

# Simulation studies in psychology: State of the art, problems, and potential solutions

---

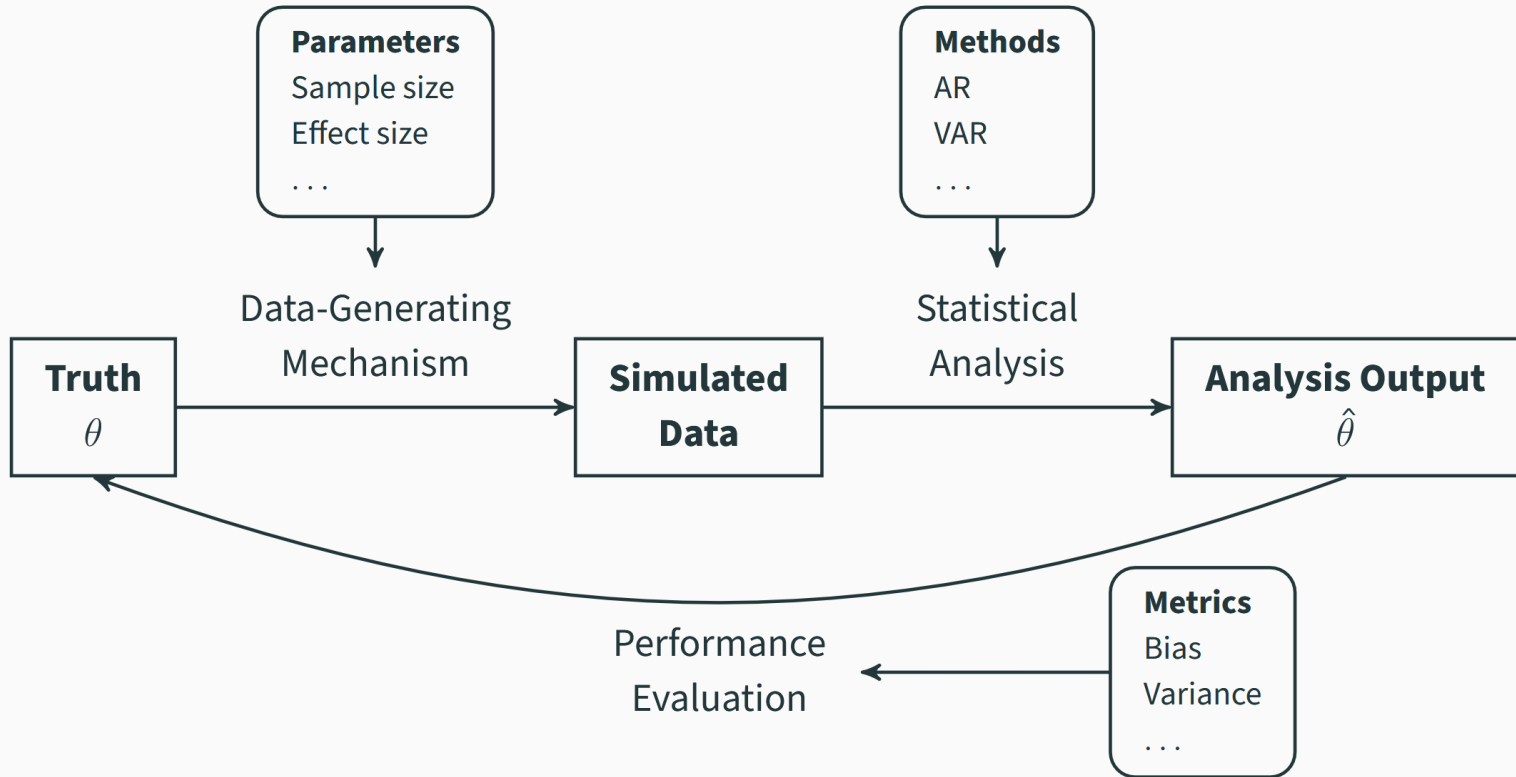
Lunch Lectures on Simulation Studies · July 2026

Björn Siepe

Psychological Methods Lab, Department of Psychology, Marburg University

# Part I: Background and Issues

# Simulation studies



# Issues in simulation studies



Marcin Wichary - Wikimedia

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

# Meta-Science on Simulation Studies

## Interdisciplinary

- (Bio-)statistics: Morris et al. (2019)
- Ecology: Williams et al. (2024)
- Psychometrics: Feinberg & Rubright (2016)

## Transparency

- 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)

## Execution Issues

- Many degrees of freedom (Pawel et al., 2024)
- Over-Optimism (e.g., Ullmann et al., 2022)
- Unclear handling of non-convergence (Pawel et al., 2025; Wünsch et al., 2025)



© 2024 American Psychological Association  
ISSN: 1082-989X

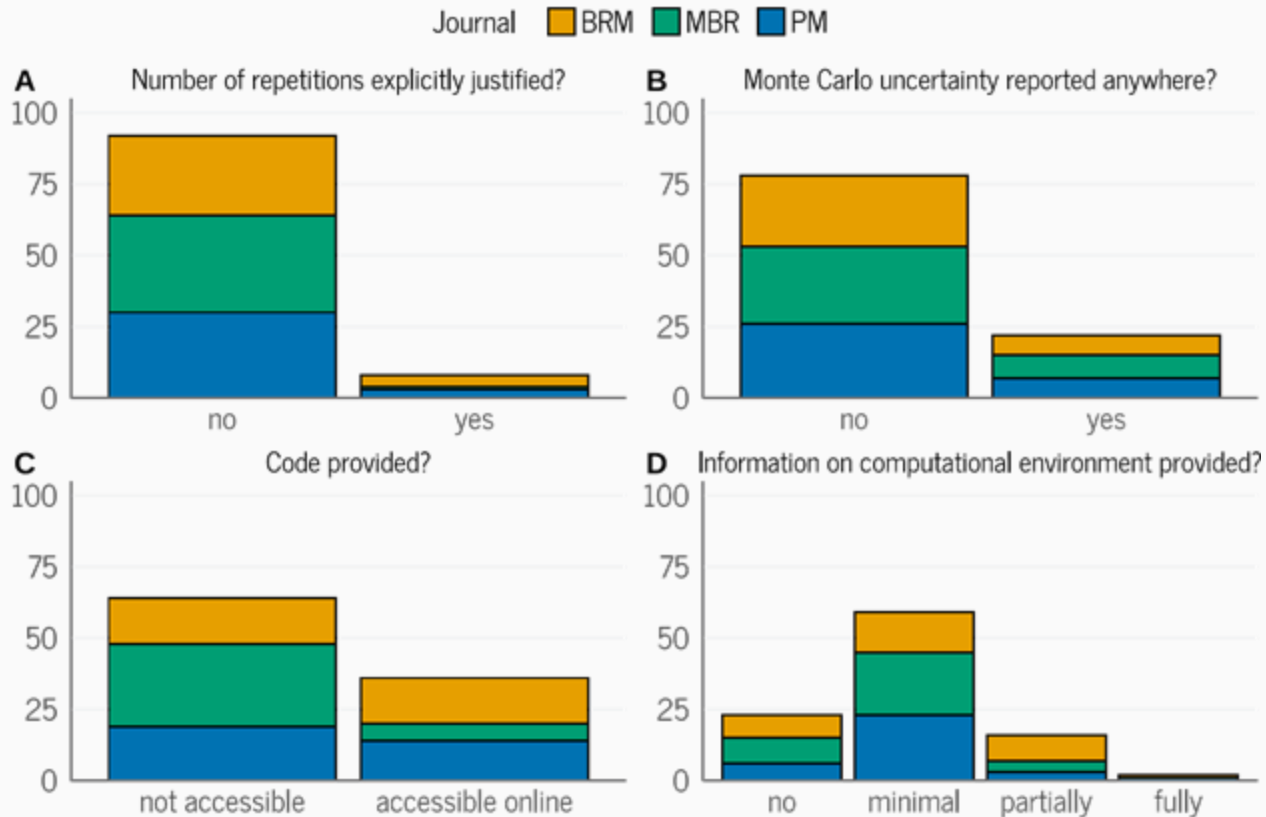
Psychological Methods

<https://doi.org/10.1037/met0000695>

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

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

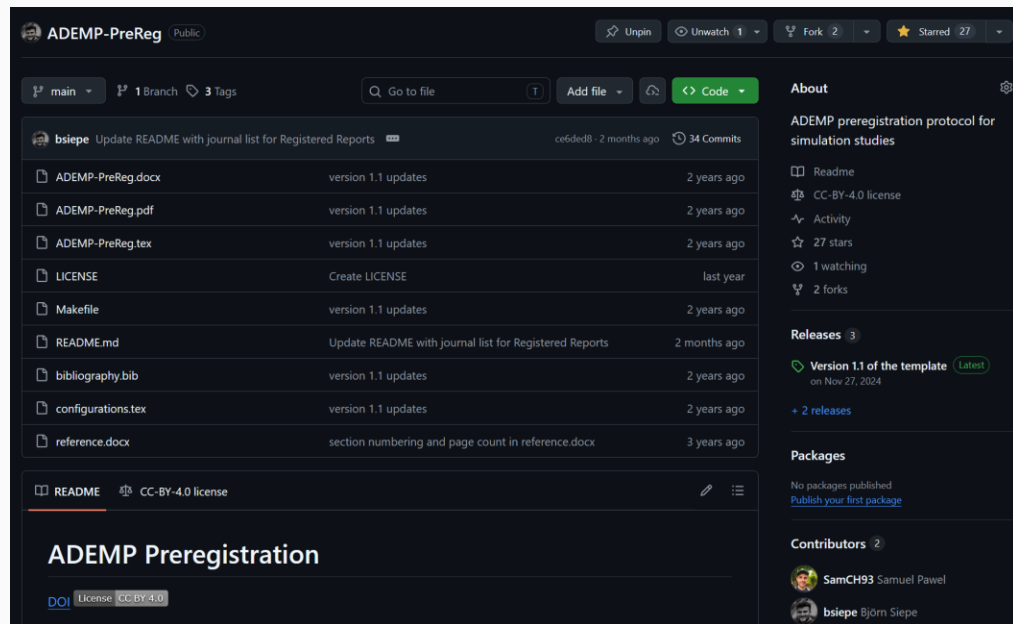
# Simulation studies in psychology



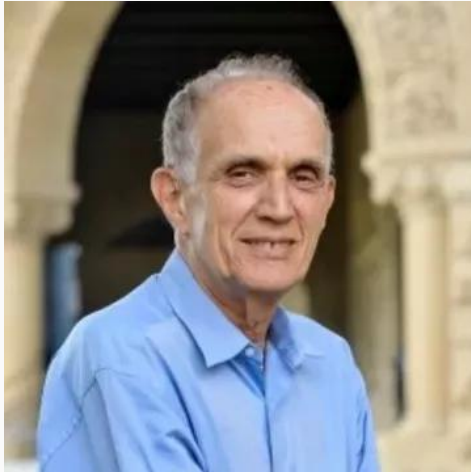
# Simulation studies in psychology

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

1. Instructions
2. General information
3. Aims
4. Data-generating mechanisms
5. Estimands & targets
6. Methods
7. Performance measures
8. Computational details



# Issues in simulation studies



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

“*[...] 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

# Potential Solutions

## Adversarial Collaboration



The Splintered Mind Blog

## Analysis Blinding




Artwork by Sandbox Studio, Chicago with Corinne Mucha

## Reporting Standards/Full Disclosure/Lab Handbook

**ADEMP-PreReg  
Template for Simulation Studies**

# Part II: To preregister or not to preregister?


# The idea of simulation protocols



Statistics  
in Medicine

Research Article

## The design of simulation studies in medical statistics

[Andrea Burton](#)  [Douglas G. Altman](#), [Patrick Royston](#), [Roger L. Holder](#)

First published: 31 August 2006 | <https://doi.org/10.1002/sim.2673> | [VIEW METRICS](#)

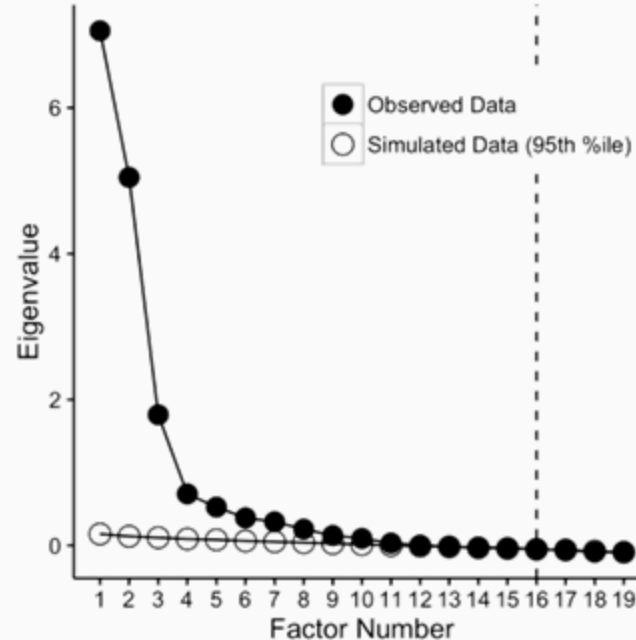
“*When planning a simulation study, it is recommended that a **detailed protocol be produced**, giving full details of how the study will be performed, analysed and reported.*”

Burton et al., 2006

# Claims based on simulation studies

## Why, when, and how to (or not to) preregister a simulation study

Björn S. Siepe<sup>† 1</sup>, František Bartoš<sup>2\*</sup>, Anne-Laure Boulesteix<sup>3,4\*</sup>,  
Daniel W. Heck<sup>1\*</sup>, Aaron Peikert<sup>5,6\*</sup>, Alexandra Sarafoglou<sup>2\*</sup>,  
Samuel Pawel<sup>7</sup>

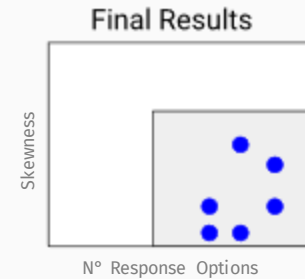


Sakaluk & Short (2016)

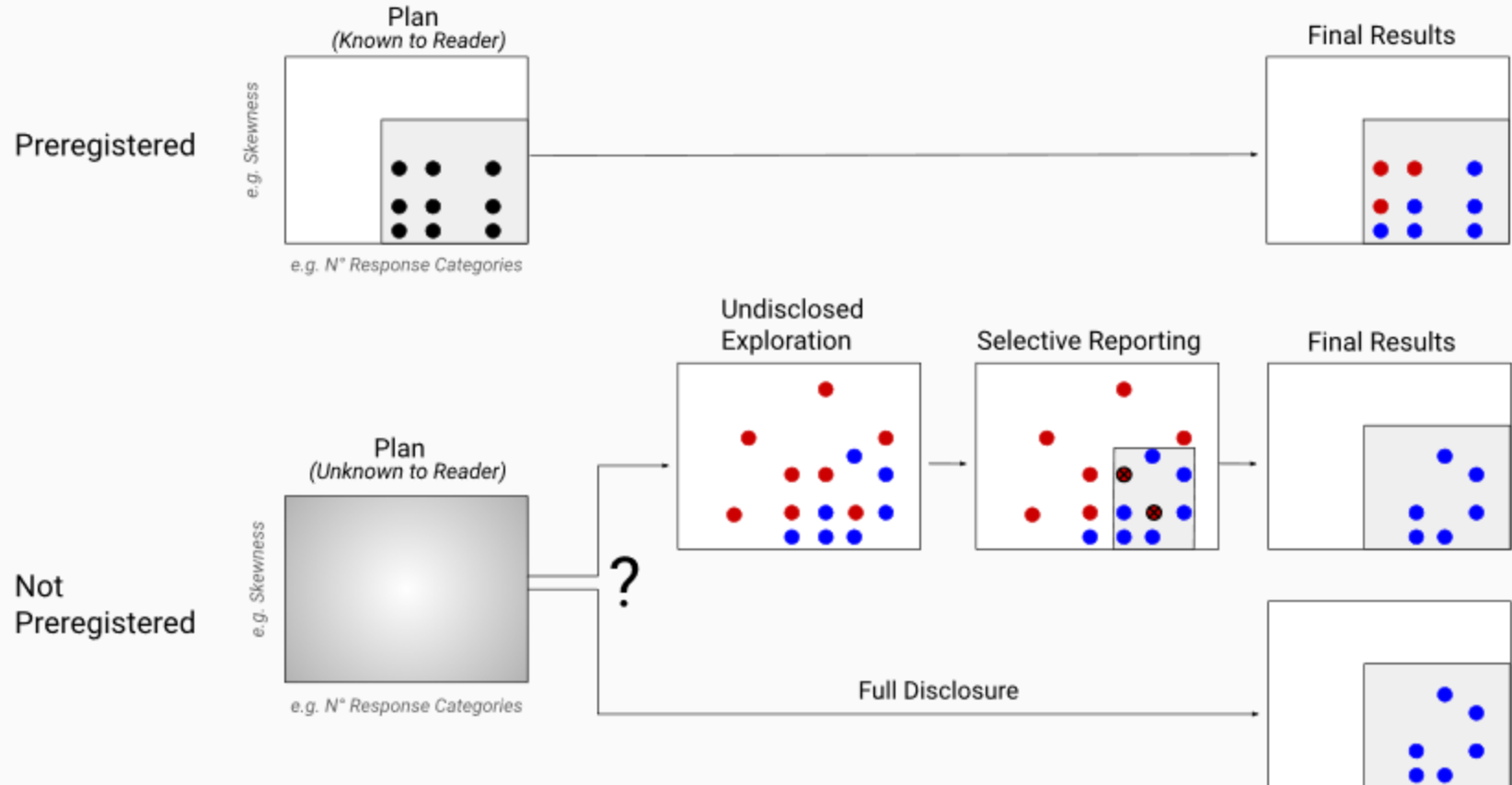
“We recommend parallel analysis for ordinal items

Hypothetical authors

# Illustration of preregistration



# Illustration of preregistration



# Theoretical advantages of preregistration



Preregistration reduces **epistemic uncertainty** about the process that produced simulation results

## **Falsificationist**

Simulation counts as evidence if it could have caught the method failing

## **Inductionist**

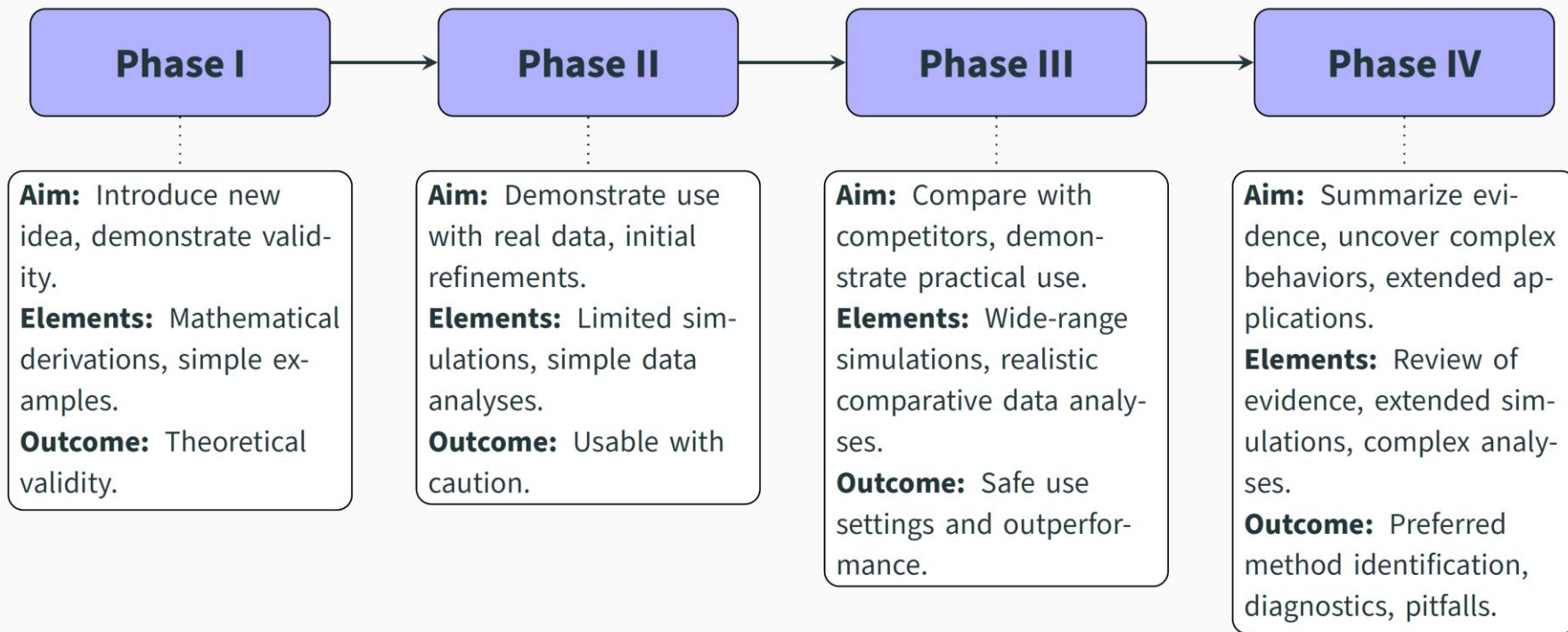
Convincing simulations need difficult conditions (catch a bad method) and easy enough conditions (let a good method succeed)

## **Predictivist**

Predicting an outcome and prespecifying a plan in advance carries more evidential weight than explaining it afterward

# When might they work?

Increasing usefulness of preregistration



# FAQ about preregistration



**Is preregistration in itself a sign of good quality?**

**What about deviations?**

**Do I need theoretical expectations?**

**When does a simulation study start?**

**What about cheating?**

No, but allows assessment thereof

Be transparent

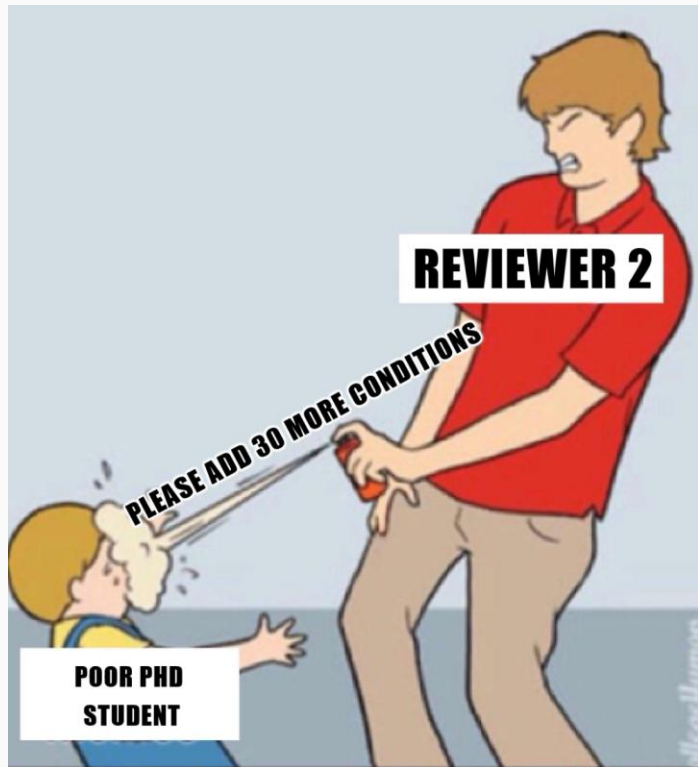
No, but can be helpful

Hard to determine, but there are practical tools

Can never be fully ruled out



# Registered Reports to the rescue



Created with ImageResizer



Centre for Open Science

# Part III: Benchmarking to the rescue?

# Incentivizing cumulative simulation research

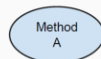
## Living Synthetic Benchmarks:

## A Neutral and Cumulative Framework for Simulation Studies

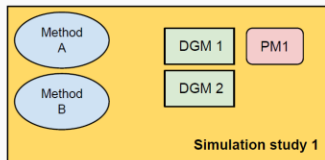
František Bartoš <sup>1</sup>, Samuel Pawel <sup>2</sup>, Björn S. Siepe <sup>3</sup>

### Separate Studies (Status Quo)

Paper 1  
(new method)



Paper 2  
(new method & simulation)



### Continuous Synthetic Benchmarking (Proposal)

Paper 3  
(new n...)

07.10.2026, Noon

František Bartoš

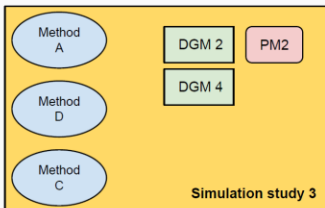
Synthetic benchmarking for simulation studies

[Join](#)



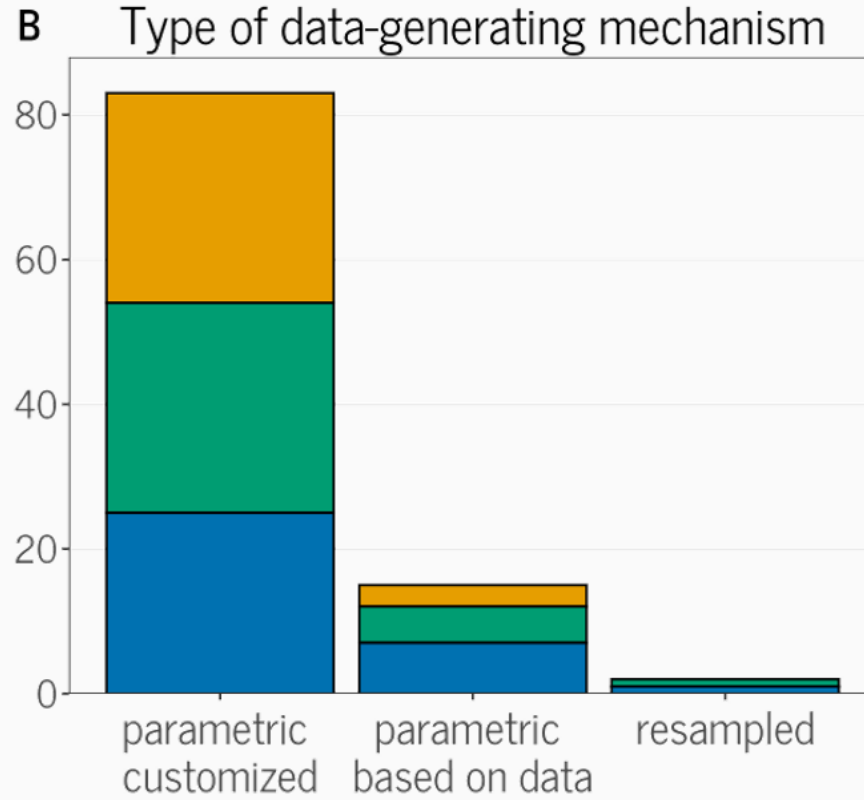
↕ Comparison not possible

Paper 4  
(new method & DGM & PM)



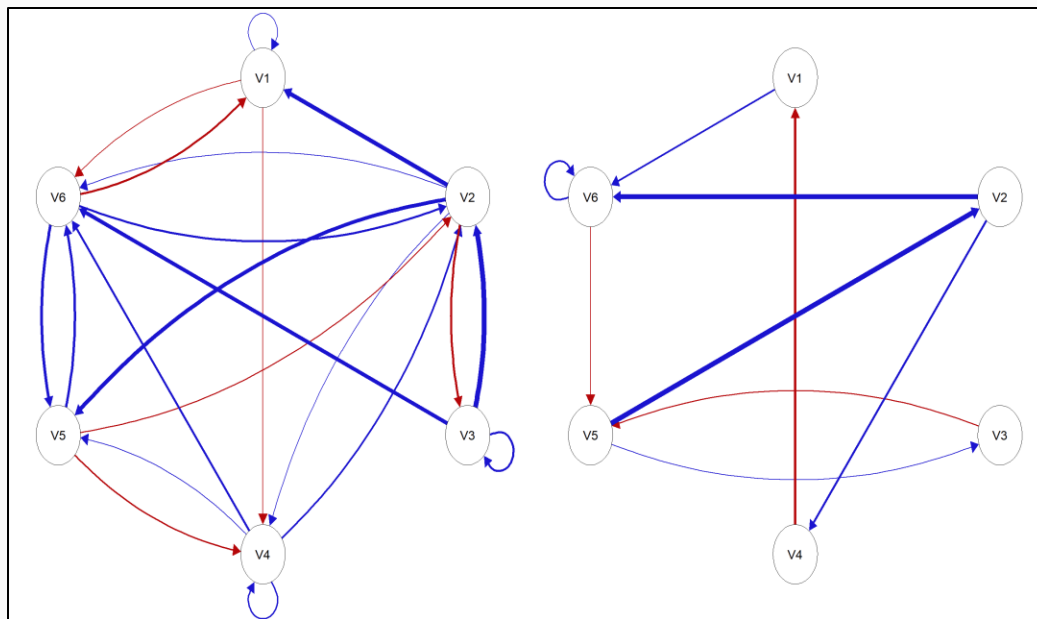
DGM: Data-Generating Mechanism  
PM: Performance Measure

# What about *real* benchmarking?



# Particular Issues in Psychology

What is a realistic data-generating mechanism?



Hard to understand when performance differences actually matter

# Real-data benchmarking

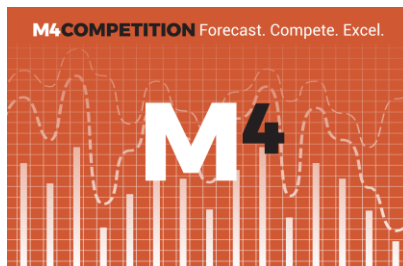
**Idea:** Application of different algorithms to real-world data

## Computer Vision/ML



Source: Steni Sebastian, Medium

## Forecasting



Source: Rob Hyndman

## Biostatistics

Robinson and Vitek *Genome Biology* (2019) 20:205  
<https://doi.org/10.1186/s13059-019-1846-5>

### EDITORIAL

## Benchmarking comes of age

Mark D. Robinson<sup>1\*</sup> and Olga Vitek<sup>2\*</sup>

Psychology has been slow to adopt benchmarking (RoCCA & Yarkoni, 2021)

# Real-data benchmarking

We now have lots of open data!

## Item Response Theory

Item Response Warehouse

## Time Series



## Cognitive Modeling

Behavior Research Methods (2023) 57:208  
<https://doi.org/10.3758/s13428-025-02717-z>

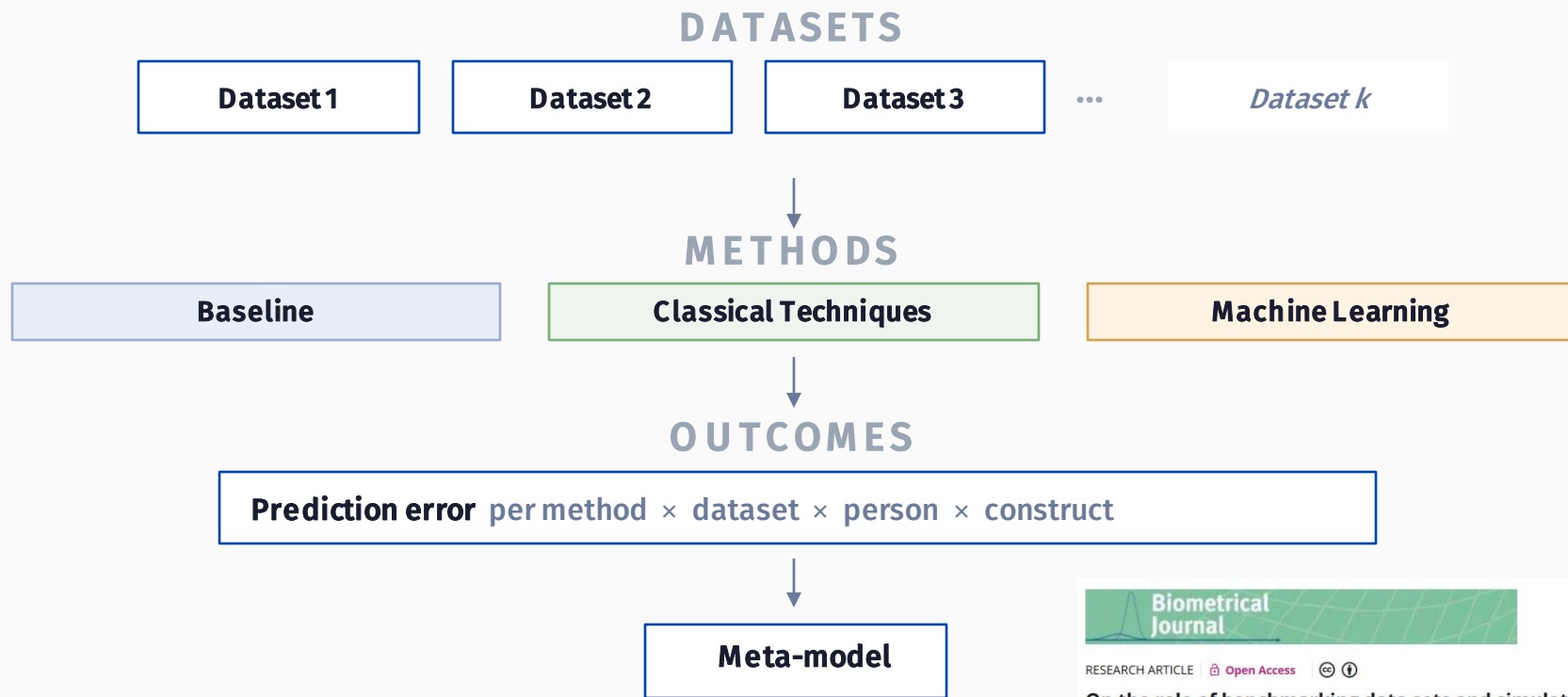
ORIGINAL MANUSCRIPT

Attentional control data collection: A resource for efficient data reuse

Julia M. Haaf<sup>1</sup> · Madlen Hoffstadt<sup>1</sup> · Sven Lesche<sup>3</sup>

Psychology is ready for benchmarking

# Benchmarking psychological methods



Methodological research has **quality & transparency issues** like any empirical field

---

Some potential solutions:

## Preregistration

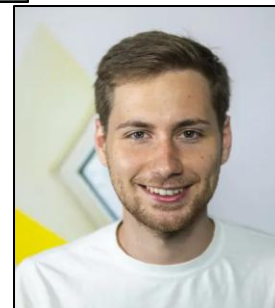
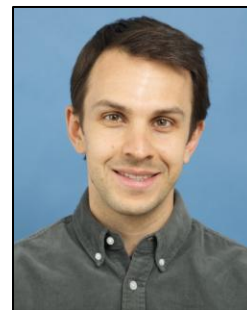
makes generalizability assessable and process transparent, **not automatically better**

## Benchmarking

closes the gap in **two ways**:

- explicit conditions
- real-world grounding

# Collaborators



# References

- Efron, B. (2001). Statistical modeling: The two cultures: Comment. *Statistical Science*, 16(3), 218-219.
- Heinze, G., Boulesteix, A. L., Kammer, M., Morris, T. P., White, I. R., & Simulation Panel of the STRATOS initiative (2024). Phases of methodological research in biostatistics-Building the evidence base for new methods. *Biometrical journal. Biometrische Zeitschrift*, 66(1), e2200222. <https://doi.org/10.1002/bimj.202200222>
- Robinson, M.D., Vitek, O. (2019). Benchmarking comes of age. *Genome Biol* 20, 205 <https://doi.org/10.1186/s13059-019-1846-5>
- Rocca, R., & Yarkoni, T. (2021). Putting Psychology to the Test: Rethinking Model Evaluation Through Benchmarking and Prediction. *Advances in Methods and Practices in Psychological Science*, 4(3), 10.1177/25152459211026864. <https://doi.org/10.1177/25152459211026864>
- Sakaluk, J. K., & Short, S. D. (2016). A Methodological Review of Exploratory Factor Analysis in Sexuality Research: Used Practices, Best Practices, and Data Analysis Resources. *Journal of Sex Research*.
- Siepe, B. S., Haslbeck, J. M. B., Kloft, M., Büchner, A., Zhang, Y., Fried, E. I., & Heck, D. W. (2025). Introducing openESM: A database of openly available experience sampling datasets. [https://doi.org/10.31234/osf.io/qfdtb\\_v1](https://doi.org/10.31234/osf.io/qfdtb_v1)

# Thank You!

---



<https://openesmdata.org>



[bjoernsiepe+openesm@gmail.com](mailto:bjoernsiepe+openesm@gmail.com)



<https://bsiepe.github.io>



[bsiepe.bsky.social](https://bsiepe.bsky.social)

## Slides & Papers



# Resources

- All icons either from PowerPoint or fontawesome
  - <https://fontawesome.com/icons/bluesky>
  - <https://fontawesome.com/icons/glasses>
  - <https://fontawesome.com/icons/code>
  - <https://fontawesome.com/icons/dna>
- Python Software Foundation. (n.d.). *The Python logo*. <https://www.python.org/community/logos/> (PSF Trademark Policy)
- Rob Hyndman - <https://robjhyndman.com/hyndsight/m4comp/>
- Steni Sebastian - <https://medium.com/@ssteni/imagenet-what-why-and-how-e5ebed04abb5>
- R Foundation for Statistical Computing. (2016). *R logo*. <https://www.r-project.org/logo/> (CC-BY SA 4.0)
- James Grellier – Own Work, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=10253357>
- Marcin Wichary from San Francisco, U.S.A. - [1]Uploaded by Partyzan\_XXI, CC BY 2.0, <https://commons.wikimedia.org/w/index.php?curid=8198235>
- Carbon (<https://carbon.now.sh/>) for pretty code screenshots
- The Splintered Mind - <https://schwitzsplinters.blogspot.com/2021/02/adversarial-collaboration.html>
- SandboxStudio Artwork - [https://www.symmetrymagazine.org/article/the-facts-and-nothing-but-the-facts?language\\_content\\_entity=und](https://www.symmetrymagazine.org/article/the-facts-and-nothing-but-the-facts?language_content_entity=und)
- Registered Reports graphic: Centre for Open Science, [https://cdn.cos.io/media/images/registered\\_reports.width-800.png](https://cdn.cos.io/media/images/registered_reports.width-800.png)

Except where otherwise noted, the content of these slides is licensed under a Creative Commons Attribution 4.0 International License. Third-party media, logos, and external graphics are property of their respective owners and are excluded from this license.