False-Positive Social Psychology: How Deviations from Preregistrations Affect False-Positive Significance Rates

By

Terry Cheng

University of California, Santa Barbara

Advisor:

Carl T. Bergstrom

University of Washington

March 14, 2022

Abstract

Numerous solutions have been proposed to address the replication crisis, in which numerous high-profile empirical research studies cannot be replicated by other research teams. One possible explanation is that researchers have the option to adjust their data analyses after viewing the results, inflating false positive rates. One popular solution is study preregistration, the practice of developing the data analysis plan before the data is collected. However, preregistrations only alleviate replication problems if researchers are held accountable to their analysis plans. Across two related studies, we explore the effectiveness of preregistration in its current form. In Study 1, we audit recent preregistered publications from a major psychology journal and observe deviations in 19 of 32 papers. In Study 2, we simulate the effects of generic deviations on the false-positive rate. We find that deviations that run more or more varied tests cause larger changes, tripling the false-positive rate in the most extreme case. We note that auditing preregistrations requires an inconsistent amount of time depending on their length and format, which we suspect contributes to the enforcement issues we observe. We suggest that researchers and journals alike adopt the asPredicted.org template for preregistrations.

Acknowledgements

The author would like to thank Carl T. Bergstrom, Shelly Lundberg, and Joe Bak-Coleman. Their unending supply of appropriate affirmations and constructive criticisms shaped this project into much more than it would have been otherwise.

Special thanks also to Angela Chikowero and Ted C. Bergstrom, who entertained and directed the author's naive enthusiasm during the project's nascent stages.

Detailed results, a Code Supplement, and all materials necessary to reproduce this document are availible at osf.io/5gpt8.

Contents

1 Introduction

The academic community has known for decades that empirical papers reporting statistically significant results are more likely to be published than those reporting null results (Sterling 1959; Rosenthal 1979). Journals are incentivised to focus on publishing novel results to grow their readerships and sell subscriptions. Combined with the direct financial or indirect career-related incentives most researchers face to publish frequently in prestigious journals (Gibson, Anderson, and Tressler 2014), it is not surprising that the published literature has been affected. Recent studies continue to show evidence of the file-drawer effect and publication bias, the tendency for researchers not to submit null results for publication and the tendency for journals not to publish the null results that are submitted, respectively (Turner et al. 2008; Camerer et al. 2016; Dwan et al. 2013).

These observed defects in the published literature are often the result of 'p-hacking,' faulty research techniques that artificially induce significance by running multiple tests and selectively reporting desirable results. Previous work suggests that researcher degrees of freedom (RDFs), the implicit optionality researchers leverage to conduct multiple tests, can seriously inflate the false-positive rate despite nominally significant p-values (Ioannidis 2005; Simmons, Nelson, and Simonsohn 2011; Huntington-Klein et al. 2021). This provides a mechanism by which the individual incentives to publish statistically significant results affect the false-positive rate of the entire published literature.

Published false-positives are difficult to retract from the academic canon, waste resources on subsequent research and policies, and lower the credibility of future publications in the field. However, well-calibrated tests will always have some probability of producing a false-positive finding. Assuming that there exists some threshold for p which strikes the

optimal balance between false-positive and false-negative findings, we can consider the excess false-positive rate caused by RDFs a negative externality generated by individual researchers optimizing the tradeoff between producing functionally significant research that advances their field and nominally significant research that advances their careers.

Numerous solutions have been proposed. Simply lowering the p-value threshold for statistical significance does reduce the probability that a given study results in a false-positive but necessarily increases the probability of a false-negative result. Additionally, tightening the significance threshold backfires by increasing the false-positive rate conditional on publication unless the underlying bias for nominally significant results is eliminated (Williams 2019).

More promising solutions include the use of registered reports and results-blind manuscript evaluation, systems where editors review and select submissions on the merit of the study design and methodology before any results are reported. This approach targets publication bias and the file-drawer effect by reducing the incentive for researchers to p-hack their studies in search of positive results (Locascio 2017). However, these systems have not been adopted outside of a small number of journals.

In contrast, study preregistrations are a strict requirement in clinical trials and have seen widespread adoption by other disciplines as well, specifically psychology and economics. To preregister a study, researchers create a timestamped document describing the data collection process and pre-analysis plan (PAP) before the study is conducted. Instead of targeting the incentive to publish positive results, the PAP prevents researchers from making biased choices by forcing them to make those choices before the data is collected. To truly remove RDFs, PAPs should contain all of the information another researcher would need to replicate the analysis.

However, some fields suffer from performative reproducibility, where "open-science

practices [have] become just another hoop to jump through, a form of virtue signaling or a smokescreen" (Buck 2021). This is likely the case in economics, where a Nobel Memorial laureate and the members of the AEA RCT registry's committee have publicly supported vague PAPs (Banerjee et al. 2020).

Even with a detailed PAP, preregistration fails to contain RDFs unless researchers are held accountable to their PAPs; if there are no consequences to deviating from a PAP, the RDFs persist. This paper explores the nature and possible effects of these deviations across two related studies.

Study 1 is a descriptive analysis of recent publications by the Journal of Social Psychology and Personality (JPSP) that identifies a set of RDFs associated with commonly observed PAP deviations. Study 2 explores the impact of these RDFs through Monte Carlo simulation.

2 Literature Review

COMPare: a prospective cohort study correcting and monitoring 58 misreported trials in real time (Goldacre et al. 2019)

A team of researchers checked the preregistrations of all new clinical trials in the top five medical journals for about six months. They counted preregistered outcomes that weren't reported in the final publication and unregistered outcomes that were silently added.

Of the 67 trials they checked, only nine abstained from outcome switching of any kind. On average, each trial reported just 58.2% of its specified outcomes and silently added 5.3 new outcomes. Overall, the results suggest that even the top medical journals routinely fail to hold researchers accountable to their PAPs.

Study 1 replicates the spirit of this paper in a new sample with an expanded scope. Instead of tracking switched outcomes only, we track all deviations from PAPs.

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant (Simmons, Nelson, and Simonsohn 2011)

In this paper, the effect of specific RDFs on the false-positive rate are estimated by testing hypotheses on a simulated dataset under the null model. Since the tested hypotheses are known to be false, any positive result can be interpreted as a false-positive.

The results are striking: simple RDFs, such as outcome switching and sample size manipulations, lead to sizable increases in the false-positive rate. The effect is especially pronounced when the RDFs are combined.

Study 2 extends the simulation work in this paper to a new set of RDFs. Additionally,

we increase the number of experiments simulated per RDF for increased precision in our Monte Carlo estimates.

3 Study 1: Common Deviations Observed in Recent JPSP Papers

3.1 Methodology

Our preregistered (aspredicted.org/cg6cd.pdf) study audited recent JPSP publications with public preregistrations. For each study in the sample, we manually compared the preregistered analysis plans to the published analysis and noted any deviations. As they were discovered, we identified common types and grouped the deviations accordingly. This occasionally involved retroactively recoding a deviation after a new type was identified.

Additionally, we tracked whether deviations were disclosed as required by the JPSP submission guidelines. We also tracked whether disclosed deviations were justified, but did not assess the validity of the justifications.

3.2 Data

The study sample consists of all JPSP publications published between August-December 2021 which claimed to have at least one preregistered component. This amounts to 98 studies across 32 papers.

While 10 papers reported the results of only a single study, most reported the results of 4-7 related studies. Often, only some of the studies would be preregistered. For example, many papers identified potential effects in an exploratory study, then replicated the results in preregistered confirmatory studies.

When designing the study, we also considered American Economic Association (AEA) journals and *Nature Human Behavior* (NHB), other journals with nominal rules about

preregistering empirical work. Upon cursory examination, we found that preregistration was not common practice in AEA journals. Preregistration was common practice in applicable NHB publications, namely those reporting trial results, but these papers are only a small subset of all NHB publications. We selected JPSP to avoid both issues.

3.3 Results

We discovered some kind of deviation in almost two-thirds of papers and one-third of studies. More often than not, discovered deviations were disclosed in the final text of the paper[1]. With one exception, the disclosures were always paired with some justification; the most common were that the deviation did not meaningfully change the results, or that the deviation led to a more valid, intuitive, or otherwise 'better' model.

Table 1: Audit results at the study and paper levels. We find that deviations are not uncommon, although most discovered deviations were disclosed.

	Study	Paper
Total	98	32
Could Not Assess	7	4
No Deviation	62	19
Deviated	29	19
Not Disclosed	10	4

[1]There is some selection bias at play here; we do not know how many undiscovered, undisclosed deviations exist in the sample.

We identified four types of deviations that captured all observations:

- Outcome, in which dependent variables were added, changed, or dropped entirely
- Covariate, in which covariates were added or dropped
- Sample, in which the sample changed. This includes unreasonable changes to the sample size, unregistered exclusion criteria, and other changes to the - sampling methodology
- Model, in which the underlying regression model or analysis methodology changed

Table 2: Discovered deviations by type

	Count
Outcome	18
Covariate	4
Sample	9
Model	8

4 Study 2: Type-I Error Rates of Common RDFs through Monte Carlo Estimation

4.1 Background

The Type-I error rate, or 'false-positive' rate commonly, quantifies how often we would expect the null hypothesis to be rejected were it actually true. In the absence of a true causal relationship between the treatment and response variables, any statistically significant result can be interpreted as a false-positive. Equivalently, false-positive rates represent how likely an experiment yields statistically significant results in the absence of a true effect; we expect 5% of these experiments to reach the standard significance threshold of $p = .05$. However, the optionality of RDFs let researchers test multiple models and select the lowest p-value from among them. Since they're choosing between the original p-value and potentially lower p-values from the alternative analysis, we expect false-positive rates to rise.

This study estimates the effects of individual RDFs using a Monte Carlo approach. For each of 50,000 simulated experiments, we conduct both the baseline analysis and the alternative analysis allowed with a specific RDF, take the p-values from each experiment, and calculate the percentage yielding significant results to estimate the false-positive rate. We attribute any difference to the RDF.

Additionally, observed p-values under a valid null model should be uniformly distributed to maintain their usual frequentist interpretations. We fit bounded kernel density estimates of our observed p-values using the `bde` R package (Santafe et al. 2015) to explore how introducing RDFs affects p-value distribution.

We test four RDFs in this study. The first three are generic examples of the Outcome,

Covariate, and Sample type deviations identified in Study 1[2]: the option to *swap outcomes* between the original outcome, an independent alternative, or a mixture of the two; the option to *add a binary covariate*; and the option to *manipulate the sample size* by adding observations and retesting until the results are significant, limited to double the original sample size. The final RDF tested is a specific form of an Outcome type deviation: the option to *drop an uncorrelated item* from an otherwise correlated composite index of responses on the Likert scale. We tested this additional RDF because it was frequently observed during Study 1.

4.2 Data Generating Process

In each simulated experiment, we generate 60 i.i.d. standard normal observations $y_1, ..., y_{60}$ to represent the experimental yield. Then, we set $cell_1, ..., cell_{30} = 0$ and $cell_{31}, ..., cell_{60} = 1$ to represent the experimental condition. We fit the model $y \sim \beta_0 + \beta_1 cell$ and report the p-value associated with β_1 as the baseline result.

For the *outcome swapping* RDF, we generate another 60 i.i.d. standard normal observations $z_1, ..., z_{60}$ and calculate a mixture variable $y^* = \frac{z+y}{2}$. We then fit $z \sim \beta_0 + \beta_1 cell$ and $y^* \sim \beta_0 + \beta_1 cell$ and extract the p-values associated with β_1. Finally, we compare the two new p-values with the baseline result and report the lowest as the result with the RDF.

For the *binary covariate* RDF, we generate an indicator column x such that $x_i = 1$ for a specified n observations per cell and $x_i = 0$ in the rest. We then fit the model $y \sim \beta_0 + \beta_1 cell + \beta_2 x$ and extract the p-value associated with β_1. Finally, we compare the new p-value with the baseline result and report the lower of the two as the result with the RDF. We simulate with covariate incidence specifications of 5, 15, and 25 observations per cell.

[2]While the form of all types of deviations varied case-by-case, Model type deviations were especially context-specific. We could not come up with adequately generic examples to test.

For the *sample size manipulation* RDF, we first check whether the baseline result is significant. If so, we report the baseline results as the result with the RDF. Otherwise, we generate a specified n additional observations of y per cell and repeat the baseline analysis until a significant result is achieved or the total sample size reaches double the original. We report the latest p-value after meeting our stopping criteria (the first significant result if one is found, otherwise the result with 60 observations per cell) as the result with the RDF. We simulate with retest interval specifications of 1, 10, and 30 observations per cell.

Simulating the *drop item from index* RDF requires implementing a different baseline analysis. As before, we set $cell_1, ..., cell_{30} = 0$ and $cell_{31}, ..., cell_{60} = 1$ to represent the experimental condition. We use the R package `faux` (DeBruine 2021) to generate four standard normal columns w, x, y, z with a specified correlation coefficient r, then generate a fifth standard normal column *drop* independent from the others. After converting the five columns to centered Binomial(7, 0.5) columns through inversion sampling, we calculate the composite column $mix = \frac{w+x+y+z+drop}{5}$ and fit the model $mix \sim \beta_0 + \beta_1 cell$. We report the p-value associated with β_1 as the baseline result.

Next, we remove the independent column from the composite to recalculate $mix = \frac{w+x+y+z}{4}$ and refit $mix \sim \beta_0 + \beta_1 cell$. We compare the new p-value with the baseline result and report the lower of the two as the result with the RDF.

4.3 Results

First, we assess the validity of the null model for our baseline analysis. About 4.8% of observations cross the standard significance threshold, which translates directly to an estimated false-positive rate of 4.8%. We expect 5% under a valid null model, so our results suggest that our baseline p-values are approximately uniformly distributed.

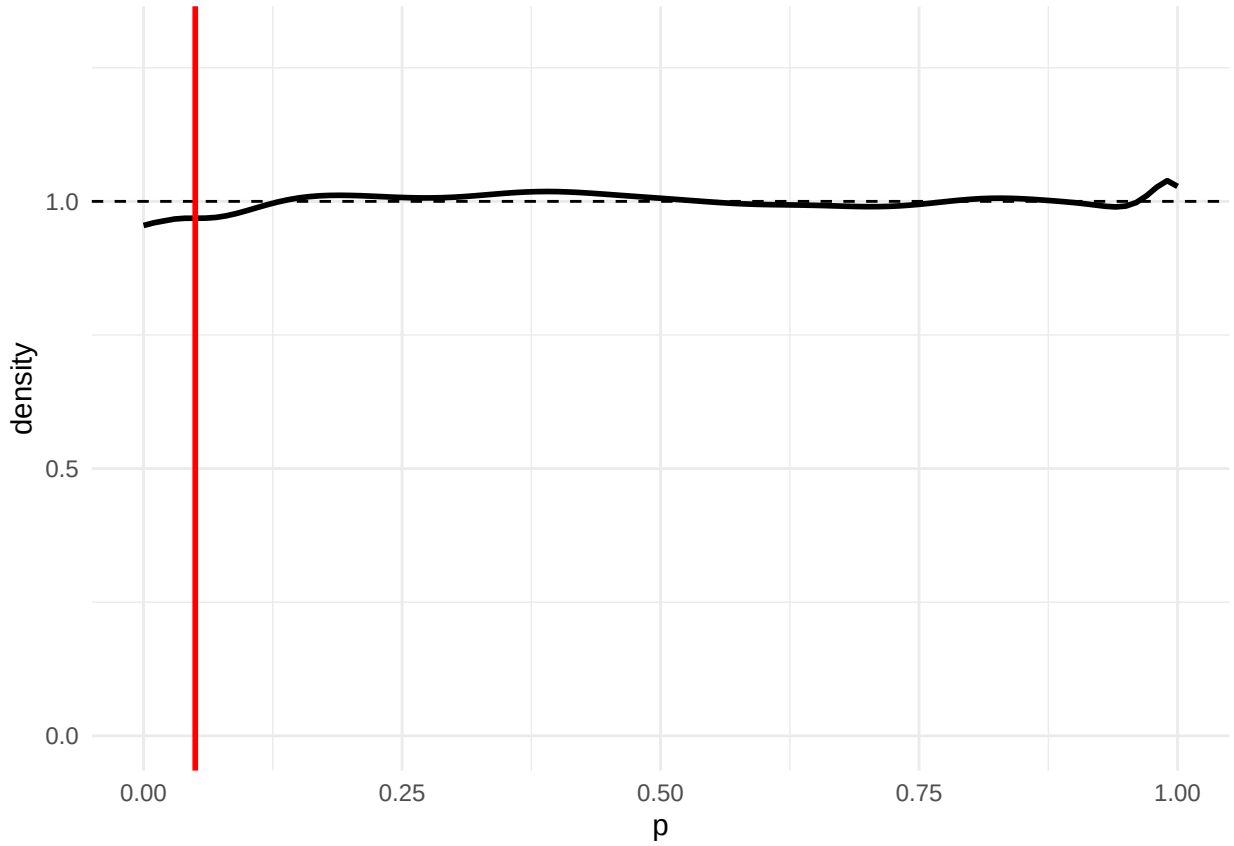

Figure 1: Bounded kernel density estimate for baseline results (black) before any RDFs are introduced. The uniform distribution (dashed) is provided for comparison. 4.8% of the probability mass is in the significant region (left of the red line $p = 0.05$). Overall, the results support the validity of the baseline null model.

We find that the effect on the false-positive rate varies significantly by RDF tested. For example, the *binary covariate* RDF adds about one-tenth of a percentage point to the false-positive rate[3], invariant across incidence specifications, while the *sample size manipulation* RDF nearly triples it when retesting after every pair of observations added. Even a single retest after doubling the sample size is enough to add three percentage points.

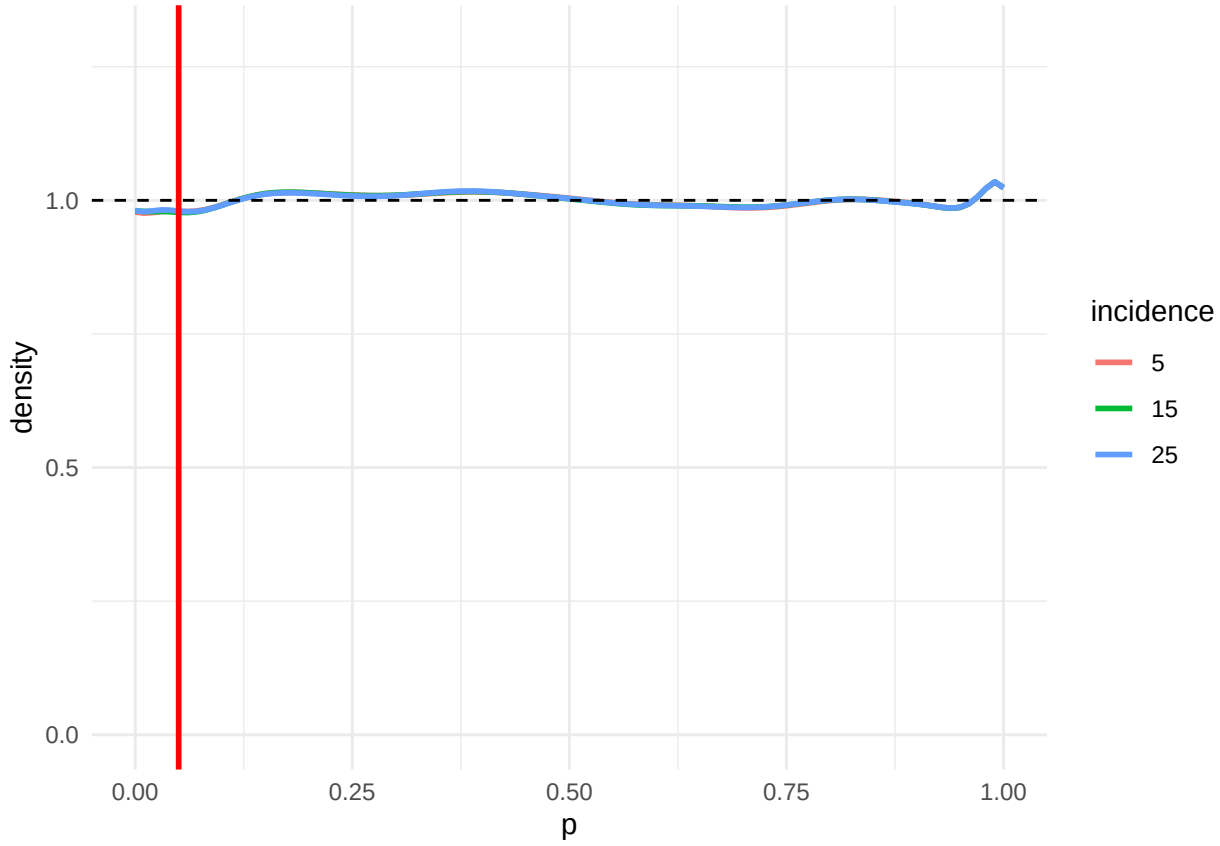

Figure 2: Bounded kernel density estimates for results after introducing the option to add a binary covariate, split by covariate incidence (color) with the uniform distribution (dashed) for comparison. 4.9% of the probability mass is in the significant region (left of the red line p = 0.05), invariant across covariate incidence specifications. Although split by how many observations of the covariate per cell are true, all results are visually indistinguishable from the baseline.

[3]It may be tempting to conclude from the nominal false-positive rate of 4.9% that the RDF had no effect. However, the p-values reported with the RDF are never greater than the baseline result, so the change caused by the RDF must be weakly positive. It follows that the very existence of a non-zero change confirms that the RDF increases the false-positive rate. Whether the effect is meaningful is a separate question.

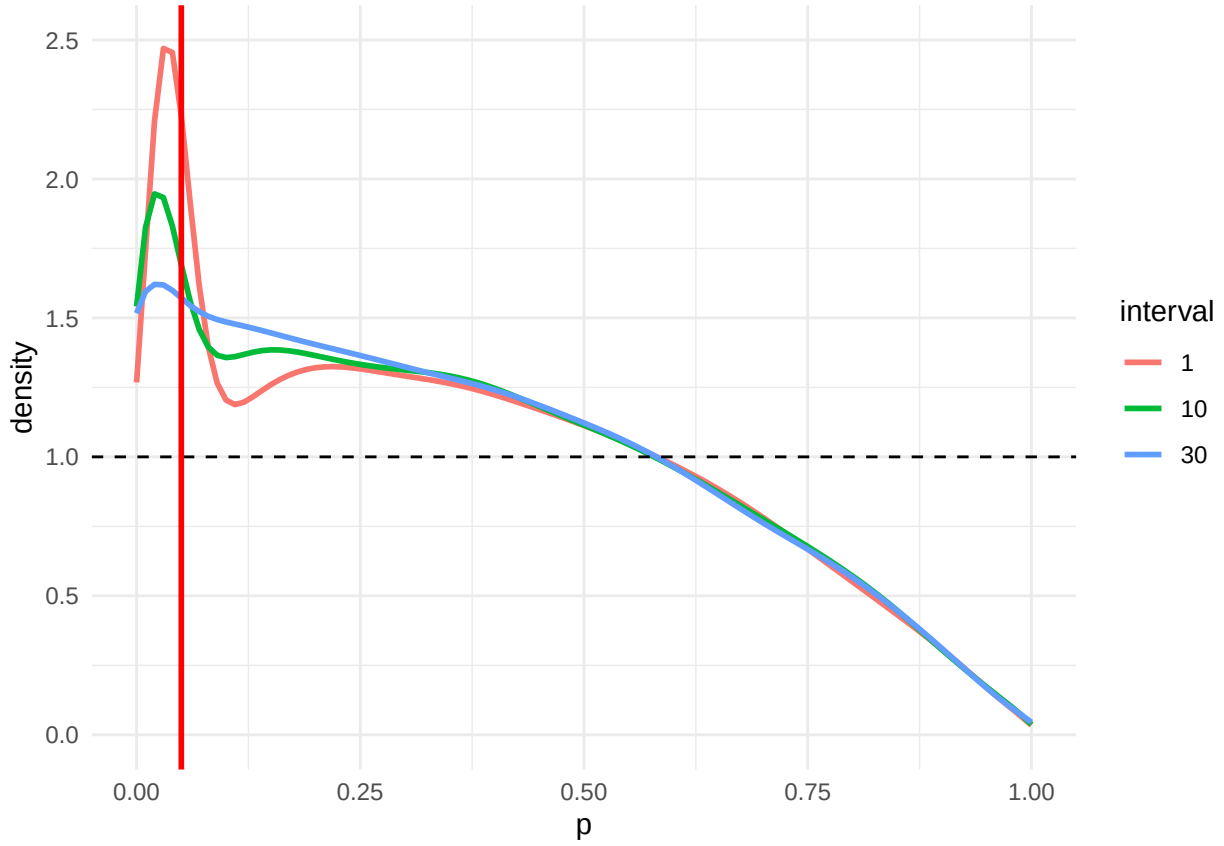

Figure 3: Bounded kernel density estimates for results with the option to add a specified number of observations per cell and retest until significant results are found or the sample size reaches double the original, split by retest interval (color) with the uniform distribution (dashed) for comparison. Much more of the probability mass is in the significant region (left of the red line $p = 0.05$), with the larger changes for lower retest interval specifications. The local peaks are a consequence of the optional stopping criteria; the simulations report the first significant result and do not try to improve them. This does not affect the final significance rates. The results suggest the null model is no longer valid.

We observe a similar spread of results between our two Outcome type RDFs. The generic *swap outcome* RDF doubles the false-positive rate while the specialized *drop item from index* RDF only adds between one and two percentage points, depending on the correlation specification.

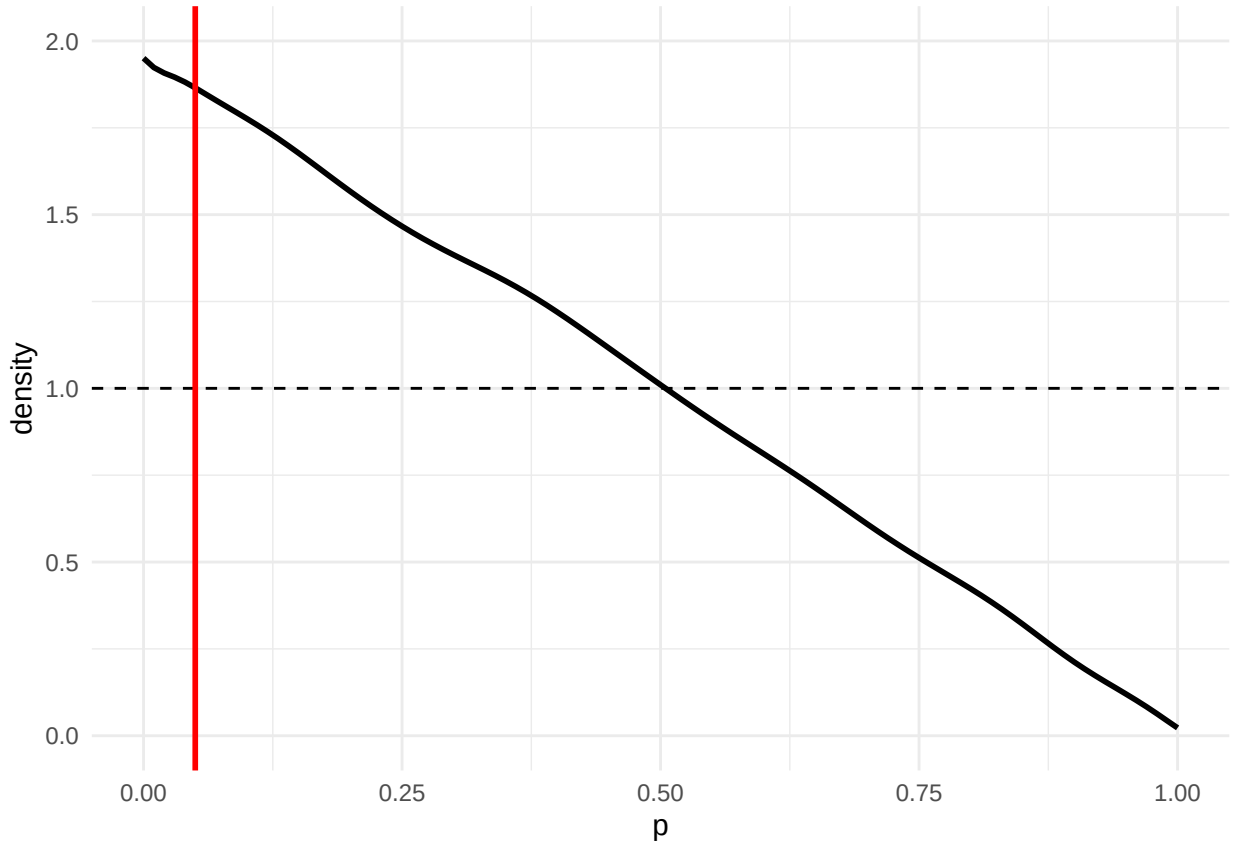

Figure 4: Bounded kernel density estimates for results after introducing the option to swap outcomes (black) with the uniform distribution (dashed) for comparison. 9.6% of the probability mass is in the significant region (left of the red line $p = 0.05$). The results suggest the null model is no longer valid.

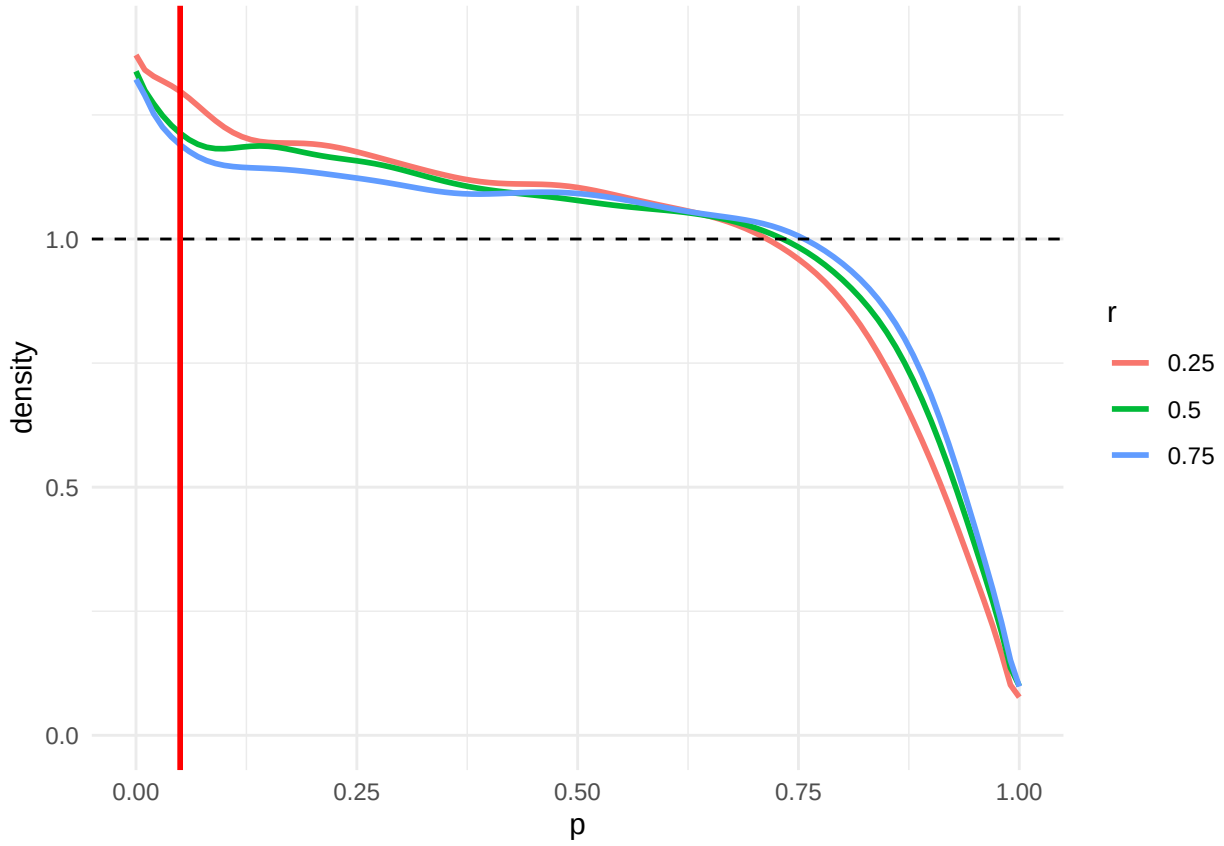

Figure 5: Bounded kernel density estimates for results after introducing the option to drop an uncorrelated column from the calculation of a composite response variable, split by the correlation between the remaining four columns (color) with the uniform distribution (dashed) for comparison. Slightly more of the probability mass is in the significant region (left of the red line p = 0.05), with slightly larger changes for lower correlation specifications. The results suggest the null model is no longer valid.

In general, the increase in the false-positive rate depends on the magnitude of the optionality inherent in the RDF tested. This intuitive trend is best illustrated by comparing the results of the *sample size manipulation* RDF by retest interval specification; it is not surprising that running more tests increases the chance of finding significant results. The results of the two Outcome type deviations provide another example; it is not surprising that an entirely new outcome is more impactful than modifying a composite index. This general principle seems to hold when comparing different RDFs as well, but these comparisons are less intuitive.

Table 3: Study 2 results by RDF, further split by parameter specification when applicable. We report the baseline false-positive rates, new false-postive rates after introducing the RDF, and the difference, all reported in exact terms

	baseline	RDF	change
Swap outcome	0.048	0.096	0.048
Add covariate, incidence = 5	0.048	0.049	0.001
Add covariate, incidence = 15	0.048	0.049	0.001
Add covariate, incidence = 25	0.048	0.049	0.001
Sample size manipulation, retest interval = 1	0.048	0.142	0.094
Sample size manipulation, retest interval = 10	0.048	0.103	0.055
Sample size manipulation, retest interval = 30	0.048	0.082	0.033
Drop item from index, correlation coefficient = 0.25	0.050	0.067	0.018
Drop item from index, correlation coefficient = 0.5	0.050	0.064	0.015
Drop item from index, correlation coefficient = 0.75	0.050	0.064	0.013

5 Discussion

Not every study needs to be preregistered. Exploratory studies should not lock themselves into an analysis plan; flexibility in the analysis is what leads to serendipitous discoveries. For example, Study 1 was not intended to confirm any hypotheses and does not gain much from being preregistered. Confirmatory studies designed to test a specific hypothesis and draw conclusions should be preregistered, as these results are expected to be reliable and replicable tests for true effects.

We recognize that there can be valid reasons to deviate from preregistrations, but the results of Study 2 highlight the importance of disclosure policies. Deviations should affect how we judge the reliability of the results. The inconsistent enforcement of disclosure policies prevent preregistration from being a consistent reliability indicator. We suspect that lowering the difficulty of auditing preregistrations could lead to consistent enforcement.

JPSP does not provide guidelines for how preregistrations should be written, so some were easier to audit than others. In the worst cases, researchers submitted a research proposal in the place of a preregistration. These 20+ page documents include research justifications, literature reviews, and other background information alongside methodology and analysis plans. Sifting through these documents added unnecessary time to the auditing process, especially since the relevant information was rarely collected in one place. While making the full research proposal publicly available is commendable in terms of transparency and openness, creating a separate and concise PAP greatly facilitates the auditing process.

In the best cases, preregistrations were hosted at aspredicted.org. This 8-item template leaves some room for background information, but is focused on defining the research methodology and analysis plan. Additionally, the template suggests a one-page limit and

enforces a strict two-page limit. Preregistrations following the template were easy to check item-by-item.

Researchers do not necessarily know where their studies will be published when preregistering their studies, so it may not be feasible for journals to require the use of a specific template. However, journals should provide guidelines about the length and breadth of the preregistrations, and researchers themselves should consider ease of auditing when writing their preregistration documents. Overall, we should not expect consistent enforcement of existing disclosure rules until the preregistrations themselves can be audited regularly.

6 Conclusion

Taken together, the results of our studies suggest that preregistration in its current form does not fully address the replicability concerns that spawned the practice. The descriptive analysis in Study 1 suggests that deviation from PAPs is not uncommon, even in a publication with nominal rules on disclosing deviations. Study 2 suggests that these deviations can severely impact the reliability of the results.

It should be noted that the false-positive rates found in Study 2 differ from false-positive discovery rates: false-positive rates are conditional on the null hypothesis, but false-positive discovery rates incorporate uncertainty about the null hypothesis. Future work assessing RDFs in empirical datasets would introduce the possibility of true effects. On the topic of future work, there are plenty of leads to follow: the right statistical framework could replace our estimated effect sizes with exact solutions, and there are always more RDFs to test.

Additionally, the publication mechanism adds another layer between false-positive discovery rates and false-positive incidence in published literature. For these and other reasons, we do not use the results of Study 2 to draw conclusions about the sample of Study 1.

Stronger enforcement of existing rules would increase the reliability of preregistered studies but cannot contain other instances of foul play. The quality of the entire published literature would benefit if the academic publication process could solve the underlying incentive mismatch between individual researchers and the broader research community.

References

Banerjee, Abhijit, Esther Duflo, Amy Finkelstein, Lawrence F. Katz, Benjamin A. Olken, and Anja Sautmann. 2020. "In Praise of Moderation: Suggestions for the Scope and Use of Pre-Analysis Plans for RCTs in Economics." Working Paper 26993. National Bureau of Economic Research. https://doi.org/10.3386/w26993.

Buck, Stuart. 2021. "Beware Performative Reproducibility." *Nature* 595 (7866): 151–51. https://doi.org/10.1038/d41586-021-01824-z.

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, et al. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science*, March. https://www.science.org/doi/abs/10.1126/science.aaf0918.

DeBruine, Lisa. 2021. *Faux: Simulation for Factorial Designs*. Zenodo. https://doi.org/10.5281/zenodo.2669586.

Dwan, Kerry, Carrol Gamble, Paula R. Williamson, and Jamie J. Kirkham. 2013. "Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias — An Updated Review." *PLoS ONE* 8 (7): 1–37. https://doi.org/10.1371/journal.pone.0066844.

Gibson, John, David L. Anderson, and John Tressler. 2014. "Which Journal Rankings Best Explain Academic Salaries? Evidence from the University of California." *Economic Inquiry* 52 (4): 1322–40. https://doi.org/10.1111/ecin.12107.

Goldacre, Ben, Henry Drysdale, Aaron Dale, Ioan Milosevic, Eirion Slade, Philip Hartley, Cicely Marston, Anna Powell-Smith, Carl Heneghan, and Kamal R. Mahtani. 2019. "COMPare: A Prospective Cohort Study Correcting and Monitoring 58 Misreported Trials

in Real Time." *Trials* 20 (1): 118. https://doi.org/10.1186/s13063-019-3173-2.

Huntington-Klein, Nick, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, et al. 2021. "The Influence of Hidden Researcher Decisions in Applied Microeconomics." *Economic Inquiry* 59 (3): 944–60. https://doi.org/10.1111/ecin.12992.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2 (8): 696–701. https://doi.org/10.1371/journal.pmed.0020124.

Locascio, Joseph J. 2017. "Results Blind Science Publishing." *Basic & Applied Social Psychology* 39 (5): 239–46. https://doi.org/10.1080/01973533.2017.1336093.

Rosenthal, Robert. 1979. "The File Drawer Problem and Tolerance for Null Results." *Psychological Bulletin* 86 (3): 638–41. https://doi.org/http://dx.doi.org/10.1037/0033-2909.86.3.638.

Santafe, Guzman, Borja Calvo, Aritz Perez, and Jose A. Lozano. 2015. *Bde: Bounded Density Estimation*. https://CRAN.R-project.org/package=bde.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359–66. https://doi.org/10.1177/0956797611417632.

Sterling, Theodore D. 1959. "Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance—or Vice Versa." *Journal of the American Statistical Association* 54 (285): 30–34. https://doi.org/10.1080/01621459.1959.10501497.

Turner, Erick H., Annette M. Matthews, Eftihia Linardatos, Robert A. Tell, and Robert Rosenthal. 2008. "Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy." *New England Journal of Medicine* 358 (3): 252–60.

https://doi.org/10.1056/NEJMsa065779.

Williams, Cole Randall. 2019. "How Redefining Statistical Significance Can Worsen the Replication Crisis." *Economics Letters* 181 (August): 65–69. https://doi.org/10.1016/j.econlet.2019.05.007.