P. Morris[4] | Ian R. White[4] | on behalf of the Simulation Panel of the STRATOS initiative

• • •

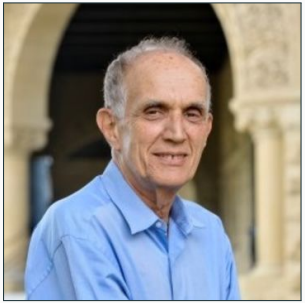
Burton et al. (2006); Skrondal (2000); Koehler et al. (2009); Strobl and Leisch (2024); Hennig (2018); Heinze et al. (2024)

Questionable research practices in simulation studies

Neutrality in simulation studies

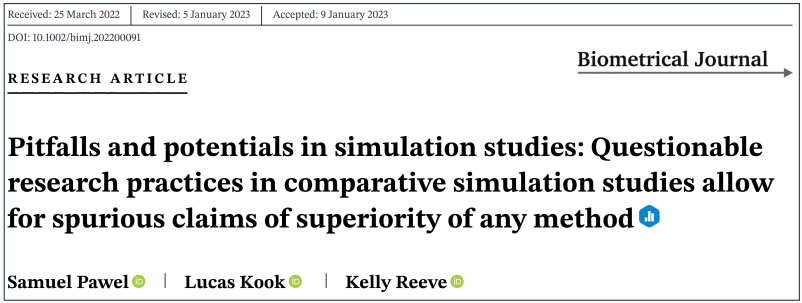
"*In fact it is* ***very difficult to run an honest simulation*** *comparison, and* ***easy to inadvertently cheat*** *by choosing favorable examples, or by not putting as much effort into optimizing the dull old standard as the exciting new challenger.*"

Brad Efron (2001)



https://statistics.stanford.edu/people/bradley-efron

Our study



Received: 25 March 2022 | Revised: 5 January 2023 | Accepted: 9 January 2023

DOI: 10.1002/bimj.202200091

Biometrical Journal

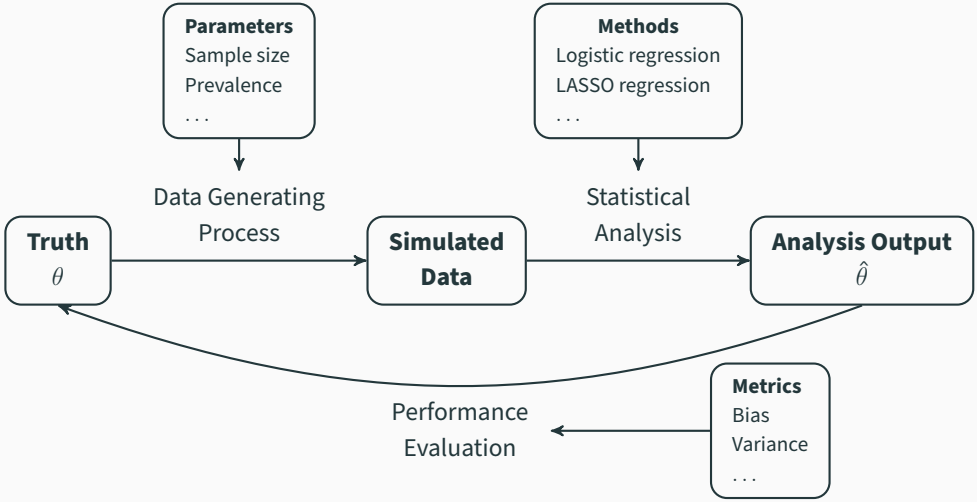
RESEARCH ARTICLE

Pitfalls and potentials in simulation studies: Questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method

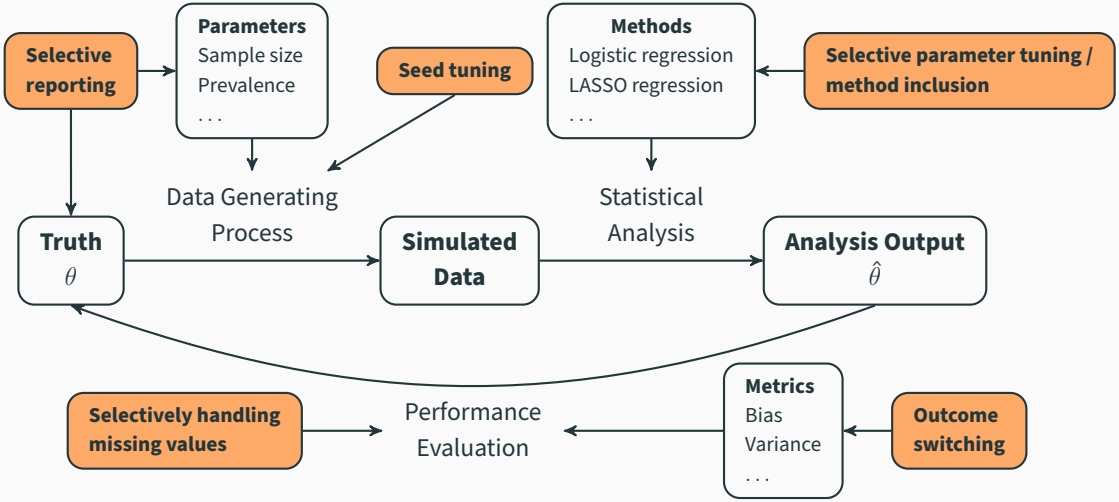
Samuel Pawel | **Lucas Kook** | **Kelly Reeve**

- Which **questionable research practices** (QRPs) exist in simulation studies?
- How can QRPs **impact the conclusions** of a study?
- How can QRPs be **addressed**?

Questionable research practices in simulation studies



Questionable research practices in simulation studies

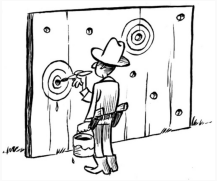


See Table 1 in doi:10.1002/bimj.202200091 for more QRPs

Questionable research practices in simulation studies

Root causes

- **Pressure to publish** novel and positive results
- **Low requirements** from journals
- **Cognitive biases** (e.g., confirmation or hindsight bias)
- **Low awareness** in scientific community



Dirk-Jan Hoek (CC-BY)

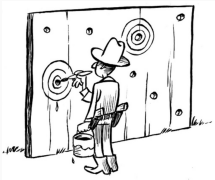
Questionable research practices in simulation studies

Root causes

- **Pressure to publish** novel and positive results
- **Low requirements** from journals
- **Cognitive biases** (e.g., confirmation or hindsight bias)
- **Low awareness** in scientific community

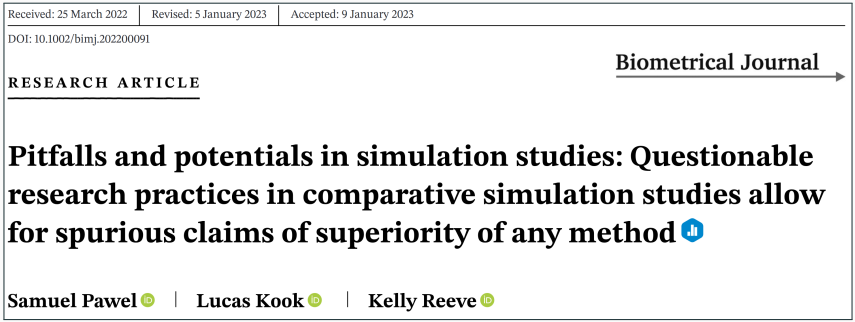
Potential consequences

- **Overoptimistic conclusions**
- **Publication bias**
- **Misinformed decisions**



Dirk-Jan Hoek (CC-BY)

QRP Illustration



Received: 25 March 2022 | Revised: 5 January 2023 | Accepted: 9 January 2023

DOI: 10.1002/bimj.202200091

Biometrical Journal

RESEARCH ARTICLE

Pitfalls and potentials in simulation studies: Questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method

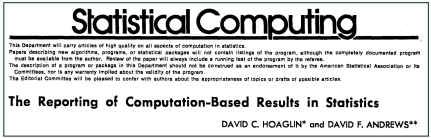
Samuel Pawel | Lucas Kook | Kelly Reeve

"*By **deliberately using several QRPs**, we were able to **present a method with no expected benefits** […] **as an improvement** over […] well-established competitors.*"

Simulation studies in Psychology

Literature Review

"*Statisticians ... often pay too little attention to their own principles of design*"(Hoaglin & Andrews, 1975)

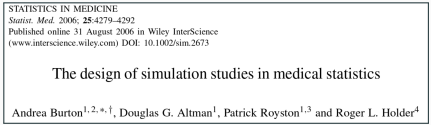


Statistical Computing

This Department will carry articles of high quality on all aspects of computation in statistics.
Papers describing new algorithms, programs, or statistical packages will not contain listings of the program, although the completely documented program must be available from the author. Review of the paper will always include a running test of the program by the referee.
The description of a program or package in this Department should not be construed as an endorsement of it by the American Statistical Association or its Committees, nor is any warranty implied about the validity of the program.
The Editorial Committee will be pleased to confer with authors about the appropriateness of topics or drafts of possible articles.

The Reporting of Computation-Based Results in Statistics

DAVID C. HOAGLIN* and DAVID F. ANDREWS**



STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:4279–4292
Published online 31 August 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2673

The design of simulation studies in medical statistics

Andrea Burton[1, 2, *, †], Douglas G. Altman[1], Patrick Royston[1, 3] and Roger L. Holder[4]

TUTORIAL IN BIOSTATISTICS

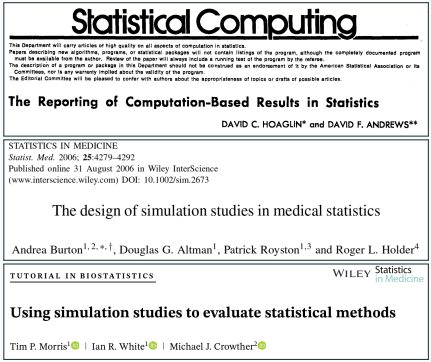
WILEY Statistics in Medicine

Using simulation studies to evaluate statistical methods

Tim P. Morris[1] | Ian R. White[1] | Michael J. Crowther[2]

Literature Review

"*Statisticians ... often pay too little attention to their own principles of design*"(Hoaglin & Andrews, 1975)



Statistical Computing

This Department will carry articles of high quality on all aspects of computation in statistics.
Papers describing new algorithms, programs, or statistical packages will not contain listings of the program, although the completely documented program must be available from the author. Review of the paper will always include a running test of the program by the referee.
The description of a program or package in this Department should not be construed as an endorsement of it by the American Statistical Association or its Committees, nor is any warranty implied about the validity of the program.
The Editorial Committee will be pleased to confer with authors about the appropriateness of topics or drafts of possible articles.

The Reporting of Computation-Based Results in Statistics

DAVID C. HOAGLIN* and DAVID F. ANDREWS**

STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:4279–4292
Published online 31 August 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2673

The design of simulation studies in medical statistics

Andrea Burton[1, 2, *, †], Douglas G. Altman[1], Patrick Royston[1, 3] and Roger L. Holder[4]

TUTORIAL IN BIOSTATISTICS

WILEY Statistics in Medicine

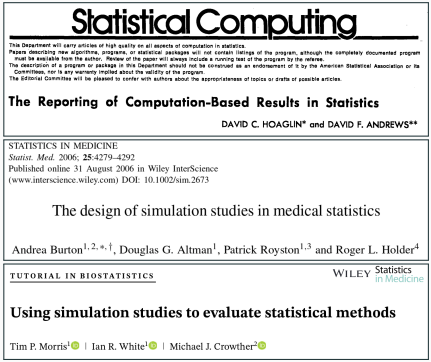
Using simulation studies to evaluate statistical methods

Tim P. Morris[1] | Ian R. White[1] | Michael J. Crowther[2]

This project:

Literature Review

"*Statisticians … often pay too little attention to their own principles of design*"(Hoaglin & Andrews, 1975)



Statistical Computing

This Department will carry articles of high quality on all aspects of computation in statistics.
Papers describing new algorithms, programs, or statistical packages will not contain listings of the program, although the completely documented program must be available from the author. Review of the paper will always include a running test of the program by the referee.
The description of a program or package in this Department should not be construed as an endorsement of it by the American Statistical Association or its Committees, nor is any warranty implied about the validity of the program.
The Editorial Committee will be pleased to confer with authors about the appropriateness of topics or drafts of possible articles.

The Reporting of Computation-Based Results in Statistics

DAVID C. HOAGLIN* and DAVID F. ANDREWS**

STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:4279–4292
Published online 31 August 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2673

The design of simulation studies in medical statistics

Andrea Burton[1, 2, *, †], Douglas G. Altman[1], Patrick Royston[1, 3] and Roger L. Holder[4]

TUTORIAL IN BIOSTATISTICS

WILEY Statistics in Medicine

Using simulation studies to evaluate statistical methods

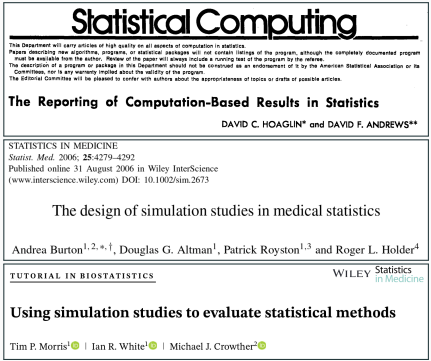
Tim P. Morris[1] | Ian R. White[1] | Michael J. Crowther[2]

This project:

- Review of **100 recent simulation studies** in psychology

Literature Review

"*Statisticians ... often pay too little attention to their own principles of design*"(Hoaglin & Andrews, 1975)



Statistical Computing

This Department will carry articles of high quality on all aspects of computation in statistics.
Papers describing new algorithms, programs, or statistical packages will not contain listings of the program, although the completely documented program must be available from the author. Review of the paper will always include a running test of the program by the referee.
The description of a program or package in this Department should not be construed as an endorsement of it by the American Statistical Association or its Committees, nor is any warranty implied about the validity of the program.
The Editorial Committee will be pleased to confer with authors about the appropriateness of topics or drafts of possible articles.

The Reporting of Computation-Based Results in Statistics

DAVID C. HOAGLIN* and DAVID F. ANDREWS**

STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:4279–4292
Published online 31 August 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2673

The design of simulation studies in medical statistics

Andrea Burton[1, 2, *, †], Douglas G. Altman[1], Patrick Royston[1, 3] and Roger L. Holder[4]

TUTORIAL IN BIOSTATISTICS

WILEY Statistics in Medicine

Using simulation studies to evaluate statistical methods

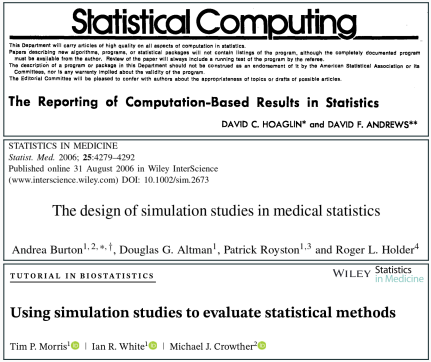
Tim P. Morris[1] | Ian R. White[1] | Michael J. Crowther[2]

This project:

- Review of **100 recent simulation studies** in psychology
- Psychological Methods, Behavior Research Methods, Multivariate Behavioral Research

Literature Review

> "*Statisticians … often pay too little attention to their own principles of design*"(Hoaglin & Andrews, 1975)



Statistical Computing

This Department will carry articles of high quality on all aspects of computation in statistics.
Papers describing new algorithms, programs, or statistical packages will not contain listings of the program, although the completely documented program must be available from the author. Review of the paper will always include a running test of the program by the referee.
The description of a program or package in this Department should not be construed as an endorsement of it by the American Statistical Association or its Committees, nor is any warranty implied about the validity of the program.
The Editorial Committee will be pleased to confer with authors about the appropriateness of topics or drafts of possible articles.

The Reporting of Computation-Based Results in Statistics

DAVID C. HOAGLIN* and DAVID F. ANDREWS**

STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:4279–4292
Published online 31 August 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2673

The design of simulation studies in medical statistics

Andrea Burton[1, 2, *, †], Douglas G. Altman[1], Patrick Royston[1, 3] and Roger L. Holder[4]

TUTORIAL IN BIOSTATISTICS

WILEY Statistics in Medicine

Using simulation studies to evaluate statistical methods

Tim P. Morris[1] | Ian R. White[1] | Michael J. Crowther[2]

This project:

- Review of **100 recent simulation studies** in psychology
- Psychological Methods, Behavior Research Methods, Multivariate Behavioral Research
- Coding of various aspects of reporting

Overview Paper

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Methods

ISSN: 1082-989X

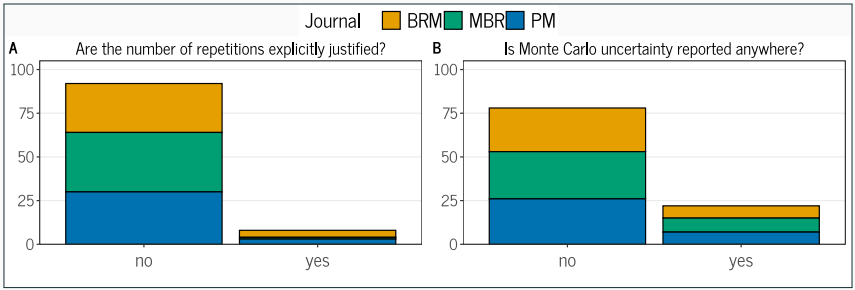
https://doi.org/10.1037/met0000695

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

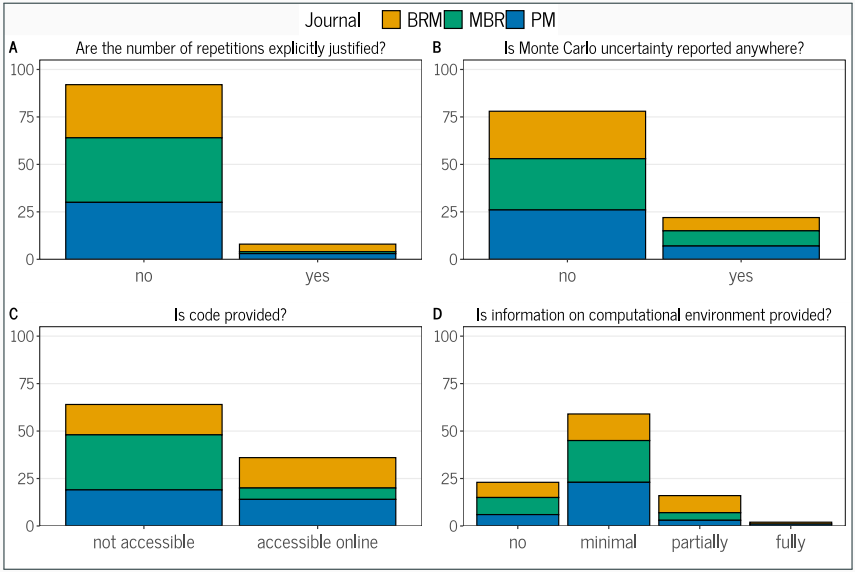
Björn S. Siepe[1], František Bartoš[2], Tim P. Morris[3], Anne-Laure Boulesteix[4, 5], Daniel W. Heck[1], and Samuel Pawel[6, 7]

Main Results

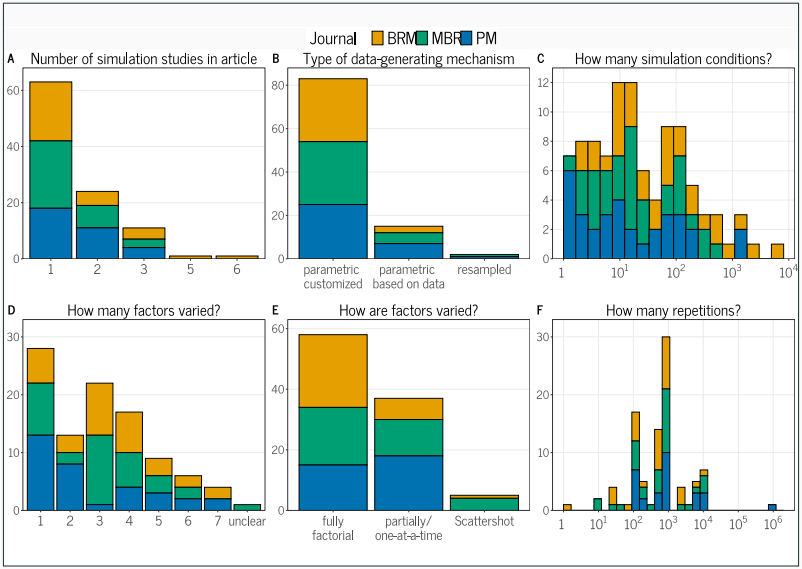
Main Results



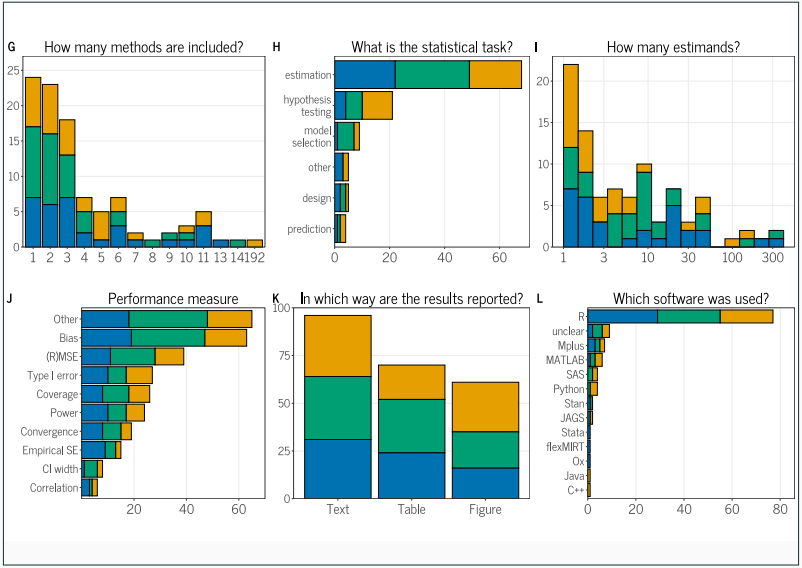
Main Results



Additional Results



Additional Results



Reporting Suggestions

Reporting Suggestions

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	$\text{E}(\hat{\theta}) - \theta$	$(\sum_{i=1}^{n_{\text{sim}}} \hat{\theta}_i / n_{\text{sim}}) - \theta$	$\sqrt{S_{\hat{\theta}}^2 / n_{\text{sim}}}$	$S_{\hat{\theta}}^2 / \text{MCSE}_*^2$
Relative bias	$\{\text{E}(\hat{\theta}) - \theta\}/\theta$	$\{(\sum_{i=1}^{n_{\text{sim}}} \hat{\theta}_i / n_{\text{sim}}) - \theta\}/\theta$	$\sqrt{S_{\hat{\theta}}^2 / (\theta^2 n_{\text{sim}})}$	$S_{\hat{\theta}}^2 / (\text{MCSE}_*^2\, \theta^2)$
Mean square error (MSE)	$\text{E}\{(\hat{\theta} - \theta)^2\}$	$\sum_{i=1}^{n_{\text{sim}}} (\hat{\theta}_i - \theta)^2 / n_{\text{sim}}$	$\sqrt{S_{(\hat{\theta}-\theta)^2}^2 / n_{\text{sim}}}$	$S_{(\hat{\theta}-\theta)^2}^2 / \text{MCSE}_*^2$
Root mean square error (RMSE)	$\sqrt{\text{E}\{(\hat{\theta} - \theta)^2\}}$	$\sqrt{\sum_{i=1}^{n_{\text{sim}}} (\hat{\theta}_i - \theta)^2 / n_{\text{sim}}}$	$\sqrt{S_{(\hat{\theta}-\theta)^2}^2 / (4 n_{\text{sim}} \widehat{\text{MSE}})}$	$S_{(\hat{\theta}-\theta)^2}^2 / (4\widehat{\text{MSE}}\ \text{MCSE}_*^2)$
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	$\text{E}(\text{CI}_{\text{upper}} - \text{CI}_{\text{lower}})$	$\sum_{i=1}^{n_{\text{sim}}} (\text{CI}_{i,\text{upper}} - \text{CI}_{i,\text{lower}})/n_{\text{sim}}$	$\sqrt{S_W^2 / n_{\text{sim}}}$	S_W^2 / MCSE_*^2
Mean of generic statistic G	$\text{E}(G)$	$\sum_{i=1}^{n_{\text{sim}}} G_i / n_{\text{sim}}$	$\sqrt{S_G^2 / n_{\text{sim}}}$	S_G^2 / MCSE_*^2

Note. Table adapted from Table 6 in Morris et al. (2019)

Potential improvements

How to address questionable research practices?

Researchers

- **Preregistered simulation protocols**
- **Adversarial collaboration**
- **Blinding** of analysis
- **Transparent reporting** (e.g., disclose non-neutrality)

How to address questionable research practices?

Researchers

- **Preregistered simulation protocols**
- **Adversarial collaboration**
- **Blinding** of analysis
- **Transparent reporting** (e.g., disclose non-neutrality)

Reviewers, journals, funders

- Encourage **simulation protocols**
- **Incentivize neutrality and transparency** in simulation studies
- **Deincentivize outperforming** state-of-the-art methods

Simulation study protocols

STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:4279–4292
Published online 31 August 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2673

The design of simulation studies in medical statistics

Andrea Burton[1, 2, *, †], Douglas G. Altman[1], Patrick Royston[1,3] and Roger L. Holder[4]

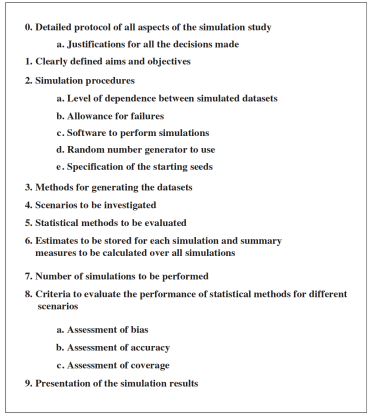
"***When planning*** *a simulation study, it is* ***recommended that a detailed protocol be produced****, giving full details of how the study will be performed, analysed and reported.*" Burton et al. (2006)

Simulation study protocols

Advantages

+ Planning and reporting
+ Transparency and replicability
+ Can be preregistered
? Less/more work

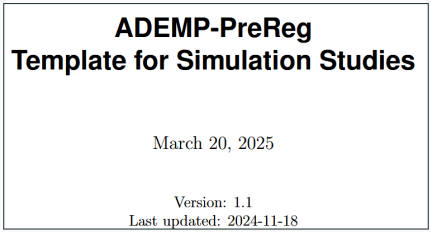
→ **How to structure protocol?**



0. Detailed protocol of all aspects of the simulation study
    a. Justifications for all the decisions made
1. Clearly defined aims and objectives
2. Simulation procedures
    a. Level of dependence between simulated datasets
    b. Allowance for failures
    c. Software to perform simulations
    d. Random number generator to use
    e. Specification of the starting seeds
3. Methods for generating the datasets
4. Scenarios to be investigated
5. Statistical methods to be evaluated
6. Estimates to be stored for each simulation and summary measures to be calculated over all simulations
7. Number of simulations to be performed
8. Criteria to evaluate the performance of statistical methods for different scenarios
    a. Assessment of bias
    b. Assessment of accuracy
    c. Assessment of coverage
9. Presentation of the simulation results

Proposal from Burton et al. (2006)

The ADEMP-PreReg template


ADEMP-PreReg
Template for Simulation Studies

March 20, 2025

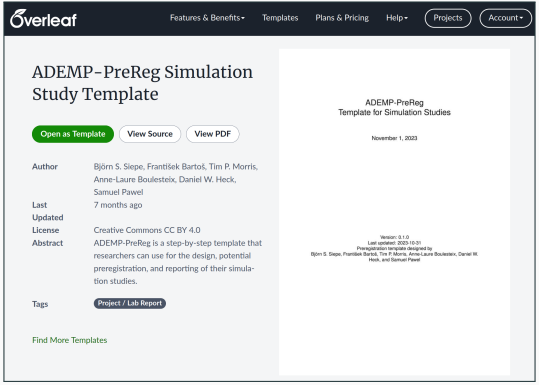
Version: 1.1
Last updated: 2024-11-18

Protocol template based on:

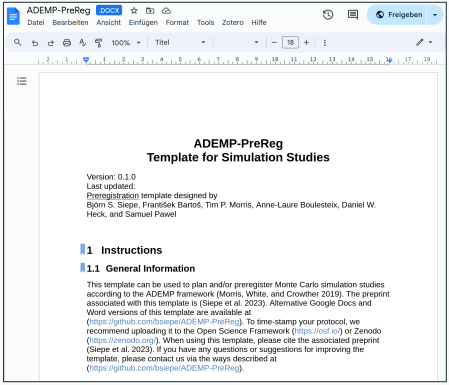
- **ADEMP structure** (Morris et al., 2019)
- **Open science** aspects
- **Reproducibility** aspects

ADEMP: doi.org/10.1214/ss/1009213726, ADEMP-PreReg: doi.org/10.31234/osf.io/ufgy6

The ADEMP-PreReg template – Different versions

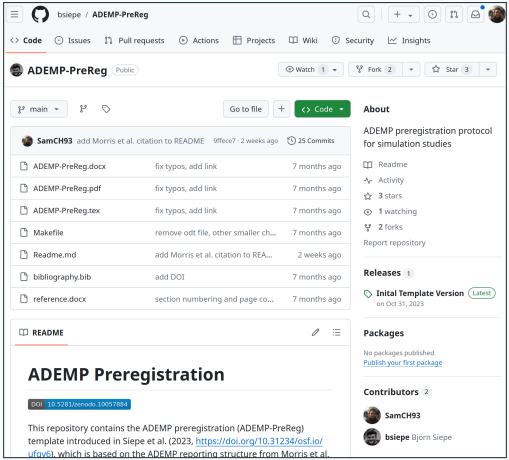


LaTeX, Overleaf



MS/Libre office, Google docs

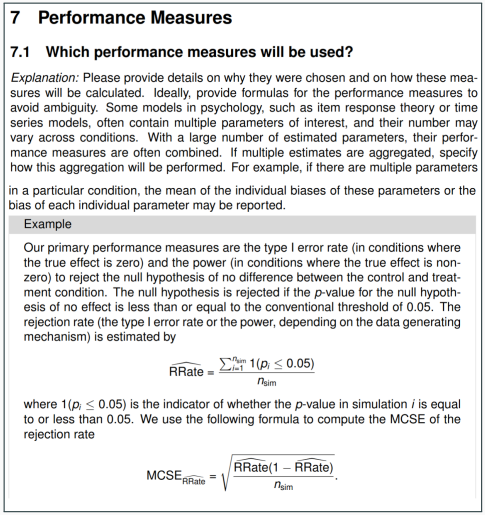
The ADEMP-PreReg template – A living document



https://github.com/bsiepe/ADEMP-PreReg

The ADEMP-PreReg template – Overview

1. Instructions
2. General information
3. **A**ims
4. **D**ata-generating mechanism
5. **E**stimands and targets
6. **M**ethods
7. **P**erformance Measures
8. Computational details



7 Performance Measures

7.1 Which performance measures will be used?

Explanation: Please provide details on why they were chosen and on how these measures will be calculated. Ideally, provide formulas for the performance measures to avoid ambiguity. Some models in psychology, such as item response theory or time series models, often contain multiple parameters of interest, and their number may vary across conditions. With a large number of estimated parameters, their performance measures are often combined. If multiple estimates are aggregated, specify how this aggregation will be performed. For example, if there are multiple parameters in a particular condition, the mean of the individual biases of these parameters or the bias of each individual parameter may be reported.

Example

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 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

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

where $1(p_i \leq 0.05)$ 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 rejection rate

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

The ADEMP-PreReg template

Purposes

The ADEMP-PreReg template

Purposes

- Blueprint for **planning**, **reporting** & **reviewing** of simulation studies

The ADEMP-PreReg template

Purposes

- Blueprint for **planning**, **reporting** & **reviewing** of simulation studies
- **Preregistration** brings multiple benefits similar to other empirical research
  - Avoid QRPs
  - Increase transparency
  - Improve informativeness

Limitations

The ADEMP-PreReg template

Purposes

- Blueprint for **planning**, **reporting** & **reviewing** of simulation studies
- **Preregistration** brings multiple benefits similar to other empirical research
  - Avoid QRPs
  - Increase transparency
  - Improve informativeness

Limitations

- Preregistration could be **faked**

The ADEMP-PreReg template

Purposes

- Blueprint for **planning**, **reporting** & **reviewing** of simulation studies
- **Preregistration** brings multiple benefits similar to other empirical research
  - Avoid QRPs
  - Increase transparency
  - Improve informativeness

doi:10.5281/zenodo.7994221

Limitations

- Preregistration could be **faked**
- May **slow down** exploratory research

Replications: What?

Reproduction

Replications: What?

Reproduction

- Checking for **computational reproducibility**
- Using the **same code** and data
- Confirms technical **correctness & transparency**

Replications: What?

Reproduction

- Checking for **computational reproducibility**
- Using the **same code** and data
- Confirms technical **correctness & transparency**

Direct Replication

Replications: What?

Reproduction

- Checking for **computational reproducibility**
- Using the **same code** and data
- Confirms technical **correctness & transparency**

Direct Replication

- Using descriptions in paper
- Same methods, new implementation
- Tests methodological **clarity**

Replications: What?

Reproduction

- Checking for **computational reproducibility**
- Using the **same code** and data
- Confirms technical **correctness & transparency**

Direct Replication

- Using descriptions in paper
- Same methods, new implementation
- Tests methodological **clarity**

Conceptual Replication

Replications: What?

Reproduction

- Checking for **computational reproducibility**
- Using the **same code** and data
- Confirms technical **correctness & transparency**

Direct Replication

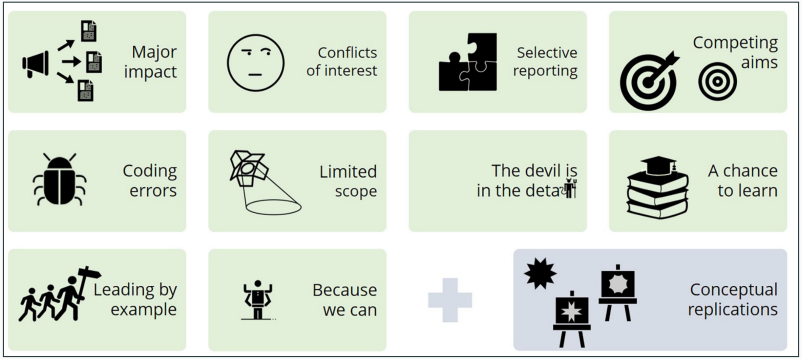
- Using descriptions in paper
- Same methods, new implementation
- Tests methodological **clarity**

Conceptual Replication

- Investigating similar underlying question
- Alternative methods or scenarios
- Tests **generalizability** of findings

Replications: Why?

Replications: Why?

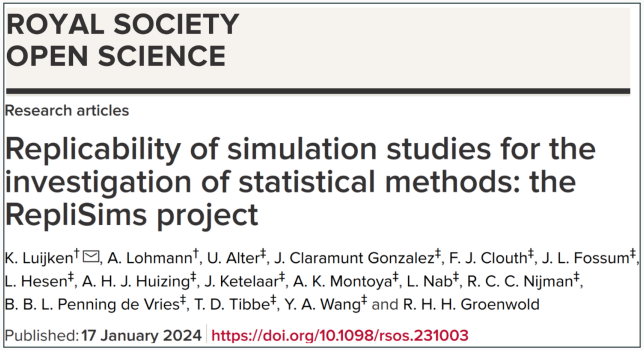


"*simulation studies face challenges similar to other experimental empirical research and hence should not be exempt from replication attempts*"

Lohmann et al. (2022)

Replications: Results?

Replications: Results?



ROYAL SOCIETY OPEN SCIENCE

Research articles

Replicability of simulation studies for the investigation of statistical methods: the RepliSims project

K. Luijken†, A. Lohmann†, U. Alter‡, J. Claramunt Gonzalez‡, F. J. Clouth‡, J. L. Fossum‡, L. Hesen‡, A. H. J. Huizing‡, J. Ketelaar‡, A. K. Montoya‡, L. Nab‡, R. C. C. Nijman‡, B. B. L. Penning de Vries‡, T. D. Tibbe‡, Y. A. Wang‡ and R. H. H. Groenwold

Published: 17 January 2024 | https://doi.org/10.1098/rsos.231003

"*the information provided in the original publication of highly cited and influential simulation studies was* ***often insufficient for complete replication***"

Luijken et al. (2024)

Replications: Results?

- **Almost Perfect Replication:** Results were **almost perfectly replicated** in three studies.

Replications: Results?

- **Almost Perfect Replication:** Results were **almost perfectly replicated** in three studies.
- **Impossible Replication:** One study provided **insufficient information** to implement any simulation scenarios.

Replications: Results?

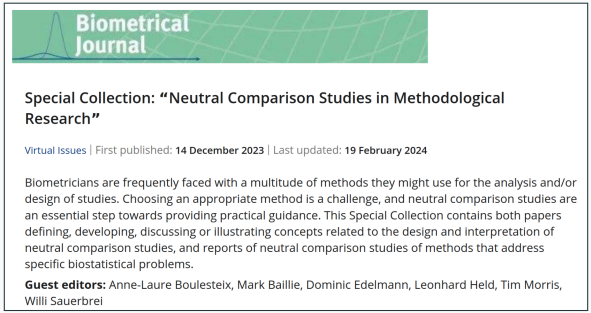
- **Almost Perfect Replication:** Results were **almost perfectly replicated** in three studies.
- **Impossible Replication:** One study provided **insufficient information** to implement any simulation scenarios.
- **Partial Replication:** Four studies with varying challenges:

Replications: Results?

- **Almost Perfect Replication:** Results were **almost perfectly replicated** in three studies.
- **Impossible Replication:** One study provided **insufficient information** to implement any simulation scenarios.
- **Partial Replication:** Four studies with varying challenges:
  - **Austin:** Parameter values misaligned with data descriptions
  - **Flora & Curran:** Overall consistency, but differences due to software environments
  - **MacKinnon et al.:** Main conclusions replicated, but one method excluded due to unclear procedures
  - **Peters et al.:** General patterns matched, but results only shown as figures made matching difficult

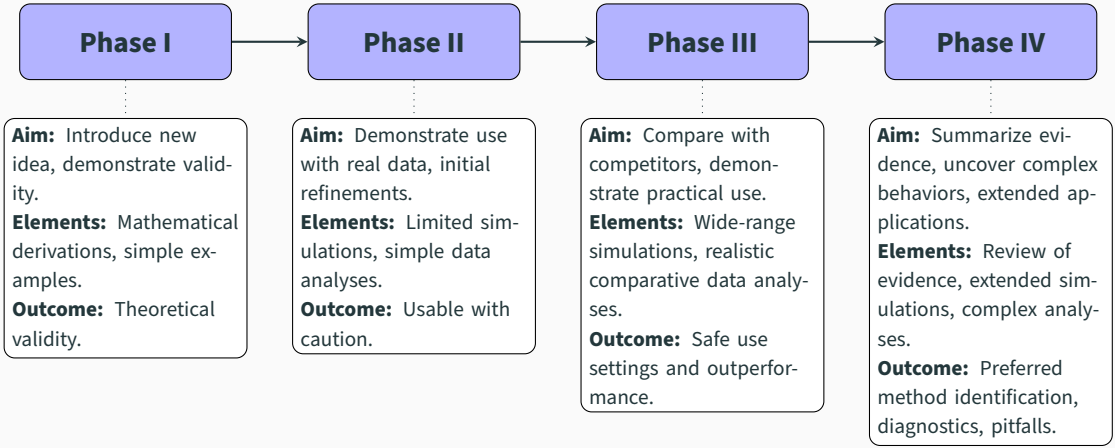
Current trends

Current trends



- Focus on “**neutral comparison studies**” (Boulesteix et al., 2013)
- Some journals adopt **reproducibility checks** (Wrobel et al., 2024)
- **Various fields** discuss how to improve methodological research (e.g., Robinson and Vitek, 2019; Van Mechelen et al., 2023; Herrmann et al., 2024)
- **Meta-research** on simulation/benchmarking studies continues

Phases of methodological research (Heinze et al., 2024)



Based on Heinze et al. (2024)

Against “one method fits all [data sets]” (Strobl and Leisch, 2024)

Against “one method fits all [data sets]” (Strobl and Leisch, 2024)

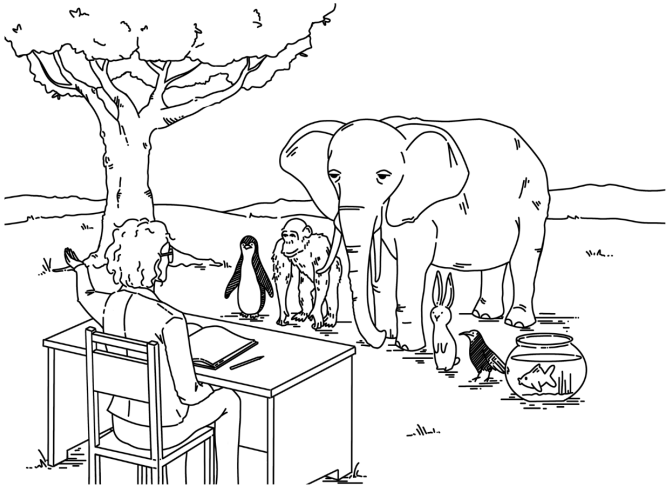
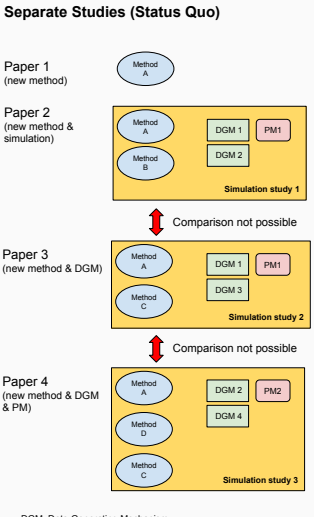
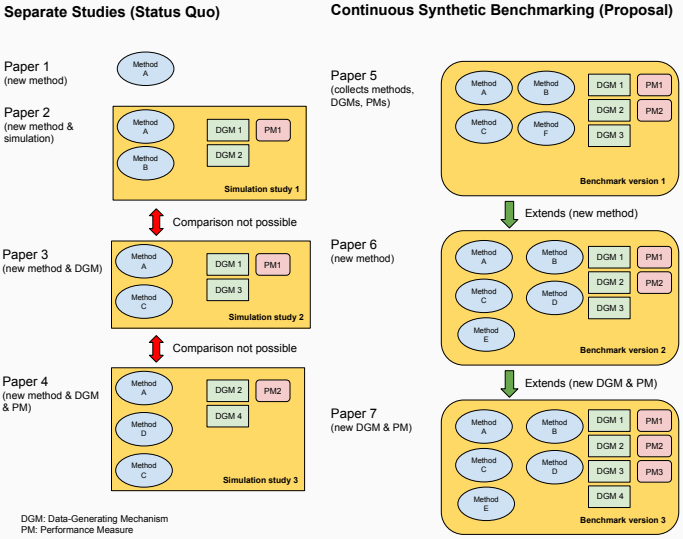


FIGURE 1 “Climb the tree”. Drawing by Alexandra Kalberer

WIP: Synthetic benchmarking

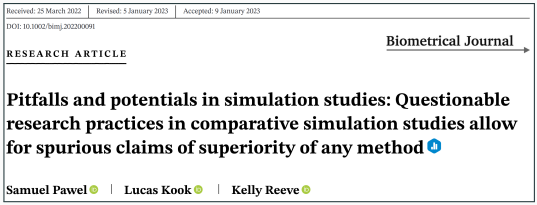


WIP: Synthetic benchmarking



Discussion

Conclusions



Received: 25 March 2022 | Revised: 5 January 2023 | Accepted: 9 January 2023

DOI: 10.1002/bimj.202200091

Biometrical Journal

RESEARCH ARTICLE

Pitfalls and potentials in simulation studies: Questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method

Samuel Pawel | Lucas Kook | Kelly Reeve

Psychological Methods

© 2024 American Psychological Association
ISSN: 1082-989X

https://doi.org/10.1037/met0000695

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

Björn S. Siepe[1], František Bartoš[2], Tim P. Morris[3], Anne-Laure Boulesteix[4, 5], Daniel W. Heck[1], and Samuel Pawel[6, 7]

- **Simulation studies** are ubiquitous in methodological research
- Simulation studies can be impacted by **questionable research practices** and misaligned **incentives**
- **Protocols** have potential to improve simulation studies
- Meta-research, discussions, and reforms needed to **increase awareness** and **improve standards**

Open questions

- Which simulation studies require which **degree of rigour**?
- How to avoid **cheating in preregistration**?
- How can journals/researchers/reviewers/communities promote **good practices**?
- Other ways to **improve** simulation studies?

A multidisciplinary collaboration

František Bartoš

Daniel W. Heck

Tim P. Morris

A.-L. Boulesteix

Anna Lohmann

Samuel Pawel

Get In Touch

- bjoern.siepe@uni-marburg.de
- https://bsiepe.github.io/

Paper & Slides

References i

Boulesteix, A.-L., Hoffmann, S., Charlton, A., and Seibold, H. (2020). A replication crisis in methodological research? *Significance*, 17(5):18–21. doi:10.1111/1740-9713.01444.

Boulesteix, A.-L., Lauer, S., and Eugster, M. J. A. (2013). A plea for neutral comparison studies in computational sciences. *PLoS ONE*, 8(4):e61562. doi:10.1371/journal.pone.0061562.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231. doi:10.1214/ss/1009213726.

Burton, A., Altman, D. G., Royston, P., and Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24):4279–4292. doi:10.1002/sim.2673.

Heinze, G., Boulesteix, A.-L., Kammer, M., Morris, T. P., and and, I. R. W. (2024). Phases of methodological research in biostatistics—building the evidence base for new methods. *Biometrical Journal*. doi:10.1002/bimj.202200222.

Hennig, C. (2018). Some thoughts on simulation studies to compare clustering methods. *Archives of Data Science, Series A*, 5(1). doi:10.5445/KSP/1000087327/24.

Herrmann, M., Lange, F. J. D., Eggensperger, K., Casalicchio, G., Wever, M., Feurer, M., Rügamer, D., Hüllermeier, E., Boulesteix, A.-L., and Bischl, B. (2024). Position: Why we must rethink empirical research in machine learning. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F., editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 18228–18247. PMLR.

Hoaglin, D. C. and 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.

Hu, L.-t. and 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.

References ii

Kirmeier, E., Eriksson, L. I., Lewald, H., Jonsson Fagerlund, M., et al. (2019). Post-anaesthesia pulmonary complications after use of muscle relaxants (popular): a multicentre, prospective observational study. *The Lancet Respiratory Medicine*, 7(2):129–140. doi:10.1016/s2213-2600(18)30294-7.

Koehler, E., Brown, E., and Haneuse, S. J.-P. A. (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.

Lohmann, A., Astivia, O. L. O., Morris, T. P., and Groenwold, R. H. H. (2022). It's time! Ten reasons to start replicating simulation studies. 2:973470. doi:10.3389/fepid.2022.973470.

Luijken, K., Lohmann, A., Alter, U., Claramunt Gonzalez, J., Clouth, F. J., Fossum, J. L., Hesen, L., Huizing, A. H. J., Ketelaar, J., Montoya, A. K., Nab, L., Nijman, R. C. C., Penning de Vries, B. B. L., Tibbe, T. D., Wang, Y. A., and Groenwold, R. H. H. (2024). Replicability of simulation studies for the investigation of statistical methods: the replisims project. *Royal Society Open Science*, 11(1). doi:10.1098/rsos.231003.

Luijken, K., Lohmann, A., Alter, U., Gonzalez, J. C., Clouth, F. J., Fossum, J. L., Hesen, L., Huizing, A. H. J., Ketelaar, J., Montoya, A. K., Nab, L., Nijman, R. C. C., de Vries, B. B. L. P., Tibbe, T. D., Wang, Y. A., and Groenwold, R. H. H. (2023). Replicability of simulation studies for the investigation of statistical methods: The replisims project. doi:10.48550/ARXIV.2307.02052. arXiv preprint.

Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102. doi:10.1002/sim.8086.

Pawel, S., Bartoš, F., Siepe, B. S., and Lohmann, A. (2024a). Handling missingness, failures, and non-convergence in simulation studies: A review of current practices and recommendations. doi:10.48550/arXiv.2409.18527. URL `https://arxiv.org/abs/2409.18527`. arXiv preprint.

Pawel, S., Kook, L., and Reeve, K. (2024b). Pitfalls and potentials in simulation studies: Questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method. *Biometrical Journal*, (e2200091):1–19. doi:10.1002/bimj.202200091.

References iii

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., and 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.

Robinson, M. D. and Vitek, O. (2019). Benchmarking comes of age. *Genome Biology*, 20(1):205. doi:10.1186/s13059-019-1846-5.

Siepe, B. S., Bartoš, F., Morris, T. P., Boulesteix, A.-L., Heck, D. W., and Pawel, S. (2024). Simulation studies for methodological research in psychology: A standardized template for planning, preregistration, and reporting. *Psychological Methods*. doi:10.31234/osf.io/ufgy6. URL `https://doi.org/10.1037/met0000695`. to appear.

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. and Leisch, F. (2024). Against the "one method fits all data sets" philosophy for comparison studies in methodological research. *Biometrical Journal*. doi:10.1002/bimj.202200104.

Ullmann, T., Beer, A., Hünemörder, M., Seidl, T., and Boulesteix, A.-L. (2022). Over-optimistic evaluation and reporting of novel cluster algorithms: An illustrative study. *Advances in Data Analysis and Classification*. doi:10.1007/s11634-022-00496-5. Advance online publication.

Van Mechelen, I., Boulesteix, A., Dangl, R., Dean, N., Hennig, C., Leisch, F., Steinley, D., and Warrens, M. J. (2023). A white paper on good research practices in benchmarking: The case of cluster analysis. *WIREs Data Mining and Knowledge Discovery*, 13(6). doi:10.1002/widm.1511.

van Smeden, M., de Groot, J. A. H., Moons, K. G. M., Collins, G. S., Altman, D. G., Eijkemans, M. J. C., and Reitsma, J. B. (2016). No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology*, 16(1). doi:10.1186/s12874-016-0267-3.

Wrobel, J., Hector, E. C., Crawford, L., McGowan, L. D., da Silva, N., Goldsmith, J., Hicks, S., Kane, M., Lee, Y., Mayrink, V., Paciorek, C. J., Usher, T., and Wolfson, J. (2024). Partnering with authors to enhance reproducibility at JASA. *Journal of the American Statistical Association*, 119(546):795–797. doi:10.1080/01621459.2024.2340557.