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

Sean Tanner

Selective reporting of results within studies, a practice known as "p-hacking," presents a pernicious threat to the validity of published social science research. This form of publication bias is particularly troubling in research that influences public policy debates and programmatic funding. While recent analyses of political science, psychology, economic, and sociology research have identified considerable publication bias, the methods used are unable to distinguish p-hacking from accidents of random sampling. Using the novel "p-curve" estimation framework, which can distinguish p-hacking from other forms of false positives, I find little evidence of p-hacking in federal research clearinghouses and three rigorous, policy-oriented journals: the Journal of Policy Analysis and Management, The Journal of Human Resources, and Education Evaluation and Policy Analysis. The results demonstrate that most of the findings in this literature are likely to be true positives. I discuss methodological and substantive divergence from the prior literature and suggest future research that can more accurately assess the cross-disciplinary veracity of social science research.

I. Introduction

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 a 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.[1] In each of these cases, the salient effect changed markedly 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 and funding decisions (Bogenschneider & Corbett, 2010; Gueron & Rolston, 2013; Haskins & Margolis, 2015).[2] To take one example, the Obama administration's $1.1 billion "Investing in Innovation Initiative" allows $50 million for each educational program that has been validated by a single randomized control trial (Haskins & Margolis, 2015, Chapter 4). Though laudable in its effort to promote high internal validity, this funding rule incentivizes p-hacking by allocating substantial resources on the basis of a single statistically significant result. 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 distinguish p-hacking in empirical social science research from other forms of publication bias with less pernicious consequences for the accuracy of published research, such as journal editors' preference for rejection of null hypotheses. This article employs a novel estimation framework, "p-curve", to assess the veracity of reported effects in three influential policy-oriented journals and two research clearinghouses maintained by the federal government. P-curve allows the analyst to distinguish amongst real effects, p-hacking, and false positives that occur by chance.

[1] The reference to these is meant to underscore the existence of controversy, not to imply that the original authors p-hacked their data.

[2] 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.

II. False Positive Social Science

Concern over dubious research practices is not new to the social sciences. Decades have passed since Leamer's (1983) foundational article on spurious causal inference in econometric work and Rosenthal's (1979) sobering illustration of the sensitivity of meta-analysis to unpublished null findings. The ensuing years have witnessed a vigorous debate on the merits of randomized and natural experiments in the production of durable scholarly knowledge, as well as a codification of Rosenthal's concerns in meta-analytic literature.[3] The hegemonic view in many disciplines is that of a "credibility revolution," wherein rigorous methods are allowing for more valid and reliable research than that produced with older methods (Angrist & Pischke, 2010).

A distressing realization has recently tempered the celebration of improved research design in empirical social science. The freedom of a researcher to choose amongst numerous specifications, variables, and sub-samples within a single study can lead to spurious results even when the method meets rigorous standards. This practice goes by many names: massaging the data, specification search, significance chasing, selective reporting. Following Simonsohn, Nelson, and Simmons (2014), the term p-hacking is used here to define attempts to generate statistically significant results from a distribution with a true null hypothesis. The following hypothetical examples illustrate, but are not exhaustive of, p-hacking: analyzing sub-groups (race, SES quintiles, gender), multiple dependent variables, or multiple estimators until a significant treatment effect is found. The operative phrase here is "until a significant treatment effect is found" as it underscores that p-hacking occurs when a researcher mines the data for significant effects, stops when they are found, and does not incorporate the full search procedure into p-value calculations. Doing so will cause p-values to cluster around significance thresholds, usually assumed to

[3] The Cochrane Handbook does not recommend Rosenthal's particular solution of calculating a fail-safe N (2008, pp. 321–322), yet opens the section on publication bias with reference to Rosenthal's concern and suggests many alternatives to address it (2008, Chapter 10).

be .05 or .1. The clustering occurs because it is easier by chance to find a p-value that crosses .05 (or .1) than one that crosses more stringent significance thresholds such as .01. In the case of a single draw from a distribution with a true null hypothesis, a researcher is five times more likely to find a p-value between .05 and 0 than a p-value between .01 and 0. Figure 4 illustrates the relative ease of crossing the .05 vs. .01 thresholds at least twice when testing multiple hypotheses. In 40 trials (hypothesis tests), a researcher is ten times as likely to find at least two p-values that cross .05 as he is to find at least two p-values that cross .01.

A number of articles provide evidence of clustered p-values in prominent sociology, political science, psychology, and economic journals. Gerber and Malhotra (2008a) find a three-fold jump in the number of p-values just under .05 in statistics taken from the *American Political Science Review* and the *American Journal of Political Science*. In a separate analysis of statistics taken from the *American Journal of Sociology* and the *Sociological Quarterly*, they find a two-fold jump in p-values that just cross the .05 threshold (2008b). Brodeur, Le, Sangnier, and Zylberberg (2013) find similar clustering around the .05 and .1 levels of significance in statistics taken from the *Quarterly Journal of Economics*, the *American Economic Review*, and the *Journal of Political Economy*. Psychology journals display the same general pattern, with "excess" p-values just below the .05 threshold in the *Journal of Experimental Psychology*, the *Journal of Personality and Social Psychology*, and *Psychological Science* (Masicampo & Lalande, 2012).

Taken together, these four analyses provide evidence of publication bias in leading social science journals across several disciplines. However, as the studies' authors acknowledge, it is difficult to distinguish amongst various forms of publication bias with the strategies employed therein. Each article analyzes the distribution of test statistics across the .05 and .1 significance thresholds, comparing the relative representation of statistics close to either side of each boundary. An overrepresentation of test statistics on the significant side of the boundary is taken as evidence of publication bias. If journal

editors favor articles that reject a null hypothesis, then one would find this empirical pattern whether or not the reported effects are real or artifacts of p-hacking. The consequences for the veracity of policy research are drastically different depending on which behavior is leading to the overrepresentation of statistically significant results. If all effects are real, journal editorial policies favoring null hypothesis rejection would censor true null results but still allow real effects into the literature. The research community would fail to learn about null effects, but would learn about true effects. However, if the effects reported are spurious artifacts of p-hacking or sampling variance, then the research community fails to learn anything, regardless of journal editors' preference for rejected null hypotheses.

An additional limitation of the previous work is that the authors included numerous p-values from each article they analyzed, leading to potential clustering of statistics at the study level and in some cases the inclusion of p-values from statistical tests that would not likely have been subjected to p-hacking. For example, though Brodeur et al. (2013) took pains to include only p-values from the main causal analyses in economics journals, they collected an average of over 50 p-values per article. Their dataset includes p-values from robustness checks, naïve OLS regressions in advance of more sophisticated designs, and the like. These analyses are not likely to suffer from p-hacking and are not the primary results of the articles. Including such p-values dilutes the proper analysis sample and will underrepresent p-hacking, as has recently been confirmed with the original analysis sample from the article that introduced the p-curve framework (Simonsohn, 2015).

III. P-curve: Assessing the Evidential Value in a Body of Literature

The introduction of p-curve by Simonsohn, Nelson, and Simmons (2014) has made it possible to distinguish true effects from those that result specifically from p-hacking or sampling variance using only the observed results that cross a particular significance level. The null hypothesis begins with 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[4]: 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 framework 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 3 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.

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

[4] This is true asymptotically.

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 $\chi^2(2k)$ under the null hypothesis that the pp-values are uniformly distributed (corresponding to no evidential value) (Fisher, 1932).

IV. Study Design

The entire study design and analytic structure has been articulated in a pre-analysis plan (PAP), which was published prior to the data collection (Tanner, 2015).[5] Several analytic decisions were altered after the PAP was published and are detailed in a deviation report. Such deviations are also denoted with "^" and an accompanying footnote in this analysis. The PAP, deviation report, data, and analysis code are all available on the author's website.[6]

Because of the well-known limitations of Fisher's method (Rosenthal, 1978, pp. 187–188)[7], the Kolmogorov-Smirnov (KS) test is the primary estimation technique used in this analysis to measure goodness of fit for a uniform distribution of p-values.[^] The KS test statistic is the maximum distance between the cumulative empirical distribution and the cumulative density of the given null distribution, which in this case is a discretized uniform distribution on the [0, .1] interval (Massey, Jr., 1951). In simulations described in appendix C, the KS test had approximately 80% power to reject a discretized uniform null with a sample of 60 studies powered at 33%. A second estimation strategy is also employed to ensure results are not local to the KS test. A simple method of estimating evidential value

[5] Note that the PAP was submitted in October, 2014.
[6] The Open Science Framework web address for this author is https://osf.io/mn8gr/
[7] As of this writing, the original p-curve authors no longer use Fisher's method as their preferred technique, but instead use Stouffer's method (see note under "Computer Programs" here: http://www.p-curve.com/Supplement/ as well as page 22 of http://p-curve.com/Supplement/Supplement_pcurve1.pdf). They note that the technique employed here, the Kolmogorov-Smirnov test, should be considered as well in larger samples of p-values than they typically use (≈20).
[^] The PAP specified Fisher's method in the author's own code and the p-curve web app 2.0.

is 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, yet it ignores variation of p-values within the high and low categories. This simple method is used as a robustness check on the primary analysis. To further ensure that the results are not local to the KS test, a subset of the data is analyzed using Stouffer's method, the estimation technique currently favored by Simonsohn, Nelson, and Simmons (2014). These results use only p-values that are less than or equal to .05[8] and are contained in appendix B. They do not meaningfully diverge from the results obtained from the KS tests.[9]

In testing multiple hypotheses, individual p-values from each KS and binomial 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) is employed here to further guard against spurious results.

V. Data Collection

This analysis evaluates 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-

[8] The precise statistical code that will reproduce Simonsohn, Nelson, and Simmons (2014) analyses exactly is not currently available for Stata and so this robustness test used the p-curve web application (version 3.01) which only admits p-values less than or equal to .05. Attempts are being made to convert the code and allow for alternate significance thresholds. When available, new results using this code will be made available on Open Science Framework website for this project: https://osf.io/9tbm8/.

[9] Stouffer's method failed to reject uniformity for the 12 p-values from moderate rating categories in the clearinghouses. The graphical evidence suggests that, given a larger sample size, the distribution of p-values would be detectably non-uniform.

related analyses. If scholarship in these sources does not contain evidential value, then the credibility of applied social science research can be seriously questioned.

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, a pre-specified data collection and estimation strategy are included in the published PAP.

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.[10] 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 include estimation of causal parameters in real data. Purely descriptive papers and modeling exercises were not included.

The studies taken into the sample were:

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*).[^]

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/

[10] The shape of p-curve relies on the assumption that all p-values are drawn from independent tests.
[^] The journal data were collected on March 9, 2015. *JPAM* was oversampled (31 articles) in order to span the editorial leadership change that began with Volume 33, Issue 4 in the fall of 2014.

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[11]

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 – they were created to serve as 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 met 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

[11] The clearinghouse data were collected on December 18, 2014. Since then, new studies have been added to both clearinghouses.

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).

The p-values analyzed in these studies are those from the tests of the main hypotheses. All such p-values that cross the .1 significance level were analyzed.[12] 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 highly 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 et al., 2013, p. 29).

The determination of "main" can be ambiguous when authors present multiple model specifications, analyze multiple dependent variables, and assess heterogeneity in treatment effects. This analysis used p-values that were mentioned in the abstract or introduction section and from specifications favored by the authors. When this did not resolve ambiguity, a single p-value was drawn at random from the list of candidates for "main" hypothesis.[13] The median number of hypotheses in each article is 11 for the journal frame and 3 for the clearinghouse frame. Table 1 includes descriptive statistics for the analysis sample.

[12] The set of robustness checks using Stouffer's method only use p-values of .05 or less and the results do not markedly differ (see footnote 9).

[13] The precise decision rules for inclusion are contained in the pre-analysis plan. The data is available online and includes how many "main" p-values were reported as well as the order that the randomly chosen p-value appeared. In hindsight, each "main" p-value should have been recorded and the analysis should have been bootstrapped to estimate variance due to which p-values were chosen.

Exact p-values, test statistics, standard errors, and degrees of freedom are not uniformly reported in this literature. Comparing p-values in the face of this variation in reporting required the use of rounding.^ Most often, p-values were reported to two decimal places. If possible, exact p-values were calculated using the test statistics, standard errors, and degrees of freedom in the relevant tables. These precise p-values were rounded to two decimal places to facilitate comparison with the format in which the majority were reported in the original studies. Similarly, it was often unclear whether tests were discrete or continuous. Given the large sample sizes in this literature, a decision was made to treat all tests as continuous.^

Table 1: Descriptive Statistics for the Analysis Sample

	Journals	**Clearinghouses**
Total # of Articles	71	75
# of Articles Reporting Effects	52	68
Statistics for those Reporting Effects		
# of Main Hypotheses- Mean	15	10
(Med)	(11)	(3)
# Reporting Precise P-values	49	51
RCTs	11	49
Sample Size Mean (Med)	304,180 (8,523)	88,484 (1,160)
Published in Journal	75[a]	45
Publishing Date Range (Med)	2014-2015 (2014)	1979-2014 (2011)

^ The PAP made no reference to rounding.

^ The PAP specified demarcating continuous and discrete tests, with separate p-curves for continuous tests and all tests.

[a] The clearinghouse sampling frame included articles published in journals, leading to more journal articles than were collected in the journal sampling frame.

VI. Results

The distribution of p-values in the sample provides strong graphical evidence that the reported effects are not artifacts of selective reporting or random sampling variance. Figure 1 shows the histogram for the sample as a whole. The bins reveal a strong right skew and thus high evidential value. There are far more p-values in the neighborhood of .01 than .05 or .1. The dashed red line is how the distribution would appear if the effects contained in the articles were purely artifacts of random sampling variance: an even distribution with 10% in each bin. It is visually clear that the distribution of p-values is neither uniform nor skewed left, which formal hypothesis tests confirm. Both the Kolmogorov-Smirnov (KS) and binomial tests reject uniformity with $p<.001$.

Figure 1: Strong Evidence in Policy Research

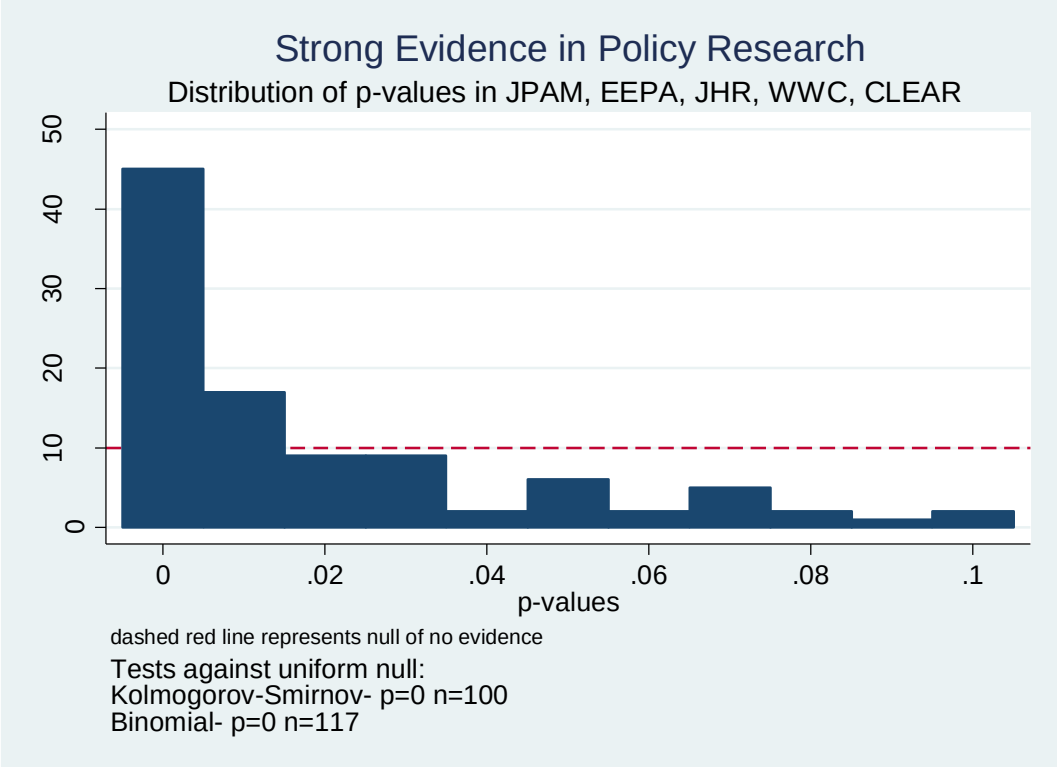
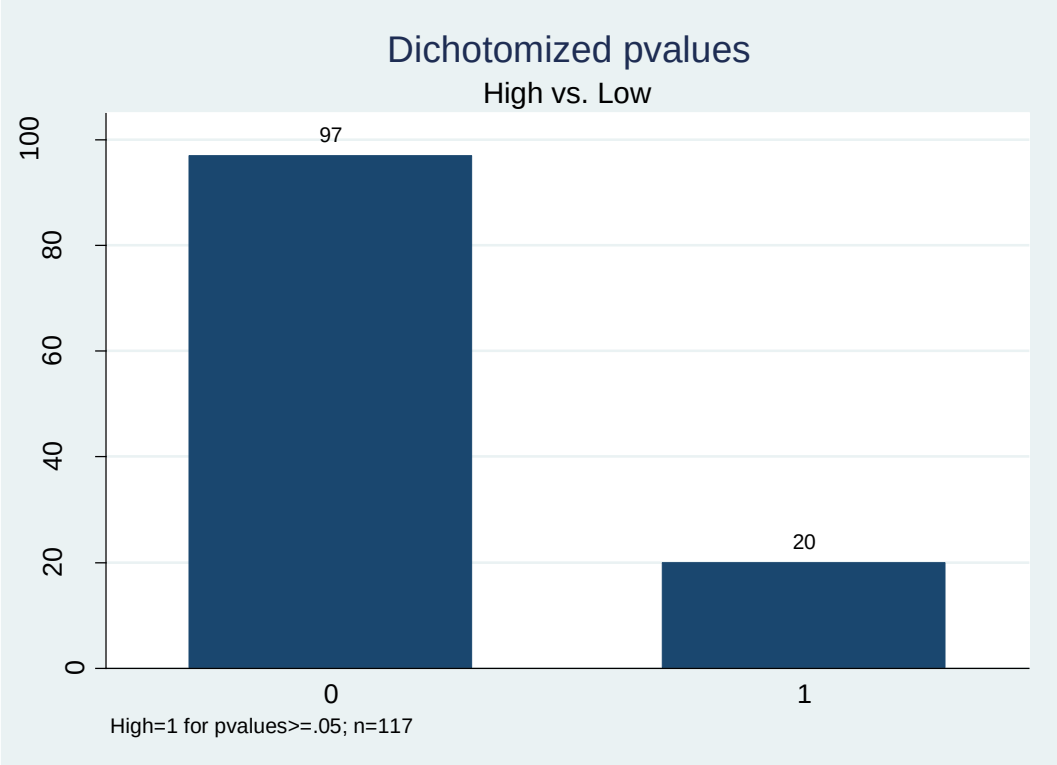


Figure 2: Robust Binomial Method Reveals Strong Evidence



These graphical patterns are broadly similar across all pre-specified subdivisions of the data: RCT vs. non-RCT, high vs moderate evidence ratings, clearinghouses vs. policy journals. Figure 5 contains the histograms for these divisions of the data. As is visually clear, the presence of p-hacking (left-skew) can be readily rejected with both KS and binomial tests. Table 2 contains the p-values for the tests against a uniform null for all groupings of the data as well as the sharpened q-values for each test after adjusting the entire set of tests[14] for the false discovery rate (Anderson, 2008; Benjamini & Yekutieli, 2001). The highest (least significant) sharpened q-value is .0256; most are less than .0001, providing compelling evidence that the results contained in this sample and all subdivision of it have "evidential value." More precisely, p-hacking can be ruled out as the sole explanation for the findings in this sample.

Table 2: P-values and sharpened q-values for the primary hypothesis tests.

Data	KS p-value	Binomial p-value	KS q-value	Binomial q-value
Entire Sample	<.0001	<.0001	<.0001	<.0001
Journals	<.0001	<.0001	<.0001	.0004
Clearinghouses	<.0001	<.0001	<.0001	<.0001
RCTs	<.0001	<.0001	<.0001	<.0001
Non-RCTs	<.0001	<.0001	<.0001	<.0001
Rating: Strong	.0004	<.0001	.002	<.0001
Rating: Moderate	.0128	.0015	.0256	.0046

[14] The sharpened q-values were calculated using all 28 p-values in the results section of the article, even the subsequent "worst case scenario" tests.

These results are reassuring for the validity of the research contained in this sample, yet several limitations are worth mentioning. 16% of studies (20) fail to report sufficient information to obtain p-values directly. 17 of those 20 report significance levels. These studies are included in the following robustness check wherein those reporting only significance levels are given the p-value of the significance level (10%=.1, 5%=.05, 1%=.01) and the three studies without readily available significance information were given a p-value of .1. This procedure will overstate the amount of clustering at significance thresholds as it assigns the right-most (highest) possible p-value to each missing entry- the worst-case p-hacking scenario. Figure 6 contains the histograms from this expanded set of p-values. By design, these graphs reveal slightly more clustering around common significance thresholds. However, they do not alter the essential finding that this group of policy research is not overly plagued by either accidental false positives or p-hacking. Most of the distributions can be distinguished from uniform with $p<.001$ even in this worst-case analysis.

The second limitation is that the sample may not be reflective of work in the larger population of policy journals and research clearinghouses. As of this writing, there are more than 20 government-supported evidence reviews that could broadly be classified as research clearinghouses (Neuhoff, Axworthy, Glazer, & Berfond, 2015, pp. 51–53). However, many of these are specific to a narrow policy issue or location such as the Best Practices Registry for Suicide Prevention[15] and the California Healthy Kids Resource Center.[16] The two used in this analysis, WWC and CLEAR, are national in scope and topically broad. WWC in particular is the eldest and most established of its peers and should provide a representative case of a broad, developed, national clearinghouse. The policy journals were also chosen as representative of rigorous policy research, yet that definition is highly subjective and excludes many policy-related articles that appear in discipline-specific social science journals. It is probable that a

[15] http://www.sprc.org/bpr
[16] http://www.californiahealthykids.org/index

different set of journals would contain a distribution of p-values with less right-skew (less credible evidence).

The third limitation of this article is that the sampling procedure collected only a single p-value from each article. This is a limitation of the p-curve framework in general, as analyzed p-values must be independent. This condition would almost certainly be violated if multiple p-values from a single article were included. The constraint of including only a single p-value from each article requires judicious decisions about what constitutes a primary hypothesis in addition to requiring a random selection in the face of multiple hypotheses that fit the definition. It could be the case that the decision rules employed here are too broad and that the p-hacking behavior in this literature is focused only on a sub-set of the tests that have been labeled as primary. If so, these results will understate the extent of p-hacking. This is a current tension in the literature on publication bias.

Finally, it should be stressed that the analysis is not a test of other forms of fraud or sophisticated p-hacking routines. If researchers purposefully altered their data, reported coefficients, or standard errors, then this analysis is based on fraudulent data and says only that the fraudulent data did not produce clustered test statistics.[17] Moreover, if researchers were p-hacking to cross .01 rather than .05 or .1, then this analysis will understate the extent of p-hacking.

VII. Conclusion

This article finds encouragingly little evidence of false-positives or p-hacking in rigorous policy research. The findings are somewhat surprising given that prior work in several prestigious scientific disciplines has suggested publication bias is a large problem. Unlike the prior work, this analysis uses a method that can disentangle p-hacking from accidents of random sampling. It also analyzes a set of p-values in each article that are more likely to have been p-hacked. Leaving aside the issue of extrapolation to a

[17] For a recent example of fraudulent research detected due to statistical abnormalities, see Simonsohn (2013).

more general body of policy scholarship, these results are a welcome sign of credibility in light of recent scandals (Bhattacharjee, 2013; Carey, 2015; Godoy, 2015).

The analysis opens up two main avenues for further exploration: one methodological and the other substantive. Methodologically, it is crucial for the p-curve framework to be able to analyze all substantively import p-values in an article, dependent or not. Research and publishing norms for laboratory experiments in psychology might be amenable to methods that require a single p-value from each article, but economics, political science, sociology, and public policy research routinely has more than ten main hypothesis tests in a single article.[18] One potential solution is to weight articles by the number of p-values they contribute, so that five legitimate p-values from a single paper don't mask five p-hacked p-values from five separate papers. The behavior of such reweighted p-curves with and without p-hacking is not currently known, yet understanding this behavior is critical to the evolution of research on publication bias. At minimum, future p-curve analyses should capture all relevant p-values from each paper and bootstrap the p-curves to estimate their variance in the face of multiple potential p-values from each paper.

Substantively, it is important to understand whether and why the policy research used in this analysis differs from research contained in the prestigious journals analyzed in the prior work. This is best addressed with a two-fold approach. First, the estimation procedures used in this analysis should be applied to a subset of the p-values collected in the prior research from economics, political science, sociology, and psychology journals. A second and ideal step would be to prospectively collect p-values from a large body of social science research across multiple disciplines and analyze the data with p-curves.

[18] This relatively simple analysis tested 7 primary hypotheses.

References

Anderson, M. L. (2008). Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, *103*(484), 1481–1495.

Angrist, J. D., & Pischke, J. (2010). The Credibility Revolution in Empirical Economics : How Better Research Design is Taking the Con out of Econometrics. *Journal of Economic Perspectives*, *24*(2), 3–30.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Royal Statistical Society. Series B (Methodological), 57*(1), 289–300.

Benjamini, Y., & Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics*, *29*(4), 1165–1188.

Bhattacharjee, Y. (2013, April 28). The Mind of a Con Man. *The New York Times*, p. MM44. New York. Retrieved from http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html?pagewanted=all

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

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

Carey, B. (2015, May 29). Study on Attitudes Toward Same-Sex Marriage Is Retracted by a Scientific Journal. *The New York Times*, p. A16. New York. Retrieved from http://nyti.ms/1FFQkzj

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

Fisher, R. A. (1932). *Statistical Methods for Research Workers* (4th ed.). Edinburgh: Oliver and Boyd.

Gerber, A., & Malhotra, N. (2008a). Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals. *Quarterly Journal of Political Science*, *3*(3), 313–326.

Gerber, A., & Malhotra, N. (2008b). Publication Bias in Empirical Sociological Research. *Sociological Methods & Research*, *37*(1), 3 –30.

Godoy, M. (2015). Why A Journalist Scammed The Media Into Spreading Bad Chocolate Science. Retrieved from http://www.npr.org/sections/thesalt/2015/05/28/410313446/why-a-journalist-scammed-the-media-into-spreading-bad-chocolate-science

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

Haskins, R., & Margolis, G. (2015). *Show Me the Evidence: Obama's Fight for Rigor and Results in Social Policy*. Washington, D.C.: Brookings Institution Press.

Higgins, J. P. T., & Green, S. (Eds.). (2008). *Cochrane Handbook for Systematic Reviews of Interventions*. Hoboken, New Jersey: Wiley-Blackwell.

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), 191–217.

Hoxby, C. (2007). Does Competition Among Public Schools Benefit Students and Taxpayers ? Reply. *American Economic Review*, *97*(5), 2038–2055.

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.). 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), 658–698.

Leamer, E. E. (1983). Let's Take the Con Out of Econometrics. *American Economic Review*, *73*(1), 31–43.

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *The Quarterly Journal of Experimental Psychology*, *65*(11), 2271–2279.

Massey, Jr., F. J. (1951). Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, *46*(253), 68– 78.

Neuhoff, B. A., Axworthy, S., Glazer, S., & Berfond, D. (2015). The What Works Marketplace Smarter Choices, (April).

Reuter, P., & Smith-Ready, J. (2002). AssessingJPAM after 20 Years. *Journal of Policy Analysis and Management*, *21*(3), 339–353.

Rosenthal, R. (1978). Combining Results of Independent Studies. *Psychological Bulletin*, *85*(1), 185–193.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641.

Rothstein, J. (2007). Does Competition Public Schools Benefit Students Among and Taxpayers? Comment. *American Economic Review*, *97*(5), 2026–2037.

Simonsohn, U. (2013). Just post it: the lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, *24*(10), 1875–88.

Simonsohn, U. (2015). Falsely Reassuring: Analyses of ALL p-values. Retrieved from http://datacolada.org/2015/08/24/41-falsely-reassuring-analyses-of-all-p-values-2/

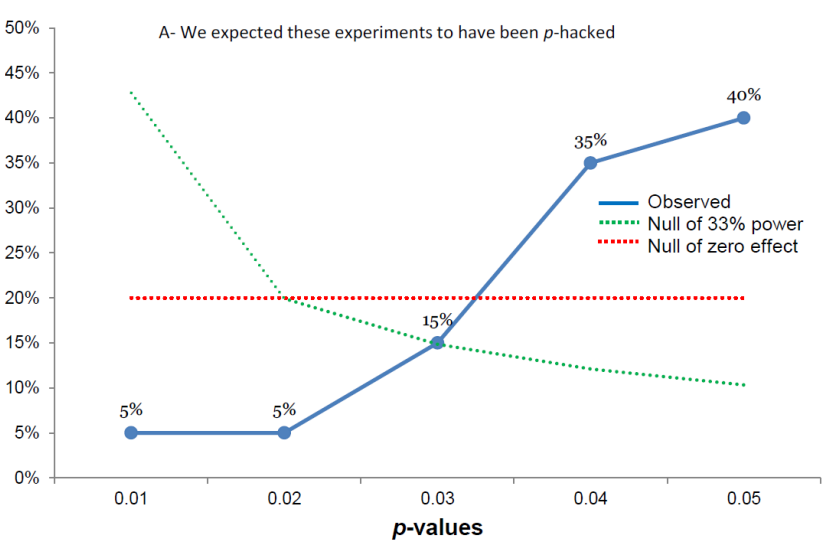
Simonsohn, U., & Nelson, L. D. (2014). Anchoring is Not a False-Positive : Maniadis , Tufano , and List ' s (2014) " Failure -to- Replicate " is Actually Entirely Consistent with the Original, 1–12.

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.

Tanner, S. (2015). False Positives and Selective Reporting in Policy. *Observational Studies*, *1*(1), 18–29.

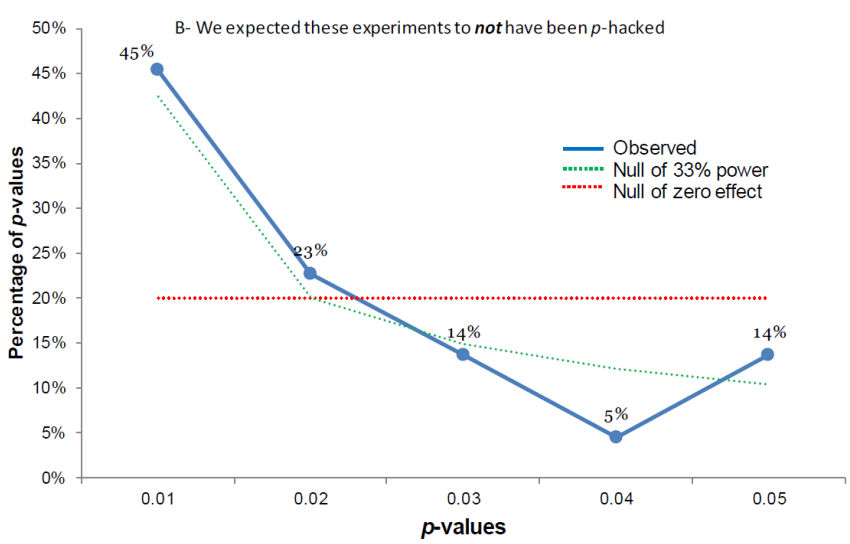
Appendix A: Supplementary Figures

Figure 3: 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 4 Crossing .05 is Easier than Crossing .01

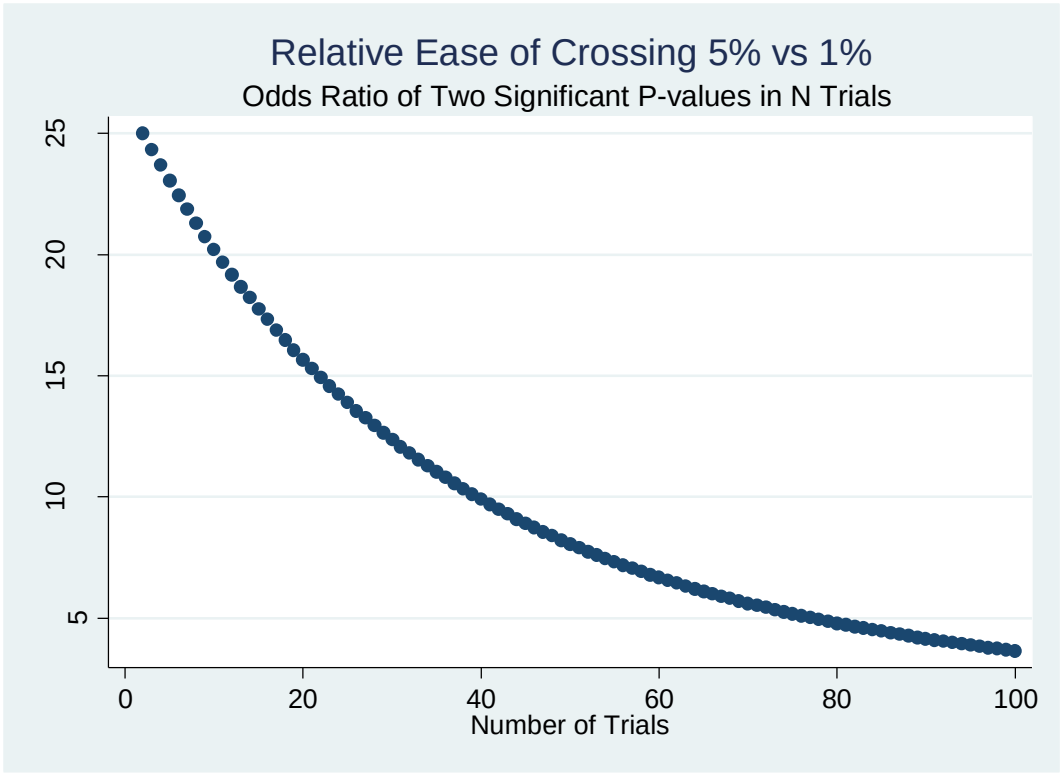


Figure 5: Strong Evidence across Multiple Partitions of the Data

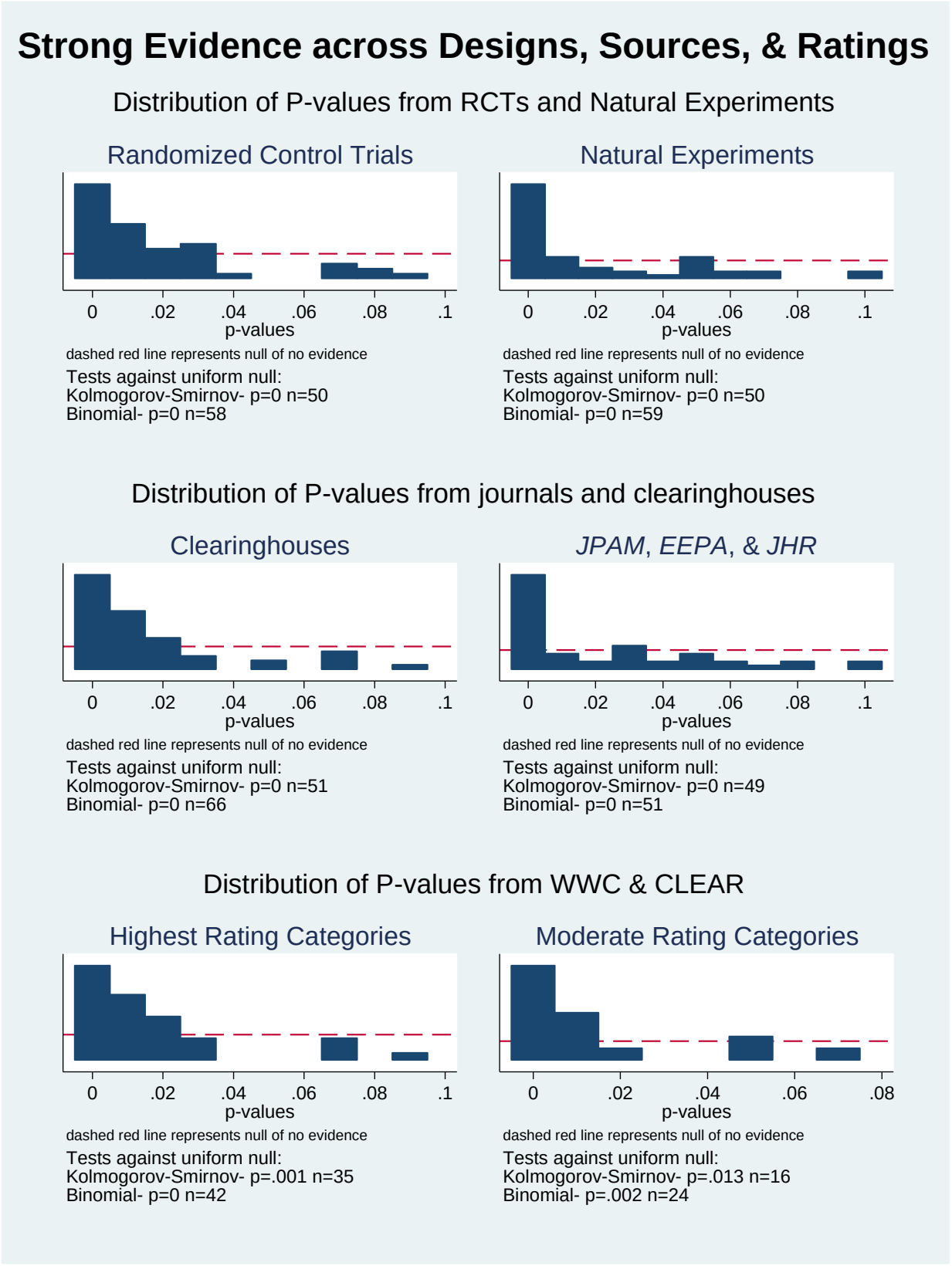
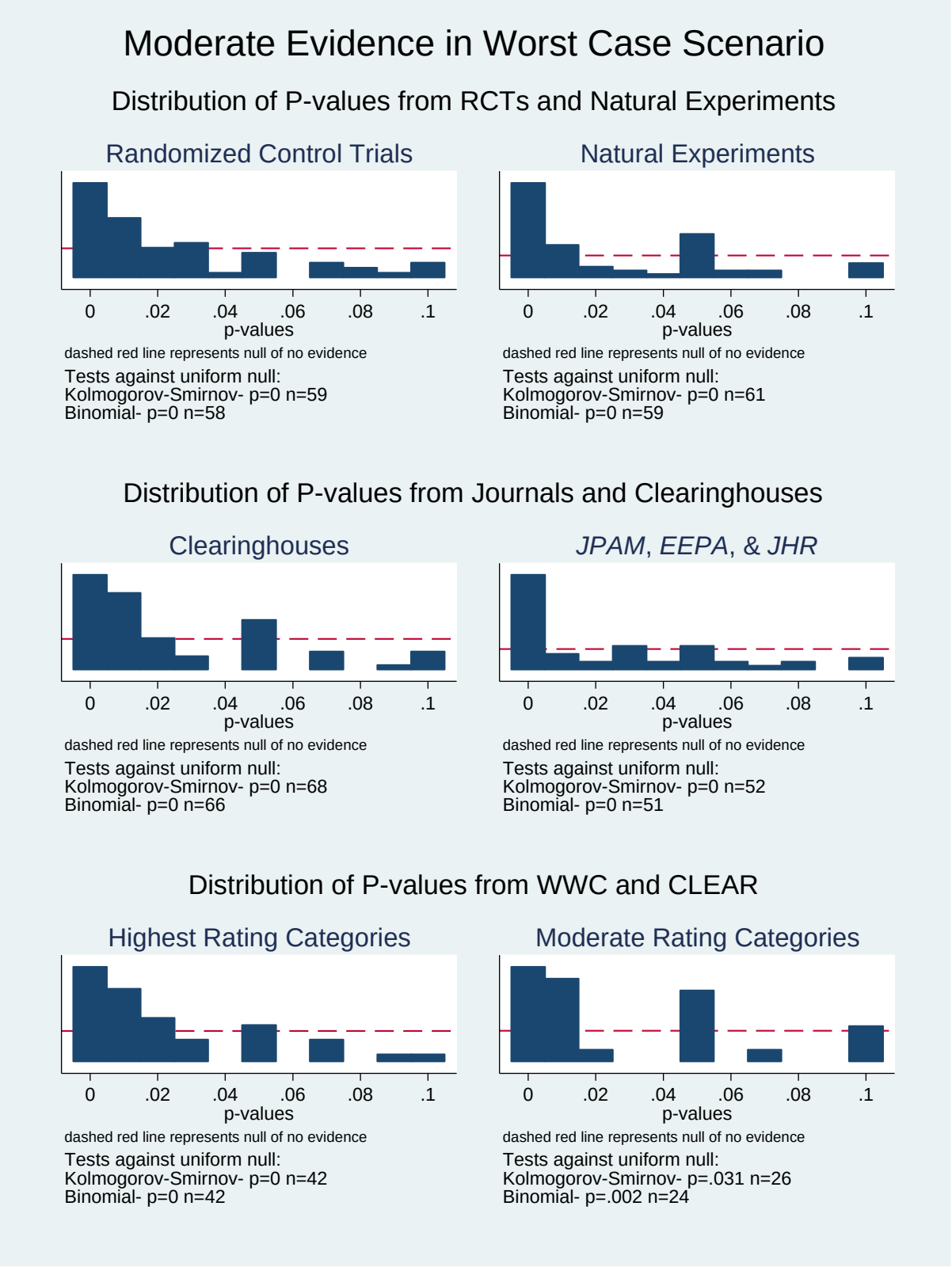
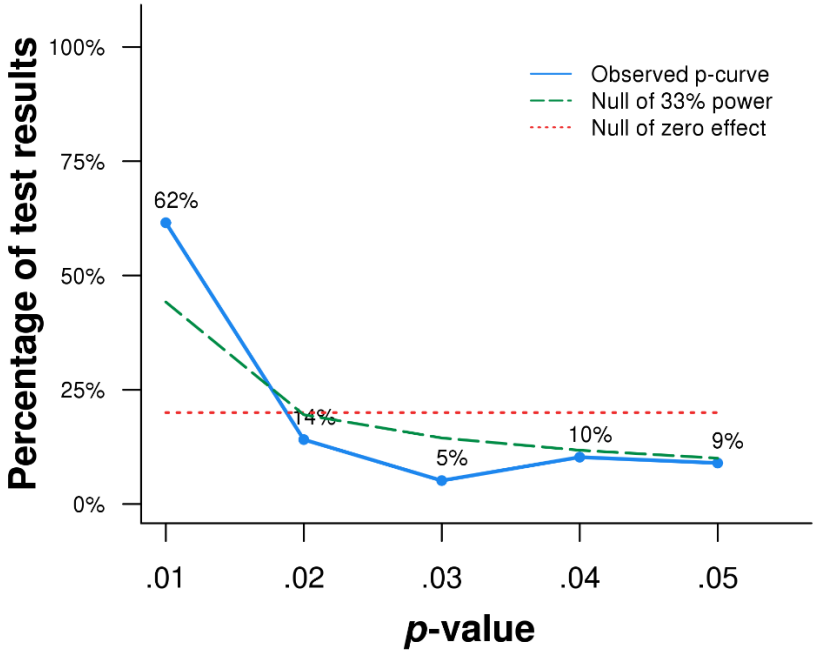


Figure 6: Moderate Evidence with Missing P-values Given Highest Values



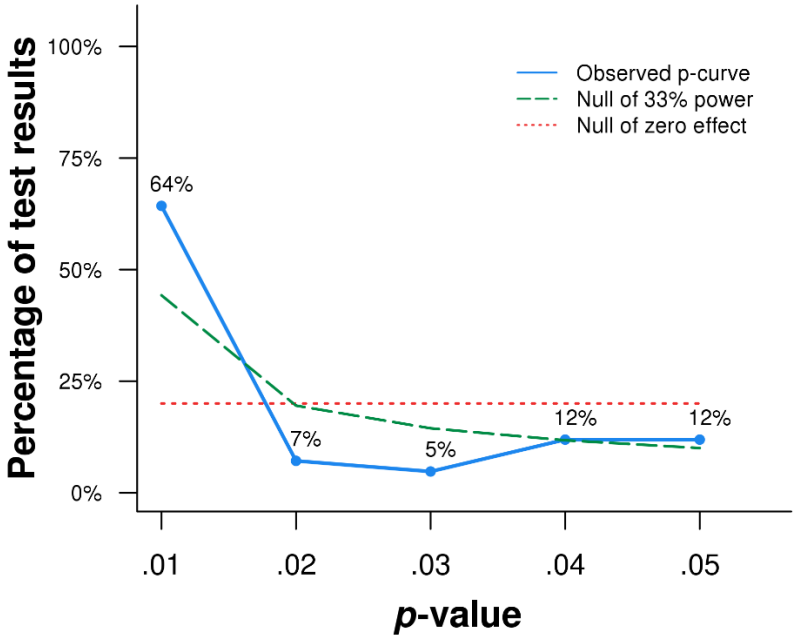
Appendix B: Results from P-Curve Web App 3.01

Figure 7: Results from P-Curve App 3.01 for all p-values <.05



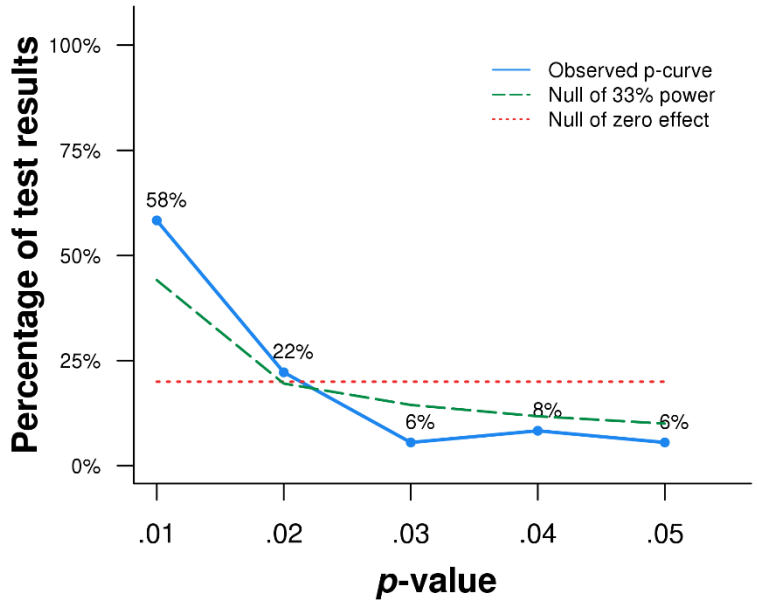
Note: Stouffer's method rejects uniformity with p<.0001 (P-Curve App 3.01)

Figure 8: Results from P-Curve App 3.01 for JPAM, EEPA, & JHR p-values <.05



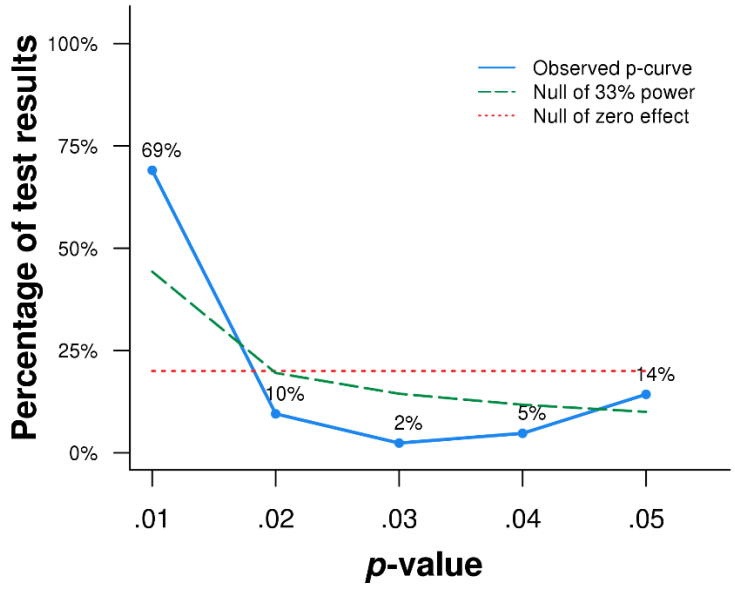
Note: Stouffer's method rejects uniformity with p=.0117 (P-Curve App 3.01)

Figure 9: Results from P-Curve App 3.01 for WWC & CLEAR p-values <.05



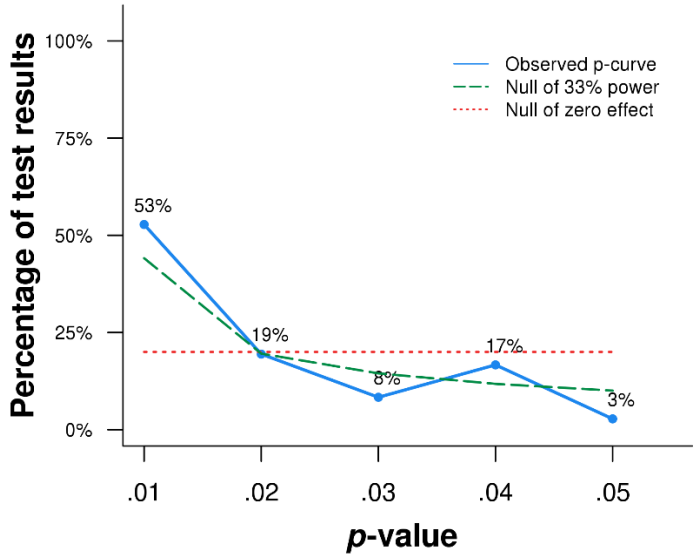
Note: Stouffer's method rejects uniformity with p<.0001 (P-Curve App 3.01)

Figure 10: Results from P-Curve App 3.01 for Natural Experiment p-values <.05



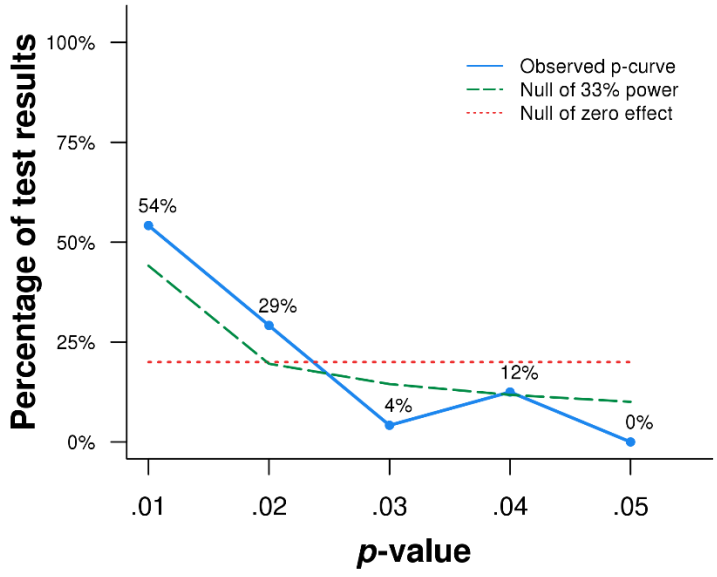
Note: Stouffer's method rejects uniformity with p=.072 (P-Curve App 3.01)

Figure 11: Results from P-Curve App 3.01 for RCT p-values <.05



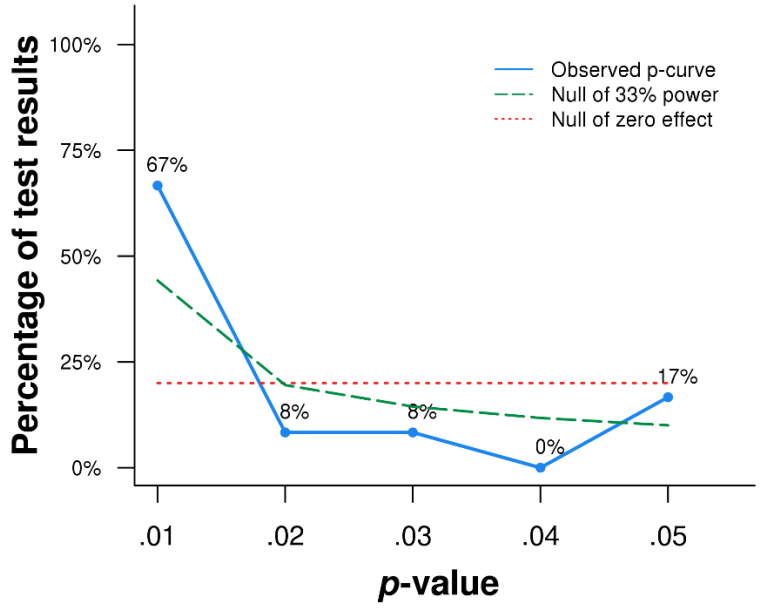
Note: Stouffer's method rejects uniformity with p<.0001 (P-Curve App 3.01)

Figure 12: Results from P-Curve App 3.01 for High Rating p-values <.05



Note: Stouffer's method rejects uniformity with p<.0001 (P-Curve App 3.01)

Figure 13: Results from P-Curve App 3.01 for Moderate Rating p-values <.05



Note: Stouffer's method fails to reject uniformity, p=.3195 (P-Curve App 3.01)

Appendix C: Simulation Description & Results

The simulations were conducted to assess the relative power of the KS test and binomial test to reject uniformity in a set of p-values drawn from studies powered at 33%. The simulation code is available at https://osf.io/pw3vr/. The **power** command in Stata 14.0 was used to generate the effects size (difference of means) that would allow for 33% power of a t test for difference of means with a 10% significance level in studies with sample size 2362- the median sample size in the studies contained in this analysis. This effect size, .0493, was set as the mean of the distribution from which the treatment group was drawn, the control group was drawn from a distribution with mean zero, and both distributions were given unit variance. For each simulated KS and binomial test, 500 "studies" were conducted by drawing 1181 observations at random from the treatment distribution, ~N (.0493, 1), and the control distribution, ~N (0, 1). For each study, the p-value from a t test for the difference of means test was recorded and rounded to two decimal places (0, .01, .02…). After 500 studies, only j significant p-values were kept, with j varying from 20 to 80 in steps of 5 (25, 30, 35…). A KS test was performed on the retained, rounded p-values. A binomial test was performed on the retained,

rounded p-values after splitting them into high (≥.05) and low (<.05) categories. The p-values from the KS and binomial tests were then recorded. This entire process was conducted 500 times for each j set of p-values, for a total of 6,500 KS and binomial tests. The results are contained in figure 14 below, which uses a locally weighted regression of significance (0 if p-value > .1, 1 otherwise) on number of studies. As can be seen, the binomial test has greater power than the KS test, though the KS test achieves 80% power near 60 studies.

Figure 14: The Binomial Test has Greater Power than the KS Test.