Marburg University

PMλ.

Improving Methodological Research in Psychology: Preregistration, Transparent Reporting, and Benchmarking

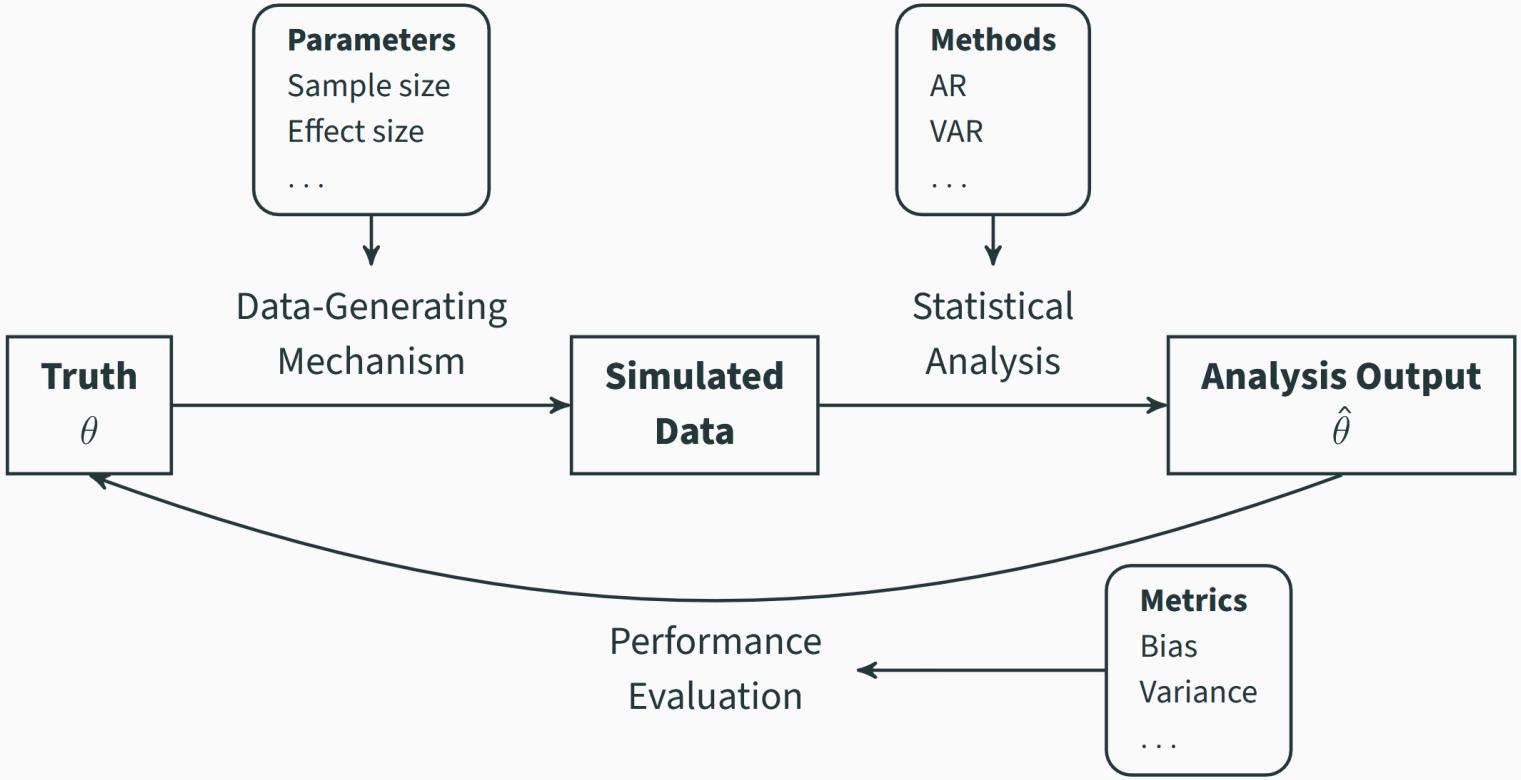
Leibniz-Institut für Bildungsverläufe · June 2026

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)

Simulation studies in psychology

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Methods

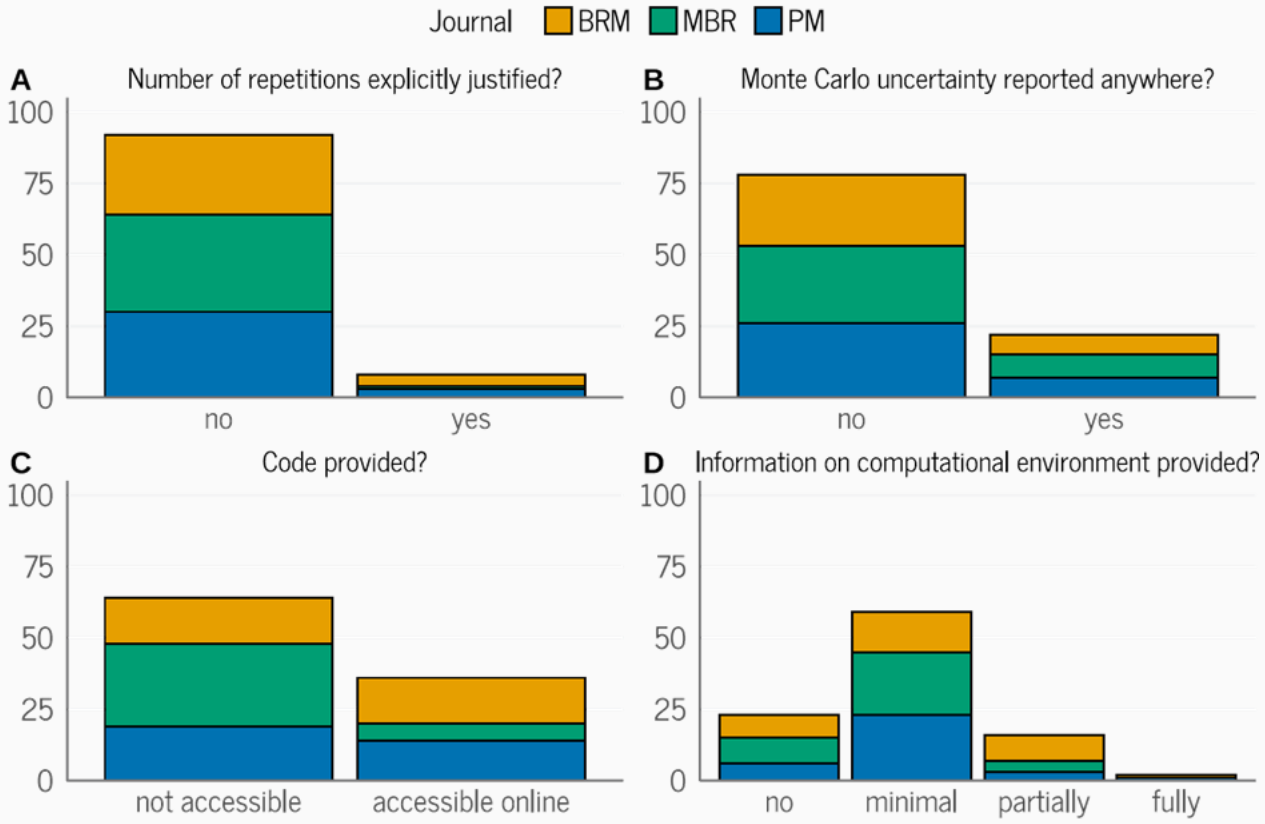
ISSN: 1082-989X

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]

Simulation studies in psychology

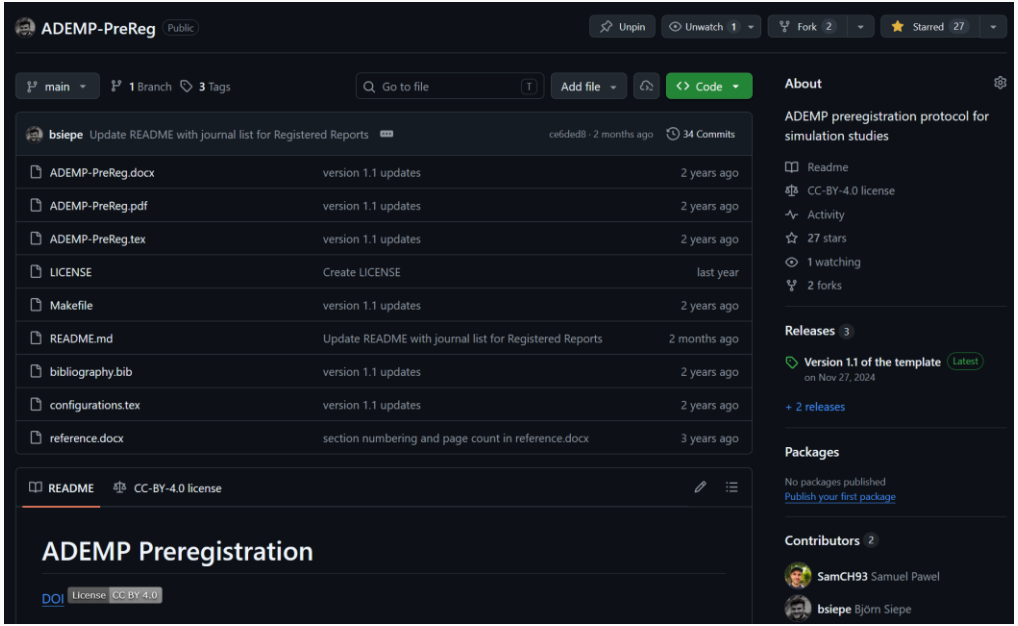


Simulation studies in psychology

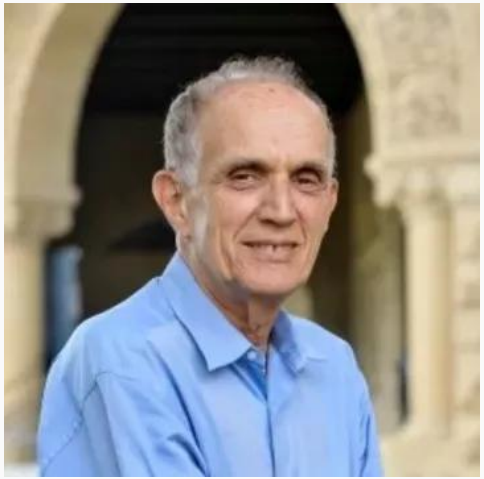
Performance measure	Definition	Estimate	MCSE	n_{sim}
Bias	$\mathrm{E}(\hat{\theta}) - \theta$	$(\sum_{i=1}^{n_{\text{sim}}} \hat{\theta}_i / n_{\text{sim}}) - \theta$	$\sqrt{S^2_{\hat{\theta}} / n_{\text{sim}}}$	$S^2_{\hat{\theta}} / \text{MCSE}^2_*$
Relative bias	$\{\mathrm{E}(\hat{\theta}) - \theta\}/\theta$	$\{(\sum_{i=1}^{n_{\text{sim}}} \hat{\theta}_i / n_{\text{sim}}) - \theta\}/\theta$	$\sqrt{S^2_{\hat{\theta}} / (\theta^2 n_{\text{sim}})}$	$S^2_{\hat{\theta}} / (\text{MCSE}^2_* \theta^2)$
MSE	$\mathrm{E}\{(\hat{\theta} - \theta)^2\}$	$\sum_{i=1}^{n_{\text{sim}}} (\hat{\theta}_i - \theta)^2 / n_{\text{sim}}$	$\sqrt{S^2_{(\hat{\theta}-\theta)^2} / n_{\text{sim}}}$	$S^2_{(\hat{\theta}-\theta)^2} / \text{MCSE}^2_*$
RMSE	$\sqrt{\mathrm{E}\{(\hat{\theta} - \theta)^2\}}$	$\sqrt{\sum_{i=1}^{n_{\text{sim}}} (\hat{\theta}_i - \theta)^2 / n_{\text{sim}}}$	$\sqrt{S^2_{(\hat{\theta}-\theta)^2} / (4 n_{\text{sim}} \widehat{\text{MSE}})}$	$S^2_{(\hat{\theta}-\theta)^2} / (4 \widehat{\text{MSE}}\, \text{MCSE}^2_*)$
Empirical variance	$\mathrm{Var}(\hat{\theta})$	$S^2_{\hat{\theta}}$	$S^2_{\hat{\theta}} \sqrt{2/(n_{\text{sim}} - 1)}$	$1 + 2(S^2_{\hat{\theta}})^2 / \text{MCSE}^2_*$
Empirical standard error	$\sqrt{\mathrm{Var}(\hat{\theta})}$	$\sqrt{S^2_{\hat{\theta}}}$	$\sqrt{S^2_{\hat{\theta}} / \{2(n_{\text{sim}} - 1)\}}$	$1 + S^2_{\hat{\theta}} / (2\text{MCSE}^2_*)$
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	$\mathrm{E}(\text{CI}_{\text{upper}} - \text{CI}_{\text{lower}})$	$\sum_{i=1}^{n_{\text{sim}}} (\text{CI}_{i,\text{upper}} - \text{CI}_{i,\text{lower}}) / n_{\text{sim}}$	$\sqrt{S^2_W / n_{\text{sim}}}$	S^2_W / MCSE^2_*
Mean of generic statistic G	$\mathrm{E}(G)$	$\sum_{i=1}^{n_{\text{sim}}} G_i / n_{\text{sim}}$	$\sqrt{S^2_G / n_{\text{sim}}}$	S^2_G / MCSE^2_*

ADEMP-PreReg

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



Issues in simulation studies



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

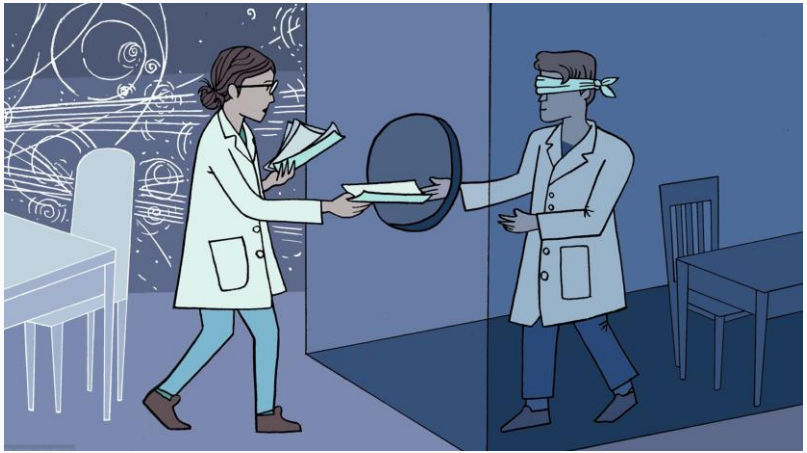
> *[...] it is very difficult to run an honest simulation comparison, and easy to inadvertently cheat by choosing favorable examples, or by not putting as much effort into optimizing the dull old standard as the exciting new challenger.*
>
> Brad Efron, 2001

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

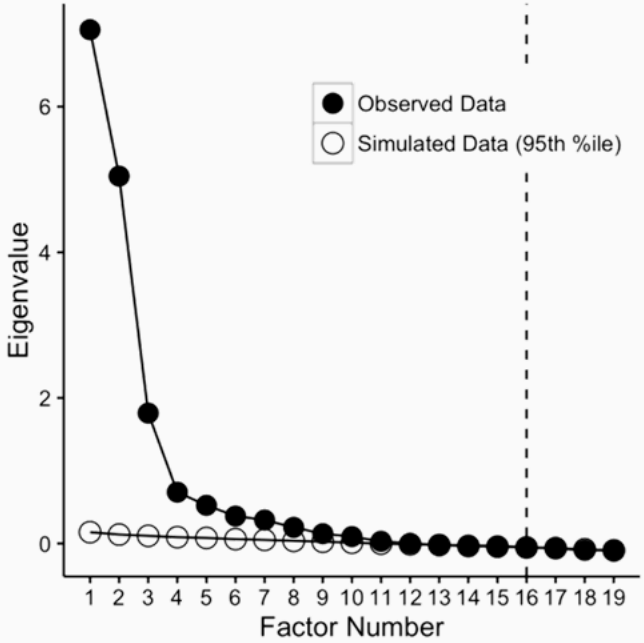
First published: 31 August 2006 | https://doi.org/10.1002/sim.2673 | VIEW METRICS

> ***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[†] [1], František Bartoš [2*], Anne-Laure Boulesteix [3,4*], Daniel W. Heck [1*], Aaron Peikert [5,6*], Alexandra Sarafoglou [2*], Samuel Pawel [7]



Sakaluk & Short (2016)

> ***We recommend parallel analysis for ordinal items***
>
> Hypothetical authors

Illustration of preregistration

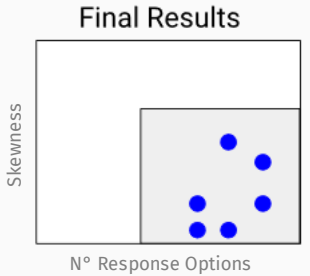
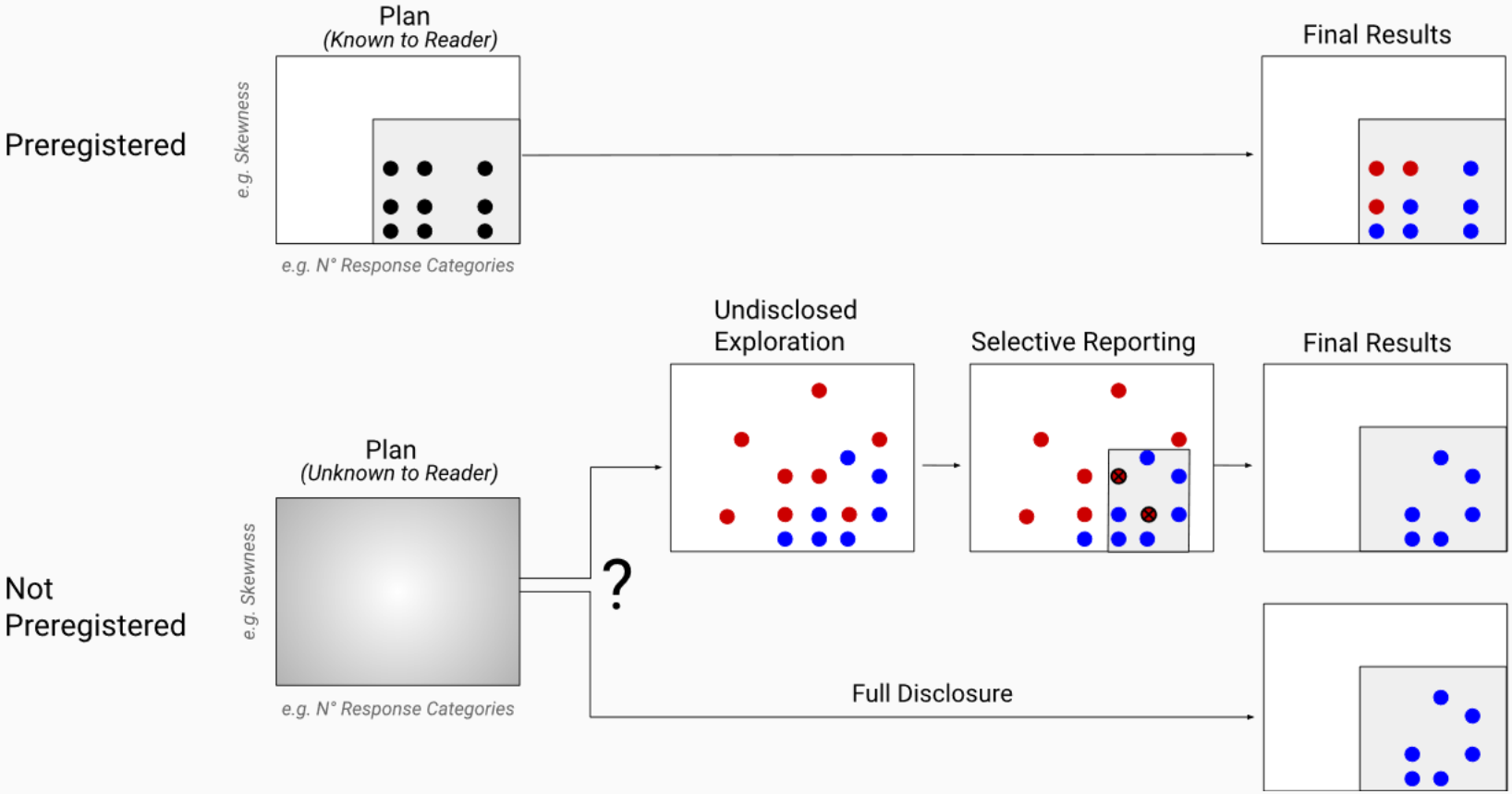


Illustration of preregistration



Theoretical advantages of preregistration

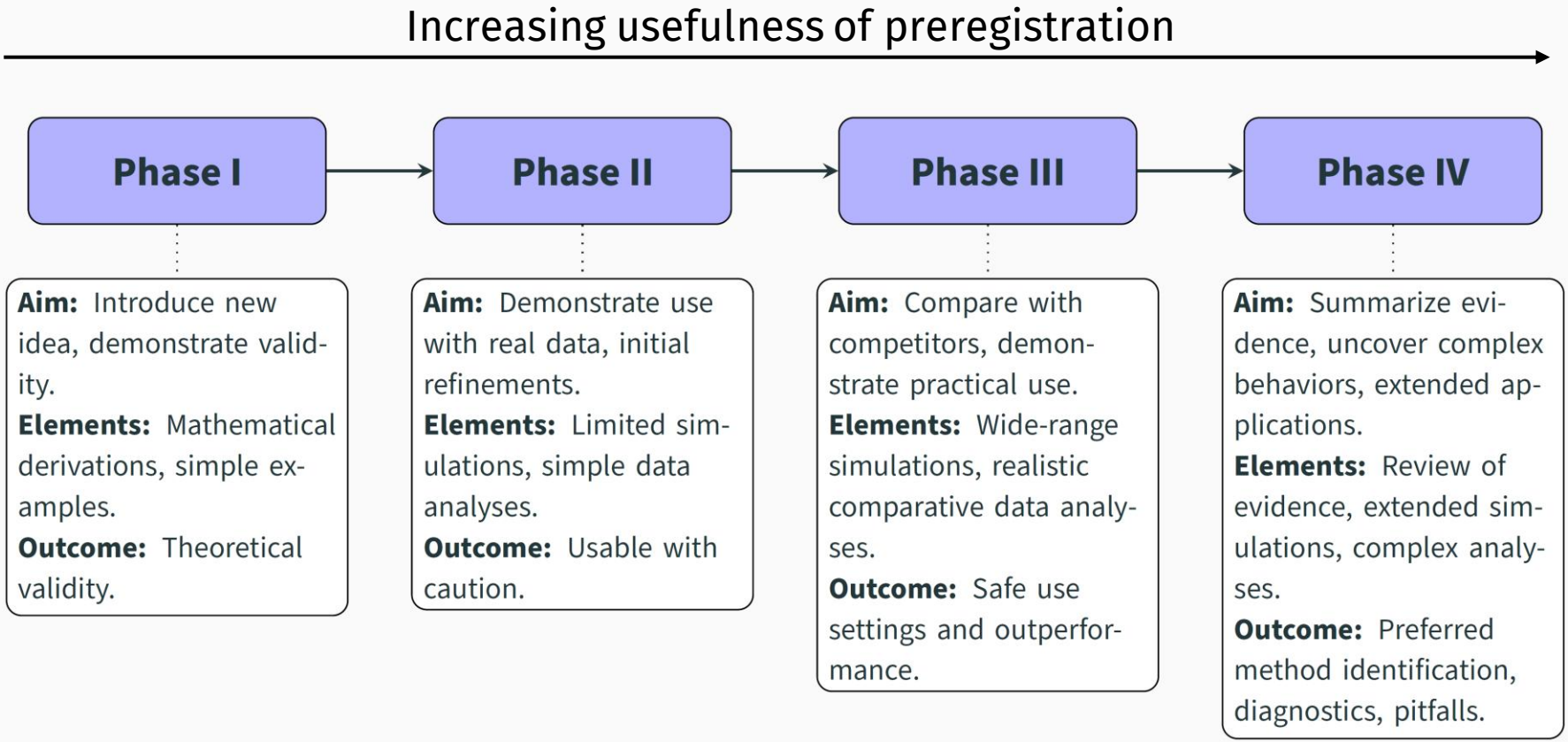
Preregistration reduces epistemic uncertainty about the process that produced simulation results

Falsificationist
Simulation only 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?



FAQ about preregistration

Is preregistration in itself a sign of good quality?	No, but allows assessment thereof
What about deviations?	Be transparent
Do I need theoretical expectations?	No, but can be helpful
When does a simulation study start?	Hard to determine, but there are practical tools
What about cheating?	Can never be fully ruled out

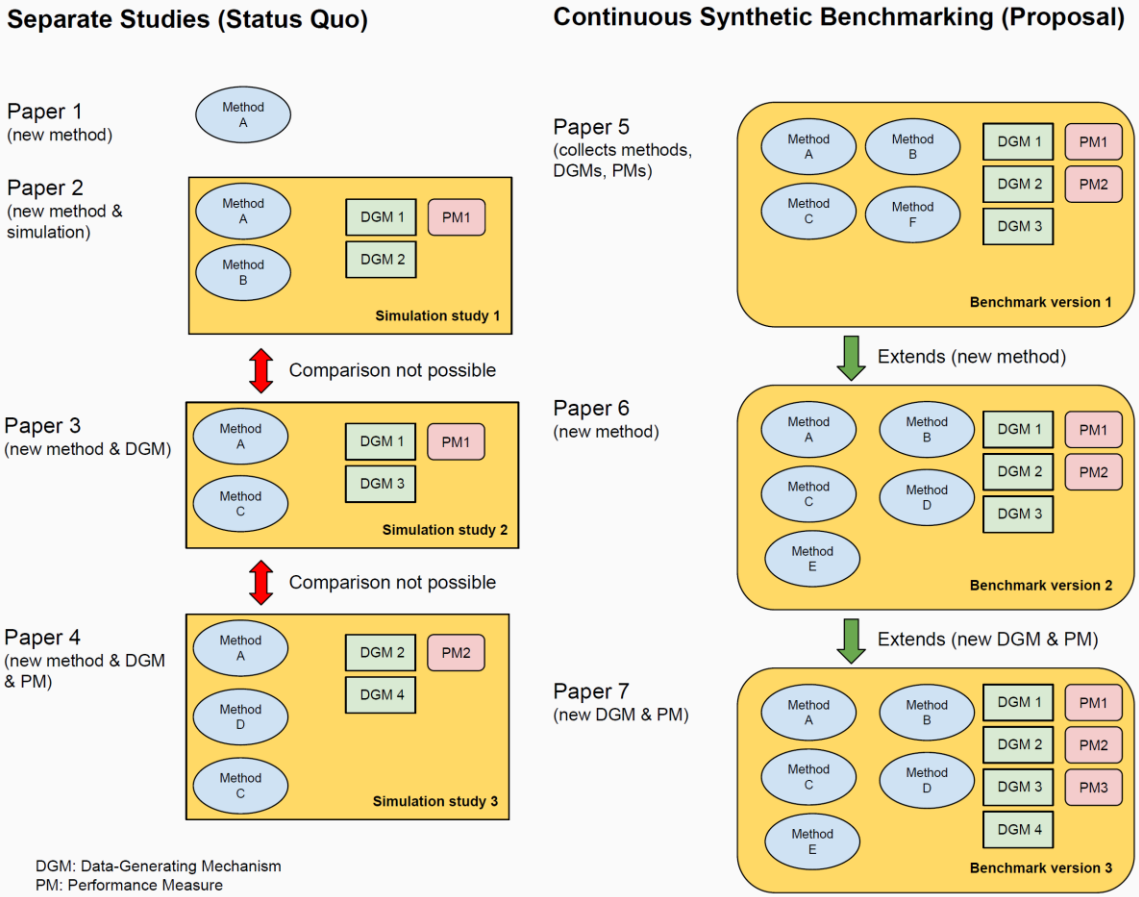
Registered Reports to the rescue

Created with ImageResizer

Centre for Open Science

Part III: Benchmarking

Incentivizing cumulative simulation research



Incentivizing cumulative simulation research

Living Synthetic Benchmarks: A Neutral and Cumulative Framework for Simulation Studies

František Bartoš [1], Samuel Pawel [2], Björn S. Siepe [3]

Synthetic benchmarking for simulation studies

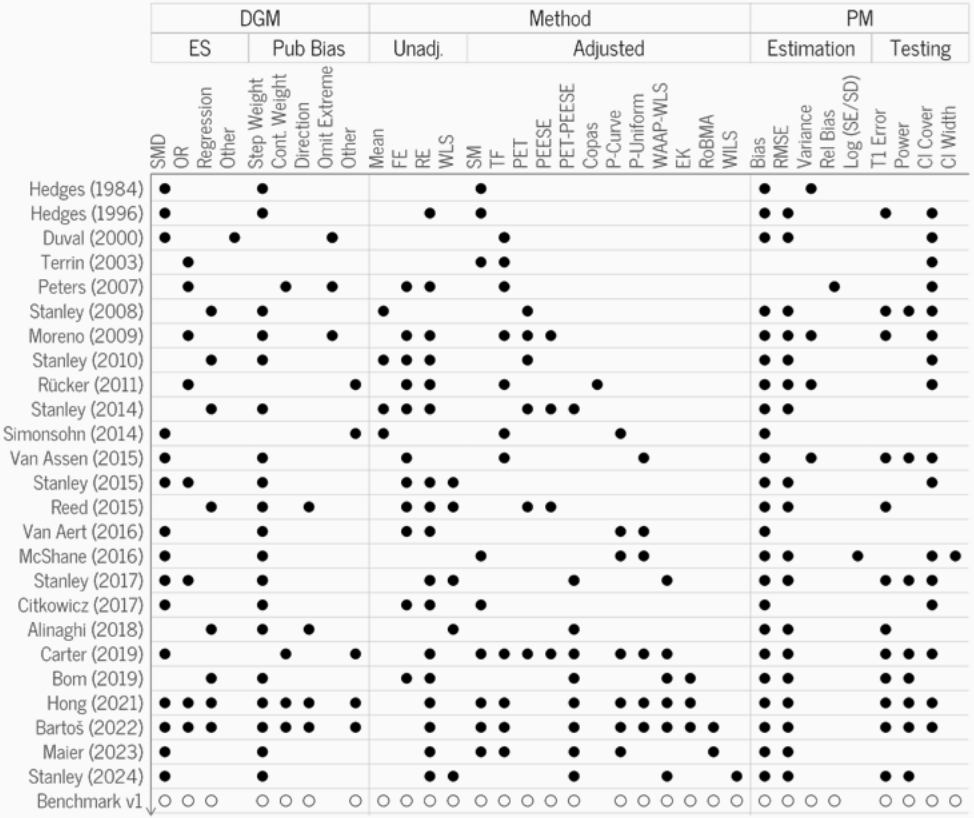
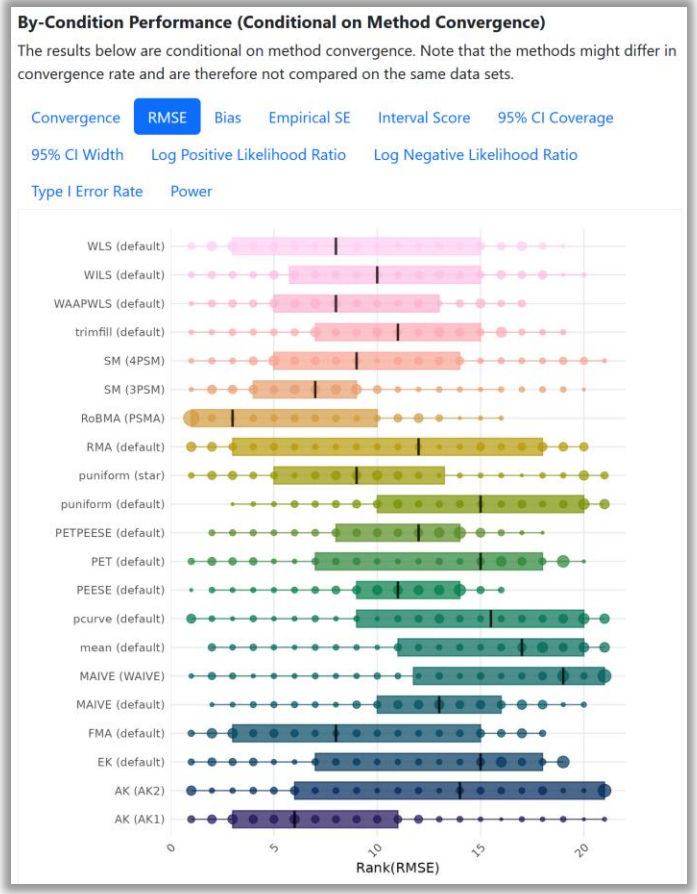


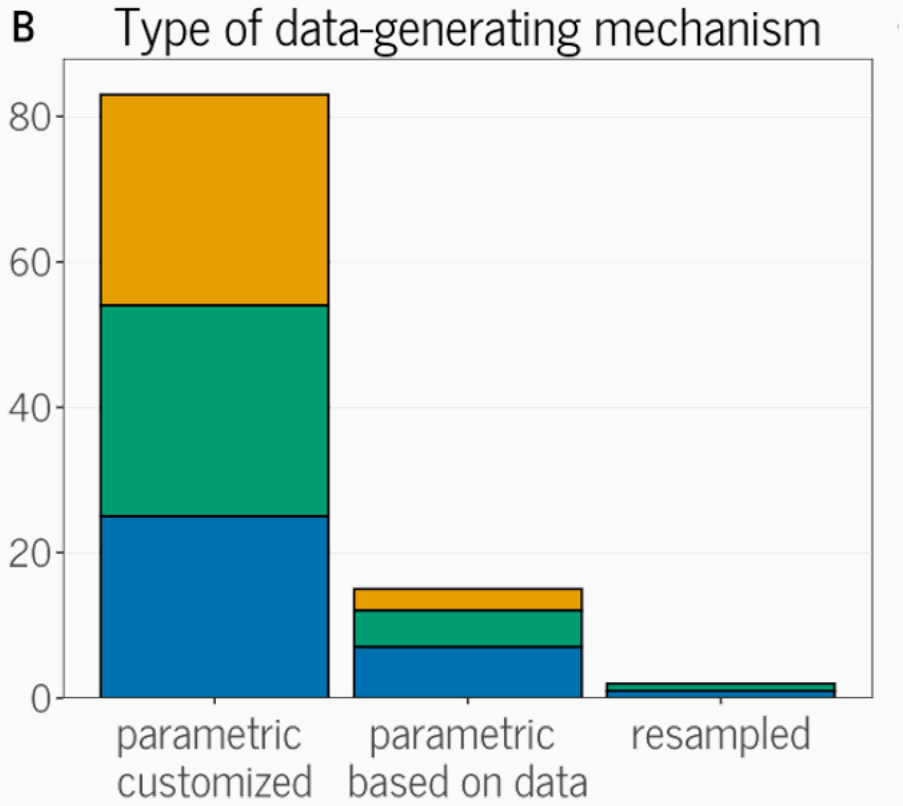
Figure 3: Timeline of major simulation studies on publication bias adjustment

Synthetic benchmarking for simulation studies

fbartos.github.io/PublicationBiasBenchmark/

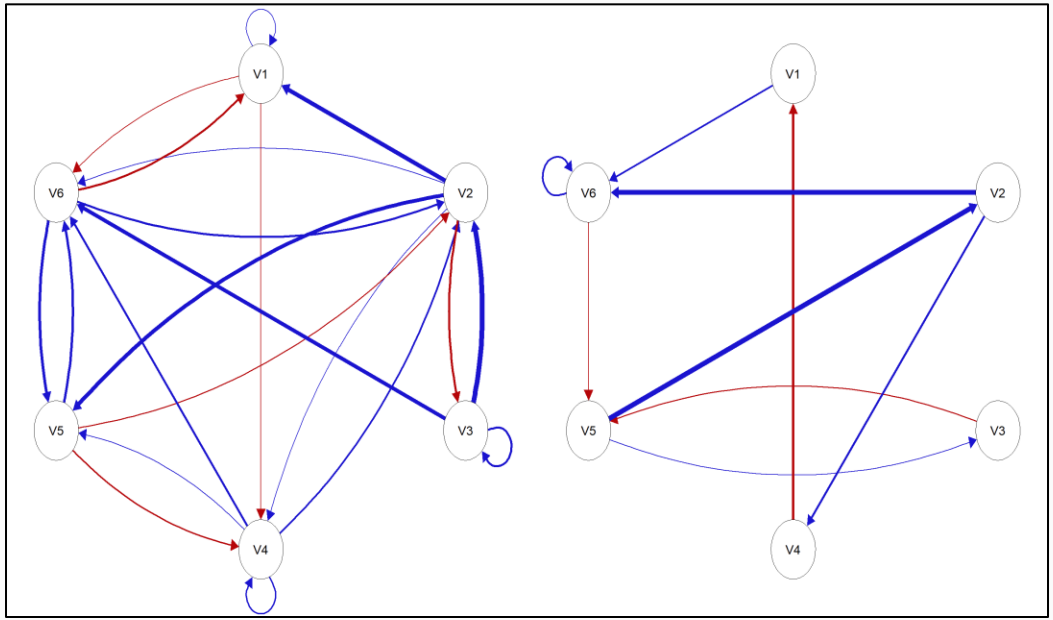


What about *real* benchmarking?



Particular Issues in Psychology

What is a realistic data-generating mechanism?



Hard to understand when performance differences actually matter

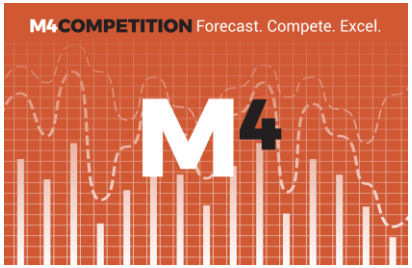
Real-data benchmarking

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[1*] and Olga Vitek[2*]

> **Psychology has been slow to adopt benchmarking (Rocca & Yarkoni, 2021)**

Introducing openESM

Database

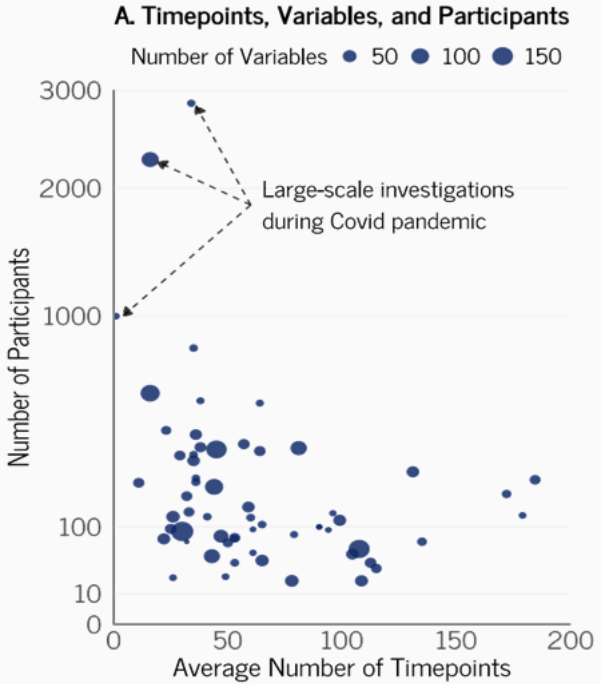
Open-source
Harmonized
Rich metadata

Scale

60 datasets
>16,000 participants
>740,000 observations

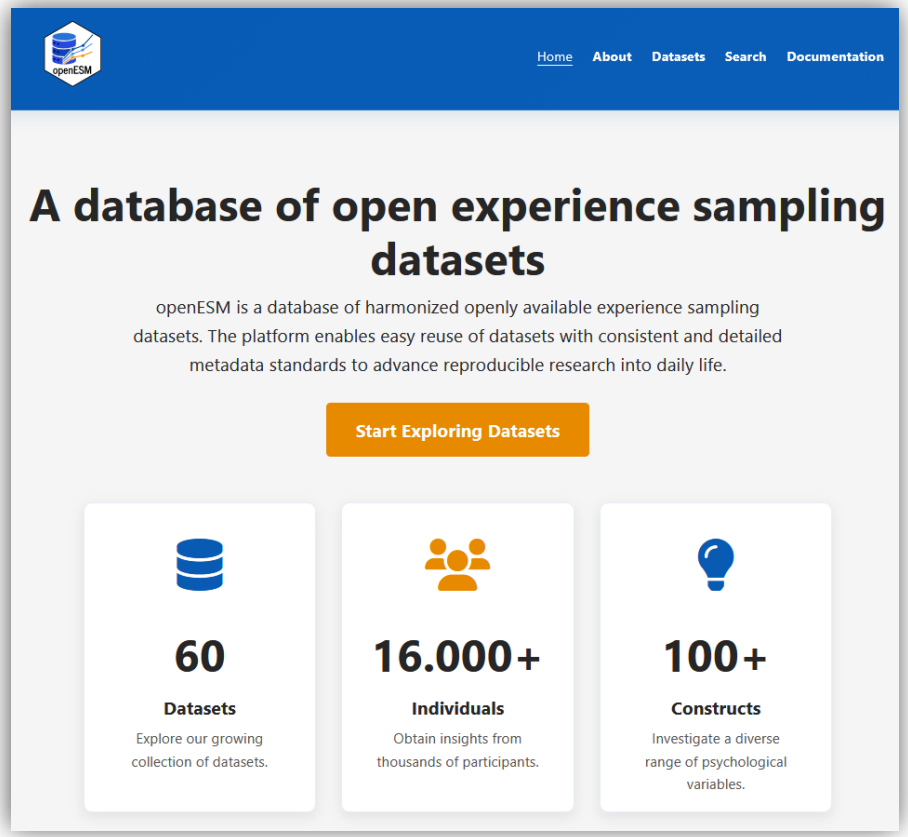
Access

Website with search
and packages

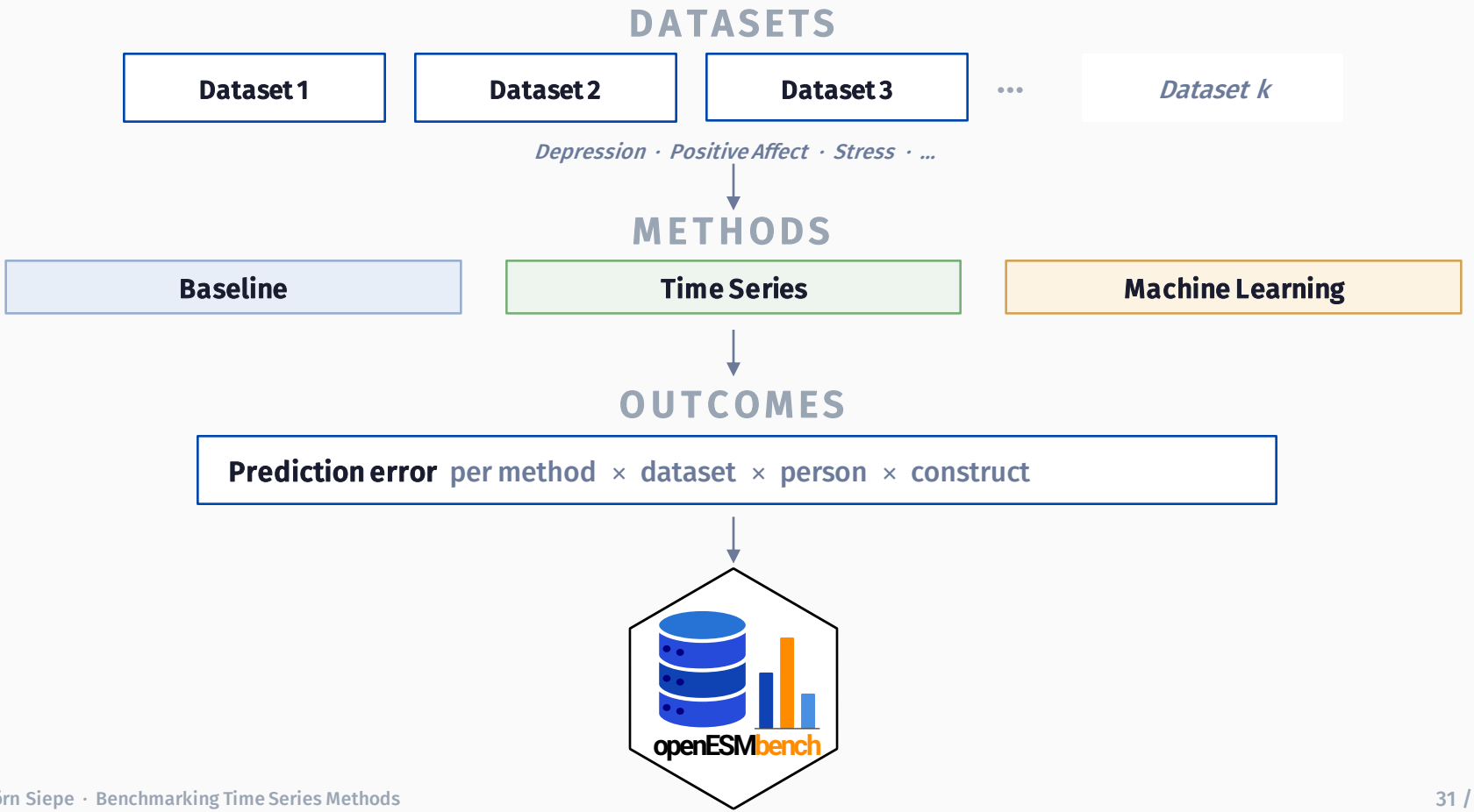


Siepe, B. S., Haslbeck, J. M. B., Kloft, M., Büchner, A., Zhang, Y., Fried, E. I., & Heck, D. W. (2025)

Live Demo


openESM
Home
About
Datasets
Search
Documentation
A database of open experience sampling datasets
openESM is a database of harmonized openly available experience sampling datasets. The platform enables easy reuse of datasets with consistent and detailed metadata standards to advance reproducible research into daily life.
Start Exploring Datasets
60
Datasets
Explore our growing collection of datasets.
16.000+
Individuals
Obtain insights from thousands of participants.
100+
Constructs
Investigate a diverse range of psychological variables.

Benchmarking psychological time series methods



Main Takeaways

Methodological research has quality issues like any empirical field

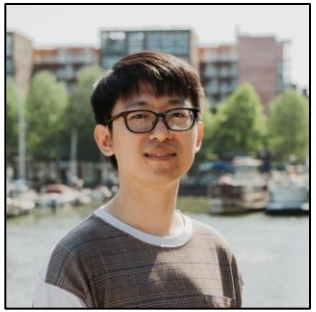
Two potential solutions:

Preregistration
makes generalizability assessable and process transparent, not automatically better

Benchmarking
closes the gap in two ways:
- explicit conditions
- real-world grounding

Item Response Warehouse

Collaborators









Thank You!

https://openesmdata.org

bjoernsiepe@gmail.com

https://bsiepe.github.io

bsiepe.bsky.social

Slides & Papers

References

Efron, B. (2001). Statistical modeling: The two cultures: Comment. Statistical Science, 16(3), 218-219.

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

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
- Sandbox Studio Artwork - 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

Resources