Quantitative Political Science Research is Greatly Underpowered

Vincent Arel-Bundock* Ryan C. Briggs† Hristos Doucouliagos‡
Marco Mendoza Aviña§ T.D. Stanley¶

October 18, 2024

Short title: Political Science Research is Underpowered
Keywords: Statistical power, publication bias, meta-analysis, p-hacking, sample size, methodology

Abstract

The social sciences face a replicability crisis. A key determinant of replication success is statistical power. We assess the power of political science research by collating over 16,000 hypothesis tests from about 2,000 articles in 46 areas of the discipline. Under generous assumptions, we show that quantitative research in political science is greatly underpowered: the median analysis has about 10% power, and only about 1 in 10 tests have at least 80% power to detect the consensus effects reported in the literature. We also find substantial heterogeneity in tests across research areas, with some being characterized by high power but most having very low power. To contextualize our findings, we survey political methodologists to assess their expectations about power levels. Most methodologists greatly overestimate the statistical power of political science research.

[1]

*Université de Montréal. vincent.arel-bundock@umontreal.ca

†University of Guelph. rbriggs@uoguelph.ca

‡Deakin University. chris.doucouliagos@deakin.edu.au

§Harvard University. mmendozaavina@fas.harvard.edu

¶Deakin University. tom.stanley1@deakin.edu.au

[1]We gratefully acknowledge funding from Canada's Social Sciences and Humanities Research Council. Replication files are available in the JOP Data Archive on Dataverse (https://dataverse.harvard.edu/dataverse/jop). The empirical analysis has been successfully replicated by the JOP replication analyst. The pre-registration for our expert survey is available at https://aspredicted.org/q3zg-4jp9.pdf

Statistical power is critical to any discipline that practices null hypothesis significance testing, such as political science. Power—or the probability that a statistical test will reject the null hypothesis when the alternative is true—must be a key consideration when researchers design a study, and when readers evaluate the credibility of published findings.

As power increases, the probability of committing a false negative error (type II) goes down. When power is low, empirical findings are less likely to replicate [Altmejd et al., 2019]. When power is low and one happens to find a statistically significant estimate, that estimate is often much greater than the "true" underlying effect, and it may well have the wrong sign [Gelman and Tuerlinckx, 2000, Gelman and Carlin, 2014, Ioannidis et al., 2017]. As political scientists grapple with the replication crisis [Baker, 2016], it is crucial to know whether our research designs have sufficient power to generate credible findings.

The power of a study is affected by a number of factors including the desired level of statistical significance, sample size, measurement variability, and the magnitude of the effects under investigation. Typically, larger effect sizes and larger samples result in higher power.[2] As such, it seems reasonable to expect that power will vary from study to study, and across scientific fields and disciplines.

In this article, we examine statistical power in political science by assembling a dataset of 16649 hypothesis tests, which are aggregated to produce 351 meta-analytic estimates, reported in 46 peer-reviewed meta-analytic articles.[3] We estimate power retrospectively by

[2]Power calculations are conducted based on assumptions regarding effect and sample size, but also by considering other factors such as measurement error. Design-based inference provides researchers with some control over the number of observations collected, yet sample size is often driven primarily by budgetary concerns rather than power calculations.

[3]We define a "meta-analysis" as a grouping of at least 5 comparable estimates which researchers have aggregated to calculate meta-analytic effects. A single meta-analytic article often reports many meta-analyses. Some meta-analyses address closely related substantive questions using slight variations on the independent and dependent variables, or using dif-

leveraging estimates of mean population effects from the meta-analyses.[4] In essence, we calculate the power of each test to detect the consensus effect reported in its literature. Our results suggest that quantitative political science research is greatly underpowered. The median research result has about 10% power to detect this effect ($\alpha = 0.05$), and only about 1 in 10 statistical tests have at least 80% power.[5] As we show below, there are some indications that power may be increasing gradually over time, but our estimates remain low, even in the later parts of our sample.

These results are both dispiriting and surprising. To contextualize them, we conducted an expert survey of political methodologists to measure their expectations about power levels in published hypothesis tests. On average, the experts in our sample believe that 66% of studies have at least 50% power, and 43% have at least 80% power. We demonstrate that these expectations are overly optimistic. On average, experts appear to overestimate the share of studies powered at the 50% level by 48 percentage points, and the share of studies powered at the 80% level by 32 percentage points. Political science research suffers from low power, and this problem should be taken seriously by empiricists and methodologists.

The rest of this paper is structured as follows. We begin by describing our research design and data, and by arguing that meta-analytic estimates of population mean effects for various research questions present the best opportunity for estimating power retrospectively. Then, we explain how we surmount challenges related to publication bias, review the data

ferent sets of estimates (e.g., experimental vs. observational).

[4]This should not be confused with *post hoc* power analysis, which we discuss and strongly discourage in the *Methods and Data* section.

[5]One potential concern is that our sample includes estimates from meta-analyses regardless of whether said meta-analysis rejected the null hypothesis. Meta-analyses where the null was not rejected might have near-zero population mean effect estimates, and this could drive down our overall estimates of power. However, we find the median level of power in the subset of studies where the meta-analytic estimate is itself significant at 0.05 level is only 0.9 percentage points higher.

we collected from meta-analysis replication files, and introduce our expert survey. Next, we propose various extensions and robustness checks to our analyses. Finally, we conclude with a discussion of low power, and propose a number of institutional, methodological, and theoretical remedies to this challenge.

Method and data

Our goal is to assess statistical power in recent quantitative political science research. In other words, we want to estimate the probability that any given statistical test will reject the null hypothesis, for a given "true" effect size. To achieve this, the ideal dataset would have a standard error along with a "true" effect size for every estimate reported in peer-reviewed journals in political science and closely adjacent fields. Given these data, one could calculate *retrospective* power for each test at some α level such as 0.05.[6] A test for which the standard error is less than the "true" effect size divided by 2.8 would have at least 80% power.[7]

Of course, in general "true" effect sizes are unknown. One can respond to this challenge in a number of ways. The most problematic approach is to use the reported effect size of each test to calculate the power of that test. This circular form of *post hoc* power calculation is rightly shunned as it reveals no new information and instead merely recapitulates p values [Hoenig and Heisey, 2001]. Perhaps worse, when there is selection on statistical significance such an approach will tend to dramatically overstate power [Gelman, 2019].

A second approach is to judge power against theoretically informed minimum effect sizes of interest. One could do this for an entire literature if the papers in it typically reported

[6] Throughout the paper, we use an α level of 0.05.

[7] Put differently, to discriminate a "true effect" from zero with 5% significance and 80% power, the effect's standard error needs to be smaller than the absolute value of the underlying effect divided by 2.8. This relationship is derived from the standard normal value that makes a 20/80% split in its cumulative distribution and the usual value of 1.96 for a significance level of 5% [Cohen and Wolman, 1965].

their minimum effect size of interest, but this is uncommon in political science. If papers report standardized effect sizes then one can use a related approach of measuring power against arbitrary but equally-sized "small" or "medium" effect sizes. However, again political scientists rarely report standardized effect sizes and so again this approach is not practical at the scale of our broad survey.

We therefore use a third approach: estimating power retrospectively using meta-analytic estimates of population mean effects for various research questions. This follows previous work in neuroscience [Button et al., 2013], economics [Ioannidis et al., 2017], psychology [Stanley et al., 2018], and ecology and evolutionary biology [Yang et al., 2022]. Because meta-analyses encompass all relevant, reported hypothesis tests for specific research areas, meta-analysis provides the best and most widely informed estimate of the population mean effect. Perhaps the only exception are estimates from preregistered multi-laboratory replications [Open Science Collaboration, 2015, Klein et al., 2018], but these are uncommon in political science. Using meta-analyses allows us to measure the power of each test against the relevant literature's most informed estimate of its effect size.

Meta-analyses

To find meta-analyses, we searched through the archives of 141 peer-reviewed journals in political science and closely adjacent fields. The list of peer-reviewed journals in the scope of our data collection came from two sources. First, we selected all journals appearing in the social science subcategories "Political Science," "Diplomacy & International Relations," and "Public Policy & Administration" of Google Scholar Metrics' top publications as of 2021. Second, we selected the 50 journals with the highest total citations for the year 2020 in the categories "political science," "international relations," and "public administration" categories of Clarivate's Journal Citation Reports. The full list of journals is printed in appendix A. We conducted full-text searches for the keyword "meta" in the archives of all these journals, ignoring all results dating from before 2000.[8] All keyword matches were checked manually to retain articles for which authors gathered data from other articles in order to run a quantitative meta-analysis.

We excluded qualitative meta-syntheses and articles in which authors combine multiple estimates from their own work, using meta-analytic techniques as a form of model averaging. We also excluded articles not focused on political topics, that is, articles where both the outcome and explanator variables are primarily the object of research in other disciplines, such as economics, criminology, management, and psychology. Ambiguous cases were reviewed by two different co-authors of the present paper. In total, 85 articles met our definition of a meta-analysis in political science.

We were able to obtain the data from 46 of those meta-analytic articles from public journal archives, research repositories, author websites, by transcribing results from published tables and figures, and by contacting authors directly.

One potential concern is that since we were able to obtain data for only 46 out of 85 eligible meta-analyses, we cannot definitively reject the possibility that non-public data would differ in terms of power or research area within political science. However, the decision of meta-analysts to make their data publicly available is likely unrelated to the power of the individual studies' estimates, since those studies were rarely conducted by the meta-analysts themselves. Consequently, although we acknowledge this limitation, it seems unlikely that this would significantly affect the overall conclusions of our analysis. Indeed, in appendix

[8]The year 2000 cutoff point is arbitrary and was chosen due to resource constraints and data availability. It is useful to note that data collection involved conducting full text searches on the full universe of articles published by a large sample of peer-reviewed journals (listed in the following sub-section). Through this process we found that very few meta-analyses had been published in political science before 2000 and that the data for these articles were generally unavailable. The median publication year of the hypothesis tests in our sample is 2010.

A, we present a balance table comparing the 46 papers in our sample to the 39 papers not in our sample. This table suggests that there are no meaningful differences between both sets of meta-analyses in terms of subfield distribution, the proportion looking at a research question related to political behavior, and the proportion published in a "top 3" political science journal (i.e., the Journal of Politics, the American Political Science Review, and the American Journal of Political Science). We do observe a slight difference in terms of the average publication date: papers in our sample were published in 2017 on average, versus 2014 for papers not in our sample. This may be due to replication archives becoming more common over time [Rainey et al., 2024].

Population mean effects

The main drawback of using meta-analyses for retrospective power analysis is that estimates of the population mean effects are based on reported results, which may have been selected based on statistical significance. Examples of such selection include the file drawer problem, reporting bias, specification searching, *p*-hacking, and the garden of forking paths.[9] Selection on statistical significance will result in there being too few statistically non-significant estimates and too many statistically significant estimates, which will inflate the size and significance of the collection of estimates reported in the literature.[10] If meta-analyses aggregate inflated estimates, they will likely produce inflated estimates of population mean effects, which will in turn lead us to overstate power [Gelman, 2019, Kvarven et al., 2020].

There is no universal agreement in the methodological literature on a best approach for

[9]The presence of publication bias is "one of the strongest findings across the sciences" [Berinsky et al., 2021, 370].

[10]As noted in the *Journal of Politics*' Pre-Registration Guidelines, "if power analyses and smallest effect sizes of interest are based on effect sizes reported in previous studies, authors should keep in mind that meta-scientific studies reliably report inflated effect sizes in the reported literature that often shrink in replication studies" [Journal of Politics, 2022].

estimating population mean effects from reported results. Thus, we use three alternative methods, which we introduce below from least to most aggressive in how they attempt to correct for publication bias.

The first method is the Unrestricted Weighted Least Squares (UWLS). The UWLS is a simple weighted average of the form $\hat{\mu}_w = \sum(1/\phi\sigma_i^2)y_i / \sum(1/\phi\sigma_i^2)$, where y_i are study-level estimates, σ_i^2 are within-study variances, and ϕ is a scaling factor. UWLS can be estimated by regressing the study-level estimates on a constant using weighted least squares with weights equal to $1/\sigma_i^2$.[11]

The second method is the Weighted Average of the Adequately Powered (WAAP), introduced by Stanley et al. [2017]. We compute the WAAP in three steps: (1) calculate the UWLS; (2) use the UWLS to estimate the power of each study; (3) calculate the UWLS again, but using only the subset of estimates that exceed 80% power.

The final method is our most aggressive strategy to unwind publication bias: the Precision-Effect Test and Precision-Effect Estimate with Standard Error (PET-PEESE) [Egger et al., 1997, Stanley and Doucouliagos, 2014, Stanley, 2017]. The intuition that motivates this approach is that selection on statistical significance will result in a positive relationship between reported effect estimates and reported standard errors. One can thus regress estimates on standard errors or standard errors squared, interpreting the intercept of this regression as the estimated effect when the standard error is equal to zero.[12] In a psychology study comparing 15 meta-analyses to multi-laboratory replications on the same research questions,

[11]UWLS produces the exact same point estimates as the more common Fixed Effects meta-analysis, but simulations suggest that the UWLS standard errors have better properties [Stanley and Doucouliagos, 2022]. Since our retrospective power calculations only require meta-analytic estimates of the population mean effects, and not their standard errors, the two approaches are functionally equivalent for our purposes.

[12]Following Stanley and Doucouliagos [2014], we first regress the effect estimates on their standard errors and a constant. If the constant from this regression is not significant at $p < 0.1$, we use it as our estimate of the population mean effect. If the constant is significant

PET-PEESE produced the least inflated estimates of the four methods tested [Kvarven et al., 2020].

Each of these three methods produces alternative estimates of population mean effects for each meta-analysis in our sample. With these in hand, we then calculate the statistical power of each estimate based on its standard error, one of our estimates of the population mean effect from the relevant meta-analysis, and an α level of 0.05. As we show below, results obtained using the three techniques are broadly consistent. For simplicity, the rest of this paper generally focuses on results obtained via UWLS, the most "generous" approach.[13]

Expert survey

We conducted an original expert survey of political methodologists in June 2022 to assess their expectations about power levels.[14] Our sample includes all authors who published in the peer-reviewed methodology journal *Political Analysis* between 2010 and 2021. In total,

at $p < 0.1$, we regress the effect estimates on the standard errors squared and a constant, using the constant from this regression as our estimate. This conditional approach is used because the linear model is better in the absence of a "true" effect, while the squared model is better in the presence of a "true" effect. All regressions in the PET-PEESE use weighted least squares, where the weights are equal to $1/\sigma^2$.

[13]We focus on the UWLS estimates because our core finding is that statistical power is generally lower than expected by political methodologists, and we want to avoid "stacking the deck" in the direction of this conclusion by using an aggressive bias correction strategy.

[14]Our analysis of the survey data was preregistered following the AsPredicted template, and the pre-analysis plan is available at `https://aspredicted.org/blind.php?x=G28_YCP`. The design was approved by two different institutional review boards from the universities of two coauthors of the present paper. See appendix C for our recruitment materials and survey instrument.

131 methodologists answered our survey, for a response rate of 27%.[15]

We asked experts to consider all hypothesis tests published in the 50 peer-reviewed journals with the highest impact factors in political science and closely adjacent fields over the past two decades. They then guessed what share of those tests had at least 50% and 80% power to reject the null at the $\alpha = 0.05$ level. To align the survey question with our empirical approach, we asked about the share of tests that had at least 50% or 80% "power to reject the null hypothesis." This guided respondents to think about power relative to likely effect sizes. For logical consistency, we exclude a few respondents who stated that more studies have at least 80% than 50% power. 42 respondents said that they did not know the answer to at least one of the questions.

Results

We begin by examining selection on statistical significance. Figure 1 shows a histogram of the absolute value of z-statistics for all hypothesis tests.[16] The distribution is right-skewed: there are many significant tests. The large spikes and humps show that many z-statistics are concentrated at or just above two conventional thresholds of statistical significance: 95% and 99%. These results resemble those found by Gerber and Malhotra [2008] for political science and by Brodeur et al. [2016; 2020] and Gorajek and Malin [2021] for economics.[17] It is difficult to explain such a distribution of z-statistics in the absence of selection on statistical significance.

Figure 1 has important implications for our study of statistical power. Since reported

[15]We emailed 478 methodologists (555 minus 77 for whom emails were undeliverable). 131 answered all of our questions, yielding a response rate of 27%.

[16]We lack information on degrees of freedom for most observations, so we calculate z-statistics instead of t-statistics. For visual clarity, we exclude z-statistics larger than 10.

[17]See especially Figure 1 by Brodeur et al. [2016]. Vivalt [2019] shows similar results for quasi-experimental impact evaluations.

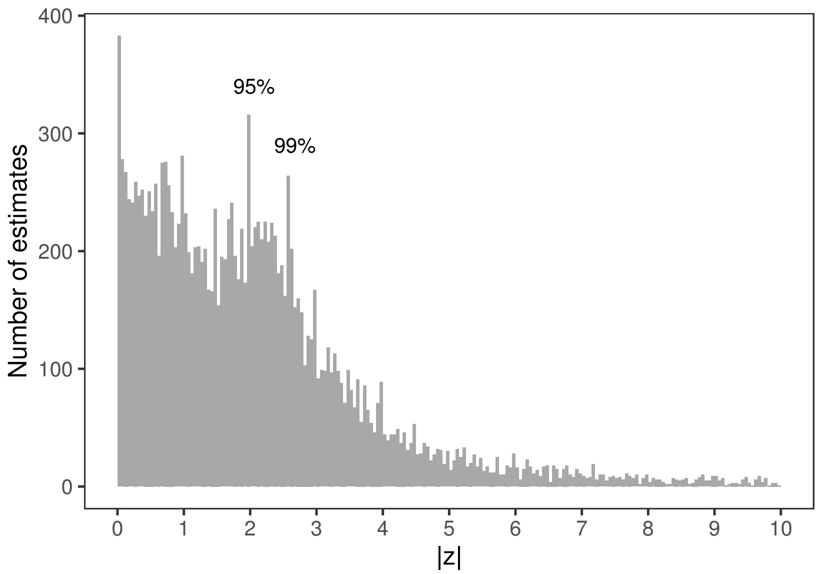

Figure 1: Distribution of z-statistics in the full sample of estimates. The dashes highlight the conventional thresholds of statistical significance of 95% and 99%.

findings are characterized by publication bias, our estimates of the "truth" and our power calculations are likely to be inflated. The UWLS results presented below should thus be interpreted as a best-case scenario for statistical power in the discipline.

Our key results are presented in Figure 2. The results for each of the three methods for estimating population effects are displayed using differently line types, which represent the share of all hypothesis tests that reach a given power level from 6% to 99%.[18]

All three techniques produce similarly shaped curves. UWLS, which addresses publication bias least aggressively, yields the highest power estimates. Yet power is low even under these generous conditions: only half of the tests reach 10% power, a fifth reach 50% power, and a tenth reach 80%. Our two other methods for estimating population mean effects result in even lower power levels.

The black overlays in Figure 2 display the results of our expert survey. As a reminder, we asked respondents to estimate the share of hypothesis tests that achieve at least 50% power

[18]The share is the proportion of all reported tests that have power at least as high as displayed on the horizontal axis.

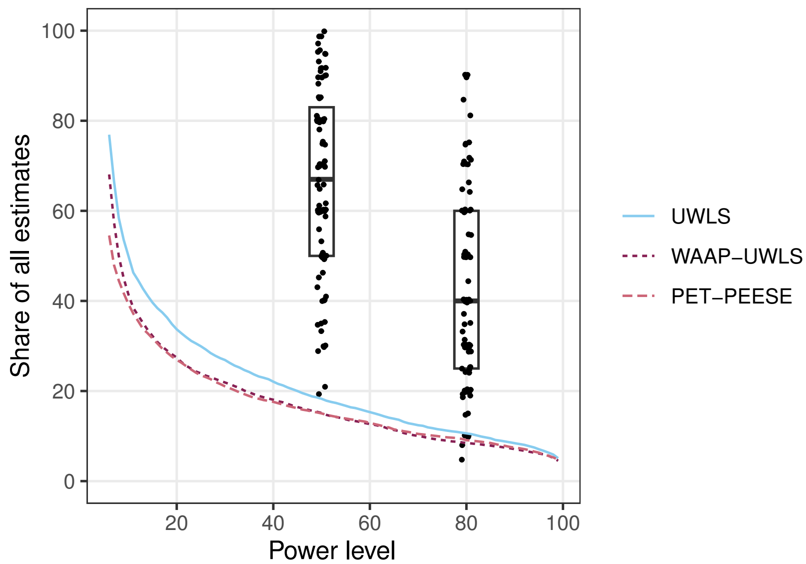

Figure 2: Retrospective power analyses and the view from political methodology. Lines represent the share of estimates by power level, using three different approaches to estimate the population mean effect. Black dots and boxplots represent the distribution of responses in the expert survey.

and at least 80% power within all papers published since 2000 in the top journals in political science and related fields. Their responses are shown in black dots and the boxplots mark the quartiles for each distribution. The curves for our power calculations barely overlap with the expert responses: methodologists appear to overestimate the power of hypothesis tests by a large margin.

The gap between expert assessment and estimated power could stem from several factors. First, methodologists could suffer from recency bias, as power appears to have increased slowly over time (see below). Second, the studies in our sample may not match the studies, journals, or subfields that survey respondents had in mind when contemplating our question. Finally, there remains considerable uncertainty in our own estimates of power. Taken together, these factors may explain part—but probably not all—of the apparent over-optimism displayed by political methodologists. In the next section, we show that power appears to have increased over time, albeit very slowly. As such, it seems possible that part of the over-optimism on the part of political methodologists may be due to recency bias.

Extensions and robustness

We conducted several complementary analyses and robustness checks. In the online appendix, we report alternative tests which confirm that our main findings are not driven by outliers. Here, we report results that refine the interpretation and add nuance to our findings.

Minimum effects of interest

For this paper we calculated power retrospectively based on meta-analytic estimates of population mean effects. In practice, however, experimentalists in political science often design their studies based on prospective power calculations anchored by the size of a "minimum effect of interest" (MEI). Our results can tell us if studies are sufficiently powered to detect consensus effects in the literature; they do not directly tell us if studies are well powered to detect the MEIs targeted by individual researchers. But even if we cannot directly characterize power with respect to MEIs, our results suggest useful informal bounds. Consider two cases.

First, if MEIs are systematically smaller than population mean effects, our analyses would overstate the power of political science. In that case, our results could be interpreted as an upper bound for the average power of the studies in our sample; the problems that we highlight in this paper would be even more worrying.

Second, if MEIs are systematically larger than meta-analytic estimates, our analyses would understate the level of power in political science research.[19]

In Figure 3 we probe the sensitivity of our conclusions to this potential issue by artificially inflating all population mean effect estimates by a factor of five and replicating our analyses.[20] After this five-fold inflation the power estimates align more closely with the expert survey responses, which gives an approximate sense of the degree of overestimation of power in the

[19]If MEIs are systematically larger than meta-analytic estimates, the vast majority of political science research should also yield null results, which we do not see in Figure 1.

[20]This analysis is inspired by Kollepara et al. [2021].

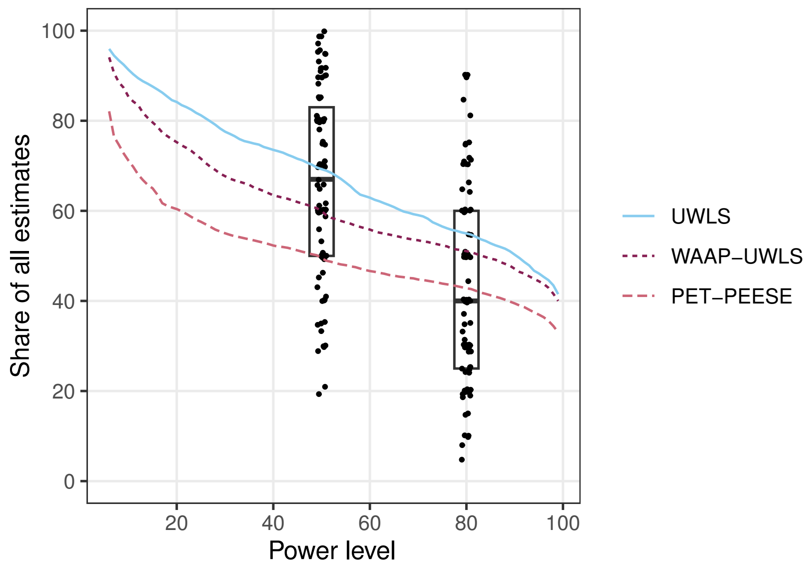

Figure 3: Share of estimates by power level, after each population mean effect estimate is inflated by a factor of 5

expert community.

Substantively, we find that one third of tests fail to reach 50% power even after the inflation. In other words, even if researchers were designing studies to target effect sizes five times larger than the consensus estimates reported in the literature, quantitative political science research would still be greatly underpowered.

Another way to circumvent the problem posed by the disconnect between population mean effects and MEIs is to shift the focus away from power, toward the magnitude and sign of point estimates.

Magnitude and sign

Low-powered studies that are subsequently filtered for statistical significance are more likely to report effects that are inflated or of the incorrect sign [Gelman and Tuerlinckx, 2000, Gelman and Carlin, 2014, Ioannidis et al., 2017]. We test for this in our data, essentially offering empirical analogues to the Type M and Type S error rates described in Gelman and Carlin [2014].

Among our 16649 hypothesis tests, 7775 are statistically significant at the 0.05 level. Among these significant effects, 15% have a different sign than the consensus UWLS estimate.[21] To see if reported estimates tend to be larger than population mean effects, we simply divide individual estimates by the corresponding UWLS estimate of the mean. In the subset of estimates that are both statistically significant and in the "correct" direction, the median estimate-to-UWLS ratio is 3.0. Thus, the significant estimates in our sample are likely to be about 3.0 times too big or wrongly signed.

These calculations are also related to the "Exaggeration Factor" which Ioannidis et al. [2017] calculate as $|\frac{\bar{\beta}-\beta^{\text{UWLS}}}{\beta^{\text{UWLS}}}| + 1$, where $\bar{\beta}$ is the average of estimates in a meta-analysis, β^{UWLS} is the UWLS estimate for that meta-analysis. The median exaggeration factor in our dataset is 1.9, which is close to what these authors found in economics.[22]

Sample composition

One potential objection relates to the composition of our dataset. In our expert survey, respondents stated expectations about power levels for all estimates published in well-ranked

[21]While low power interacted with significance filtering can lead to these "errors," they do not depend in any way on how we calculated power. These results come simply from comparing the meta-analytic population mean effect to each significant coefficient within the meta-analysis.

[22]Note that this exaggeration factor is the same quantity reported in Ioannidis et al. [2017], but that it is not equivalent to the Type M error of Gelman and Carlin [2014]. One important difference is that the exaggeration factor is computed based on the full set of estimates, whereas Type M error is a characteristic of the subset of statistically significant estimates. Another difference is that unlike for Type M (see Figure 2 of Gelman and Carlin [2014]), there is no simple mechanical relationship between power and the exaggeration ratio. This makes comparisons across disciplines difficult, since such comparisons will depend on rates of selective reporting, bias in the estimate of UWLS, and other factors.

journals over the past two decades. However, our sample includes only a small share of those estimates because few studies ever get aggregated in meta-analyses (see appendix A for the full list of meta-analyses in the sample), and the estimates are not all drawn from highly-ranked journals.[23]

To address this concern, we asked our experts if they believed estimates in meta-analyses to be better- or worse-powered than all other estimates. 75% of them think that studies included in meta-analyses (in our sample) have about the same power or higher power than other studies (out of our sample). This should make one comfortable with our comparison of expert survey estimates of power in political science and our power results drawn from meta-analyses in Figure 2. This also implies that our approach of generalizing from tests in meta-analyses to those not in meta-analyses is either accurate or overstates the power of hypothesis tests in political science in general. Appendix C reports further details about our survey and this question.

Heterogeneity across research questions

Figure 2 merges all our estimates together, but there might be variation in power levels across meta-analyses. We examine this by graphing the median power within each of the 351 meta-analyses in our dataset. Figure 4 reveals substantial heterogeneity. While most research areas have median power below 10%, a substantial share has median power above 80%. This finding is consistent with previous assessments of power in neuroscience [Button et al., 2013, Nord et al., 2017] and economics [Ioannidis et al., 2017].[24]

One factor which may explain the high level of heterogeneity in power across study areas

[23] As we noted above, results from other disciplines suggest that it is not obvious that publications in higher ranked journals should be better powered than others. Investigating this question would be an interesting path for future research.

[24] Using a somewhat different method, Szucs and Ioannidis [2017] find similarly low power in psychology, cognitive neuroscience, and medically-oriented neuroscience.

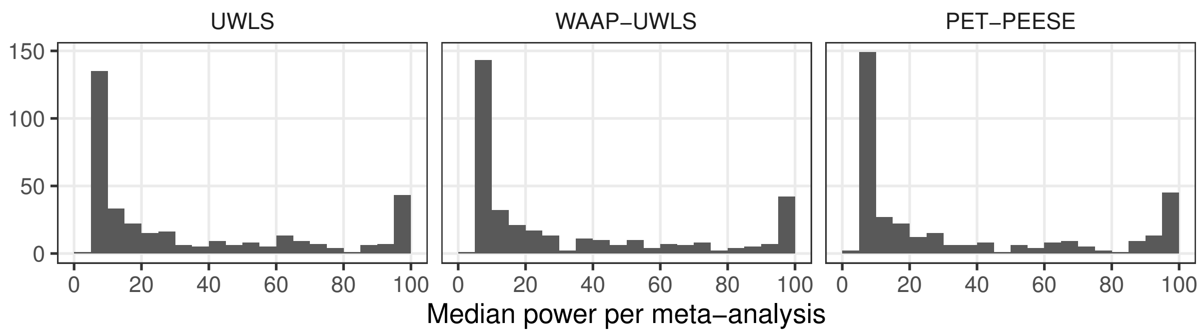

Figure 4: Histogram of median power per meta-analysis, using three different approaches to estimate the population mean effect.

is that different journals may apply different evaluation standards to research. For example, if "top" journals seek to publish studies with better statistical power (which is unclear), this might account for part of the heterogeneity that we observe. Unfortunately, we do not have the data required to test this idea empirically in political science. However, prior research in economics suggests that studies in top-tier journals do *not* consistently exhibit higher statistical power [Askarov et al., 2023], and that highly selective venues may in fact fall prey to a "winner's curse" phenomenon [Young et al., 2008, Costa-Font et al., 2013].

Power by subfield

We now turn to the question of whether power varies across subfields. To do so, we coded each meta-analytic articles in our sample into crude subfields of: American politics, comparative politics, international relations, political economy, and public administration.[25] Table 1 shows the number of individual estimates, meta-analytic estimates, and meta-analysis articles we found in each subfield. The smallest number of estimates is in American politics, whereas International Relations and Political Economy lead in terms of meta estimates and individual estimates, respectively. Given our data collection protocol, we cannot claim that sample is balanced, but it has reasonable coverage across some of the disciplines major fields.

[25]Two coauthors independently coded the subfields and any disagreements were resolved through a discussion that reached consensus.

Table 1: Number of estimates and articles by subfield

	Estimates	Meta estimates	Articles
American Politics	685	25	7
Comparative Politics	2150	13	11
International Relations	4438	264	5
Political Economy	8045	18	11
Public Administration	1331	31	12

Figure 5 shows two ways of looking at the distribution of power by subfield. First, the dotted lines show the distribution of median power, where the median is taken for each of our 351 meta-analytic estimates. Second, the solid lines show the distribution of median power, where the median is taken at the level of our 46 meta-analytic articles. Since tests within a meta-analytic article tend to be conducted using similar data and methods, it arguably makes more sense to look at the article-based estimates (solid line) in order to avoid basing our conclusions on a few meta-analytic articles which could have outlying results.

Overall, Figure 5 shows that (a) power is low in all subfields, and (b) there is substantial variation *within* subfields. Of course, we have to be careful when interpreting these graphs, to avoid making an underpowered comparison about power. Nevertheless, our impressionistic sense is that the problem of lower power is not confined to any one particular subfield. Instead, there appears to be low power across the discipline.

Power over time

We examine how power has changed over time based on the publication year of each test. To reduce noise, we group each test into five-year bins from 1990 until 2020 and then calculate mean and median UWLS power per bin.[26]

Power has increased over time but remains low. Mean power rose from 16% in the first period to 28% in the last. Median power is lower and has increased less, rising from 7% to 12%. The gap between mean and median power reveals that some of the increase in mean

[26] We do not examine tests before 1990 or after 2020 as the data are sparse.

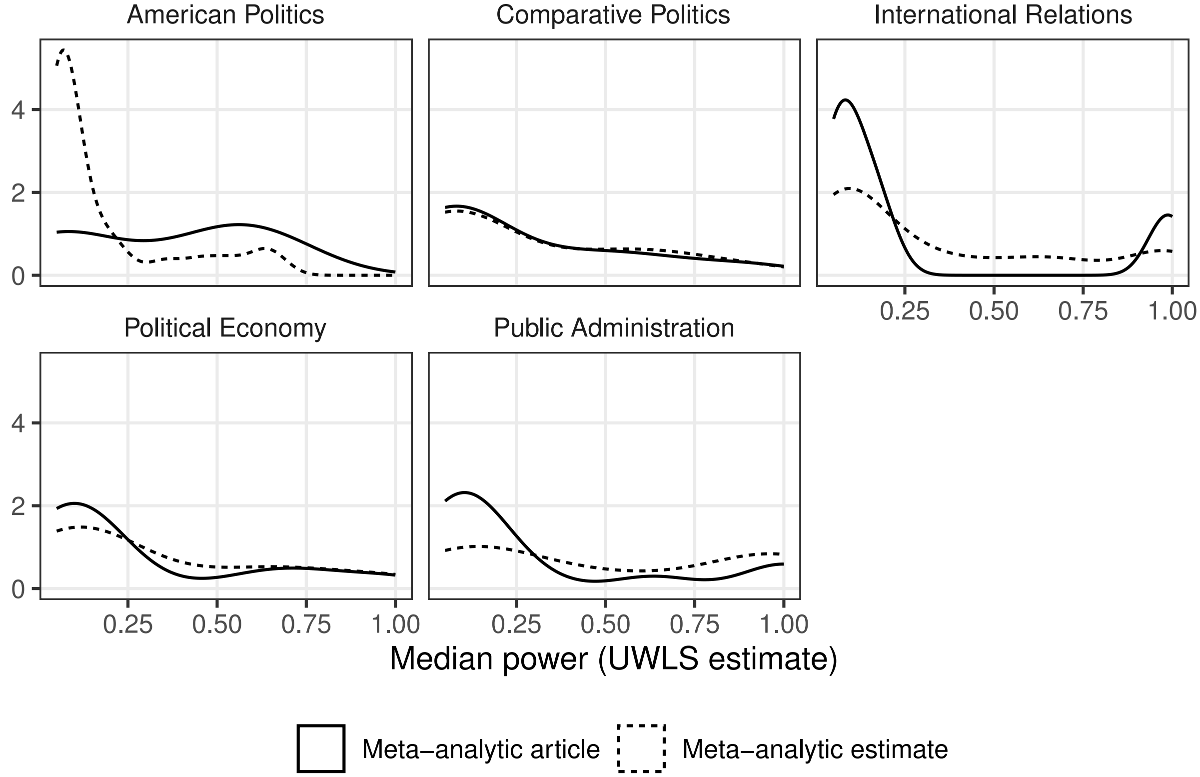

Figure 5: Densities of median power per subfield.

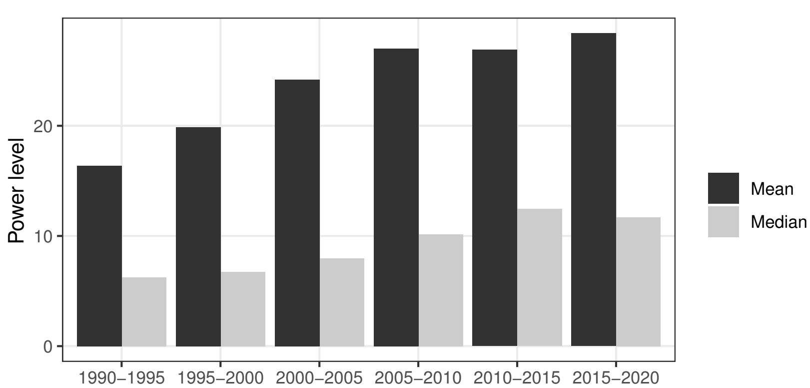

Figure 6: Power over time (UWLS).

power is being driven by increases in power at the top end of the distribution rather than a distribution-wide shift.

Why is power low and what can we do about it?

While this paper has demonstrated that power is low, the sources of this problem remain less clear. Perhaps the most useful diagnostic fact is that most of the variation in power in our sample occurs across rather than within meta-analyses, with some having either very high or very low median power (see Figure 4). Differences of this size across areas of the discipline are more likely due to differences in the magnitude and stability of the effects under study rather than sample sizes. Put simply, some research areas may be studying larger and more consistent treatment effects than others. If many of the effects that political scientists study are very small, then it is not surprising to find that significant estimates often have the "wrong" sign, and that significant and "correctly-signed" effects are inflated.

At least four broad categories of interventions could be helpful in addressing these issues: increasing precision, limiting selection on significance, facilitating the evaluation of published research, and relying on theory development and qualitative methods. These interventions are not new, but they are still too rare in political science so their value bears repeating [Nyhan, 2015, Malhotra, 2021, Williamson et al., 2022].

First, interventions that directly increase the precision of statistical estimates can be useful. One obvious way forward would be to assemble bigger datasets, perhaps by incentivizing team-based data collection efforts. In doing so, we should be mindful that large data collection efforts are expensive, and that our quest for higher power could reinforce the discipline's existing resource inequities and hinder more speculative research agendas. In addition to collecting more data, researchers can also increase the precision of their estimates by leveraging research designs and measurement strategies that prioritize precision such as pre-post, within-subject, or repeated measures experimental designs and the use of indexes or scales combining multiple measures rather than single-item outcomes [Clifford et al., 2021, Broockman et al., 2017, Hainmueller et al., 2014, Ansolabehere et al., 2008].

Second, interventions that limit the various forms of selection on significance are important, as they ensure that low powered studies do not introduce bias when they enter the literature. Pre-registration and registered reports[27] can be very useful here [Nosek et al., 2018]. While not always possible, registered reports can limit selection on significance both during data analysis and during the publication process. More political science journals should experiment with them. More also needs to be done to reduce selection on statistical significance and the file-drawer problem [Laitin, 2013].[28] A necessary shift in this direction involves disciplinary norms: political science needs to do more to value null findings. Researchers need to develop the skills required to properly frame and analyze null findings [Williamson et al., 2022], using tools such as sensitivity analyses and equivalence tests [Rainey, 2014].

Third, interventions that allow us to better judge the quality of already published research are important both for interpreting past work and for checking quality going forward. Data and code sharing are already common in political science [King, 2003], though we could do more to share full code pipelines, from data acquisition and cleaning to analysis [Nosek et al., 2015]. Less common but very useful are replications of past work, and more could be done to promote and value replications [Camerer et al., 2018].[29]

[27]A registered report is reviewed at the design stage and papers are conditionally accepted before data is gathered. The *Journal of Experimental Political Science* has been accepting pre-registered reports since its creation, and the *Journal of Politics* is currently piloting this approach. We note, however, that *Comparative Political Studies* piloted pre-registered reports but soon abandoned this practice [Findley et al., 2016]

[28]The *Political Studies Review* publishes the Null Hypothesis section, which is dedicated to research notes reporting a null finding. More journals should follow suit.

[29]For instance, the Institute for Replication (I4R) "works to improve the credibility of science by systematically reproducing and replicating research findings in leading academic

More generally, many subfields of political science still have a culture that rewards single-authored publications and lone researchers pursuing more or less standalone research programs. Moreover, editors and peer reviewers often enforce a requirement of simultaneous theoretical and empirical innovation in research articles. We worry that this one-off incentive structure may lead to a proliferation of theories, which are all to be tested using single-use and often low powered research designs. Many of the problems noted in this paper would be lessened if journal editors, peer reviewers, and promotion committees did more to incentivize researchers to work in teams on shared questions tested using replicated and collaborative empirical studies.[30] We are hopeful that this would help us develop more durable and cumulative knowledge of politics.

Finally, we must accept that it is sometimes impossible to conduct an adequately powered hypothesis test. For example, researchers may simply not be able to design a well-powered study to detect small differences between a limited number of states.[31] Moreover, standard practices like estimating models with interaction terms to test conditional theories or treatment effect heterogeneity can often do more harm than good, since it can dramatically increase required sample sizes [Gelman, 2018]. In such cases, a useful response is to turn to theory: derive new implications and test them where data are richer. Alternatively, researchers could draw on qualitative or mixed-method approaches, using interviews, ethnographies, or archival work to shed light on causal mechanisms or heterogeneity [Small, 2011, Gerring, 2017]. In our view, these methods are often better entry points than underpowered quantitative research designs.

journals." See `https://i4replication.org/`.

[30] See the *Metaketa Initiative* [Dunning et al., 2019].

[31] Indeed, Doucette [2023] conducts Monte Carlo simulations to study the statistical power of over 1100 studies on the effects of democracy using cross-national data. The author concludes that these studies generally lack the power to detect anything but very strong homogeneous effects.

Conclusion

Statistical power is an important and neglected aspect of quantitative political science research. This paper has shown that power in political science is typically very low, that it is improving only slowly, and that this issue should be taken seriously by political scientists and methodologists. The problem of low power combined with the selection on statistical significance that occurs in our literature means that published and significant effects are likely to be much larger in print than in reality. Our best estimate is that published and statistically significant results in political science are around three times larger than the true effect under study.

Low power poses a fundamental challenge to researchers in political science. We must resist the temptation of business as usual, and avoid commiting the *what does not kill my statistical significance makes it stronger* fallacy [Gelman, 2017].[32] Instead, our research community must address the problems of low power and selection on significance with institutional, methodological, and theoretical remedies.

[32]One example of this fallacy would be to argue that a point estimate gives strong support for one's theory because it achieved statistical significance *despite* a small sample.

Acknowledgments

We thank Stephen Ansolabehere, Elissa Berwick, Aaron Erlich, Patrick Fournier, Bart Harvey, Andrew Philips, Cyrus Samii, as well as the software authors who made this work possible (see our Software Bibliography, appendix F) and all researchers who shared data with us (see appendix A). Sara Duran-Bettancour, Khaoula El Khalil, and Aden Morton-Ferguson provided research assistance. We would also like to thank participants at talks given at the University of Toronto, University of Zurich, American University, MAER-Net, and APSA 2022 for feedback.

References

Adam Altmejd, Anna Dreber, Eskil Forsell, Juergen Huber, Taisuke Imai, Magnus Johannesson, Michael Kirchler, Gideon Nave, and Colin Camerer. Predicting the replicability of social science lab experiments. *PLOS ONE*, 14(12):e0225826, 2019.

Stephen Ansolabehere, Jonathan Rodden, and James M Snyder. The strength of issues: Using multiple measures to gauge preference stability, ideological constraint, and issue voting. *American Political Science Review*, 102(2):215–232, 2008.

Zohid Askarov, Alex Doucouliagos, Hristos Doucouliagos, and T. D. Stanley. Selective and (mis)leading economics journals: Meta-research evidence. *Journal of Economic Surveys*, 2023.

Monya Baker. Reproducibility crisis. *Nature*, 533(26):353–66, 2016.

Adam J. Berinsky, James N. Druckman, and Teppei Yamamoto. Publication biases in replication studies. *Political Analysis*, 29(3):370–384, 2021.

Abel Brodeur, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, 8(1):1–32, 2016.

Abel Brodeur, Nikolai Cook, and Anthony Heyes. Methods matter: P-hacking and publication bias in causal analysis in economics. *American Economic Review*, 110(11):3634–60, 2020.

David E. Broockman, Joshua L. Kalla, and Jasjeet S. Sekhon. The design of field experiments with survey outcomes: A framework for selecting more efficient, robust, and ethical designs. *Political Analysis*, 25(4):435–464, 2017.

Katherine S. Button, John Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S.J. Robinson, and Marcus R. Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376, 2013.

Colin F Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A Nosek, Thomas Pfeiffer, et al. Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature human behaviour*, 2(9):637–644, 2018.

Scott Clifford, Geoffrey Sheagley, and Spencer Piston. Increasing precision without altering treatment effects: Repeated measures designs in survey experiments. *American Political Science Review*, 115(3):1048–1065, 2021.

J Cohen and BB Wolman. Handbook of clinical psychology, 1965.

Joan Costa-Font, A. McGuire, and T. D. Stanley. Publication selection in health policy research: The winner's curse hypothesis. *Health Policy*, 109:78–87, 2013.

Jonathan Stavnskær Doucette. What can we know about the effects of democracy using cross-national data? *QOG Working Paper Series*, (11), 2023.

Thad Dunning, Guy Grossman, Macartan Humphreys, Susan D Hyde, Craig McIntosh, and Gareth Nellis. *Information, accountability, and cumulative learning: Lessons from Metaketa I.* Cambridge University Press, 2019.

Matthias Egger, George Davey Smith, Martin Schneider, and Christoph Minder. Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109):629–634, 1997.

Michael G Findley, Nathan M Jensen, Edmund J Malesky, and Thomas B Pepinsky. Can results-free review reduce publication bias? the results and implications of a pilot study. *Comparative Political Studies*, 49(13):1667–1703, 2016.

Andrew Gelman. The "what does not kill my statistical significance makes it stronger" fallacy. https://statmodeling.stat.columbia.edu/2017/02/06/not-kill-statistical-significance-makes-stronger-fallacy/ Visited 2023-03-30., 2017.

Andrew Gelman. You need 16 times the sample size to estimate an interaction than to estimate a main effect. https://statmodeling.stat.columbia.edu/2018/03/15/need-16-times-sample-size-estimate-interaction-estimate-main-effect/ Visited 2023-04-06., 2018.

Andrew Gelman. Don't calculate post-hoc power using observed estimate of effect size. *Annals of Surgery*, 269(1):e9–e10, 2019.

Andrew Gelman and John Carlin. Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–651, 2014.

Andrew Gelman and Francis Tuerlinckx. Type S error rates for classical and bayesian single and multiple comparison procedures. *Computational Statistics*, 15(3):373–390, 2000.

Alan Gerber and Neil Malhotra. 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, 2008.

John Gerring. Qualitative methods. *Annual review of political science*, 20:15–36, 2017.

Adam Gorajek and Benjamin A. Malin. Comment on "star wars: The empirics strike back". Technical report, Federal Reserve Bank of Minneapolis, 2021. URL https://doi.org/10.21034/sr.629.

Jens Hainmueller, Daniel J Hopkins, and Teppei Yamamoto. Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political analysis*, 22(1):1–30, 2014.

John M Hoenig and Dennis M Heisey. The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55(1):19–24, 2001.

John Ioannidis, T.D. Stanley, and Hristos Doucouliagos. The power of bias in economics research. *The Economic Journal*, 127(605):F236–F265, 2017.

Journal of Politics. Pre-registration guidelines, 2022. URL https://www.journals.uchicago.edu/journals/jop/pre-registration.

Gary King. The future of replication. *International Studies Perspectives*, 2003.

Richard A. Klein, Michelangelo Vianello, Fred Hasselman, Byron G. Adams, Reginald B. Adams Jr., Sinan Alper, Mark Aveyard, Jordan R. Axt, Mayowa T. Babalola, and Štěpán Bahnik. Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490, 2018.

Pratyush K. Kollepara, Alexander F. Siegenfeld, Nassim Nicholas Taleb, and Yaneer Bar-Yam. Unmasking the mask studies: why the effectiveness of surgical masks in preventing respiratory infections has been underestimated. *Journal of Travel Medicine*, 28(7):taab144, Oct 2021. ISSN 1195-1982, 1708-8305. doi: 10.1093/jtm/taab144.

Amanda Kvarven, Eirik Strømland, and Magnus Johannesson. Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, 4 (4):423–434, 2020.

David D Laitin. Fisheries management. *Political Analysis*, 21(1):42–47, 2013.

Neil Malhotra. Threats to the scientific credibility of experiments: Publication bias and p-hacking. In James N. Druckman and Donald P. Green, editors, *Advances in Experimental Political Science*, pages 354–368. Cambridge University Press, 2021. doi: 10.1017/9781108777919.025.

Camilla L Nord, Vincent Valton, John Wood, and Jonathan P Roiser. Power-up: a reanalysis of 'power failure' in neuroscience using mixture modeling. *Journal of Neuroscience*, 37(34): 8051–8061, 2017.

Brian A Nosek, George Alter, George C Banks, Denny Borsboom, Sara D Bowman, Steven J Breckler, Stuart Buck, Christopher D Chambers, Gilbert Chin, Garret Christensen, et al. Promoting an open research culture. *Science*, 348(6242):1422–1425, 2015.

Brian A Nosek, Charles R Ebersole, Alexander C DeHaven, and David T Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11): 2600–2606, 2018.

Brendan Nyhan. Increasing the credibility of political science research: A proposal for journal reforms. *PS: Political Science & Politics*, 48(S1):78–83, 2015.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

Carlisle Rainey. Arguing for a negligible effect. *American Journal of Political Science*, 58 (4):1083–1091, 2014.

Carlisle Rainey, Harley Roe, Qing Wang, and Hao Zhou. Data and code availability in political science publications from 1995 to 2022, 2024.

Mario Luis Small. How to conduct a mixed methods study: Recent trends in a rapidly growing literature. *Annual review of sociology*, 37:57–86, 2011.

T.D. Stanley. Limitations of pet-peese and other meta-analysis methods. *Social Psychological and Personality Science*, 8(5):581–591, 2017.

T.D. Stanley and Hristos Doucouliagos. Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, 5(1):60–78, 2014.

T.D. Stanley and Hristos Doucouliagos. Harnessing the power of excess statistical significance: Weighted and iterative least squares. *Psychological Methods.*, Online ahead of print 12 May 2022. DOI: 10.1037/met0000502, 2022.

T.D. Stanley, Hristos Doucouliagos, and John Ioannidis. Finding the power to reduce publication bias. *Statistics in Medicine*, 36(10):1580–1598, 2017.

T.D. Stanley, Evan C. Carter, and Hristos Doucouliagos. What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, 144(12):1325–1346, 2018.

Denes Szucs and John Ioannidis. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, 15(3): e2000797, 2017.

Eva Vivalt. Specification searching and significance inflation across time, methods and disciplines. *Oxford Bulletin of Economics and Statistics*, 81(4):797–816, 2019.

Scott Williamson, Andrea Dillon, Jens Hainmueller, Dominik Hangartner, Michael Hotard, David D. Laitin, Duncan Lawrence, and Jeremy Weinstein. Learning from null effects: A bottom-up approach. *Political Analysis*, pages 1–9, 2022.

Yefeng Yang, Alfredo Sánchez-Tójar, Rose E O'Dea, Daniel WA Noble, Julia Koricheva, Michael D Jennions, Timothy H Parker, Malgorzata Lagisz, and Shinichi Nakagawa. Publication bias impacts on effect size, statistical power, and magnitude (type m) and sign (type s) errors in ecology and evolutionary biology. *BMC Biology*, 21(71), 2022.

N. S. Young, J. P. A. Ioannidis, and O. Al-Ubaydli. Why current publication practices may distort science. *PLoS Medicine*, 5:134–145, 2008. doi: 10.1371/journal.pmed.0050201.

Biographical statement

Vincent Arel-Bundock is an Associate Professor of Political Science at the Université de Montréal, in Montréal, Canada.

Ryan C. Briggs is an Associate Professor of Political Science and Development Studies at the University of Guelph, in Guelph, Canada.

Marco M. Aviña is a PhD student in the Department of Government at Harvard University, in Cambridge, Massachusetts, USA.

TD Stanley is a Meta-Analysis Professor at the Deakin Business School, in Victoria, Australia.

Chris Doucouliagos is a Professor of Economics at the Deakin Business School, in Victoria, Australia.

A Data collection

Our data collection proceeded in three main steps:

1. Build a list of journals to search for meta-analyses.
2. Identify meta-analyses published since 2000 in each of those journals.
3. Collect replication materials for each of these meta-analyses.

Our search of journals resulted in 85 entries fitting our definition of a meta-analytic paper. We were able to find public data repositories with complete information for 21 of these papers, and we extracted estimates and standard errors from the tables or figures of another 7 papers. For the remaining 32 papers, we contacted authors directly through email to request access to their replication materials. In total, authors from 4 of these papers replied positively to our request. This resulted in a complete sample of estimates and standard errors for 32 meta-analytic papers (21 + 7 + 4). We then merged this dataset with data for 14 separate meta-analytic papers collected by one of the coauthors of the present paper for a separate project. Overall, this resulted in a dataset of 46 (32 + 12) papers. In total, we assembled a dataset with 16649 point estimates and standard errors grouped within 351 meta-analyses.

The list of every journal we surveyed for meta-analyses and the list of every meta-analytic article included in our analysis follow. Appendix B provides details on our data cleaning.

Journals surveyed for meta-analyses

Administration and Society; African Affairs; American Journal of International Law; American Journal of Political Science; American Political Science Review; Annals of the American Academy of Political and Social Science; Annual Review of Political Science; Australian Journal of Public Administration; British Journal of Political Science; British Journal of Politics and International Relations; Bulletin of the Atomic Scientists; Cambridge Review of International Affairs; Canadian Public Administration; Canadian Public Policy; Citizenship Studies; Climate Policy; Common Market Law Review; Communist and Post-Communist Studies; Comparative Political Studies; Comparative Politics; Conflict Management and Peace Science; Contemporary Economic Policy; Cooperation and Conflict; Critical Policy Studies; Democratization; Electoral Studies; Emerging Markets Finance and Trade; Environment and Planning C: Government and Policy; Environmental Politics; Ethics and International Affairs; European Journal of International Law; European Journal of International Relations; European Journal of Political Economy; European Journal of Political Research; European Union Politics; Geopolitics; Global Environmental Politics; Global Governance; Global Policy; Globalizations; Governance: An International Journal of Policy Administration and Institutions; Human Rights Quarterly; Human Service Organizations Management Leadership and Governance; International Affairs; International Affairs; International Interactions; International Journal of Public Administration; International Journal of Transitional Justice; International Organization; International Peacekeeping; International Political Sociology; International Public Management Journal; International Relations; International Review of Administrative Sciences; International Security; International Studies Perspectives;

International Studies Quarterly; International Studies Review; JCMS: Journal of Common Market Studies; Journal of Accounting and Public Policy; Journal of Accounting and Public Policy; Journal of Comparative Policy Analysis; Journal of Conflict Resolution; Journal of Democracy; Journal of European Integration; Journal of European Public Policy; Journal of European Social Policy; Journal of Homeland Security and Emergency Management; Journal of Peace Research; Journal of Policy Analysis and Management; Journal of Politics; Journal of Public Administration Research and Theory; Journal of Public Policy; Journal of Social Policy; Journal of Strategic Studies; Journal of the Japanese and International Economies; Latin American Politics and Society; Lex localis – Journal of Local Self-Government; Local Government Studies; Local Government Studies; Millennium: Journal of International Studies; New Political Economy; Nonprofit Management and Leadership; Pacific Review; Party Politics; Perspectives on Politics; Philosophy and Public Affairs; Policy and Politics; Policy and Politics; Policy and Society; Policy and Society; Policy Sciences; Policy Studies; Policy Studies Journal; Policy Studies Journal; Political Analysis; Political Behavior; Political Communication; Political Geography; Political Psychology; Political Research Quarterly; Political Studies; Political Theory; Politics; Politics and Society; Post-Soviet Affairs; PS: Political Science and Politics; Public Administration; Public Administration and Development; Public Administration Review; Public Choice; Public Management Review; Public Money and Management; Public Opinion Quarterly; Public Performance and Management Review; Public Personnel Management; Public Policy and Administration; Regulation and Governance; Review of International Organizations; Review of International Political Economy; Review of International Studies; Review of Policy Research; Review of Public Personnel Administration; Review of World Economics; Science and Public Policy; Security Dialogue; Security Studies; Social Policy and Administration; Social Science Quarterly; Socio-Economic Review; Studies in Comparative International Development; Studies in Conflict and Terrorism; Survival; Terrorism and Political Violence; American Review of Public Administration; Third World Quarterly; Transylvanian Review of Administrative Sciences; VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations; West European Politics; World Economy; World Politics

Meta-analyses included in the sample

Ahmadov, Anar K. 2014. "Oil, Democracy, and Context: A Meta-Analysis." *Comparative Political Studies* 47(9): 1238–67.

Arceneaux, Kevin, and David W. Nickerson. 2009. "Who Is Mobilized to Vote? A Re-Analysis of 11 Field Experiments." *American Journal of Political Science* 53(1): 1–16.

Askarov, Zohid, Hristos Doucouliagos, Martin Paldam, and T.D. Stanley. 2021. "Rewarding Good Political Behavior: US Aid, Democracy, and Human Rights." *European Journal of Political Economy*: 102089.

Awan, Sahar, Germà Bel, and Marc Esteve. 2020. "The Benefits of PSM: An Oasis or a Mirage?" *Journal of Public Administration Research & Theory* 30(4): 619–35.

Balliet, Daniel. 2010. "Communication and Cooperation in Social Dilemmas: A Meta-Analytic Review." *Journal of Conflict Resolution* 54(1): 39–57.

Balliet, Daniel, Joshua M. Tybur, Junhui Wu, Christian Antonellis, and Paul AM Van Lange. 2018. "Political Ideology, Trust, and Cooperation: In-Group Favoritism among Republicans and Democrats during a US National Election." *Journal of Conflict Resolution* 62(4): 797–818.

Barnhart, Joslyn N., Robert F. Trager, Elizabeth N. Saunders, and Allan Dafoe. 2020. "The Suffragist Peace." *International Organization* 74(4): 633–70.

Belle, Nicola, and Paola Cantarelli. 2017. "What Causes Unethical Behavior? A Meta-Analysis to Set an Agenda for Public Administration Research." *Public Administration Review* 77(3): 327–39.

Bhatti, Yosef, Jens Olav Dahlgaard, Jonas Hedegaard Hansen, and Kasper M. Hansen. 2019. "Is Door-to-Door Canvassing Effective in Europe? Evidence from a Meta-Study across Six European Countries." *British Journal of Political Science* 49(1): 279–90.

Blair, Graeme, Darin Christensen, and Aaron Rudkin. 2020. "Do Commodity Price Shocks Cause Armed Conflict? A Meta-Analysis of Natural Experiments." *American Political Science Review* 115(2): 709–116.

Broderstad, Troy Saghaug. 2018. "A Meta-Analysis of Income and Democracy." *Democratization* 25(2): 293–311.

Burke, Brian L., Spee Kosloff, and Mark J. Landau. 2013. "Death Goes to the Polls: A Meta-Analysis of Mortality Salience Effects on Political Attitudes." *Political Psychology* 34(2): 183–200.

Colagrossi, Marco, Domenico Rossignoli, and Mario A. Maggioni. 2020. "Does Democracy Cause Growth? A Meta-Analysis (of 2000 Regressions)." *European Journal of Political Economy* 61: 101824.

Dinesen, Peter Thisted, Merlin Schaeffer, and Kim Mannemar Sønderskov. 2020. "Ethnic Diversity and Social Trust: A Narrative and Meta-Analytical Review." *Annual Review of Political Science* 23(1): 441–65.

Ding, Fangda, Jiahuan Lu, and Norma M. Riccucci. 2021. "How Bureaucratic Representation Affects Public Organizational Performance: A Meta-Analysis." *Public Administration Review* 81(6): 1003–18.

Doucouliagos, Chris, and Mehmet Ali Ulubaşoğlu. 2006. "Economic Freedom and Economic Growth: Does Specification Make a Difference?" *European Journal of Political Economy* 22(1): 60–81.

Efendic, Adnan, Geoff Pugh, and Nick Adnett. 2011. "Institutions and Economic Performance: A Meta-Regression Analysis." *European Journal of Political Economy* 27(3): 586–99.

Eshuis, Jasper et al. 2021. "The Effect of the EU-Brand on Citizens' Trust in Policies: Replicating an Experiment." *Public Administration Review* 81(4): 776–86.

Gerrish, Ed. 2016. "The Impact of Performance Management on Performance in Public Organizations: A Meta-Analysis." *Public Administration Review* 76(1): 48–66.

Green, Donald P., and Alan S. Gerber. 2019. *Get Out the Vote: How to Increase Voter Turnout.* Brookings Institution Press.

Greenberg, David H., Charles Michalopoulos, and Philip K. Robin. 2006. "Do Experimental and Nonexperimental Evaluations Give Different Answers about the Effectiveness of Government-Funded Training Programs?" *Journal of Policy Analysis & Management* 25(3): 523–52.

Heimberger, Philipp. 2020. "Does Economic Globalization Affect Government Spending? A Meta-Analysis." *Public Choice* 187(3): 349–74.

———. 2021. "Corporate Tax Competition: A Meta-Analysis." *European Journal of Political Economy* 187(349–374).

Heinemann, Friedrich, Marc-Daniel Moessinger, and Mustafa Yeter. 2018. "Do Fiscal Rules Constrain Fiscal Policy? A Meta-Regression-Analysis." *European Journal of Political Economy* 51: 69–92.

Holbein, John B., Marcos A. Rangel, Raeal Moore, and Michelle Croft. 2021. "Is Voting Transformative? Expanding and Meta-Analyzing the Evidence." *Political Behavior*: 1–30.

Homberg, Fabian, Dermot McCarthy, and Vurain Tabvuma. 2015. "A Meta-Analysis of the Relationship between Public Service Motivation and Job Satisfaction." *Public Administration Review* 75(5): 711–22.

Houck, Shannon C., and Lucian Gideon Conway. 2019. "Strategic Communication and the Integrative Complexity-Ideology Relationship: Meta-Analytic Findings Reveal Differences Between Public Politicians and Private Citizens in Their Use of Simple Rhetoric." *Political Psychology* 40(5): 1119–41.

Incerti, Trevor. 2020. "Corruption Information and Vote Share: A Meta-Analysis and Lessons for Experimental Design." *American Political Science Review* 114(3): 761–74.

Kalla, Joshua L., and David E. Broockman. 2018. "The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments." *American Political Science Review* 112(1): 148–66.

Lau, Richard R., Lee Sigelman, and Ivy Brown Rovner. 2007. "The Effects of Negative Political Campaigns: A Meta-Analytic Reassessment." *Journal of Politics* 69(4): 1176–1209.

Li, Quan, Erica Owen, and Austin Mitchell. 2018. "Why Do Democracies Attract More or Less Foreign Direct Investment? A Metaregression Analysis." *International Studies Quarterly* 62(3): 494–504.

Lu, Jiahuan. 2016. "The Philanthropic Consequence of Government Grants to Nonprofit Organizations." *Nonprofit Management & Leadership* 26(4): 381–400.

———. 2018. "Fear the Government? A Meta-Analysis of the Impact of Government Funding on Nonprofit Advocacy Engagement." *American Review of Public Administration* 48(3): 203–18.

Lu, Jiahuan, Weiwei Lin, and Qiushi Wang. 2019. "Does a More Diversified Revenue Structure Lead to Greater Financial Capacity and Less Vulnerability in Nonprofit Organizations? A Bibliometric and Meta-Analysis." *VOLUNTAS: International Journal of Voluntary & Nonprofit Organizations* 30(3): 593–609.

Matthes, Jörg et al. 2019. "A Meta-Analysis of the Effects of Cross-Cutting Exposure on Political Participation." *Political Communication* 36(4): 523–42.

Merkle, Jessica S., and Michelle Andrea Phillips. 2018. "The Wage Impact of Teachers Unions: A Meta-Analysis." *Contemporary Economic Policy* 36(1): 93–115.

Munzert, Simon, and Sebastian Ramirez-Ruiz. 2021. "Meta-Analysis of the Effects of Voting Advice Applications." *Political Communication* 38(6): 1–16.

O'Brochta, William. 2019. "A Meta-Analysis of Natural Resources and Conflict." *Research & Politics* 6(1): 2053168018818232.

Owen, Erica, and Quan Li. 2020. "The Conditional Nature of Publication Bias: A Meta-Regression Analysis." *Political Science Research & Methods*: 1–11.

Philips, Andrew Q. 2016. "Seeing the Forest through the Trees: A Meta-Analysis of Political Budget Cycles." *Public Choice* 168(3): 313–41.

Schwarz, Susanne, and Alexander Coppock. 2021. "What Have We Learned About Gender From Candidate Choice Experiments? A Meta-Analysis of 67 Factorial Survey Experiments." *Journal of Politics*: 40.

Trinn, Christoph, and Thomas Wencker. 2020. "Integrating the Quantitative Research on the Onset and Incidence of Violent Intrastate Conflicts." *International Studies Review* (1): 115–39.

Walter, Nathan, Jonathan Cohen, R. Lance Holbert, and Yasmin Morag. 2020. "Fact-Checking: A Meta-Analysis of What Works and for Whom." *Political Communication* 37(3): 350–75.

de Wit, Arjen, and René Bekkers. 2017. "Government Support and Charitable Donations: A Meta-Analysis of the Crowding-out Hypothesis." *Journal of Public Administration Research & Theory* 27(2): 301–19.

Yesilyurt, Filiz, and M Ensar Yesilyurt. 2019. "Meta-Analysis, Military Expenditures and Growth." *Journal of Peace Research* 56(3): 352–63.

Zhang, Jiasheng, Wenna Chen, Nicolai Petrovsky, and Richard M. Walker. 2021. "The Expectancy-Disconfirmation Model and Citizen Satisfaction with Public Services: A Meta-Analysis and an Agenda for Best Practice." *Public Administration Review* 82(1): 147–59.

Table 1: Balance table

		In sample (N=46)		Not in sample (N=39)	
		Mean	Std. Dev.	Mean	Std. Dev.
Prop. political behavior		0.3	0.5	0.5	0.5
Prop. top journal		0.1	0.3	0.1	0.3
Average publication year		2017.4	4.1	2014.6	5.6
		N	Pct.	N	Pct.
Subfield	American Politics	8	17.4	6	15.4
	Comparative Politics	11	23.9	11	28.2
	International Relations	5	10.9	3	7.7
	Political Economy	10	21.7	8	20.5
	Public Administration	12	26.1	11	28.2

B Data cleaning

We gathered estimates and standard errors for all hypothesis tests included in the data files or published tables and figures of the 46 meta-analytic articles identified during data collection. When standard errors were missing, we calculated them from related information when available (e.g. from confidence intervals or variances).

We filtered out hypothesis tests based on a series of criteria, beginning with a missing estimate or standard error. We dropped observations containing invalid computations or implausible values; these often involved pure zero estimates or pure zero standard errors (both probably the result of data entry mistakes). We excluded standard errors smaller than 10^{-10}. While this was an arbitrary tolerance level, using different thresholds did not make a substantial difference. We checked for mistakes in the data such as p-values reported as standard errors, and we computed transformations when necessary (e.g., deriving partial correlations from t-statistics and degrees of freedom). We also checked whether excluding estimates with absolute values less than or equal to 10^{-5} made a difference (it did not). We visually examined funnel plots to identify variables in need of transformation or visually odd patterns in some meta-analyses, which were then double-checked for accuracy. Lastly, we dropped all observations drawn from meta-analyses with fewer than 5 aggregated hypothesis tests.

As a check on the faithfulness of our cleaning process, we cross-validated our cleaned dataset against authors' replication scripts when available. We ran meta-analyses using our cleaned dataset rather than the raw data originally found online or provided to us by the authors. We were generally able to produce results that matched what authors describe in their original meta-analytic articles, though in some cases we had fewer observations due to our filtering criteria. In some rare cases, we were unable to back out how authors computed their results from their raw data without their replication scripts.

All in all, our dataset of estimates and standard errors contains 16649 rows with unique identifiers for hypothesis tests, meta-analyses, and meta-analytic articles.

C Expert survey

Our analysis of the survey data was preregistered following the AsPredicted template, and the pre-analysis plan is available at `https://aspredicted.org/blind.php?x=G28_YCP`. The design was approved by two different institutional review boards from the universities of two coauthors of the present paper.

Below, we present the following: the measures we took to achieve ethical standards; the full survey questionnaire; and an assessment of whether or not tests in meta-analysis are biased according to our expert sample.

Ethics and transparency in research

The survey had only 3 questions and took only a few minutes to complete. We obtained consent via a web form. The participants were not paid. The participant pool was comprised only of people who published in *Political Analysis*, and in order to keep the survey short (to increase the response rate) we did not collect any demographic data. The participant pool most likely did not include people that would typically be considered vulnerable or marginalized, given that it was mainly comprised of PhD political scientists. Surveying this group of respondents also eases any concerns about consent being meaningful and fully informed, as they are experts in methodology and understand what a survey is and how it works and to what they are agreeing. The survey did not differentially benefit or harm specific groups. We did not confront any ethical challenges related to our survey.

Recruitment email template

Dear Professor [LAST NAME],

We write to invite you to participate in a survey to gauge expectations around levels of statistical power in political science research. The survey consists of only 3 questions and should take about 2 minutes of your time.

We are emailing you because you published in Political Analysis since 2010, and so are working in political methodology and are part of our expert sample. The survey will close in one month. If you would like to do the survey, please click this link to see the consent form and take the survey

Or copy and paste the URL below into your internet browser: [URL]

If you click this link, you will be removed from our mailing list and we will not email you about this again

Please do not hesitate to contact us if you have any questions or concerns by sending an email to [EMAIL].

This project has been reviewed by the [IRB] for compliance with federal guidelines for research involving human participants ([CERTIFICATE NUMBER]).

Thank you,
[AUTHORS NAMES]

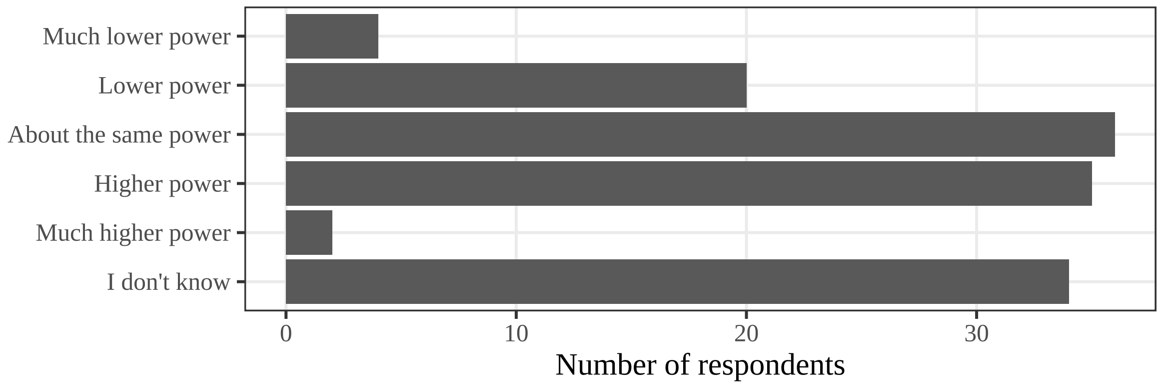

Figure C.1: Power of tests in meta-analyses compared to those not in meta-analyses

Questionnaire

Our questions focus on statistical power in political science research. As a reminder, higher statistical power means a better chance to detect an effect, if in fact an effect exists.

For the two questions on this page, consider all hypothesis tests reported in the 50 peer-reviewed journals with the highest impact factors in political science, international relations, and public administration over the past two decades.

What percent of the tests in these journals do you believe had at least **80% power** to reject the null with a significance level of 0.05? [Slider 0-100; Don't know]

What percent of the tests in these journals do you believe had at least **50% power** to reject the null with a significance level of 0.05? [Slider 0-100; Don't know]

Our final question is about statistical power in the subset of research articles that end up being included in meta-analyses. Not every published study or hypothesis test ends up being included in a meta-analysis. In some cases, for example, there are not enough comparable estimates to conduct a meta-analysis. For this question, we are **not** interested in the meta-analyses themselves, but rather in the individual hypothesis tests that are included, aggregated, and summarized in meta-analyses.

Do you think that the individual hypothesis tests that end up being included in meta-analyses are likely to have lower, equal, or higher power than those that do not end up included in meta-analyses? [Much lower power; Lower power; About the same power; Higher power; Much higher power; Don't know]

Are tests in meta-analyses biased?

In our expert survey we asked the respondents if they thought that hypothesis tests in meta-analyses were likely to have lower or higher power than tests that do not end up in meta-analyses. This helps us understand if our strategy of starting with meta-analyses produces a biased picture of power in the overall discipline, and it helps us link the respondent's guesses about power to our core results. Respondents do not generally think that tests in meta-analyses will have lower power.

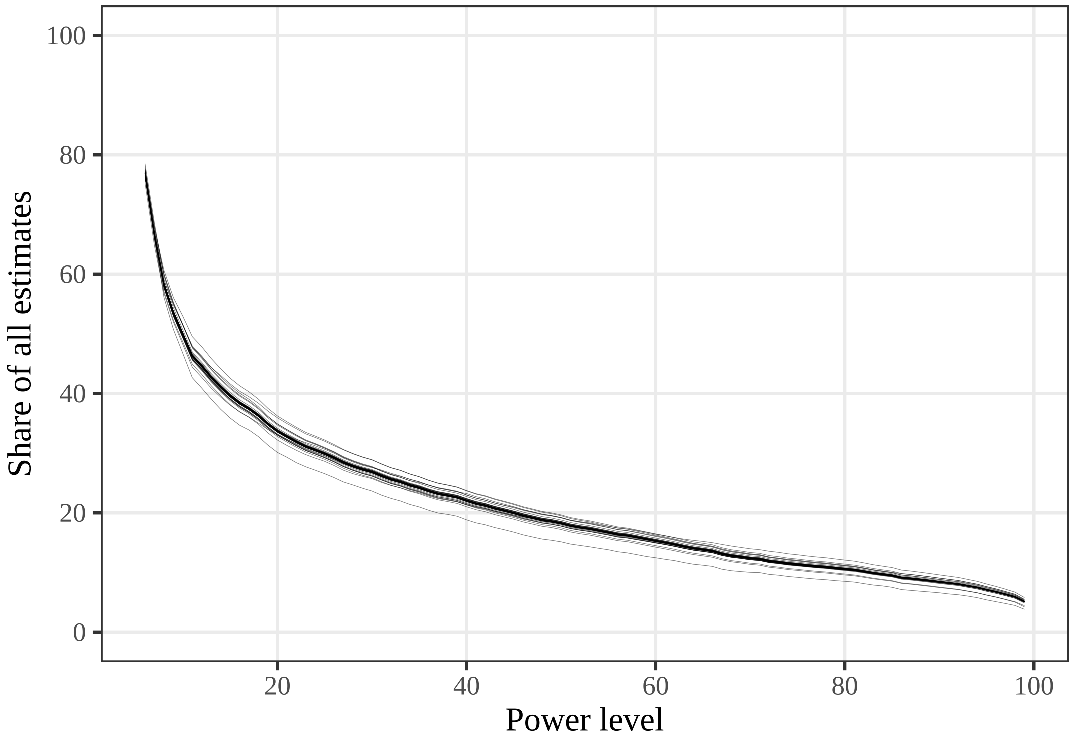

Figure D.1: Share of estimates by power level, excluding one full meta-analytic paper at a time. Unrestricted weighted least squares estimates.

D Outliers

One potential concern is that our findings may be driven by atypical observations or atypical meta-analyses.[33]

To examine the sensitivity of the main results to which meta-analyses we include, we reproduce the UWLS portion of Figure 2 while sequentially withholding one of the 46 meta-analytic articles at a time.[34] The lines are plotted with high transparency, so the dark black line emerges only from over-plotting. The result that power is very low does not rely on the inclusion of any specific meta-analytic paper. To ensure that our main results are not driven by extreme observations, we replicate our core findings from Figures 1 and 2 while excluding outliers identified with Tukey's "fences."

For each meta-analysis, we first compute the interquartile range of estimates and standard

[33]One atypical article in our dataset contributes 259 of our 351 meta-analyses (though only 3445 of our 16649 hypothesis tests). Median power (UWLS) overall is 0.1; excluding this paper brings median power to 0.09. An alternative approach to guard against such problems is to proceed in two steps: (1) take the median estimated power for each of the 46 meta-analytic papers, and (1) take the median of medians. Doing this gives us an estimate of 11%.

[34]As a reminder, individual articles can contain more than one meta-analysis. This is thus a stricter test than withholding one of the 351 meta-analyses at a time.

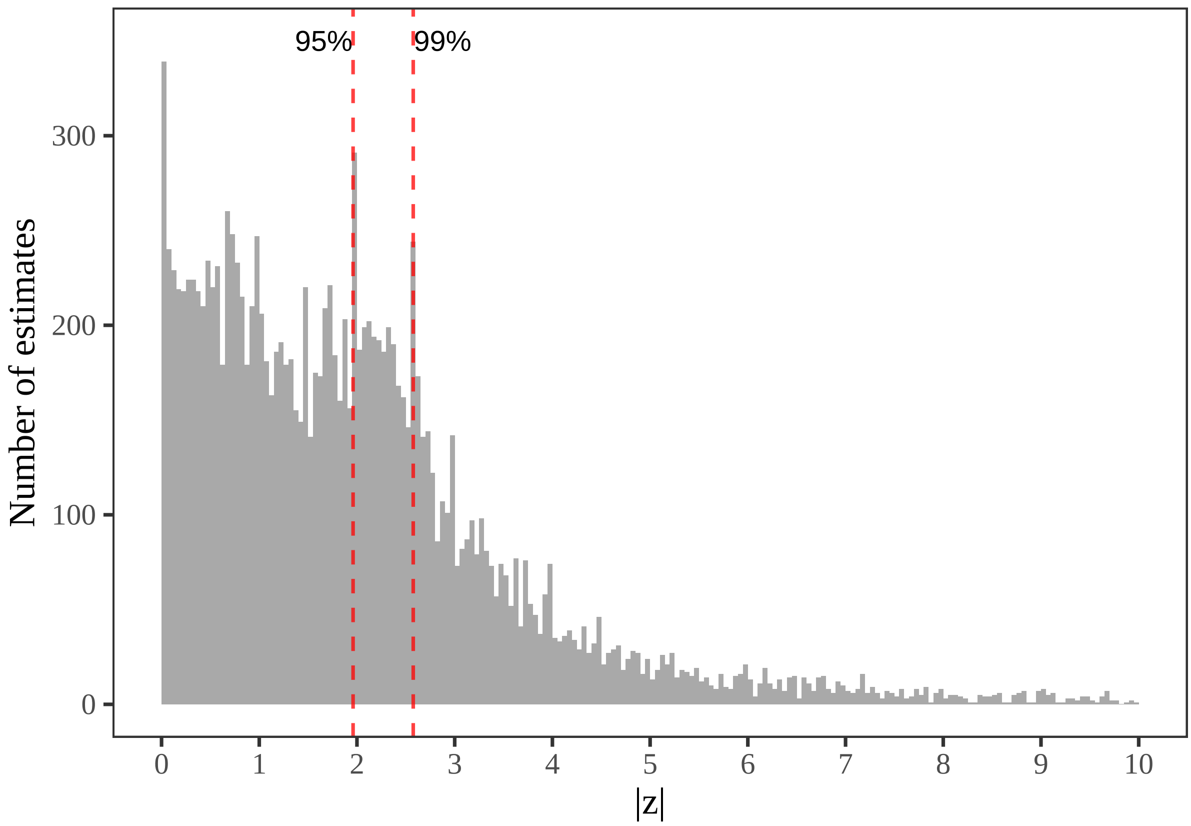

Figure D.2: Distribution of Z statistics in the full sample of estimates, excluding observations with outlying estimates or outlying standard errors.

errors. Then, we categorize an estimate or standard error as an outlier if it is 1.5 interquartile range above the 75th percentile or below the 25th percentile of estimates or standard errors within that meta-analysis.

To exclude influential observations, we calculate DFBeta statistics for each meta-analysis and reproduce our main figure when we exclude observations with DFBeta scores above 2 divided by the square root of the number of estimates in the relevant meta-analysis [Belsley et al., 1980].

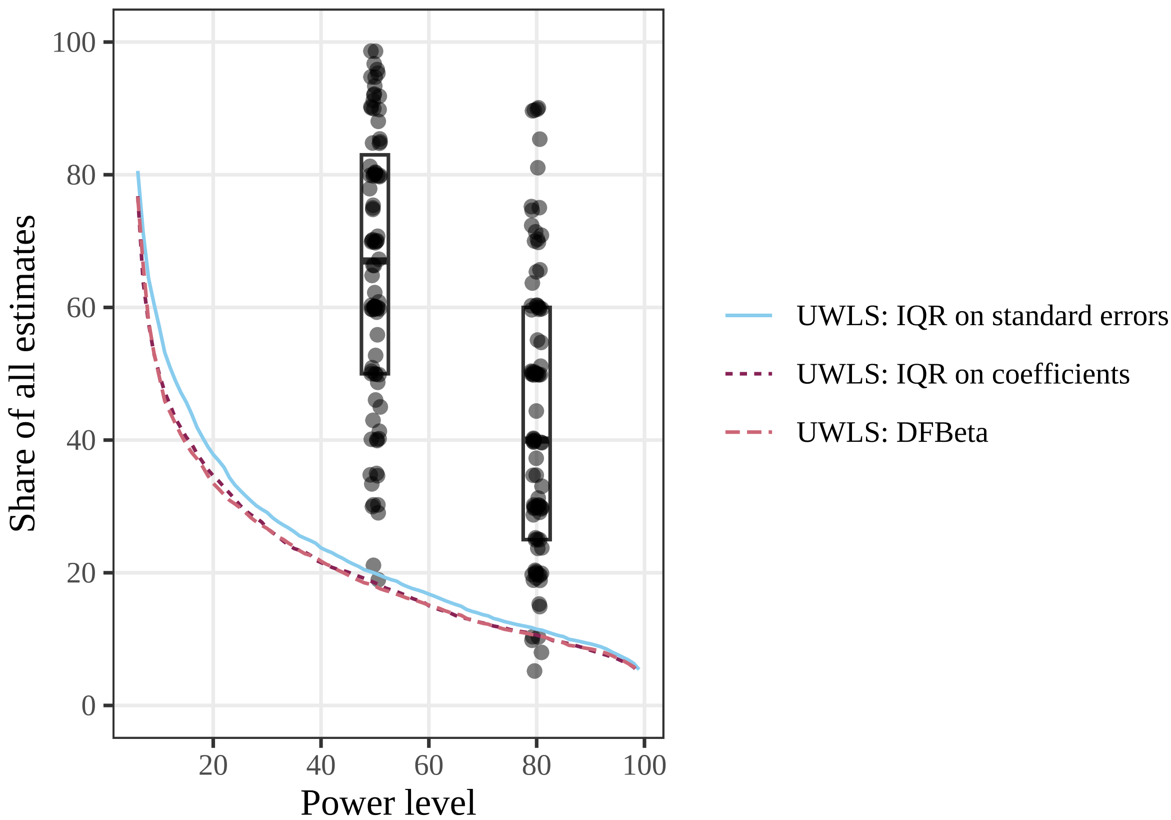

Figure D.3: Share of estimates by power level, excluding outliers.

E Andrews and Kasy (2019)

Here we implement the publication bias correction method of Andrews and Kasy [2019] and show that our results are substantively similar when using it. Our implementation is based on code from the authors' GitHub repository.

Andrews and Kasy's method (AK) essentially estimates publication probabilities for different spans of the z-score space. One must specify the cut points in this space in advance, and the method is somewhat sensitive to this choice. We choose a minimal but fairly agnostic approach, setting cut points to -1.96 and 1.96. This means that results are essentially re-weighted depending on estimated publication probabilities for negative and significant results, positive and significant results, and null results. We also use a normal distribution and allow the publication probabilities to be asymmetric. The latter is important as for some literatures a positive result is more or less likely to be selected by the publication process than a negative result.

When applying the AK method to our data, we encountered a major obstacle: the authors' maximum likelihood methods seem better suited for large samples, and they often fail to converge for meta-analyses with fewer than 10 estimates (and with much of the data in Trinn and Wencker [2021]). When we exclude these estimates and reproduce Figure 2 using the AK method to estimate population mean effects we find the results are substantively similar to those that we present in the text. The AK method produces somewhat higher power than UWLS, but our literature still seems under-powered.

We do not report the AK results in the full text as they do not use the full dataset.

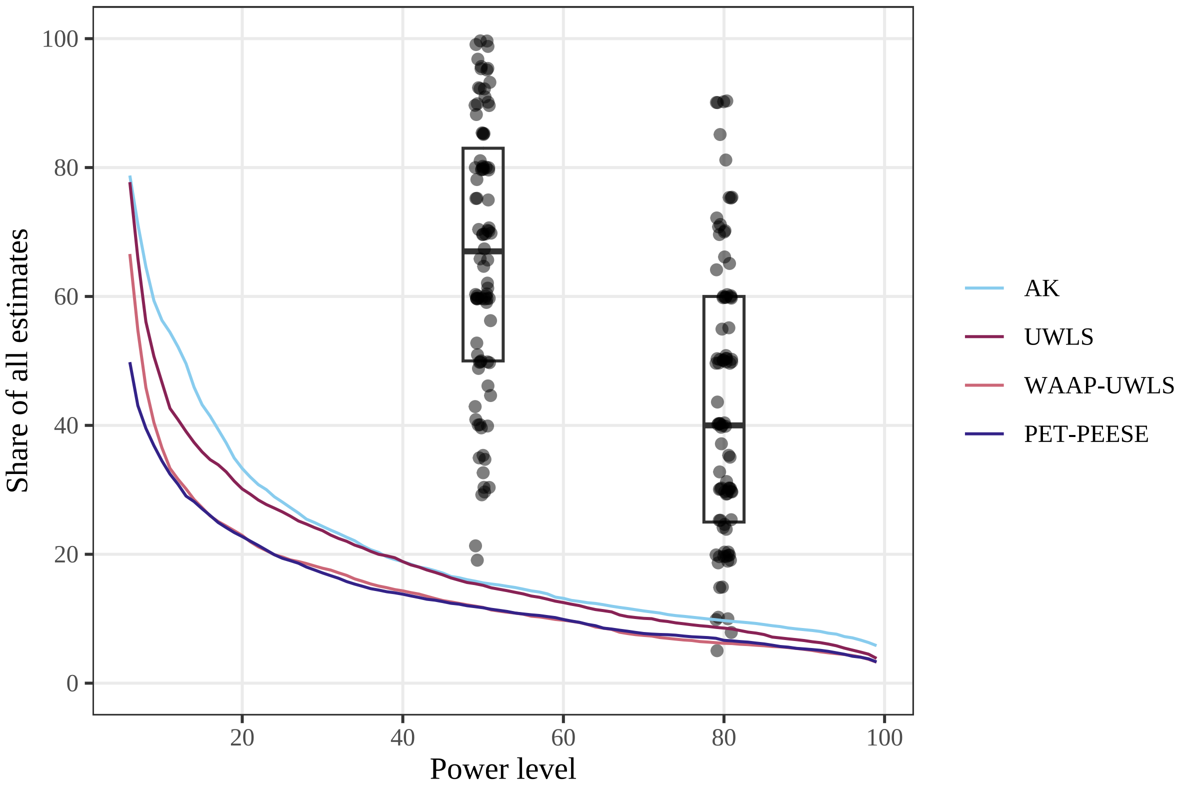

Figure E.1: Retrospective power analyses and the view from political methodology (run on a subset of the full data)

F Software bibliography

Arel-Bundock, Vincent. 2022. *modelsummary: Summary Tables and Plots for Statistical Models and Data: Beautiful, Customizable, and Publication-Ready.* https://vincentarelbundock.github.io/modelsummary/.

Barrett, Malcolm. 2021. *ggokabeito: Okabe-Ito Scales for ggplot2 and ggraph.* https://github.com/malcolmbarrett/ggokabeito

Blair, Graeme, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. 2019. "Declaring and Diagnosing Research Designs." *American Political Science Review* 113 (3): 838–59. https://declaredesign.org/paper.pdf.

Cameron, Allan, and Teun van den Brand. 2022. *geomtextpath: Curved Text in ggplot2.* https://allancameron.github.io/geomtextpath/.

Dowle, Matt, and Arun Srinivasan. 2021. *data.table: Extension of data.frame.* https://rdatatable.gitlab.io/data.table/

Lang, Michel. 2017. "checkmate: Fast Argument Checks for Defensive R Programming." *The R Journal* 9 (1): 437–45. https://doi.org/10.32614/RJ-2017-028.

Müller, Kirill. 2020. *here: A Simpler Way to Find Your Files.* https://here.r-lib.org/

Murdoch, Duncan. 2020. *tables: Formula-Driven Table Generation.* https://r-forge.r-project.org/projects/tables/.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Timm, Andrew. 2019. *retrodesign: Tools for Type S (Sign) and Type M (Magnitude) Errors.* https://github.com/andytimm/retrodesign.

Vaughan Davis, Matt Dancho (2022). *furrr: Apply Mapping Functions in Parallel using Futures.* https://github.com/DavisVaughan/furrr.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham H, Jennifer Bryan (2022). *readxl: Read Excel Files.* https://readxl.tidyverse.org.

Wickham, Hadley, Evan Miller, and Danny Smith. 2022. *haven: Import and Export SPSS, Stata and SAS Files.* https://haven.tidyverse.org/

Wickham Hadley 2021. *conflicted: An Alternative Conflict Resolution Strategy.* https://conflicted.r-lib.org, https://github.com/r-lib/conflicted.

Xie, Yihui. 2022. *xaringan: Presentation Ninja.* https://github.com/yihui/xaringan.

Zeileis, Achim, Susanne Köll, and Nathaniel Graham. 2020. "Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R." *Journal of Statistical Software* 95 (1): 1–36. https://doi.org/10.18637/jss.v095.i01.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with knitr::kable() + %>%.* https://haozhu233.github.io/kableExtra/