Journal of Behavioral
Public Administration
Vol 4(2), pp. 1-17
DOI: 10.30636/jbpa.41.239

Research Article

Everything hacked? What is the evidential value of the experimental public administration literature?

Dominik Vogel[*], Chengxin Xu[†]

Abstract: The rise of behavioral public administration provides new perspectives – especially from a psychological point of view – to understand public administration theories and the growing interest in using experiments to enhance the internal validity of empirical studies. However, psychology and other social sciences are undergoing a replication crisis where experimental results often do not replicate. One reason for the limited replicability is the publication bias sparked by journals' preference for significant effects and the resulting incentive to create significant results. This study employs a meta-analytical approach to examine the evidential value of experimental evidence in public administration. It uses the p-curve method to test whether this body of research is dominated by selectively reporting significant results, e.g., by omitting insignificant results or engaging in practices to obtain significant results (i.e. p-hacking). The analysis includes 172 statistically significant findings published in top public administration journals and shows that the distribution of p values of these findings is right-skewed. Such a distribution indicates that the experimental public administration research contains evidential value, which means it is not solely the result of selective reporting of significant results. It therefore is unlikely that the field is dominated by research that is based on null effects that have been hacked to reach significance or where underpowered studies on null effects have been selectively reported based on their significance. Although the analysis shows a good sign, we discuss important practices to further strengthen the validity and reliability of experimental methods in public administration.

Keywords: *P*-hacking, Publication bias, Evidential value, Questionable research practices, Experiments

Supplements: Open data, Open materials, Preregistered

The rise of behavioral public administration (Jilke, Meier, & Van, 2018) sparked public administration scholars' interest in the theory and methods used in psychology. Among other things, this fostered the use of experimental methods in public administration research, which rapidly became widely used. The advantages of an experimental approach to public administration, especially the ability to make causal claims, have been widely discussed. Also, critics raised their concerns about the limitations of experimental research. The critics, for example, argue that especially vignette and lab experiments have limited external validity, that they are not driven by theory, or that they fail to consider contextual factors such as politics and institutions (Bertelli & Riccucci, 2020; Hassan & Wright, 2020).

However, one potential threat to the value of experiments in public administration has been seldom discussed so far (for an exception, see Hassan & Wright, 2020); this is not just adapting the experimental method from psychology but also transferring the so-called replication crisis to public administration. The replication or credibility crisis in psychology was sparked by the findings of the Open Science Collaboration (2015). The

* University of Hamburg, Germany
† Seattle University, USA
Address correspondence to Dominik Vogel at (dominik.vogel-2@uni-hamburg.de.)

project attempted to replicate 100 findings published in three major psychology journals. Surprisingly, only 39 of the 100 original results were replicated. In the aftermath of this finding, psychologists and other social scientists cast serious doubts about the major findings in psychology and other disciplines. Additional large-scale replication projects sparked skepticism when central predictions by prominent theories such as ego depletion effect (Hagger et al., 2016), power posing (Davis, Papini, Rosenfield, Roelofs, Kolb, Powers, & Smits, 2017), or terror management theory (Klein et al., 2019) could not be replicated.

Scholars identified multiple mechanisms explaining how the results that were backed by dozens and dozens of significant findings did not hold when large-scale replication projects tried to confirm them. The most important ones are publication bias and *p*-hacking. Publication bias describes the common practices that journals prefer publishing significant effects and novel findings, while rejecting insignificant findings and replication attempts (Simmons, Nelson, & Simonsohn, 2011). As researchers want and need to publish, publication bias creates an incentive to produce statistically significant and novel findings. Too often, these incentives lead to researchers engaging in questionable research practices, especially in *p*-hacking, which describes practices intended to turn insignificant into significant results. This includes conducting underpowered studies, collecting additional data until results are significant, excluding observations, conducting additional analyses, and many more (Earp & Trafimow, 2015).

Publication bias and *p*-hacking together can create literature that overwhelmingly consists of significant effects but does not build on a true effect. Instead of falsely concluding that there is an effect when actually there is none 5% of the time (type I error rate), the amount of false-positive findings is much higher (Simmons et al., 2011).

Experimental public administration research might be less prone to suffering from low replicability since many researchers quickly adopted the practices that were developed to increase replicability, such as preregistration and publishing code and data (Munafò, Nosek, Bishop, Button, Chambers, Percie du Sert, Simon-sohn, Wagenmakers, Ware, & Ioan-nidis, 2017). Various scholars within public administration emphasized the importance of replicability and replications. Perry (2017), for example, discusses his decision to sign up *Public Administration Review* for the transparency and openness guidelines (TOP) and encouraged replications. Walker et al. (2017) introduce a special issue on replications and discuss the state of replications in public administration. They conclude that "replication is an essential part of the scientific process that can help produce a sounder knowledge base for our field" (Walker et al., 2017, p. 1231). They extended their effort by providing a best practice approach to replications in public administration (Walker, Brewer, Lee, Petrovsky, & van Witteloostuijn, 2019). Finally, Zhu et al. (2019) discuss several ways public administration can address a potential "methodology crisis" (p. 296) and emphasize that public administration researchers can learn from the replication crisis.

Such attention to the practices and mistakes that lead to low replicability resulted in an environment where many researchers (and reviewers) are also aware of and try to avoid them. However, there are limits to that too. Although some public administration journals now encourage (e.g., *Public Administration Review*) or even require (*Journal of Public Administration Research and Theory*) the publication of data and code, such practices have been rarely applied in the past. Furthermore, preregistration is not widely used, no journal offers registered reports (Chambers, 2019; Nosek & Lakens, 2014), and for a long time, publication of null findings and replications was discouraged.

Therefore, on the one hand, it is vital to insist on open science practices and, on the other hand, monitor the development of public administration research. Such practices are especially important in the case of experimental research since it is specifically prone to questionable research practices. One way to do so is by conducting and publishing replication studies (Walker et al., 2019). However, it is impossible to replicate all public administration findings. Psychologists and statisticians, therefore, suggested changes to the ways we draw inference from data, for example, by requesting that the common significance threshold should be lowered to .005 (Benjamin et al., 2018) or that researchers should justify their significance criteria (Lakens et al., 2018). They also developed methods to assess the credibility of published research without replicating it. The most prominent of these methods is the *p*-curve method (Simonsohn et al., 2014, 2015), which uses published studies' *p* values to assess the literature's credibility.

This article uses the *p*-curve method to test the evidential value of the experimental public administration literature. A body of research contains evidential value "if it is not solely the result of *p*-hacking and selective reporting of significant effects" (Vogel & Homberg, 2020, p. 2). Analyzing 113 published studies, we find that

the experimental public administration literature contains evidential value, and overall, to a reasonable degree, we can be confident in it.

Assessing Evidential Value

As already mentioned, it is very difficult to assess how credible a body of literature is. One way to do this is through (large-scale) replications. We appreciate that more and more public administration journals encourage replications. However, it seems impossible to replicate everything that gets published. Therefore, we need additional tools that allow us to judge the findings we see in the literature. One set of tools are those regularly used in meta-analysis to test for publication bias (e.g., Trim and Fill, PET-PEESE). However, simulation studies show that these tools do not perform well (Carter, 2019).

Psychologists as well as statisticians therefore developed new tools that aim to assess published research. The most widely applied tool is the *p*-curve method (Simonsohn et al., 2014, 2015). It addresses a specific aspect of credibility the authors call "evidential value" (Simonsohn et al., 2014, p. 535). With evidential value they describe the absence of selective reporting. Selective reporting summarizes different practices resulting in a misrepresentation of the conducted research. It comprises the reporting of significant results while hiding insignificant results and various practices to produce significant results when they have been insignificant in the first place. This is called *p*-hacking. *P*-hacking can be done by collecting additional data or conducting additional analyzes, e.g., with different estimators, with and without covariates, with and without outliers (Carbine, Lindsey, Rodeback, & Lar-son, 2019, p. 33).

A set of studies that is the result of selective reporting does not contain evidential value. Such a set of studies does not help us to learn something about the tested hypotheses because they only provide a restricted or even biased view on empirical tests. To truly get knowledge from a set of empirical studies, we need to see all conducted tests and they cannot be modified in order to yield significant results. If this is (predominantly) the case, we can state the studies contain evidential value.

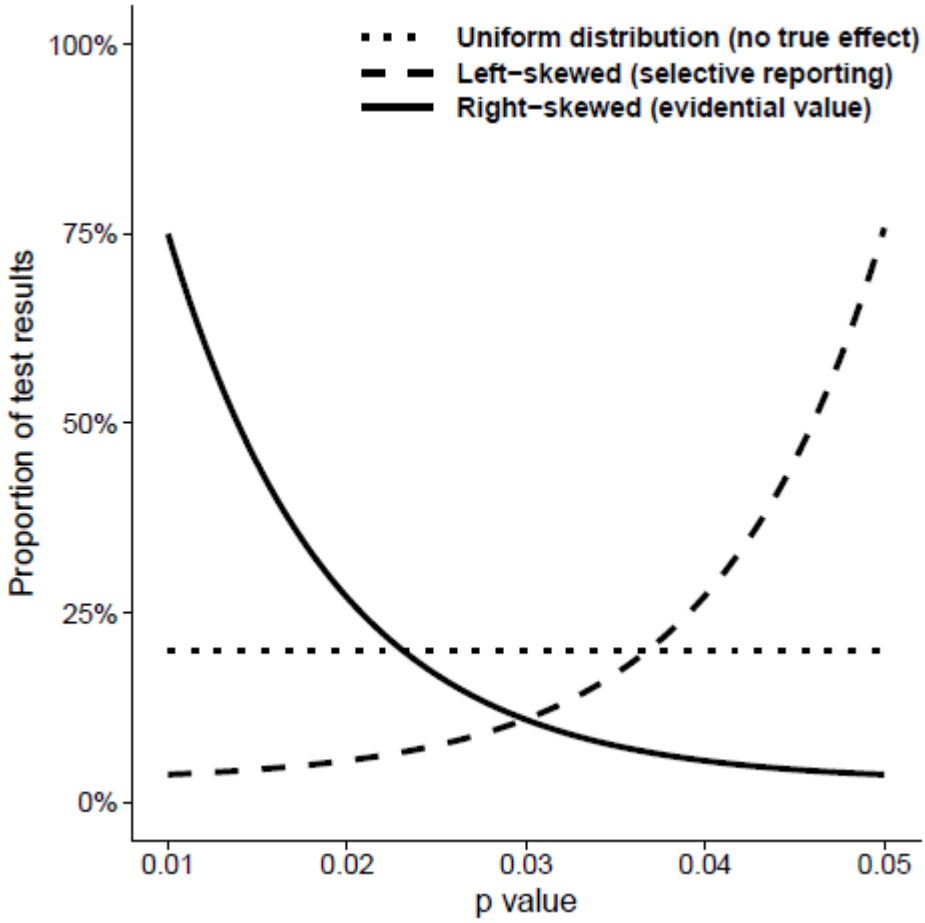
Such an assessment can be done for a set of studies testing a specific effect (see, e.g., Shariff, Willard, Andersen, & Norenzayan, 2016; Vogel & Homberg, 2020). If the studies contain evidential value, we can learn something from the empirical results about the effect they test. The assessment can also be done for a more abstract set of studies, like we do in this article. Analyzing the experimental public administration literature regarding their evidential value tells us if this research – overall – is based on selective reporting and *p*-hacking or if selective reporting can be ruled-out and the literature provides a solid base for further research. If we conclude that the experimental public administration literature contains evidential value, we can put more trust into this stream of research. Such an approach was previously used, for example, to assess the evidential value of the psychophysiological literature (Carbine et al., 2019).

The assessment made by the *p*-curve method is based on the *p* values reported in the published literature. It uses the fact that the distribution of *p* values follows a predictable pattern. If a series of tests is based on a non-existing effect, i.e., if the null hypothesis is true, the *p* values are uniformly distributed. All *p* values are equally likely and five percent are below the 0.05 threshold. If the tests are performed on a true effect, the distribution depends on the statistical power of the tests: the more power, the more right-skewed the distribution is. Hence, small *p* values are much more likely than bigger ones.

The interesting element here is that the predictability of the distribution of *p* values also holds for the subset of significant *p* values ($p < .05$). If the analyzed tests are based on a null effect (i.e., there is no true effect), the distribution of significant *p* values is uniform (see dotted line in Figure 1). If the tests are performed on a true effect, the distribution of significant *p* values is the more right-skewed the higher the statistical power is (see solid line in Figure 1).

Figure 1

Potential Uniform, Left-skewed, and Right-skewed Distributions of *P* Values. Figure Based on Carbine et al., 2019



The distribution of significant *p* values, however, changes when there is *p*-hacking or selective reporting. Obtaining additional data or running additional tests to reach statistical significances when there is no true effect results in a left-skewed distribution of *p* values (see dashed line in Figure 1). This can also happen when there is strong *p*-hacking on a true effect and the statistical power of the tests is low. When the studies overall have a high statistical power, mild *p*-hacking does not change the shape of the *p*-curve. The *p*-curve method therefore cannot exclude *p*-hacking; it tests whether the results are dominated by such practices.

Assuming that all significant effects get ultimately published or at least that there is no biasing selection effect, analyzing the significant effects enable an unbiased assessment of the literature. Focusing on the subset of significant results therefore has the advantage that it should not be affected by an unknown amount of unpublished insignificant results that are kept in the "file-drawer" due to journals and researchers' preference to publish significant results. Unlike conventional meta-analysis, the *p*-curve method does not try to quantify the amount of unpublished literature but just focuses on the significant effects and draws inference from that subsample. The described predictable distribution of *p* values makes such an approach possible.

The *p*-curve method therefore tests if the distribution of significant *p* values is right-skewed. If the distribution is right-skewed it indicates that the analyzed studies are based on a true effect and predominantly based on selective reporting. If the distribution is not right-skewed it indicates that there is no evidential value. A left-skewed distribution is indicative for intense *p*-hacking. The *p*-curve method therefore can be seen as a mild test of a set of studies. Finding no evidential value is clearly a bad sign for the tested findings. Concluding that there

is evidential value is a first good sign but it does not indicate that there is no selective reporting and *p*-hacking. It is just not so dominant that it entirely drives the results. Simonsohn et al. (2014) give a more detailed introduction to the *p*-curve method and the rationale behind the distribution of *p* values. Vogel and Homberg (2020) provide an application of the method to pubic administration research and introduce the method in detail.

Literature Search

With this article, we want to test whether the experimental public administration literature contains evidential value, i.e., if it is based on something more than publication bias and *p*-hacking. To do so, we first identified all the studies that use an experimental design and were published in one of the ten core public administration journals with the highest impact factor (InCites Journal Citation Reports, 2018). This includes *The American Review of Public Administration*, *Governance*, *International Public Management Journal*, *International Review of Administrative Sciences*, *Journal of Public Administration Research and Theory*, *Local Government Studies*, *Public Administration*, *Public Administration Review*, *Public Management Review*, and *Review of Public Personnel Administration*. We used the full time span available in Web of Science (Core Collection) up until 2020-03-18 – the day of the database query. We limited the sample to those who studies that applied a design that fully randomized participants to receive or not receive a treatment. The search procedure as well as the analytical approach were preregistered at the Open Science Framework: https://doi.org/10.17605/OSF.IO/TXQ43.[1] Because of its strong focus on experimental studies we post-hoc added the *Journal of Behavioral Public Administration* (*JBPA*) to the analysis. The inclusion of *JBPA* did not alter the overall conclusions. An overview of the results excluding *JBPA* is presented in Appendix C.

We initially identified 376 articles through the database search and 43 in *JBPA (N = 416)*. After carefully assessing them and excluding those who did not use a fully randomized experimental design (n = 222) or where a *p* value could not be calculated because the required information was missing (n = 3) or the applied method did not provide *p* values (n = 1), we were left with a set of 190 articles with at least one experimental study (see Appendix A). Since the *p*-curve method only uses the subset of significant results, because only this subset can be assumed to be published without a biasing selection effect, we excluded 60 articles with insignificant effects. The remaining 130 article in total report 172 experiments with significant findings regarding Hypothesis 1 (see below). These 172 effects are the sample we used in the subsequent *p*-curve analysis.

The included studies cover a wide range of different research designs and topics. Topic-wise the full range of public administration researchers' interests is covered, e.g., citizen satisfaction, motivation, performance management, decision-making, transparency, red tape, and many more. Design-wise the sample includes between-groups as well as within-person designs, conjoint-experiments, and others. Overall, a third of the studies uses a two-group design (33.7 %). 11.0 % use three groups, 25.6 % four groups, 5.2 % five groups, 13.4 % six groups, and 11.1 % more than six groups. The median sample size for the included effects is 495.5 (M = 1572.6, SD = 6236.0). 79.7 % of the studies conducted a survey experiment, while 12.2 % are field experiments, and 8.1 % lab experiments.

Extracting *P* Values

To be analyzed with the *p*-curve method, the selected *p* values must fulfill three requirements: they need to test the hypothesis of interest, they have to have a uniform distribution under the null, and they need to be statistically independent of other *p* values (Simonsohn et al., 2014, p. 542). One consequence of this is that often only one *p* value can be selected per study. To be as objective as possible, we did not decide on our own what the most important hypothesis of a study is. Instead, we always selected the first hypothesis presented in a study and required that it was tested using a fully randomized experimental approach. This approach also avoided giving articles with many hypotheses too much weight. If articles consisted of multiple studies performed to test Hypothesis 1, we included all of them since the *p* values are independent from each other. However, when multiple tests (e.g., different operationalizations of the dependent variable or different kinds of tests for the same relationship) where performed to test the first hypothesis, we only included the first significant test.

In case the respective article did not provide all the information necessary to calculate an exact *p* value, we used a multi-step approach to nevertheless be able to include the study in the analysis. The measures ranged

from contacting the authors to ultimately deducting a *p* value based on the indicated significance level. The detailed information about the extracted *p* values are provided in the so-called "*p*-curve disclosure table" in the supplementary files.

Results

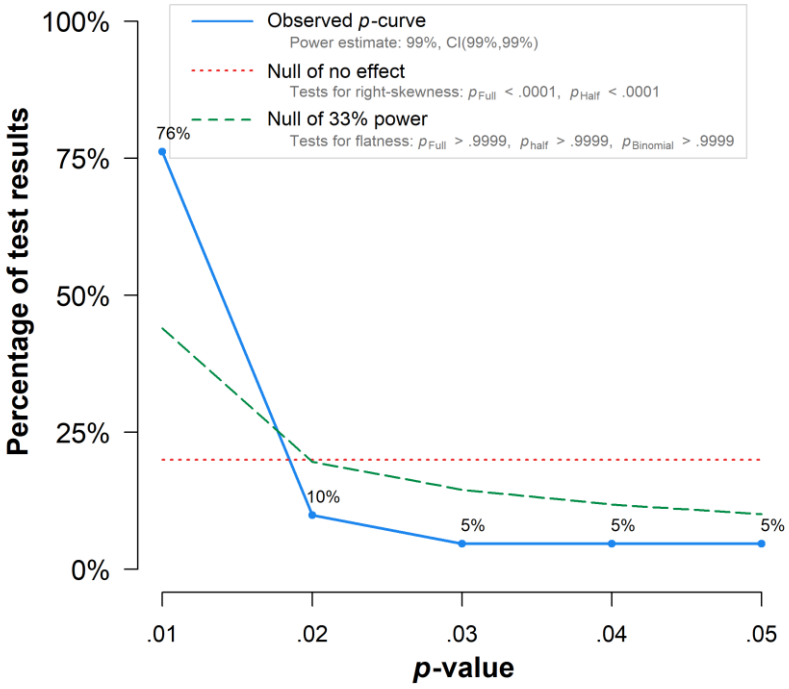
We used the R-based *p*-curve app (v. 4.06, http://www.p-curve.com) developed by Simonsohn et al. (2014) to apply the *p*-curve method to the selected experimental public administration studies. The full data and code are available at https://doi.org/10.17605/OSF.IO/MJ7XF.

The result of the *p*-curve analysis is displayed in Figure 2. The distribution of *p* values (*p*-curve) is shown as a blue solid line. The red dotted line represents the distribution that can be expected if the analyzed effects would be based on null effects.

The solid line clearly shows a right-skewed distribution, resembling what can be expected from the distribution of *p* values that contains evidential value. There are much more small *p* values than large ones, and 76 % of them are smaller than .01. The distribution is also more skewed than what is to be expected when the studies have low statistical power. The green dotted line shows how a distribution would look like when the studies were only based on 33 % power.

Figure 2

The *P*-curve for 172 Statistically Significant Effects Presented in Experimental Public Administration Studies



Note: The observed *p*-curve includes 172 statistically significant ($p < .05$) results, of which 150 are $p < .025$.
There were no non-significant results entered.

Simonsohn et al. (2014, 2015) also developed statistical tests to assess whether the distribution of *p* values is right-skewed. The results are presented in Table 1. The *p*-curve app provides two tests, the binomial test and

the more advanced continuous test. In the first row, we see the results of the test for right-skewedness. Both tests strongly reject the null hypothesis that the *p*-curve is not right-skewed. We can therefore conclude that the analyzed studies contain evidential value. In the second row, it is tested whether the evidential value is inadequate, which means that it is based on very low power. Both tests do not reject the null hypothesis that the *p*-curve is more right-skewed than it would be if it were based on 33 % power. The estimated overall power of the sample is 99 %. However, it has to be noted that there is strong criticism on the approach the *p*-curve app uses to estimate the underlying power (Brunner & Schimmack, 2020). We therefore do not further discuss the results.

Table 1
The Results of the *P*-curve Analysis

	Binomial Test *(Share of results p<.025)*	**Continuous Test** *(Aggregate with Stouffer Method)*	
		Full *p*-curve (*p*'s<.05)	**Half *p*-curve (*p*'s<.025)**
1) Studies contain evidential value. *(Right skew)*	$p<.0001$	$Z=-41.07$, $p<.0001$	$Z=-41.67$, $p<.0001$
2) Studies' evidential value, if any, is inadequate. *(Flatter than 33% power)*	$p=.9999$	$Z=28.97$, $p>.9999$	$Z=37.75$, $p>.9999$
Statistical Power			
Power of tests included in *p*-curve *(correcting for selective reporting)*	Estimate: 99% 90% Confidence interval: (99%, 99%)		

For some studies we needed to use the indicated significance level as the best way to calculate a *p* value. In some cases, the indicated level was 0.05, and hence, we assumed a *p* value of 0.049. To exclude that this affected the analysis, we conducted a robustness test excluding the studies where the *p* value was based on the significance level (n =12). The results in Appendix B show that the proportion of *p* values between .04 and .05 is reduced from 5 % to 4 %, resulting in a *p*-curve that is again clearly right-skewed. Excluding the studies published in the *Journal of Behavioral Public Administration*, which was not considered in the preregistration, also does not alter the results (see Appendix C).

We also conducted separate *p*-curve analysis for each journal. For four journals (*The American Review of Public Administration*, *Governance*, *Local Government Studies*, and *Review of Public Personnel Administration*), the results are of limited validity because there were less than ten studies for each of them. We nevertheless report them together with the results for the remaining six journals in Table 2. The analyses reveal no deviations from the overall results. For all journals – even the ones with less than ten effects – the continuous test indicates a right-skewed distribution of *p* values. The less advanced binomial test fails to confirm a right-skewed distribution for three journals with a very small number of studies (n=3 or 4). This should not be seen as an indicator of lacking evidential value. The *p*-curve analysis also does not find the evidence to be inadequate, i.e., none of the *p*-curves is flatter than what could be expected when the studies were conducted with less than 33 % power.

To test for potential issues in specific sub-fields of public administration, we manually coded all effects and assigned them to one of ten different sub-fields. Two sub-fields only contain one (financial management) or two (governance and networks). We did not perform a *p*-curve analysis on them because the results would be meaningless. For the remaining eight sub-fields, separate analyses were carried out. The results are presented in Table 3. Again, we did not find indications of selective reporting in specific sub-fields of public administration research.

Table 2
The Results of one Separate *P*-curve Analysis Per Journal

Journal	n	Binomial Test	Continuous Test	
		(Share of results p<.025)	*(Aggregate with Stouffer Method)*	
			Full *p*-curve	**Half *p*-curve**
			(*p*'s<.05)	**(*p*'s<.025)**
Evidential value (Right skew)				
ARPA	3	p=0.500	Z=-5.91, p<0.001	Z=-7.94, p<0.001
Governance	3	p=0.500	Z=-3.47, p<0.001	Z=-5.43, p<0.001
IPMJ	25	p<0.001	Z=-16.66, p<0.001	Z=-15.94, p<0.001
JBPA	21	p<0.001	Z=-10.89, p<0.001	Z=-10.21, p<0.001
JPART	31	p=0.005	Z=-13.83, p<0.001	Z=-16.21, p<0.001
LGS	4	p=0.312	Z=-7.64, p<0.001	Z=-9.10, p<0.001
PA	22	p<0.001	Z=-16.78, p<0.001	Z=-16.40, p<0.001
PAR	40	p<0.001	Z=-20.78, p<0.001	Z=-20.30, p<0.001
PMR	17	p<0.001	Z=-14.89, p<0.001	Z=-14.43, p<0.001
RoPPA	6	p=0.016	Z=-10.36, p<0.001	Z=-9.86, p<0.001
p-curve flatter than 33% power				
ARPA	3	p=0.642	Z=4.31, p>0.999	Z=6.67, p>0.999
Governance	3	p=0.638	Z=1.90, p=0.971	Z=4.82, p>0.999
IPMJ	25	p=0.998	Z=12.22, p>0.999	Z=14.92, p>0.999
JBPA	21	p=0.992	Z=7.24, p>0.999	Z=10.01, p>0.999
JPART	31	p=0.704	Z=8.81, p>0.999	Z=14.27, p>0.999
LGS	4	p=0.741	Z=5.51, p>0.999	Z=7.55, p>0.999
PA	22	p=0.994	Z=11.95, p>0.999	Z=14.40, p>0.999
PAR	40	p>0.999	Z=15.36, p>0.999	Z=19.20, p>0.999
PMR	17	p=0.997	Z=10.53, p>0.999	Z=12.49, p>0.999
RoPPA	6	p>0.999	Z=7.42, p>0.999	Z=8.11, p>0.999

Notes: ARPA = The American Review of Public Administration, IPMJ = International Public Management Journal, JPART = Journal of Public Administration Research and Theory, LGS = Local Government Studies, PA = Public Administration, PAR = Public Administration Review, PMR = Public Management Review, RoPPA = Review of Public Personnel Administration. There are no results for the International Review of Administrative Sciences because there were no significant effects in the data.

Table 3
The Results of One Separate *P*-curve Analysis Per Public Administration Sub-field

Sub-field	n	Binomial Test *(Share of results p<.025)*	Continuous Test *(Aggregate with Stouffer Method)* Full *p*-curve (*p*'s<.05)	Half *p*-curve (*p*'s<.025)
Evidential value (Right skew)				
Citizen Participation & Voting	5	p=0.500	Z=-5.69, p<0.001	Z=-7.19, p<0.001
Decision-Making	21	p<0.001	Z=-13.96, p<0.001	Z=-13.82, p<0.001
Gender, Diversity, & Equal Treatment	10	p=0.172	Z=-8.07, p<0.001	Z=-10.16, p<0.001
HRM	38	p<0.001	Z=-18.15, p<0.001	Z=-17.76, p<0.001
Marketing & Communication	13	p<0.001	Z=-13.65, p<0.001	Z=-12.43, p<0.001
Performance Management & Performance Data	59	p<0.001	Z=-25.99, p<0.001	Z=-26.01, p<0.001
Red Tape & Bureaucracy	8	p=0.035	Z=-7.62, p<0.001	Z=-7.67, p<0.001
Trust, Accountability, & Transparency	15	p=0.004	Z=-13.71, p<0.001	Z=-14.79, p<0.001
p-curve flatter than 33% power				
Citizen Participation & Voting	5	p=0.448	Z=3.59, p>0.999	Z=5.89, p>0.999
Decision-Making	21	p=0.992	Z=10.40, p>0.999	Z=13.30, p>0.999
Gender, Diversity, & Equal Treatment	10	p=0.584	Z=5.03, p>0.999	Z=8.67, p>0.999
HRM	38	p=0.993	Z=12.67, p>0.999	Z=16.47, p>0.999
Marketing & Communication	13	p>0.999	Z=9.88, p>0.999	Z=11.19, p>0.999
Performance Management & Performance Data	59	p>0.999	Z=18.66, p>0.999	Z=23.43, p>0.999
Red Tape & Bureaucracy	8	p=0.935	Z=5.19, p>0.999	Z=6.81, p>0.999
Trust, Accountability, & Transparency	15	p=0.957	Z=9.65, p>0.999	Z=12.73, p>0.999

Discussion and Conclusion

Analyzing 172 significant effects published in 130 articles, we find that the experimental public administration literature contains evidential value, which means that it is not solely based on selective reporting. Hence, we can have a certain amount of confidence in the evidence we – as a community – produced. This is a good sign for public administration research because it is a first indicator that the discipline has learned from the mistakes that were made in other fields, especially in psychology.

However, it also does not mean that the experimental public administration literature is free of *p*-hacking. We can only conclude that selective reporting (including *p*-hacking) does not substantially drive the results. It

is unlikely that the field is dominated by research that is based on null effects that have been hacked to reach significance or where underpowered studies on null effects have been selectively reported based on their significance. Nevertheless, this might be the case for some studies.

We see three factors that contribute to the overall positive result: First, we see a widespread effort to conduct high-powered studies that are able to reliably detect an effect if it is present and reduce the danger of type I errors. This effort has been intensified in the past years. Second, there seems to be an awareness among the (experimental) public administration community about the practices that led to the replication crisis in psychology. Third, we see more and more efforts to preregister experiments and publish replications. We also found that the number of insignificant findings is substantial. We identified 60 insignificant findings – 25 % of all considered effects. This indicates that publishing null results – at least for the first hypothesis – is quite common in public administration research. An environment in which publishing null results is common is certainly beneficial for public administration because it avoids a widespread publication bias. However, we only focused on the first hypothesis, and it is unclear if publishing articles that present solely null results is just as common. If we want to achieve an unbiased picture of what researchers find out about public administrations, papers consisting of null results need to be as common as papers that find support for all or some hypotheses.

At the same time researchers often choose to lower the significance threshold to $p < .1$ to claim support for their hypotheses. Given that statisticians recently argued that even a significance level of .05 might be too high (Benjamin et al., 2018), this practice needs to be assessed critically. Especially if it is applied selectively to effects researchers have an interest to present as significant.

So, what should researchers, editors, and reviewers take away from this study? We should not conclude that everything is fine, and we, therefore, should not change anything. In our opinion, the results show that we learned from the mistakes others made and should continue to do so. There are still many practices developed in other fields that can further improve the replicability and validity of (experimental) public administration research. Registered reports, splitting the review process into two parts and judge a study before the data is collected, is one such practice (Chambers, 2019; Nosek & Lakens, 2014). So far, no public administration journal is offering registered reports.

The *p*-curve method is a valuable instrument that helps to test if a field of study is affected by very serious issues. It is, however, not the bar we should set for ourselves. Instead, a right-skewed distribution of *p* values should be seen as a minimum standard. As a whole, experimental public administration literature has passed this hurdle. This means, we can focus on additional "hurdles". These have been discussed widely in recent years (Perry, 2017; Vogel & Homberg, 2020; Zhu et al., 2019) and include, for example, making research more transparent (open data, open code, open materials, open peer-review, open access, …), increasing external validity of experimental studies, improving theorizing, applying more diverse methods, and many more.

With regard to further research, we want to encourage researchers to use the *p*-curve method to address other literatures. We think it is especially valuable to assess the literature on specific theories or effects. Previous work has shown that a lot can be learned from "*p*-curving" such research (Carbine et al., 2019; Lakens, 2017; Vogel & Homberg, 2020). We also believe that the issue of statistical power deserves more attention, especially since the *p*-curve method's power estimations have been strongly criticized (Brunner & Schimmack, 2020).

In sum, we can conclude that the experimental public administration research is on a good path to create credible evidence. But there is no reason to lay back and lower efforts to improve our research practices and make it more rigor. Nevertheless, we should extend efforts to further increase the reliability and credibility of our research. This includes rigorous application of open science principles such as preregistration, registered reports, open data, open code, and open materials. We also want to highlight that the reporting of experimental results needs to substantially improve in order to give readers the chance to validate the results and enable meta-analytical analysis. Various attempts were made to encourage researchers to do so (e.g., James, Jilke, & Van Ryzin, 2017; Vogel & Homberg, 2020).

That said, we also want to be clear of the limitations of the presented *p*-curve analysis. First of all, our results do not mean that the analyzed results are true in the sense of confirming the underlying theories. The *p*-curve method only tests for signs of selective reporting and *p*-hacking. It cannot test if a study is a valid assessment of a theory and thereby confirm the theory. More importantly, as the *p* value can be inflated by enlarging the sample size, statistically significant findings may speak little to the true effect of the experimental treatment. After all, this is a question statistical tools are never able to answer. It is the responsibility of authors, reviewers,

editors, and the readers to assess whether a theory is plausible and whether an experiment is a valid assessment of the theory.

We decided to conduct a very broad assessment of the experimental public administration literature. This way, we can say something about the field in general, and with regard to the sub-field analysis, we also gained knowledge about sub-fields of experimental public administration research. However, it is not impossible that other subfields using different methods or studies on a certain theory might be affected by selective reporting and *p*-hacking. We did not find indications of that during our analysis, and only it is difficult to get a definitive answer, even with more detailed analysis. Additionally, it is challenging for the *p*-curve method to detect *p*-hacking if a body of research is heavily influenced by it (Simonsohn et al., 2014). Finally, the *p*-curve method assumes that there is no bias that affects which significant effects are ultimately published (Simonsohn et al., 2014), which might be up for debate.

Notes

1. There was an error in the preregistration that concerned the search term for the literature search (accidentally replacing the "research" in "journal of public administration research and theory" with "review"). We fixed this error before conducting the literature search.

Acknowledgments

The authors would like to thank Sebastian Jilke for encouraging us to start this project together. We also want to thank Jessica Steenbock for her indispensable help with collecting and managing the data.

References

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … Johnson, V. E. 2018. Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Bertelli, A. M., & Riccucci, N. M. 2020. What is behavioral public administration good for? *Public Administration Review*, puar.13283. https://doi.org/10.1111/puar.13283

Brunner, J., & Schimmack, U. 2020. Estimating population mean power under conditions of heterogeneity and selection for significance. *Meta-Psychology*, *4*(4). https://doi.org/10.15626/MP.2018.874

Carbine, K. A., Lindsey, H. M., Rodeback, R. E., & Larson, M. J. 2019. Quantifying evidential value and selective reporting in recent and 10-year past psychophysiological literature: A pre-registered P-curve analysis. *International Journal of Psychophysiology*, *142*, 33–49. https://doi.org/10.1016/j.ijpsycho.2019.06.004

Carter, E. C. 2019. *Deep learning for robust meta-analytic estimation* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/zad47

Chambers, C. 2019. What's next for registered reports? *Nature*, *573*(7773), 187–189. https://doi.org/10.1038/d41586-019-02674-6

Davis, M. L., Papini, S., Rosenfield, D., Roelofs, K., Kolb, S., Powers, M. B., & Smits, J. A. J. 2017. A randomized controlled study of power posing before public speaking exposure for social anxiety disorder: No evidence for augmentative effects. *Journal of Anxiety Disorders*, *52*, 1–7. https://doi.org/10.1016/j.janxdis.2017.09.004

Earp, B. D., & Trafimow, D. 2015. Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00621

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., De Ridder, D. T. D., Dewitte, S., … Zwienenberg, M. 2016. A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*(4), 546–573. https://doi.org/10.1177/1745691616652873

Hassan, S., & Wright, B. E. 2020. The behavioral public administration movement: A critical reflection. *Public Administration Review*, *80*(1), 163–167. https://doi.org/10.1111/puar.13130

James, O., Jilke, S. R., & Van Ryzin, G. G. (Eds.). 2017. Appendix—recommended reporting guidelines for experiments in public management, a checklist. In *Experiments in Public Management Research* (pp. 509–511). Cambridge University Press. https://doi.org/10.1017/9781316676912.026

Jilke, S., Meier, K. J., & Van Ryzin, G. G. 2018. Editorial. *Journal of Behavioral Public Administration*, *1*(1). https://doi.org/10.30636/jbpa.11.9

Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C., Nosek, B. A., Chartier, C. R., Christopherson, C. D., Clay, S., Collisson, B., Crawford, J., Cromar, R., Vidamuerte, D., Gardiner, G., Gosnell, C., Grahe, J., Hall, C., Joy-Gaba, J., Legg, A. M., Levitan, C., … Ratliff, K. 2019. *Many labs 4: Failure to replicate mortality salience effect with and without original author involvement*. PsyArXiv. https://doi.org/10.31234/osf.io/vef2c

Lakens, D. 2017. *Professors are not elderly: Evaluating the evidential value of two social priming effects through P-curve analyses* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/3m5y9

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., … Zwaan, R. A. 2018. Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171. https://doi.org/10.1038/s41562-018-0311-x

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS Medicine*, *6*(7), e1000100. https://doi.org/10.1371/journal.pmed.1000100

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. 2017. A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 0021. https://doi.org/10.1038/s41562-016-0021

Nosek, B. A., & Lakens, D. 2014. Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. https://doi.org/10.1126/science.aac4716

Perry, J. L. 2017. Practicing what we preach! Public administration review promotes transparency and openness. *Public Administration Review*, *77*(1), 5–6. https://doi.org/10.1111/puar.12705

Shariff, A. F., Willard, A. K., Andersen, T., & Norenzayan, A. 2016. Religious priming: A meta-analysis with a focus on prosociality. *Personality and Social Psychology Review*, *20*(1), 27–48. https://doi.org/10.1177/1088868314568811

Simmons, J. P., Nelson, L. D., & Simonsohn, U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. 2014. P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. https://doi.org/10.1037/a0033242

Simonsohn, U., Simmons, J. P., & Nelson, L. D. 2015. Better p-curves: Making p-curve analysis more robust to errors, fraud, and ambitious p-hacking, a reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General*, *144*(6), 1146–1152. https://doi.org/10.1037/xge0000104

Vogel, D., & Homberg, F. 2020. P-Hacking, p-Curves, and the PSM–performance relationship: Is there evidential value? *Public Administration Review*, puar.13273. https://doi.org/10.1111/puar.13273

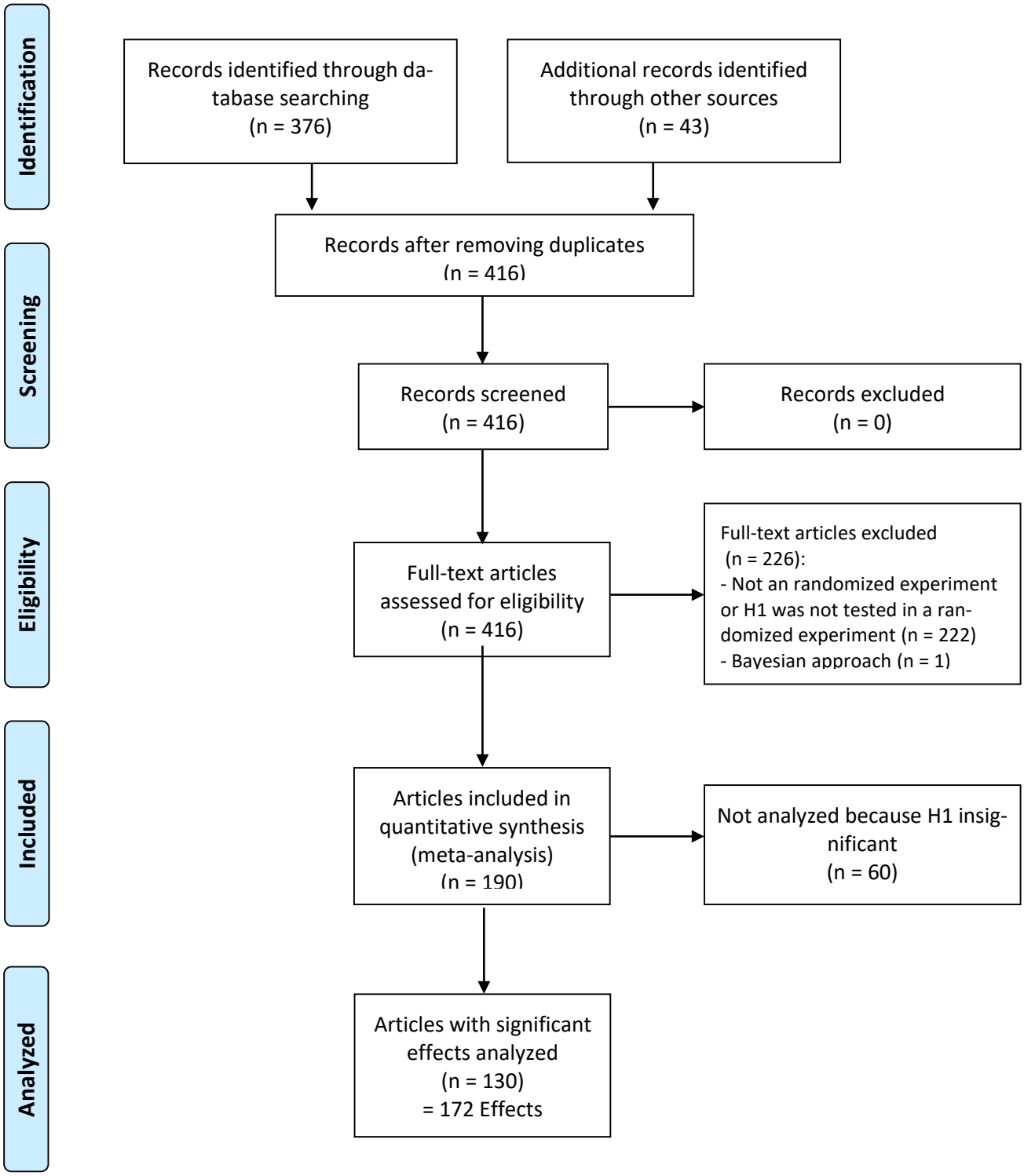
Walker, R. M., Brewer, G. A., Lee, M. J., Petrovsky, N., & van Witteloostuijn, A. 2019. Best practice recommendations for replicating experiments in public administration. *Journal of Public Administration Research and Theory*, *29*(4), 609–626. https://doi.org/10.1093/jopart/muy047

Walker, R. M., James, O., & Brewer, G. A. 2017. Replication, experiments and knowledge in public management research. *Public Management Review*, *19*(9), 1221–1234. https://doi.org/10.1080/14719037.2017.1282003

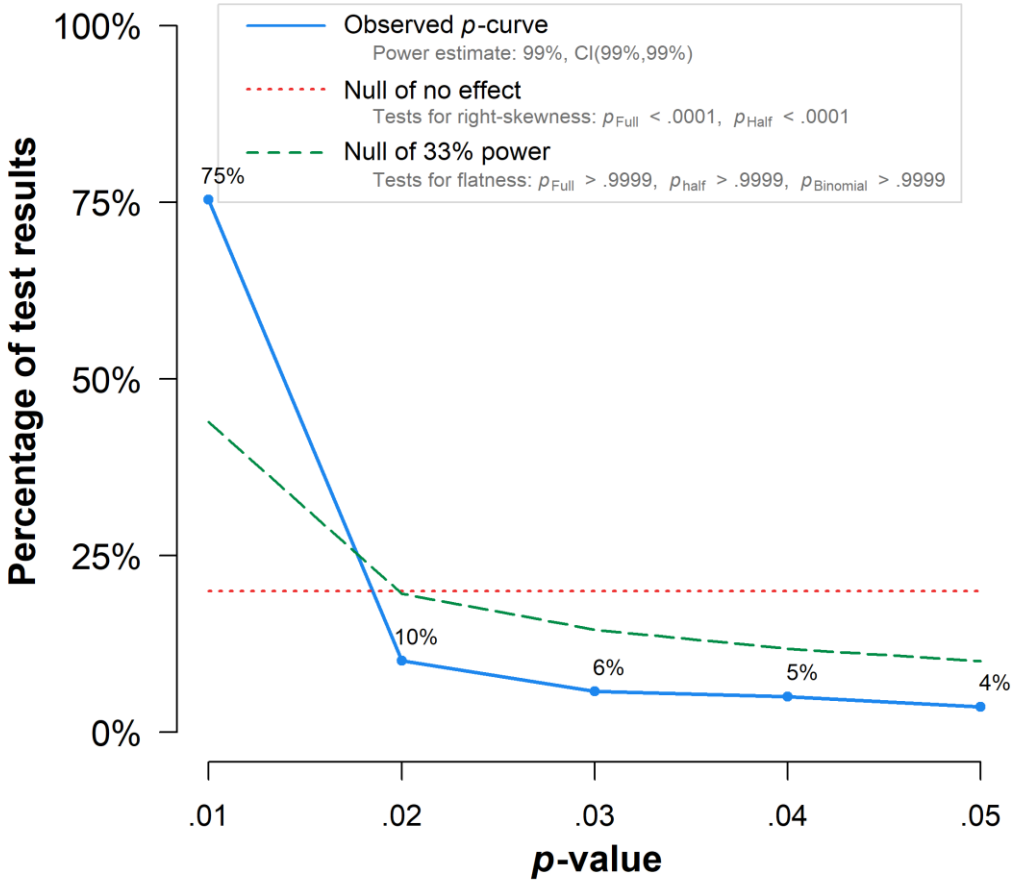
Zhu, L., Witko, C., & Meier, K. J. 2019. The public administration manifesto II: Matching methods to theory and substance. *Journal of Public Administration Research and Theory*, *29*(2), 287–298. https://doi.org/10.1093/jopart/muy079

Appendix

Appendix A.
The Search Flow: Results of the Search for Articles Based on Liberati et al. (2009)



Appendix B:
Robustness check excluding effects that are based on significance level

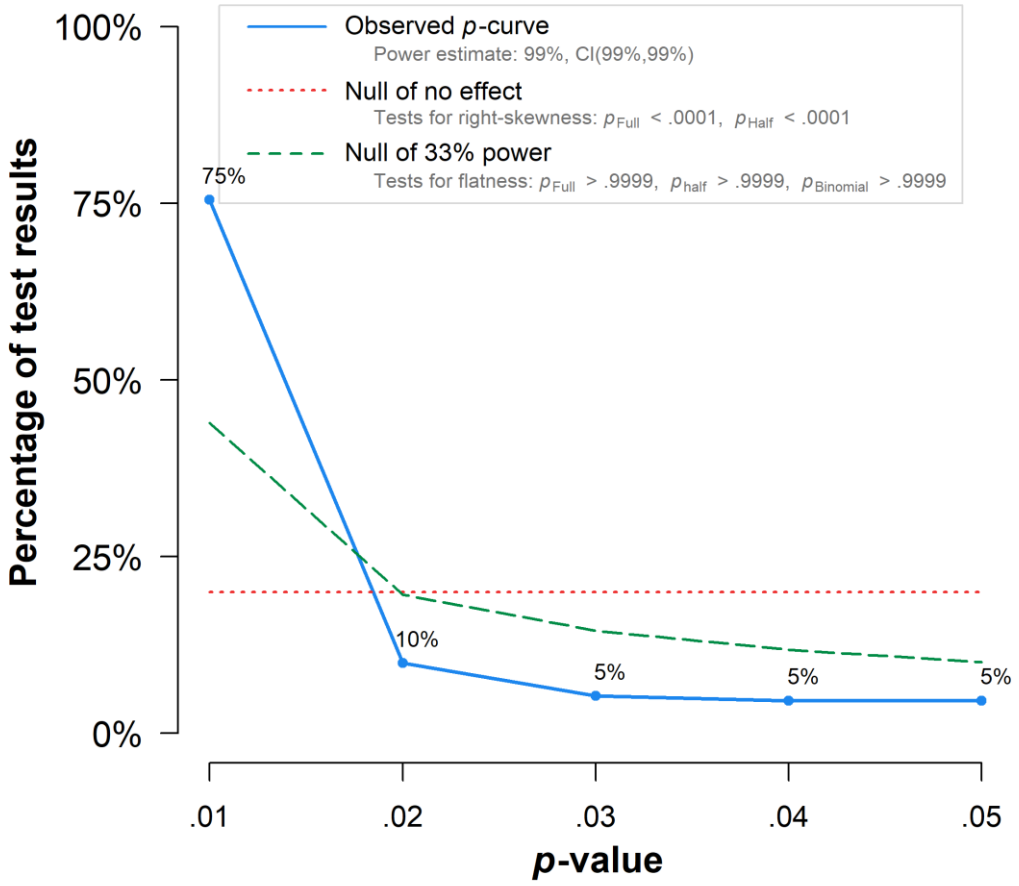


Note: The observed *p*-curve includes 138 statistically significant ($p < .05$) results, of which 120 are $p < .025$.
There were no non-significant results entered.

Results of the *P*-curve Analysis for the Robustness Check (Excluding Effects that are Based on Significance Level)

	Binomial Test *(Share of results p<.025)*	**Continuous Test** *(Aggregate with Stouffer Method)*	
		Full *p*-curve (*p*'s<.05)	**Half *p*-curve (*p*'s<.025)**
1) Studies contain evidential value. *(Right skew)*	$p<.0001$	$Z=-40.03$, $p<.0001$	$Z=-40.8$, $p<.0001$
2) Studies' evidential value, if any, is inadequate. *(Flatter than 33% power)*	$p=.9999$	$Z=28.81$, $p>.9999$	$Z=36.50$, $p>.9999$
	Statistical Power		
Power of tests included in *p*-curve *(correcting for selective reporting)*	Estimate: 99% 90% Confidence interval: (99%, 99%)		

Appendix C:
Robustness check excluding studies published in the Journal of Behavioral Public Administration



Note: The observed *p*-curve includes 151 statistically significant ($p < .05$) results, of which 131 are $p < .025$. There were no non-significant results entered.

Results of the *P*-curve Analysis for the Robustness Check
(Excluding Studies Published in the Journal of Behavioral Public Administration)

	Binomial Test *(Share of results p<.025)*	**Continuous Test** *(Aggregate with Stouffer Method)*	
		Full *p*-curve (*p*'s<.05)	**Half *p*-curve (*p*'s<.025)**
1) Studies contain evidential value. *(Right skew)*	$p<.0001$	$Z=-39.77$, $p<.0001$	$Z=-40.70$, $p<.0001$
2) Studies' evidential value, if any, is inadequate. *(Flatter than 33% power)*	$p=.9999$	$Z=28.22$, $p>.9999$	$Z=36.59$, $p>.9999$
	Statistical Power		
Power of tests included in *p*-curve *(correcting for selective reporting)*	Estimate: 99% 90% Confidence interval: (99%, 99%)		