Critical Metascience

Preregistration, Severity, and Deviations

Mark Rubin

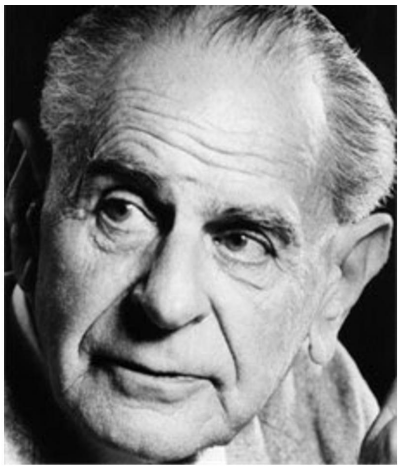
Published September 7, 2024

Citation

Rubin, M. (2024, September 7). Preregistration, Severity, and Deviations. *Critical Metascience*. https://doi.org/10.59350/a1ghn-eas67

Keywords

P-hacking, Error Statistics, Preregistration, Popper, Critical Rationalism



Karl Popper

Deborah Mayo

Preregistration Distinguishes Between Exploratory and Confirmatory Research?

Previous justifications for preregistration have focused on the distinction between "exploratory" and "confirmatory" research. However, as I discuss in this recent presentation, this distinction faces unresolved questions. For example, the distinction does not appear to have a formal definition in either statistical theory or the philosophy of science. In addition, critics have questioned related concerns about the "double use" of data and "circular reasoning" (Devezer et al., 2021; Rubin, 2020, 2022; Rubin & Donkin, 2022; Szollosi & Donkin, 2021; see also Mayo, 1996, pp. 137, 271-275; Mayo, 2018, p. 319).

Confirmatory	Exploratory	Problem
Hypothesis testing	Descriptive research	But most would agree that an unplanned hypothesis test is not "confirmatory"
Result-independent	Result-dependent	But then is a preregistered decision tree that "depends" on the current results "exploratory"?
Strong theory	Weak theory	But then a preregistered test based on weak theory is "exploratory"?
Planned	Unplanned	But then planning to undertake a result-dependent exploratory analysis based on weak theory makes it "confirmatory"?
Prediction	Postdiction	But if a researcher's prediction is a "mere guess," is it "confirmatory"?
Hypothesis-testing	Hypothesis-generating	But what if a researcher views their results and then retrieves a hypothesis from the literature that is confirmed by them? They didn't generate the hypothesis. So, is their test "confirmatory"?

Preregistration Improves the Transparent Evaluation of Severity

Lakens and colleagues provide a more coherent justification for preregistration based on Mayo's (1996, 2018) error statistical approach (Lakens, 2019, 2024; Lakens et al., 2024; see also Vize et al., 2024). Specifically, Lakens (2019) argues that "preregistration has the goal to allow others to transparently evaluate the capacity of a test to falsify a prediction, or the severity of a test" (p. 221).

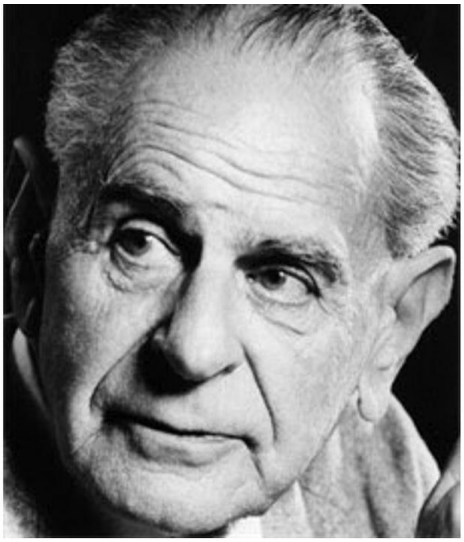
A hypothesis passes a severe test when there is a high probability that it would not have passed, or passed so well, if it was false (Mayo, 1996, 2018). A test procedure's error probabilities play an important role in evaluating severity. In particular, "pre-data, the choices for the type I and II errors reflect the goal of ensuring the test is capable of licensing given inferences

severely" (Mayo & Spanos, 2006, p. 350). For example, a test procedure with a nominal pre-data Type I error rate of $\alpha = 0.05$ is capable of licensing specific inferences with a minimum "worst case" severity of 0.95 (i.e., $1 - \alpha$; Mayo, 1996, p. 399).

Importantly, "biasing selection effects" in the experimental testing context (e.g., *p*-hacking) can lower the capability of a test procedure to license inferences severely by increasing the error probability with which the procedure passes hypotheses. From this error statistical perspective, preregistration allows a more transparent evaluation of the capability of a test procedure to perform severe tests. In particular, preregistration reveals a researcher's *planned* hypotheses, methods, and analyses and enables a comparison with their *reported* hypotheses, methods, and analyses in order to identify any biasing selection effects in the experimental testing context that may increase the test procedure's error probabilities and lower its capability for severe tests.

What Type of Severity?

Mayo's (1996, 2018) error statistical conceptualization of severity is not the only one out there! Other types of severity have been proposed by Bandyopadhyay and Brittan (2006), Hellman (1997, p. 198), Hitchcock and Sober (2004, pp. 23-25), Horwich (1982, p. 105), Lakatos (1968, p. 382), Laudan (1997, p. 314), Popper (1962, 1983), and van Dongen et al. (2023). Furthermore, preregistration may not facilitate the transparent evaluation of these other types of severity. In my article, I illustrate this point by showing that, although preregistration can facilitate the transparent evaluation of Mayoian severity, it does not improve the transparent evaluation of Popperian severity.



Karl Popper

Deborah Mayo

I show that a valid measurement of Popperian severity can be made using a potentially *p*-hacked result, a potentially HARKed hypothesis, and potentially biased background knowledge. In addition, I show that Popper's "requirement of sincerity" can be transparently evaluated

during a public critical rational discussion among scientists. Preregistration does not facilitate a transparent evaluation in either case because neither evaluation requires knowledge of the researcher's planned approach or unreported biasing selection effects.

Preregistration When Deviations are Allowed

I also argue that a preregistered test procedure that allows deviations does not provide a more transparent evaluation of Mayoian severity than a non-preregistered procedure. In particular, I consider deviations that are intended to maintain or increase the validity of a test procedure in light of unexpected issues that arise in particular samples of data (e.g., a violation of the assumption of homogeneity). I argue that a test procedure that allows these sample-based validity-enhancing deviations in its implementation will suffer an unknown inflation of its Type I error rate due to the forking paths problem (Gelman & Loken, 2013, 2014). Consequently, the test procedure will have an unknown reduction of its capability to license inferences with Mayoian severity.

I conclude that preregistration does not improve the transparent evaluation of severity in Popper's philosophy of science or when deviations are allowed.

The Article

Rubin, M. (2024). Preregistration does not improve the transparent evaluation of severity in Popper's philosophy of science or when deviations are allowed. *arXiv*. https://doi.org/10.48550/arXiv.2408.12347

References

Bandyopadhyay, P. S., & Brittan, G. G. (2006). Acceptibility, evidence, and severity. *Synthese, 148*, 259-293. https://doi.org/10.1007/s11229-004-6222-6

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. Department of Statistics, Columbia University. Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*, 460. http://dx.doi.org/10.1511/2014.111.460

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Ozge Buzbas, E. (2021). The case for formal methodology in scientific reform. *Royal Society Open Science, 8*(3). https://doi.org/10.1098/rsos.200805

Hellman, G. (1997). Bayes and beyond. *Philosophy of Science, 64*, 191–221. https://doi.org/10.1086/392548

Hitchcock, C., & Sober, E. (2004). Prediction versus accommodation and the risk of overfitting. *British Journal for the Philosophy of Science*, *55*(1), 1-34. https://doi.org/10.1093/bjps/55.1.1

Horwich, P. (1982). *Probability and evidence*. Cambridge University Press.

Lakatos, I. (1968). Changes in the problem of inductive logic. *Studies in Logic and the Foundations of Mathematics*, 51, 315-417. https://doi.org/10.1016/S0049-237X(08)71048-6

Lakens, D. (2019). The value of preregistration for psychological science: A conceptual analysis. *Japanese Psychological Review, 62*(3), 221–230. https://doi.org/10.24602/sjpr.62.3_221

Lakens, D. (2024). When and how to deviate from a preregistration. *Collabra: Psychology*, 10(1): 117094. https://doi.org/10.1525/collabra.117094

Lakens, D., Mesquida, C., Rasti, S., & Ditroilo, M. (2024). The benefits of preregistration and Registered Reports. *Evidence-Based Toxicology, 2:*1, Article 2376046, https://doi.org/10.1080/2833373X.2024.2376046

Laudan, L. (1997). How about bust? Factoring explanatory power back into theory evaluation. *Philosophy of Science, 64*(2), 306-316. https://doi.org/10.1086/392553

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. University of Chicago Press.

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars.* Cambridge University Press.

Mayo, D. G., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *The British Journal for the Philosophy of Science, 57*(2), 323-357. https://doi.org/10.1093/bjps/axl003

Popper, K. R. (1962). *Conjectures and refutations: The growth of scientific knowledge.* Routledge.

Popper, K. R. (1983). *Realism and the aim of science: From the postscript to the logic of scientific discovery.* Routledge.

Rubin, M. (2020). Does preregistration improve the credibility of research findings? *The Quantitative Methods for Psychology, 16*(4), 376–390. https://doi.org/10.20982/tqmp.16.4.p376

Rubin, M. (2022). The costs of HARKing. *British Journal for the Philosophy of Science, 73*(2), 535-560. https://doi.org/10.1093/bjps/axz050

Rubin, M. (2024). Inconsistent multiple testing corrections: The fallacy of using family-based error rates to make inferences about individual hypotheses. *Methods in Psychology, 10,* Article 100140** https://doi.org/10.1016/j.metip.2024.100140

Rubin, M., & Donkin, C. (2022). Exploratory hypothesis tests can be more compelling than confirmatory hypothesis tests. *Philosophical Psychology.* https://doi.org/10.1080/09515089.2022.2113771

Szollosi, A., & Donkin, C. (2021). Arrested theory development: The misguided distinction between exploratory and confirmatory research. *Perspectives on Psychological Science, 16*(4), 717-724. https://doi.org/10.1177/1745691620966796

van Dongen, N., Sprenger, J., & Wagenmakers, E. J. (2023). A Bayesian perspective on severity: Risky predictions and specific hypotheses. *Psychonomic Bulletin & Review*, *30*(2), 516-533. https://doi.org/10.3758/s13423-022-02069-1

Vize, C., Lynam, D., Miller, J., & Phillips, N. L. (2024, May 28). On the use and misuses of preregistration: A reply to Klonsky (2024). *PsyArXiv*. https://doi.org/10.31234/osf.io/g7dn2