Observational Studies 1 (2015) 18-29 Submitted 10/14; Published 12/14

Evidence of False Positives in Research Clearinghouses and Influential Journals: An Application of P-Curve to Policy Research

Sean Tanner **stanner@berkeley.edu**
Goldman School of Public Policy
University of California, Berkeley
Berkeley, CA 94720-7320

Abstract

This article presents a pre-analysis plan for analyzing the evidential value in a selection of policy research taken from scholarly journals and two research clearinghouses run by the federal government. The analysis will collect p-values from selected studies and estimate the evidential value that they represent using the newly introduced p-curve. This article outlines a precise data collection routine, a set of decision rules for including p-values in the analysis sample, and exact hypothesis tests to be used.

Keywords: selective reporting, causal inference, public policy, labor economics, education, transparency.

1. Introduction

This pre-analysis plan (PAP) provides an outline for analyzing the evidential value in a selection of policy-oriented research using the p-curve estimation strategy. PAPs, also referred to as study protocols, are generally suggested as guards against false-positives in randomized control trials (Chalmers & Altman, 1999, p. 491). Publishing a PAP before analyzing data leaves a record of the intended data analysis, whether in a peer-reviewed journal or a formal trial registry such as that run by the United States National Institutes of Health[1]. Advocates of such pre-registration argue that this published record helps ameliorate the risk of false positives in two ways. First, it reduces the file-drawer problem whereby meta-analyses are biased upward (in magnitude) because they fail to include unpublished null results (Rosenthal, 1979). Second, PAPs diminish the threat of selective reporting of results within studies by explicitly detailing the hypotheses to be tested and the estimation procedures for doing so (Casey, Glennerster, & Miguel, 2012, pp. 17761778)

Though PAPs are now part of standard research procedures in medical trials (De Angelis et al., 2004), they are still used only sparingly in social science field experiments[2] and almost never in observational work[3]. The use of PAPs can be contentious, particularly

[1] https://clinicaltrials.gov/

[2] At the time of this writing, the American Economic Association's registry, https://www.socialscienceregistry.org/, contains 244 registered studies, but only 54 of those had analysis plans on file.

[3] To the best of the author's knowledge, this is the first registered PAP for a purely descriptive, observational study

in observational work in which ex-post data analysis can reveal unexpected, scientifically meritorious findings (Pearce, 2011)[4]. Proponents of PAPs contend that their use does not preclude such analysis, but merely defines which analyses were planned ex-ante and which were done ex-post (The PLOS Medicine Editors, 2014, p. 2). What to make of the divergence in ex-ante and ex-post analyses is an ongoing issue of concern (Dwan et al., 2014). Despite the controversy over PAPs, it is especially important in this case to articulate the data collection and analysis plan in advance in order to avoid iteratively searching through data and model specifications until results are found that conform to the researcher's prior hypotheses[5]. As the results of the study for which this PAP is created will bear directly on the need for PAPs in policy research, publication of a protocol lends needed transparency.

2. Overview

Recent work has uncovered a troubling pattern in social science research: statistically significant results can often be attributed to selective reporting of results within studies, or p-hacking (Ioannidis, 2005; Simonsohn, Nelson, & Simmons, 2014). P-hacking occurs when a researcher, consciously or not, adjusts his model specifications or analysis sample in numerous ways until he finds a significant coefficient on the explanatory variable(s) of interest. This practice is given numerous labels, including "significance chasing," "specification search," "selective reporting," and "massaging the data." Implicit and explicit accusations of such behavior have generated heated controversy over numerous research questions, including which studies merit inclusion in meta-analysis (Krueger, Hanushek, & Rice, 2002), how to define an African-American sub-sample from survey data (Howell, Wolf, Campbell, & Peterson, 2002; Krueger & Zhu, 2004), and how to dichotomize streams based on their width (Hoxby, 2007; Rothstein, 2007) to name but a few. In each of these cases, the salient effect disappeared when an alternative, equally defensible coding choice was employed, leading to partisan disagreements rather than scientific advancement.

The specter of p-hacking is worrisome in applied social science research, the results of which are relied on by policy makers in their deliberations (Bogenschneider & Corbett, 2010; Gueron & Rolston, 2013).[6] The aforementioned examples of p-hacking controversies were taken from research on the impacts of class size, school vouchers, and school competition on student achievement. Similar lists could be generated for policy-related research on labor, crime, health, and the environment. If p-hacking plagues such research, then the information social scientists provide to the policy community is unreliable. It is therefore of first order importance to investigate the reliability of policy-related scholarship. Until recently, however, it was difficult to detect p-hacking in empirical social science research.

The introduction of p-curve by Simonsohn, Nelson, and Simmons (2014) has made it possible to distinguish true effects from those that result from p-hacking using only the observed results that cross a particular significance level. The null hypothesis begins with

[4] In a satirical blog post, William Easterly imagines that a PAP would have prevented Columbus from reporting the discovery of America:
http://www.nyudri.org/2012/10/15/if-christopher-columbus-had-been-funded-by-gates/

[5] Full disclosure: the author is associated with the group of scholars working toward greater transparency in social science. The author has not received any funding for this research.

[6] This is not a claim that policy makers rely *sufficiently* on academic research or that they rely on it as heavily as social scientists would prefer.

an assumption that all null hypotheses are true, which would create a uniform distribution of p-values, where the relative frequency of p-values between .05 and .04 is the same as the relative frequency of p-values between .03 and .02, which is also the same as the relative frequency of p-values between .02 and .01, and so on. The choices of bin width and significance thresholds do not change the uniformity of relative frequencies so long as the bin width is constant across to support of observed p-values that cross a particular significance threshold.

False null hypotheses (real effects) generate a right-skewed distribution of p-values, with more frequent values in the range .01 to .02 than .04 to .05. The greater the statistical power of individual studies, the more right-skewed a p-curve that results from those studies. There is no natural distribution of p-values that is left-skewed: either the distribution is uniform (all null hypotheses are true) or is right-skewed (at least some null hypotheses are false). A left-skewed distribution of p-values can only occur under p-hacking, whereby the researcher keeps altering the analysis until a significant p-value is found. This creates clustering of p-values around the chosen significance threshold and a left-skewed distribution.

This estimation strategy has been used to detect spurious results in the Journal of Personality and Social Psychology (Simonsohn et al., 2014) and support the evidence in a foundational set of studies on anchoring (Simonsohn & Nelson, 2014). Figure 1 (in the appendix) shows the distribution of p-values in a set of studies suspected to be plagued by p-hacking and another set suspected to contain evidence. The research suspected to have been p-hacked generated a left-skewed distribution, whereas the research suspected to contain evidence generated a right-skewed distribution.

This analysis will evaluate the *evidential* value of recent policy-related scholarship in influential academic journals and research clearinghouses. Using the definition adopted by Simonsohn et al. (2014, p. 3), a set of findings in a body of research contains evidential value when p-hacking can be ruled out as their sole explanation. The sources were chosen as they are believed to be repositories of rigorous, policy-related analyses. If scholarship in these sources does not contain evidential value, then the credibility of applied social science research can be seriously questioned.

3. Data Collection Guidelines

Applying p-curve to a set of studies requires selecting the studies and the sub-set of p-values from them to include in the analysis sample. Because this process is itself open to the p-hacking that p-curve is meant to uncover, Simonsohn et al. (2014) recommended that a study selection rule be set in advance of the data collection that is precise enough to allow an independent researcher using it to get the same set of studies. Once the studies have been selected, the p-values contained therein need to be extracted for the analysis sample. Because most studies include multiple p-values, a careful selection routine must also be implemented for the p-values themselves. In addition to being replicable and set in advance of data collection, the selection routine must: (1) test the hypothesis of interest, (2) have a uniform distribution under the null, and (3) be statistically independent of other p-values in the p-curve.

4. Data Collection Plan

The unit of analysis is conceptually a study but operationally a p-value, as usually only one p-value can be used from each study. The sampling frame of studies was chosen to fit three primary criteria: (1) the perception of rigor, (2) policy-relevance, and (3) recency. These criteria allow for an analysis of the current state of prestigious, applied social science research and are features of research onto which policy makers have explicitly placed high value (Bogenschneider & Corbett, 2010, p. 35). The sampling frame is also narrow enough to allow for a census of studies that meet the following rules. All studies chosen must include estimation of causal parameters in real data. Purely descriptive papers and modeling exercises will not be included. The studies are likely to include a mix of observational/natural experiment designs and randomized control trials. As is made clear in the hypothesis list, separate p-curve analyses will be done for these two categories of research designs.

The studies taken into the sample will be:

A. The 20 most recent articles in the *Journal of Policy Analysis and Management (JPAM)*, *The Journal of Human Resources (JHR)*, and *Education Evaluation and Policy Analysis (EEPA)* that estimate causal parameters in real data.

B. The entire group of studies included in the What Works Clearinghouse's (WWC) Single Study Review section in the following categories that meet the clearinghouse's standards of evidence (with and without qualifications): http://ies.ed.gov/ncee/wwc/.

   a. Dropout Prevention
   b. Early Childhood Education
   c. Postsecondary Education
   d. School Choice
   e. Teacher Incentives

C. The entire group of causal studies included in the Clearinghouse for Labor Evaluation and Research (CLEAR) database that have been labeled high or moderate strength. http://clear.dol.gov/study_database

The journals were chosen as representative of influential, policy-oriented scholarly research. *JPAM* is the sole journal published by the Association for Public Policy Analysis and Management and is widely considered to be the flagship journal in policy research. *JHR* and *EEPA* were both identified as having authors and readership similar to *JPAM* (Reuter & Smith-Ready, 2002).

The two clearinghouses were chosen for similar reasons and because they are supposed to be repositories for high quality research. WWC is maintained by the Department of Education's Institute for Educational Sciences as a "central and trusted source of scientific evidence for what works in education to improve student outcomes" (Institute of Educational Sciences, Department of Education, 2014). As is typical of research clearinghouses, the WWC assigns standards ratings to individual studies. The studies used in this analysis must meet the WWC standards "without reservation" or "with reservation." In order to meet the WWC standards without reservation, "study participants must have been placed

into each study condition through random assignment or a process that was functionally random," whereas the rating "meets standards with reservation" is applied to studies whose participants were not placed into study conditions through random assignment but can demonstrate baseline equivalence (Institute of Educational Sciences, Department of Education, n.d.). The ratings used by CLEAR, high causal evidence and moderate causal evidence, are determined in a similar manner and correspond to the WWC ratings of without reservation and with reservation, respectively (Department of Labor, 2014).

These sampling frames are not entirely independent, as a study might be published in one of the three journals and also listed in one of the clearinghouses. P-values that appear in more than one category (i.e. in a journal and a clearinghouse) will be counted only once in each analysis for which they meet the inclusion criteria.

The p-values analyzed from the studies will be those from the tests of the main hypotheses stated by the authors in either the abstract or introduction. All p-values that cross the .1 significance level will be analyzed. This level of significance is justified for two reasons. First, it is common in this literature to call attention to marginally significant results for which the p-value falls between .05 and .1. Such results are publishable by most standards of policy-oriented social science research. It is therefore equally plausible that a researcher would p-hack the data until p-values cross the .1 or the .05 levels. Second, there is some evidence that significance levels cluster around both the .1 and .05 levels (Brodeur, Lé, Sangnier, & Zylberberg, 2013, p. 29).

The determination of "main" can be ambiguous for several reasons, which will be dealt with through the following rules:

A. Multiple Model Specifications
   Often authors will report multiple model specifications, for example by including a regression table with increasingly abundant covariates. In these cases, the main specification will be determined by searching for key words such as favored or preferred specification.

B. Multiple Dependent Variables
   Policy research commonly focuses on several outcome variables, such as educational attainment, wage, and program participation (SNAP, Medicaid, WIC). In these cases, the dependent variables mentioned in the abstract will be chosen. If that criterion does not resolve ambiguity, then p-values from each equation will be collected and a single one from that set will be included in the analysis sample by random assignment (see below).

C. Heterogeneous Treatment Effects
   Social science theory and practical policy experience often suggest that an effect should be heterogeneous across some demographic variable, such as the impact of a negative income tax on men vs. women. Simonsohn et al. articulate rules for dealing with simple versus attenuated effects (2014, pp. 2022) that rely on a model of p-hacking that is eminently reasonable for psychological laboratory experiments, but may not be plausible in applied policy research.[7] The most plausible form of p-hacking with regard

[7] The relevant issue is whether a paper is publishable if a simple effect is not significant but an interaction effect is. Simonsohn et al. state that such a paper would not be publishable and therefore potentially

to heterogeneous treatment effects in this domain is when a researcher searches across various socioeconomic or regional variables until a significant interaction is found. In the case of either multiple bivariate interactions (treatment x gender; treatment x Caucasian; etc.) or multivariate interactions (treatment x gender x Caucasian), the p-values from groups mentioned in the abstract will be chosen. If that criterion does not resolve ambiguity, then one p-value will be chosen by random assignment (see below).

If more than one main hypothesis is identified, a single p-value will be randomly chosen by assigning a value to the order in which the p-value is referenced in the paper (1st, 2nd,...Kth), then using the random number generating service Random.org (Haahr, 2014) to generate a random integer between 1 and K. The p-value whose order corresponds to the randomly chosen integer will be included in the primary analysis sample. The publicly posted data set will identify ambiguities and how determination of the included p-value was made ("in abstract," "by random assignment," "authors' preferred model," etc.).

If precise p-values are not reported, precise p-values will be recalculated given the test statistics reported in the articles.

P-curve is not well-defined for discrete tests (Simonsohn et al., 2014, p. 27). P-values from discrete tests will be included, but a separate p-curve will be generated for each hypothesis that does not include discrete tests as a robustness check.

5. Hypotheses

The following null hypotheses will be tested using the data. For each hypothesis, the distribution of p-values will be analyzed for right-skew (evidential value) and for left-skew (p-hacking). This leads to four potential outcomes: (1) skewed right but not skewed left (evidential value), (2) skewed left but not skewed right (p-hacking), (3) neither right nor left skewed (uniform distribution- no evidential value but no p-hacking), (4) both right and left skewed (U-shaped- a mix of evidential value and p-hacking).

Hypothesis I is the primary hypothesis of this study as it is the most direct and general analysis of the evidential value in policy scholarship. Hypotheses I.a, I.b, II, II.a, II.b, and III are secondary as they are each specific to a particular sub-domain of policy research. It is unlikely that hypotheses I.c and I.d will be feasible. If they are, then they should be considered secondary hypotheses at the level of I.a, I.b, II, II.a, II.b, and III.

I. Overall, policy research contains evidential value.

   The entire set of p-values in the sample will be analyzed using p-curve. Sub-components of this hypothesis are:

   a. Randomized control trials contain evidential value

      Only p-values from randomized control trials will be analyzed. This includes studies exploiting a random lottery even when subjects are endogenously selected onto the list from which the lottery is held.

subject to p-hacking of the simple effect (2014, p. 21, note 17). In policy research, a paper would likely be publishable if an interesting interaction effect was significant regardless of (or especially when) the simple effect was not.

b. Non-experimental research contains evidential value

Only p-values from non-experimental research (all studies not included in I.a) will be analyzed.

c. Research with published pre-analysis plans or study protocols contain evidential value

It is unlikely that many studies will fit this category, as PAPs are still rare in this field. If at least 20 studies fit this category, then hypotheses I.c and I.d will be tested.

d. Research without published pre-analysis plans or study protocols contain evidential value See hypothesis I.c.

II. Research clearinghouses contain evidential value

The entire set of p-values in the sample gathered from WWC and CLEAR will be analyzed using p-curve.

a. High-quality designation indicates evidential value

Only p-values from studies designated by WWC as meeting standards *without qualification* and CLEAR as having *high causal evidence* will be analyzed using p-curve.

b. Medium-quality designation indicates evidential value

Only p-values from studies designated by WWC as meeting standards *with qualification* and CLEAR as having *moderate causal evidence* will be analyzed using p-curve.

The preceding sub-hypotheses are divided in such a way that the grouping will correspond strongly to the grouping in hypotheses I.a and I.b. The two sets of sub-hypotheses will not be completely similar, however, as the journal articles are not included in II.a and II.b and the clearinghouses label some experiments as second-tier rather than first-tier due to high attrition, lack of correction for multiple hypothesis testing, and other threats to strong causal inference (Department of Labor, 2014; Institute of Educational Sciences, Department of Education, n.d.).

III. Policy-oriented journals contain evidential value

The entire set of p-values taken from *JPAM*, *JHR*, and *EEAP* will be analyzed using p-curve.

In addition to tests for evidential value in these domains, it might be of interest to know whether one domain has greater evidential value than another. Unfortunately, the exact testing procedure for this question has not been formalized. If such a procedure becomes available between the submission of this PAP and the final analysis, the author will employ it and detail the research strategy in a deviation report (see section VII).

6. Estimation Procedure and Statistical Power

The logic of p-curve is somewhat intuitive: real effects should generate lower p-values more often than higher p-values. This logic underlies the estimation strategy favored by Simonsohn et al. (2014). In the first stage, each significant p-value is transformed into the probability of observing such a value at least as extreme if the null were true. For continuous tests, this is merely the p-value divided by the significance threshold chosen by the p-curve analyst (for example .1). This transformed value is labeled the pp-value by Simonsohn et al. (2014). The second stage aggregates the pp-values using Fisher's method in which -2 multiplied by the sum of the natural log of k uniform distributions is distributed 2(2k) under the null hypothesis that the pp-values are uniformly distributed (corresponding to no evidential value) (Fisher, 1932). In addition to performing these calculations with the author's own code, the analysis will use the p-curve web application, http://www.p-curve.com/app2/, to construct the p-curve graphs and simultaneously test for evidential value and p-hacking as a check on the coding.

Because of the well-known limitations of Fisher's method (Rosenthal, 1978), a second estimation strategy will be employed to ensure results are not specific to Fisher's method. A simple method of estimating evidential value would be to dichotomize p-values as low ($0 < p < .05$) and high ($.05 < p < .1$) and submit the data to a binomial test with a uniform null (50% low). This procedure is resistant to extreme values, simple, and transparent. However, it is inefficient as it ignores variation of p-values within the high and low categories. This simple method will be used as a robustness check on all hypotheses, but will not be the primary focus of the analysis.

In testing seven hypotheses, individual p-values from each test cease to have their intended meaning and underestimate the risk of false positives. Consequently, the method of controlling the false discovery rate (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001) will be employed to further guard against spurious results.

Simonsohn et al. (2014) provide results from simulations that shed light on the potential power of p-curve to uncover evidential value in a set of studies. For studies with a fixed sample size of 20 subjects each and power of 33%, p-curve (Fisher's method) will conclude that a set of 20 studies contains evidential value 85% of the time (see figure 2 in the appendix). As the primary hypothesis in the study will be tested on nearly 100 p-values, statistical power will not likely be an impediment to inference. Sub-hypotheses will almost certainly have more than 20 p-values in each test and so will also not likely be plagued by low statistical power.

7. Dissemination of Deviations, Analyses, and Reproductions

Deviations from this pre-analysis plan will be dated and posted on the author's website (https://sites.google.com/site/patrickseantanner/). Justifications for each deviation will be given in the document.

Data and statistical software code will be posted on the author's website upon completion of the analysis. The raw Excel sheet that contains all the data together with the Stata do-file should allow individuals to exactly reproduce the analysis. The data will include notes on determining which p-values were included and why, so that an individual

can easily assess the sensitivity of results to the author's coding choices. No special access permissions will be needed to view or download the data and code- it will be publicly available.

All publications resulting from the analysis will reference this PAP and any deviations from it. Links to this PAP and any resulting conference presentations and publications will be added to the author's website.

References

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), 289300.

Benjamini, Y., and Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics*, 29(4), 11651188.

Bogenschneider, K., and Corbett, T. (2010). *Evidence-Based Policymaking: Insights from Policy-Minded Researchers and Research-Minded Policymakers*. New York: Routledge.

Brodeur, A., Lé, M., Sangnier, M., and Zylberberg, Y. (2013). *Star Wars: The Empirics Strike Back*, Discussion Paper Series, Forschungsinstitut zur Zukunft der Arbeit, No. 7268).

Casey, K., Glennerster, R., and Miguel, E. (2012). Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan. *Quarterly Journal of Economics*, 127(4), 17551812.

Chalmers, I., and Altman, D. G. (1999). How can medical journals help prevent poor medical research? Some opportunities presented by electronic publishing. *Lancet*, 353(9151), 4903.

De Angelis, C., Drazen, J. M., Frizelle, F. A., Haug, C., Hoey, J., Horton, R., Van Der Weyden, M. B. (2004). editorials Clinical Trial Registration: A Statement from the International Committee of Medical Journal Editors. *New England Journal of Medicine*, 351(12), 12501251.

Department of Labor, U. S. (2014). CLEAR CAUSAL EVIDENCE GUIDELINES, VERSION 1.1 (pp. 111).

Dwan, K., Altman, D. G., Clarke, M., Gamble, C., Higgins, J. P. T., Sterne, J. a C., Kirkham, J. J. (2014). Evidence for the selective reporting of analyses and discrepancies in clinical trials: a systematic review of cohort studies of clinical trials. *PLoS Medicine*, 11(6), e1001666.

Fisher, R. A. (1935). *Design of Experiments*, Edinburgh: Oliver & Boyd.

Gueron, J. M., & Rolston, H. (2013). *Fighting for Reliable Evidence*. New York: Russell Sage Foundation.

Haahr, M. (2014). RANDOM.ORG - Company Information. Retrieved September 24, 2014, from http://www.random.org/company/

Howell, W. G., Wolf, P. J., Campbell, D. E., & Peterson, P. E. (2002). School vouchers and academic performance: results from three randomized field trials. *Journal of Policy Analysis and Management*, 21(2), 191217.

Hoxby, C. (2007). Does Competition Among Public Schools Benefit Students and Taxpayers? Reply. *American Economic Review*, 97(5), 20382055.

Institute of Educational Sciences, Department of Education, U. S. (n.d.). *What Works Clearinghouse Procedures and Standards Handbook (Version 3.0).*

Institute of Educational Sciences, Department of Education, U. S. (2014). *About Us: What Works Clearinghouse.* Retrieved September 24, 2014, from http://ies.ed.gov/ncee/wwc/aboutus.aspx

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8).

Krueger, A. B., Hanushek, E. A., & Rice, J. K. (2002). *The Class Size Debate.* (M. Lawrence & R. Rothstein, Eds.) (p. 102). Washington, D.C.: Economic Policy Institute.

Krueger, A. B., & Zhu, P. (2004). Another Look at the New York City School Voucher Experiment. *American Behavioral Scientist*, 47(5), 658698.

Pearce, N. (2011). Registration of protocols for observational research is unnecessary and would do more harm than good. *Occupational and Environmental Medicine*, 68(2), 868.

Reuter, P., & Smith-Ready, J. (2002). Assessing JPAM after 20 Years. *Journal of Policy Analysis and Management*, 21(3), 339353.

Rosenthal, R. (1978). Combining Results of Independent Studies. *Psychological Bulletin*, 85(1), 185193.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638641.

Rothstein, J. (2007). Does Competition Public Schools Benefit Students Among and Taxpayers? Comment. *American Economic Review*, 97(5), 20262037.

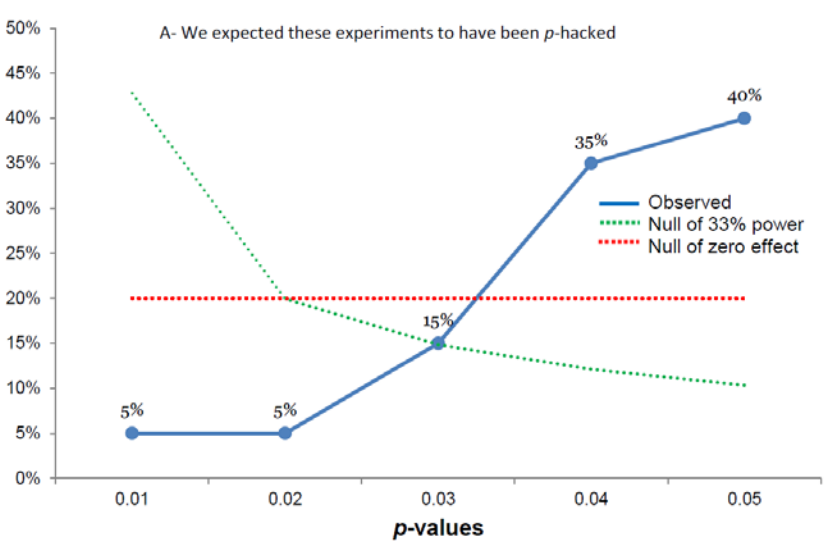
Simonsohn, U., and Nelson, L. D. (2014). Anchoring is Not a False-Positive: Maniadis, Tufano, and List's (2014) "Failure-to-Replicate" is Actually Entirely Consistent with the Original. Working paper (April 27, 2014). Available at http://dx.doi.org/10.2139/ssrn.2351926

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology. General*, 143(2), 53447.

The PLOS Medicine Editors. (2014). Observational studies: getting clear about transparency. *PLoS Medicine*, 11(8), e1001711.

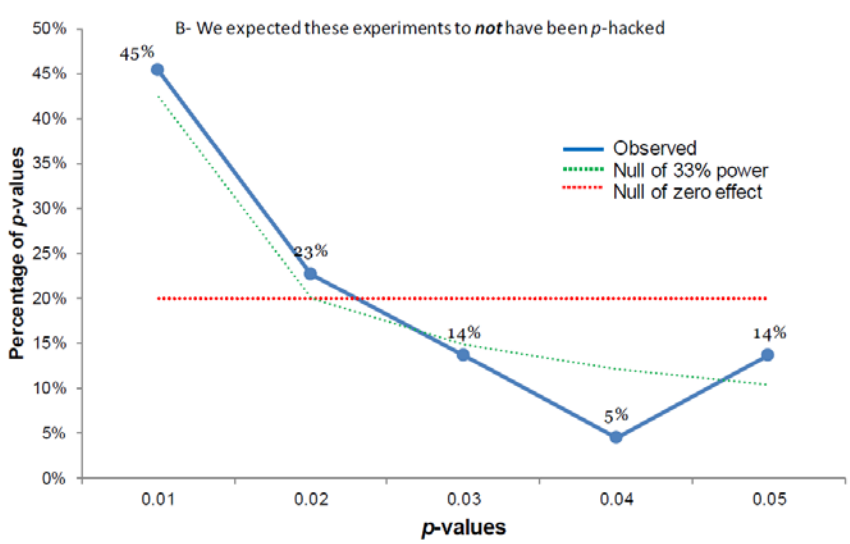
Appendix

Figure 1: P-curves with and without p-hacking



Statistical Inference	Results
1) Studies contain evidential value *(right-skewed)*	$x^2(40)=18.3$, p=.999
2) Studies lack evidential value (flatter than 33%)	$x^2(40)=82.5$, $p<.0001$
3) Studies lack evidential value and were intensely *p*-hacked? *(left-skewed)*	$x^2(40)=58.2$, $p=.031$

The observed p-curve includes 20 significant p-values, an additional 3 were p>.05
Of those 20 p-values, 3 are p<.025, binomial test for right-skew: p>.999,left -skew: p=.0013

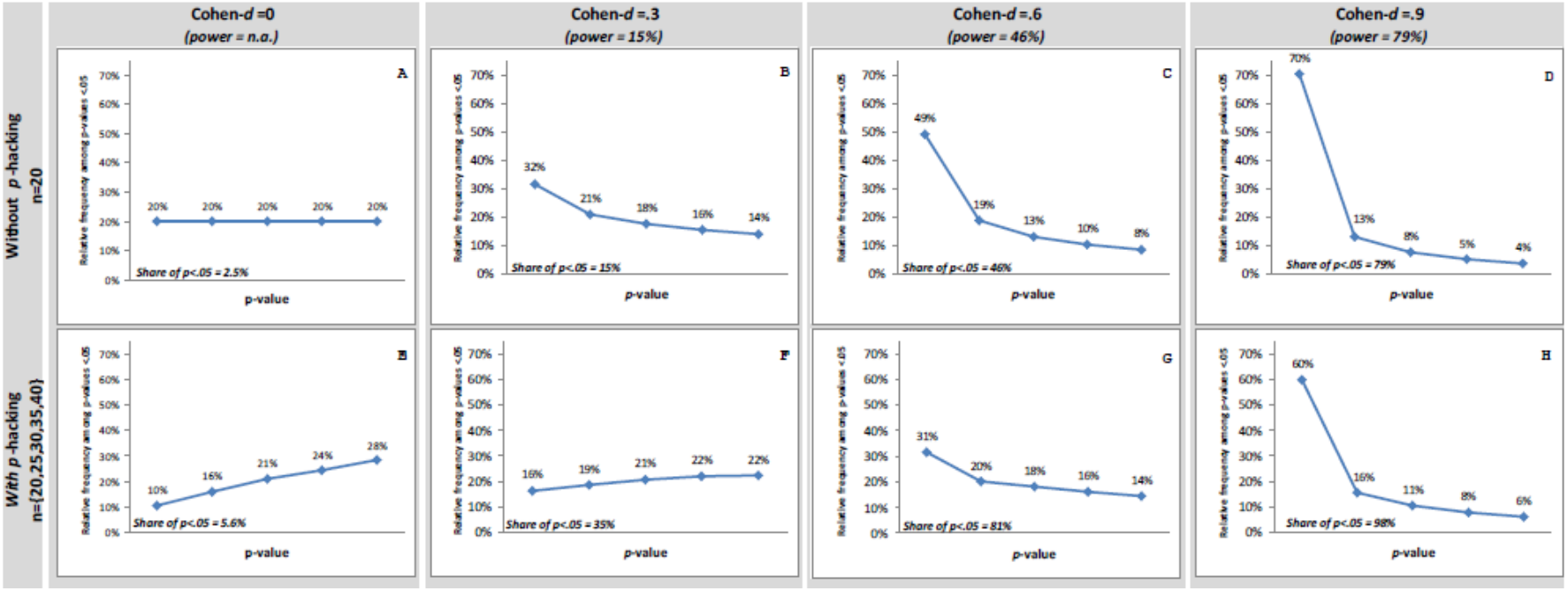


Statistical Inference	Results
1) Studies contain evidential value *(right-skewed)*	$x^2(44)=94.2$, p<.0001
2) Studies lack evidential value (flatter than 33%)	$x^2(44)=43.2$, $p=.507$
3) Studies lack evidential value and were intensely *p*-hacked? *(left-skewed)*	$x^2(44)=27.2$, $p=.978$

The observed p-curve includes 22 significant p-values, an additional 3 were p>.05
Of those 22 p-values, 16 are p<.025, binomial test for right-skew p=.026, for left-skew p=.991.

Note. The shape of p-curves with and without p-hacking. Reprinted from Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General,* 143(2), 534–47. Copyright 2014 American Psychological Association. Reprinted with permission.

Figure 2: P-curves with and without p-hacking for various real effect sizes.



Note. The power of p-curve to detect evidential value with and without p-hacking. Reprinted from Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General,* 143(2), 534–47. Copyright 2014 American Psychological Association. Reprinted with permission.