Analysis Preregistration

Authors: Beth Clarke, Felix D. Schönbrodt, Hannah Perfecto, Simine Vazire, Angelika M. Stefan

This release marks the state of our project before we begin the main analyses and acts as an analysis preregistration for the planned simulation study.

Background

The skew of p-curves has often been used to make claims about the existence of p-hacking and publication bias in a literature (Simonsohn et al., 2014). However, a body of criticism suggests that the diagnostic properties of the p-curve methodology may be poor (e.g., Ulrich & Miller, 2015; Morey & Davis-Stober, 2025). In this study, we aim to provide a more nuanced view: Instead of using the p-curve shape to make claims about whether p-hacking occurred or not, we aim to investigate what real-world conditions would be able to generate a p-curve of a specific observed shape. Specifically, we will use a simulation-based approach to identify what combination of p-hacking severity, prior probability of a research hypothesis being true, proportion of p-hacking researchers, and true effect sizes would be necessary to produce a p-curve that is very close to an observed p-curve.

In this study, we will focus on two observed target p-curves. These are based on the data collected in a metascientific project led by @beth099, @hperfecto, and colleagues who recorded the key p-value(s) in 2600 experiments in 842 articles in the *Journal of Social and Personality Psychology: Attitudes and Social Cognition* and *Social Psychological and Personality Science*. The specific *p*-curves we will focus on are: the first timeframe that was studied at *Social Psychological and Personality Science* (2010 – 2012) and the final timeframe that was studied at *Social Psychological and Personality Science* (2016 – 2019). While @beth099, @hperfecto, and colleagues plan to present p-curves using 15 bins (Clarke et al., in prep.), our study will use p-curves of the same data with 5 equally-spaced bins ranging from 0 to 0.05.

Goals of this preregistration

At its heart, this is an exploratory project. We are primarily interested in investigating the sets of simulation parameters that lead to a good fit to the target p-curves. We have no specific hypotheses as to what real-world scenarios may have generated the observed p-curves.

The purpose of this preregistration is two-fold: First, we want to preregister the core expectation that motivated this project before analysing the target p-curves. Second, we want to limit our analytical flexibility in the planned analyses to avoid hindsight bias and motivated reasoning.

Core expectation

There will be multiple simulated scenarios that yield a similarly good fit to the observed p-curves. We do not expect to identify a single best-fit scenario that is much better than all other simulated scenarios, and instead we expect that there will be multiple constellations of simulation conditions that can reproduce all relevant patterns in the observed p-curve.

This prediction was inspired by Appendix E in Stefan & Schönbrodt (2023) and was one of the main motivators for the current study.

Limiting analytic flexibility

With this time-stamped release of the repository, we preregister both the database of simulated p-curves and the analysis approach. By doing so, we remove our analytical flexibility to add simulation conditions that we haven't considered so far (but may yield a better fit) and we commit to specific model fit indices to assess the fit (see Methodology section below; as recommended in Crüwell & Evans, 2021).

Familiarity with the data

All analyses will be conducted by @nicebread and @astefan1 based on the current status of the code in this Github project. At the time of preregistration, neither of them have seen the two p-curves that will be the main target of analyses in this study (i.e., the p-curves from @beth099, @hperfecto, and colleagues' project). This means that @nicebread and @astefan1 are unaware of whether the target p-curves are right-skewed, left-skewed, not skewed at all, or have multiple modes.

@nicebread and @astefan1 used three existing p-curves from the literature to develop the simulation code. The results of the simulation-based analyses of these three "reference" p-curves are not subject to this preregistration but may be presented in the planned publication. The reference p-curves stem from the following studies:

- Simonsohn et al. (2014): 22 significant p-values from studies in social and personality psychology
- Wetzels et al. (2011): 593 significant p-values from studies in experimental psychology
- Sotola (2023): 163 significant p-values from replication studies across psychology

Methodology

Simulation database

So far, in this project, we have generated a large database of simulation results that cover a wide range of possible real-world scenarios. This database will form the basis of our analyses of the two target p-curves.

The simulation database can be reproduced by calling the master script 00-MAKE.R. This script calls two core simulation scripts that contain further details about the simulations: sim_multDV_scenarios.R and sim_optionalStopping.R. These simulation scripts in turn rely on custom R functions that are also part of this repository and can be installed as a package called `fitPCurve` (for more information on reproducibility, see the repository landing page). The simulation results can be found in the folder `simulations/sim-results`. The tables in this section also provide a summary of all simulated scenarios.

Simulation conditions

Across a total of 85,426 simulations, we varied: the proportion of p-hackers, the p-hacking strategy, the severity of p-hacking, the proportion of true research hypotheses, and the mean and variance in population effect sizes. We have grouped these simulations into four "possible worlds":

- "Worst possible world": All research hypotheses are false (i.e., the null is always true) and a non-zero proportion of researchers p-hack (the proportion of p-hackers and the severity and strategy of the p-hacking varies). Total simulations: 705 (360 with multiple DVs p-hacking + 345 with optional stopping)

- "Realistic world": A non-zero proportion of researchers p-hack (the proportion of p-hackers and the severity and strategy of the p-hacking varies). Some of them are investigating true non-zero effects (with effect sizes that can vary between studies), others are studying null effects. Total simulations: 84,600 (43,200 with multiple DVs p-hacking + 41,400 with optional stopping)
- "Perfect world": Nobody p-hacks. Some researchers are investigating true non-zero effects (with effect sizes that vary between studies), others are studying null effects. Total simulations: 120
- "Flat world": All research hypotheses are false (i.e., the null is always true), but nobody p-hacks. Note that this is only a single condition mainly used as a baseline and sanity check. Because of this, we may not report it in the main text of the manuscript). Total simulations: 1

To model p-hacking (in the simulations that include p-hacking), we modeled two strategies separately (i.e., we did not model a combination of multiple p-hacking strategies): 1) analysing multiple dependent variables, and 2) optional stopping. Within each of these strategies, we modeled the following conditions:.

Multiple Dependent Variables

	“Worst possible world”	**“Realistic World”**
Number of DVs:	2, 5, 10, 30, 70, 150	2, 5, 10, 30, 70, 150
Correlation between DVs:	0, 0.3, 0.6, 0.9	0, 0.3, 0.6, 0.9
Effect size d:	0	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8

Heterogeneity (SD of d):	0	0, 0.2, 0.4
Proportion hacker:	0.1, 0.25, 0.5, 0.75, 1	0.1, 0.25, 0.5, 0.75, 1
Proportion H1 true:	0	0.1, 0.2, 0.3, 0.5, 0.7
Reporting strategy (as defined in Stefan & Schönbrodt, 2023):	smallest,smallest significant, first significant	smallest, smallest significant, first significant

Optional Stopping

	“Worst possible world”	**“Realistic World”**
Minimum sample size:	8, 22, 50, 177, 444	8, 22, 50, 177, 444
Maximum sample size:	22, 50, 177, 444, 1063	22, 50, 177, 444, 1063
Step size:	1, 5, 10, 50, 100	1, 5, 10, 50, 100
Effect size d:	0	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8
Heterogeneity (SD of d):	0	0, 0.2, 0.4
Proportion hacker:	0.1, 0.25, 0.5, 0.75, 1	0.1, 0.25, 0.5, 0.75, 1
Proportion H1 true:	0	0.1, 0.2, 0.3, 0.5, 0.7

“Perfect World” (no p-hacking)

Effect size d:	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8
Heterogeneity (SD of d):	0, 0.2, 0.4
Proportion hacker:	0
Proportion H1 true:	0.1, 0.2, 0.3, 0.5, 0.7

“Flat World” (no p-hacking)

Effect size d:	0
Heterogeneity (SD of d):	0
Proportion hacker:	0
Proportion H1 true:	0

All simulations were conducted for two-sided independent-samples t-tests, effect sizes refer to Cohen's d. Furthermore, all simulations were conducted using 10,000 iterations. This was based on initial exploratory analyses of the Monte Carlo error (cf. scripts compute_MCE.R and eval_MCE.qmd). All simulations were conducted with a seed of 12345, which will also be used in the future for any random process in our analyses.

Evaluating fit

Each simulated scenario yields a different “expected” p-curve that can, in a next step, be compared to the observed target p-curve. Our procedure for finding the conditions that create the best-fitting p-curves will be as follows:

(1) We will compute G^2 statistics for each simulated condition to evaluate the fit between the expected p-curve in the simulation scenario and the observed target p-curve. G^2, the multinomial likelihood ratio chi-square statistic, is commonly used for fitting models based on binned data, and is defined as two times the difference between the maximum possible multinomial log likelihood and the model-implied (here: simulated) multinomial log likelihood (Ratcliffe & Childers, 2015). Smaller values of G^2 are associated with a better model fit.

(2) Using G^2, we will identify a set of best-fitting simulation scenarios in each "world" (see previous section). Importantly, in this step, we do not plan to identify a single, best-fitting model/simulation, but instead, our goal is to identify clusters of good fitting spaces within each simulated "world" and qualitatively interpret the parameter combinations that led to a good fit. Our interpretations will be supported by visualizations such as p-curve plots and heat maps. Note that in addition to simulation parameters, we will also interpret the simulated "size of the file drawer", that is, the proportion of p-values that did not turn out significant in the simulated scenario and are hence not part of the p-curve. This value is a corollary of the simulated p-curve and helps to critically evaluate whether a set of simulation parameters may be realistic.

(3) Robustness check: We will check whether the identified well-fitting model spaces from (2) are sensitive to the choice of fit index by also calculating Chi-squared and RMSE.

(4) Comparing fit between simulated "worlds": Can we make a statement about which "world" is more or less likely to have produced the target p-curve (based on our simulations)? To answer this question, we will convert the G^2 statistic of the best fitting model in each "world" into a BIC score that also takes model complexity (i.e., the number of free parameters in the simulation that determines the flexibility of the world-model) into account (Smith & Ratcliff, 2022). By comparing BIC values between "worlds", we will be able to make tentative conclusions about whether better observed fit in one world or another can be explained through model complexity alone.

A script containing each step of the fit evaluation procedure can be found in the script eval_simResults_new.qmd.

References

- Clarke, B., Perfecto, H., Park, A.B., Gonzalez, F., O'Donnell, M., Schiavone, S.R., Bottesini, J.G., Nelson, L.D., Vazire, S. (in prep). *P-curving social and personality psychology experiments: Changes from 2006 to 2019.* Manuscript draft.

- Crüwell, S., & Evans, N. J. (2021). Preregistration in diverse contexts: A preregistration template for the application of cognitive models. *Royal Society Open Science*, *8*(10), 210155. https://doi.org/10.1098/rsos.210155
- Morey, R. D., & Davis-Stober, C. P. (2025). On the poor statistical properties of the P-curve meta-analytic procedure. *Journal of the American Statistical Association*, 1–19. https://doi.org/10.1080/01621459.2025.2544397
- Ratcliff, R., & Childers, R. (2015). Individual differences and fitting methods for the two-choice diffusion model of decision making. *Decision*, *2*(4), 237–279. https://doi.org/10.1037/dec0000030
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*(2), 534–547. https://doi.org/10.1037/a0033242
- Sotola, L. (2023). How can I study from below, that which is above?: Comparing replicability estimated by z-curve to real large-scale replication attempts. *Meta-Psychology,* 7. https://doi.org/10.15626/MP.2022.3299
- Smith, P. L., & Ratcliff, R. (2022). Modeling evidence accumulation decision processes using integral equations: Urgency-gating and collapsing boundaries. *Psychological Review*, *129*(2), 235–267. https://doi.org/10.1037/rev0000301
- Stefan, A. M., & Schönbrodt, F. D. (2023). Big little lies: A compendium and simulation of p-hacking strategies. *Royal Society Open Science, 10*(2), 220346. https://doi.org/10.1098/rsos.220346
- Ulrich, R., & Miller, J. (2015). p-hacking by post hoc selection with multiple opportunities: Detectability by skewness test?: Comment on Simonsohn, Nelson, and Simmons (2014). *Journal of Experimental Psychology: General, 144*(6), 1137–1145. https://doi.org/10.1037/xge0000086
- Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. –J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science, 6*(3), 291–298. https://doi.org/10.1177/1745691611406923