Challenges that Proprietary Research Poses for Meta-Analysis

Donald P. Green, Thomas Leavitt and Daniel Markovits

To Appear in Oxford Handbook of Engaged Methodological Pluralism in Political Science (Vol 1)

Abstract

The social sciences in general and political science in particular have seen marked improvements in research transparency. Data-sharing, pre-analysis plans, and results-blind reviewing have helped improve the trustworthiness of research findings, which in turn has facilitated credible research syntheses. At the same time, the rapid growth of proprietary research conducted by or on behalf of private entities presents a new set of challenges. Proprietary research seldom adheres to best practices in terms of open data, pre-registration, or pre-analysis plans; moreover, results from an undisclosed set of studies may be selectively released to the public in order to advance the interests of the research sponsors. This chapter addresses the problems of inference that arise when meta-analysts attempt to synthesize a research literature that comes exclusively or partly from proprietary sources. The authors' approach invokes a model of how proprietary organizations choose to disclose their results. Uncertainty over the parameters of this model implies that proprietary research should be down-weighted. This broad analytic argument is illustrated with reference to two experimental literatures on voter mobilization, one involving Facebook's banner ads and another involving handwritten letters to voters. The chapter concludes by outlining procedural changes that can decrease the threat selective reporting poses to inference.

Introduction

During the past decade, the techniques of experimental political science — already popular in the academic world — have spread to a network of businesses, non-governmental organizations, and campaign organizations. Entities as diverse as social media companies and advocacy groups staffed by volunteers have carried out randomized controlled trials (RCTs). These organizations operate outside the system of ethical and resource constraints that limit the scope and timing of experiments conducted by academics. While the results of these private experiments may sometimes be closely guarded, certain findings are released to the public. When they are, they often make quite a splash, such as the effectiveness of large-scale Facebook banner ads on voter turnout (Bond et al., 2012). Given the unique advantages of this type of research, access to large-scale, well-funded, private research on topics important to political science would seem to be a windfall for academics.

The challenge, however, is that the splashy nature of certain privately generated results may be why they are public in the first place. That the existence of some experiments may be concealed and others publicized based on the treatment effect estimates they produce raises the problem of selective reporting, which is the focus of this chapter. Any specific study that is concealed because of its findings is a loss to researchers in and of itself. Selective reporting also poses a broader challenge to meta-analysis, a statistical method that aggregates distinct studies of broadly similar treatments. We will show that suspicions about hidden results from the same organization can undermine the value accorded to meta-analyses that include selectively released findings. Specifically, given the potential for selective reporting, we should accord less weight to proprietary studies when their results are disclosed publicly.

In conventional meta-analysis, individual studies are often weighted by their precision (perhaps with a random effect adjustment) and nothing else. Meta-analyses under this weighting scheme may be distorted when results are selectively reported. Consider a social media company that tests the efficacy of its ads 100 times an five times finds a positive, statistically significant result (as we would expect if the null hypothesis were true given conventional standards of significance). A simple meta-analysis of these findings could report a positive treatment effect even though none exists. In addition to generating a misleading estimate, a meta-analysis of proprietary studies could produce an unjustified *confidence* in the pooled results, especially since meta-analyses often feature authoritative language about what the meta-analysis tells us about the population from which the studies were drawn. In other words, concealing results of studies with unfavorable results distorts meta-analytic conclusions, in the same way that dropping unfavorable data points distorts a single study.

To be sure, so-called "file drawer" problems (Rosenthal, 1979) exist in academia as well. When researchers find a substantively large and statistically significant effect, they may be more to have their work accepted by journals or conferences. Convincing empirical demonstrations of publication bias may be found across a range of fields including political science (Franco et al., 2014; Gerber and Malhotra, 2008; Gerber et al., 2001). However, the relatively insulated nature of academia has allowed for the development of professional norms around reporting results, and it appears that over time fewer experiments have been relegated to the file drawer. These disclosure norms are especially potent in sub-fields where experimental research is most frequently conducted. For example, between working papers, conferences and the journal submission process, some record exists of most academic experiments on voter turnout. In addition, the growing use of public-facing pre-analysis plans means that even if results are not publicized, some record of the experiment exists.

However, as these disclosure norms have mitigated one problem, another has emerged in its place. Proprietary research[1] exists in a different world. No single professional network binds together all practitioners; no set of ethical rules governs research; and economic self-interest is far more immediate than for academic researchers. Consider the case of a not-for-profit campaign organization publicly identified with a specific technique to increase voter turnout. After a carefully executed RCT, the organization finds a null effect. While academic researchers in a similar situation might be disappointed by a null, their careers are unlikely to be damaged more by publicizing the result than by concealing it. For the organization, however, a public null might lead to an

[1] We define proprietary research as wholly conducted by private organizations that have no obligation to release data to the public and are not cooperating with academics. Any cases where proprietary researchers contact academics after results are known to collaborate on dissemination still qualify as proprietary research; so too do cases when proprietary organizations retain a veto about the publication or public release of findings.

evaporating stream of donations, forcing layoffs or scaled back activities. With no pre-registration of the study or professional norms to the contrary, the results may be attributed to bad luck and quietly swept under the rug. Even the mere risk of selective disclosure of results should affect how researchers interpret findings from proprietary organizations (Green and Gerber, 2016).

In this chapter we consider how the potential for selective reporting of proprietary data challenges meta-analysts. We begin with a theoretical analysis of reporting bias. In this analysis, we make four key points: first, we introduce a new model of selective reporting by proprietary organizations, which differs from extant models that focus primarily on the academic publishing process. Second, we show the importance of accounting for nonstatistical uncertainty — based on a behavioral model of proprietors — in meta-analyses. In particular, the greater the statistical uncertainty of a given study, the less weight that study receives in the meta-analysis. Third, we show that large experiments by selective reporters help to mitigate concerns about selective reporting. The research community should be especially skeptical of splashy results from small experiments. Fourth, we show that in some cases researchers can draw on extrinsic information to establish that no conducted but unreported experiments exist. Intuitively, it would seem that the absence of any actual concealing of results would mitigate concerns about hidden bias; yet we show that reducing nonstatistical uncertainty comes primarily from ex-ante commitment mechanisms to report results rather than ex-post assessments of whether unreported experiments exist.

To illustrate the importance of these theoretical considerations, we analyze a pair of empirical examples. First, we consider Facebook's proprietary research on the efficacy of banner ads encouraging voting. Next, we turn our attention to research by academics and private organizations on the question of whether sending postcards to voters increases turnout. These examples provide a useful contrast in that we can ask Facebook, a single organization, how many experiments it has run, but we can do no such thing for the unknown number of organizations capable of conducting postcard experiments. Nevertheless, in both cases, the lack of a specifically ex-ante commitment to reporting results decreases the weight accorded to the results disclosed by proprietary organizations.

We therefore conclude with suggestions for proprietary organizations themselves to implement rules mitigating reporting bias. We also reflect on what academic researchers should look for when assessing the potential for selective reporting. This process also involves thinking about the interaction between other forms of manipulation of results — notably p-hacking. Finally, we discuss how reporting biases may be shaped by the broader institutional context. We note that when lives are on the line, the Food and Drug Administration (FDA) steps in to run later stage drug trials so as to prevent selective reporting. This form of an ex-ante commitment to reporting increases the information content of experiments. We argue that a unified skeptical framework can be used to properly assess results reported by proprietary and academic entities alike.

Why Models of Selective Reporting Matter for Meta-Analysis

We begin by outlining challenges to meta-analysis already identified in existing literature and the links between these challenges and selective reporting. There are two precedents in the existing literature on meta-analysis for thinking about non-statistical uncertainty and meta-analysis: the "file drawer" problem we discussed previously and the challenge of integrating observational results.

To properly account for selective reporting of experimental results, we distinguish between what we term statistical uncertainty and nonstatistical uncertainty. The former characterizes uncertainty induced by a physical randomization or sampling process that generates experimental results. The latter, by contrast, reflects a proprietor's strategic or taste-based behavior — whether and how to report an experimental result —upon observing whatever results an experiment generates.

The existence of publication bias favoring statistically significant results is well established. This phenomenon has been detected empirically based on the relationship between sample size and effect size. As the logic of a power analysis clarifies, for a study to detect a significant effect with a small sample size, the true average treatment effect (ATE) must be large. If reported effect size diminishes as the sample size grows, this is a sign of publication bias in favor of significant findings (Andrews and Kasy, 2019). This pattern is in fact observed in published political science research (Gerber et al., 2001) as well as research that straddles political science and other disciplines (Paluck et al., 2021). Similar logic underlies diagnostic approaches to identifying publication bias (see "filter graphs" of Duval and Tweedie 2000 and Franco et al. 2014, who used a grant database).

Given concerns about bias in a wide array of social scientific fields, statisticians have proposed techniques for accounting for it in meta-analysis. Duval and Tweedie (2000) suggest that examining the distribution of published studies can provide insight into the suppressed treatment effect estimates, which can allow for a re-weighted meta-analysis. McCrary et al. (2016) propose a simpler approach of requiring higher levels of statistical significance to adjust the Type 1 Error probability back to traditional levels of 5% while accounting for the possibility of file drawer bias.

While existing research has addressed how to account for file drawer bias toward either statistically significant or positive results (Cools et al., 2021; Hedges, 1984; Hedges and Olkin, 1985) and uncertain observational bias (for an especially clear example of this, see Thompson et al., 2011), no scholarship to our knowledge considers the selective reporting problem for integrating proprietary and non-proprietary research. We think of this problem as akin to the challenge posed by observational research (insofar as we seek to integrate a set of findings with known and unknown degrees of bias) but with its own distinct characteristics. Unlike academics, proprietary researchers are more likely to be concerned with the direction of an effect rather than its statistical significance. Proprietary researchers can also choose — unlike academics writing for peer-reviewed outlets — to report overall treatment effects without vital contextual information about their study. While academics may avoid data sharing, proprietary researchers can share results with journalists without so much as a standard error or sample size. The integration of proprietary and non-proprietary experiments is also related to the problem posed by publication bias insofar as we are concerned with a "suppressed" set of findings that may be concealed by proprietary organizations. Our overall framework accounts for both of statistical and nonstatistical uncertainties in the context of selective reporting.

Proprietary Selective Reporting and Meta-Analysis

As the basic set-up to our argument, we envision a set of experiments conducted on random samples from a common target population of interest. Under the usual assumptions, the difference-in-means estimator applied to each experiment is unbiased for the true average effect in the target population. In addition, we follow standard practice by modeling each experiment's standardized estimator — constructed by subtracting from the estimator its expected value and then dividing this difference by the estimator's variance — as an approximately standard Normal distribution. The estimators are independent because which experimental units are randomized into treatment and control depend only on each experiment's specific randomization process. Because each estimator is unbiased for the target population's average effect, the unknown mean of each experiment's Normal distribution is identical. However, each experiment's variance is not since it depends on features specific to each experiment, such as the size of the experiment and the numbers of treated and control units. Experiments with smaller variances (and, hence, higher precisions) typically receive more weight in a meta-analysis.

Given the statistical models of each distribution described previously, we let the joint likelihood function be the product of each distribution's probability density. This joint likelihood function permits likelihood-based inference (e.g., maximum likelihood estimation) as well as Bayesian inference when combined with a prior distribution on the relevant parameters. We omit the technical details here, but it is straightforward to define a prior distribution on only the average causal effect in the target population, which is equal to each distribution's mean. The variance parameters of each distribution can be accounted for by simply plugging in their consistently estimated variances and then acting as if these estimates are their true variances.

If we knew that the proprietor of each experiment's data were equally likely to report every possible result that an experiment could produce, then inference of the target population's average effect would be straightforward. In practice, however, proprietors may be more likely to report positive results than null or negative results. To capture this process, we propose the following behavioral model and associated sensitivity analysis: we partition the possible values of the estimator in a given experiment into non-positive and positive values. If the realized estimate happens to be non-positive, then the proprietor reports this result with probability $\gamma_{\leq 0}$. By contrast, if the estimate happens to be positive, then the reporting probability is $\gamma_{>0}$.

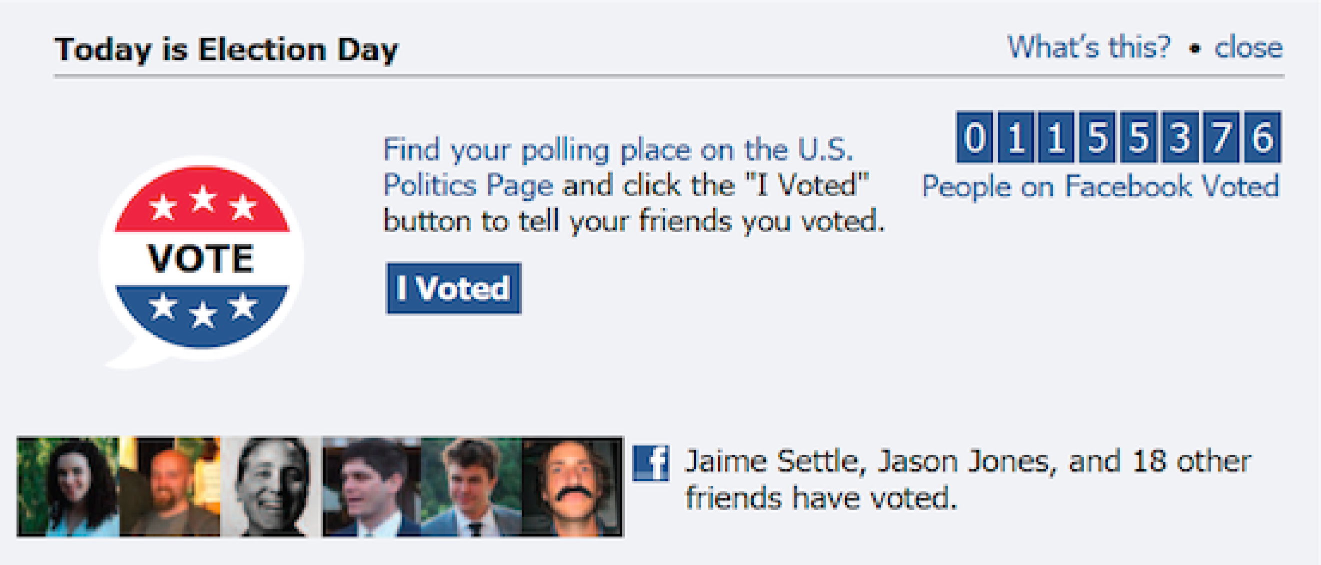
We then propose a sensitivity analysis that depends on one parameter, which is the ratio of $\gamma_{>0}$ to $\gamma_{\leq 0}$. When this ratio is equal to 1, no selective reporting exists. That is, a proprietor is equally likely to report a positive and nonpositive result. When this ratio is greater than 1, a proprietor is more likely to report a positive than a nonpositive result. In principle, this ratio could also be between 0 and 1, but we do not focus on this case since it is unlikely that proprietors would be more likely to report a nonpositive result than a positive one.

This behavioral model is obviously an oversimplification. It is perhaps more plausible to think of other, finer partitions of the space of possible estimates along with reporting probabilities that are not the same for all nonpositive or all positive values. It turns out, though, that our simplified model is helpful so long as we are careful to interpret $\gamma_{\leq 0}$ and $\gamma_{>0}$ as the marginal reporting probabilities over all nonpositive and over all positive reporting probabilities. Even if the reporting probability is not truly $\gamma_{\leq 0}$ for every nonpositive estimate and $\gamma_{>0}$ for every positive estimate, a weighting scheme that uses $\gamma_{\leq 0}$ and $\gamma_{>0}$ is sufficient to adjust for selective reporting.

The sensitivity analysis we propose formalizes our problem about the unknown degree of selective reporting. Insofar as our inferences about the average effect change across different degrees of selective reporting, then the unknown degree of selective reporting implies an additional layer of uncertainty, what we defined above as nonstatistical uncertainty. In some cases, uncertainty induced by the degree of selective reporting can be so severe as to render a result from even a seemingly well-executed experiment uninformative about the causal target. We now introduce our first empirical example — the case of Facebook's "I Voted" widget — to make two points: The first is the importance of ex-ante commitment devices to report the results of an experiment, whatever they end up being. The second is that, in large experiments, valuable information can still be had even in the presence of unknown degrees of selective reporting.

Facebook Advertising

During the 2010 Congressional elections, Bond et al. (2012) conducted an experiment on Facebook in which a randomly selected subset of Facebook users were exposed to a social message encouraging them to vote. This treatment consisted of a banner at the top of a user's News Feed encouraging that user to vote. The message included a link for information about local polling stations; a clickable "I voted" button with a counter showing the number of other Facebook users who already voted; and a display of the profile photos of up to six of the user's randomly selected Facebook friends who clicked the "I voted" button, along with the total number of Facebook friends who did so.



Today is Election Day

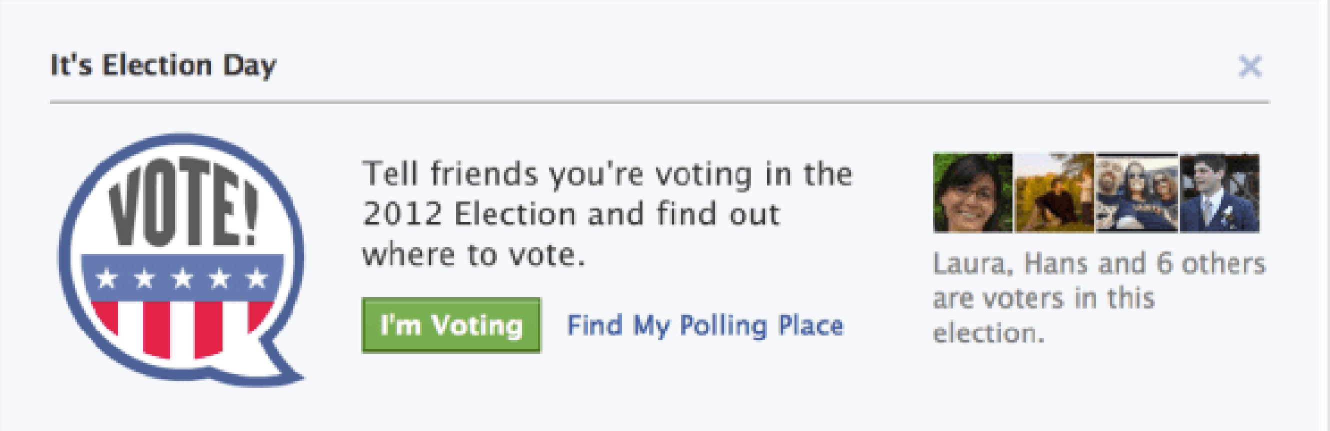
Find your polling place on the U.S. Politics Page and click the "I Voted" button to tell your friends you voted.

0 1 1 5 5 3 7 6 People on Facebook Voted

Jaime Settle, Jason Jones, and 18 other friends have voted.

Figure 1: The "Social Message" Treatment in Facebook's 2010 Voter Turnout Experiment (Bond et al., 2012, 296)

Among users in thirteen states (Arkansas, California, Connecticut, Florida, Kansas, Kentucky, Missouri, Nevada, New Jersey, New York, Oklahoma, Pennsylvania, and Rhode Island) with publicly available voting records, Bond et al. (2012) compared actual voter turnout among Facebook users who received the treatment in Figure 1 with users randomly assigned to the control group who did not receive any message at the top of their news feeds. The estimated average effect was 0.0039 percentage points with an estimated standard error of 0.0017 percentage points. This positive and significant result suggests that Facebook's social message increased turnout by roughly 60,000 votes (Bond et al., 2012, 297).



It's Election Day

Tell friends you're voting in the 2012 Election and find out where to vote.

Laura, Hans and 6 others are voters in this election.

Figure 2: The "Social Message" Treatment in Facebook's 2012 Voter Turnout Experiment (Jones et al., 2017, e0173853)

A follow-up study for the 2012 Presidential election by Jones et al. (2017) also reports a positively significant direct effect of Facebook's "social message" treatment on voter turnout. The treatment in this experiment, depicted in Figure 2, is almost identical to the treatment in the 2010 experiment. Jones et al. (2017) estimate an average effect of 0.0024 percentage points and a standard error of 0.001 percentage points. This estimated effect pertains only to Facebook users in the same thirteen states as the 2010 study, where voting records are readily available. This estimated effect in the 2012 study is slightly weaker than that of the 2010 experiment. Nevertheless, the sign of estimated effects are identical in both experiments, and the magnitudes of estimated effects are similar, despite the fact that "get-out-the-vote" messages typically yield smaller effects during high-stakes elections due to saturation of mobilization efforts from many sources" (Jones et al., 2017, e0173851). Because of the greater number of Facebook users in 2012 compared to 2010, the smaller estimated average effect in 2012 implies a greater number of Facebook users who voted due to Facebook's message (90,000 in 2012 compared to 60,000 in 2010).

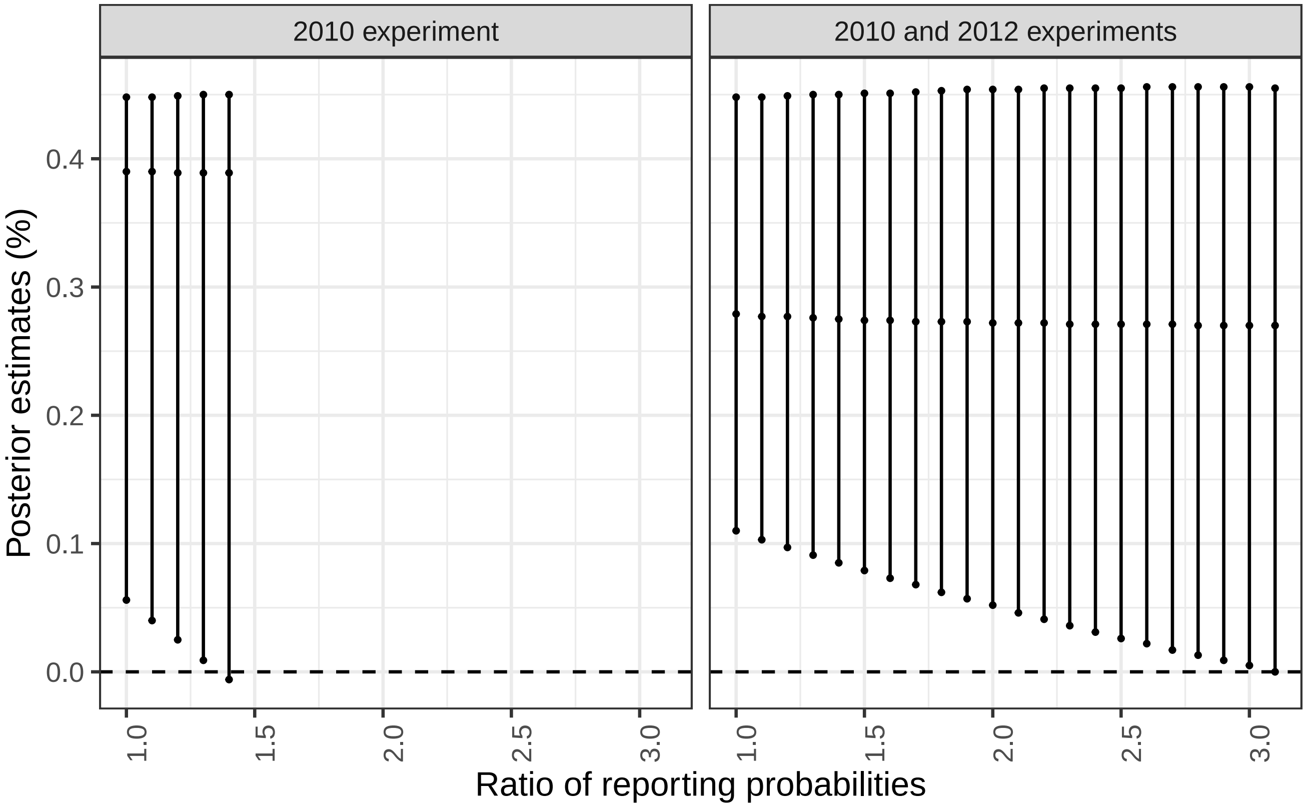


Figure 3: 95% Posterior Intervals under Uniform Prior and Increasing Ratios of Reporting Probabilities

Figure 3 shows the posterior intervals formed after reading these results. The left pane of the Figure illustrates the inference about the average treatment effect after reading the 2010 experiment. The right pane shows the inference drawn after reading both the 2010 and 2012 experiments. When the ratio of reporting probabilities is 1.0 (i.e., no selective reporting), the addition of the second study narrows the posterior distribution. In other words, when selective reporting is assumed to be absent, the accumulation of evidence leads to more precise inferences. But as we move from left to right along the horizontal axis of either the left or right pane, our uncertainty about the average effect increases. The greater the threat of selective reporting, the wider the posterior intervals. Our overall uncertainty is therefore greater than our statistical uncertainty conditional on any given assumption about the degree of selective reporting.

The Importance of Ex-Ante Commitment Devices

We now draw on this Facebook example to illustrate the importance of ex-ante commitments to report results whatever they end up being. Facebook, of course, is a proprietary organization. What we infer about online political mobilization from these two experiments depends in part on assumptions about Facebook's degree of selective reporting. Why, for example, did Facebook only report results of experiments for the 2010 and 2012 elections? Why not elections after 2012? One reason is that no such experiments were conducted after 2012. Another possibility is that the results in experiments after 2012 were less favorable and Facebook sought to keep a lid on these unflattering results.

To get to the bottom of this question, we reached out to Facebook researchers directly and learned that no other experiments were conducted. This Facebook example is instructive because, in a certain sense, Facebook did everything exactly right: through direct communication with Facebook we learned that the two experiments it reported were the only two it actually conducted. That is, Facebook behaved in a model way by reporting the result of every experiment it conducted.

However, despite this model behavior, nonstatistical uncertainty remains. The concern is that Facebook happened to conduct experiments that produced positive results. We do not know how Facebook would have behaved had an experimental result turned out to be nonpositive.

In more formal terms, the extrinsic information we gathered from Facebook provided us with the total number of experiments and the proportion of which that were reported. Under the assumption that Facebook's selective reporting probabilities are independent and identical across its two experiments, we can derive a likelihood function for Facebook's reporting probability of a positive result. We are able to narrow the plausible range of values for this reporting probability. The problem, though, is that because there are no experiments that produced a nonpositive result, we are unable to learn about the reporting probabilities of a nonpositive result. If we had reliable prior information about Facebook's probability of reporting a nonpositive result, then we gain information about the degree of selective reporting (the ratio of reporting a positive relative to nonpositive result). However, without prior information about Facebook's probability of reporting a nonpositive result, the absence of unreported results informs us about only the reporting probability of positive results, not the ratio of reporting a positive relative to nonpositive result.

This problem can work against the interests of Facebook when the true effect of its "I voted" widget is indeed positive. Note that extrinsic information about the absence of unreported results would be informative about the degree of selective reporting if Facebook had reported a positive and nonpositive result (as opposed to two positive results). But the probability of observing a positive and nonpositive results is less likely when the true effect is positive. Consequently, if Facebook were to report the results of its experiments, whatever they happened to be, uncertainty about selective reporting will be higher when the true effect is positive (which is more likely to yield only positive experimental results compared to when there is no effect). In short, exemplary behavior can work against Facebook when the true effect is large and positive. The lack of a way to commit to non-selective reporting before any experiments are actually conducted makes mitigating nonstatistical uncertainty difficult even in this case of model behavior by a proprietary organization.

Both of Facebook's experiments are quite large, which in turn provides potentially valuable information. We have focused thus far on how the absence of a mixture of positive and nonpositive results limits our ability to rule out selective reporting. This concern is offset to some extent by the fact that the degree of selective reporting is less relevant in large experiments, which is a point to which we now turn.

Large Experiments Don't Grow on Trees

Uncertainty over selective reporting may be less of a concern, all else equal, in large relative to small experiments. The key insight is captured by the notion that large, well-powered experiments do not grow on trees. So long as results from large experiments are reported, there is at least some information we can extract from them. The same is not necessarily true for small experiments.

To lay out this point, we first define two types of reporters: partial (or selective) reporters and impartial (or non-selective) reporters.

- **Partial reporter**: a partial reporter would report some possible estimates that an experiment could produce with greater probability than other possible estimates.
- **Impartial reporter**: an impartial reporter would report every possible estimate an experiment could produce with the same probability.

A key insight that emerges is that all reporters are asymptotically impartial. What this means in practical terms is that a large, reported experiment by even the most partial reporter leads to learning about the causal effect of interest. The intuition for this claim is relatively straightforward: suppose that a partial reporter will report causal estimates that fall in one interval with a high probability and estimates that fall in another interval with a low probability. Also suppose that the causal estimator is consistent for the true average effect, which will be true in a randomized experiment. Under these conditions, for some interval around the true average effect, there will be some size of the experiment such that, if the experiment is at least as large as this size, then the estimator will lie in the interval around the truth with probability 1. The interval around the truth can always be chosen to be small enough so that it is a subset of the interval in which the proprietor reports the estimate with a high or low probability. Therefore, when the experiment is large enough, the proprietor will report every estimate that an experiment could produce with the same probability. That is, with a large enough size of the experiment, the partial reporter is now an impartial reporter.

The previous logic is based on asymptotic reasoning. For an actual experiment of a finite size, we cannot actually know if the experiment's size is large enough to render a potentially partial reporter an impartial one. Thus, partial reporting is always at least somewhat of a concern. An ideal way to address this concern is by finding a way to identify the degree of selective reporting. Yet in the absence of such an identification strategy, it is straightforward to assess the sensitivity of meta-analytic inferences to increasingly severe forms of selective reporting. All else equal, when experiments are small, their results are more sensitive to departures from impartial reporting.

We now turn to our second empirical example, the so-called "postcards-to-voters" studies, which provides a useful contrast to the two Facebook experiments. There are two key analytic distinctions between the Facebook and "postcards-to-voters" experiments: The first is that, for the former, we know who could have conducted the studies and can go about gathering information about whether other studies were conducted. For "postcards-to-voters" experiments, many different organizations could have conducted such studies, which means that the task of figuring out whether there were any unreported studies is intractable. Second, the broader range of "postcards-to-voters" experiments underscores the crucial role of the assumption that experimental populations are sampled from the same superpopulation. This assumption, which is made in the Facebook case means that one very large experiment can provide enough information about this superpopulation effect to render selective reporting less important. Yet if the target of interest is a weighted combination of all experiments actually conducted without the assumption of a common superpopulation, then selective reporting matters more: There could always be a large, unreported experiment with a result that differs entirely from the results of large, reported experiments.

Postcards to Voters

Among the many electioneering tactics to have been evaluated experimentally in recent years is the practice of sending postcards to voters—especially handwritten postcards (as opposed to mass printed equivalents). Cheaper and more flexible than tactics such as door-to-door canvassing (in that volunteers need not live near the targets of their intervention), postcard-writing campaigns also present few barriers to researchers. Ease of entry means that a vast array of academic and proprietary research has been published or publicized in recent years. The flexibility of this research presents new problems compared to Facebook research. In the latter case, only a single organization was capable of carrying out the intervention and evaluation; by contrast, potentially hundreds of organizations could muster a postcard writing campaign of sufficient size to support an evaluation.[2] Given the sheer number of possible organizations using this tactic, it is no longer possible for researchers to make an inquiry or two to discern whether any other undisclosed research exists.

Instead, we must draw inferences based on the studies we find in the public domain. We draw the distinction between two types of studies: proprietary studies that were selectively publicized and proprietary studies that were never publicized.[3] Because we cannot assess the reporting status of experiments of which we have no knowledge, we take a slightly different approach to assessing reporting status. Having contacted other researchers as well as organizations that we know conduct Get Out the Vote (GOTV) experiments, we have access to twenty-three studies. Of these we code as "public" those accessible to prospective meta-analysts through publicly accessible records: searches on Google Scholar or Google that turn up studies described in on-line newspapers, conference papers, and the like as of Summer of 2022. We code as "non-public" those studies made accessible to us in private correspondence but unavailable through public channels.

While our collection of studies is unlikely to be comprehensive, having access to a set of reported and unreported findings allows us to more carefully examine degrees of selective reporting—unlike the Facebook case, we can estimate these selective reporting parameters, $\gamma_{\leq 0}$ and $\gamma_{>0}$.

[2] The low costs and flexibility of handwritten postcard interventions means an even broader selection of groups could implement this tactic than comparable efforts like door-to-door canvassing.

[3] We are aware of no pre-registered studies in this area that meet current standards of transparency in terms of data-sharing.

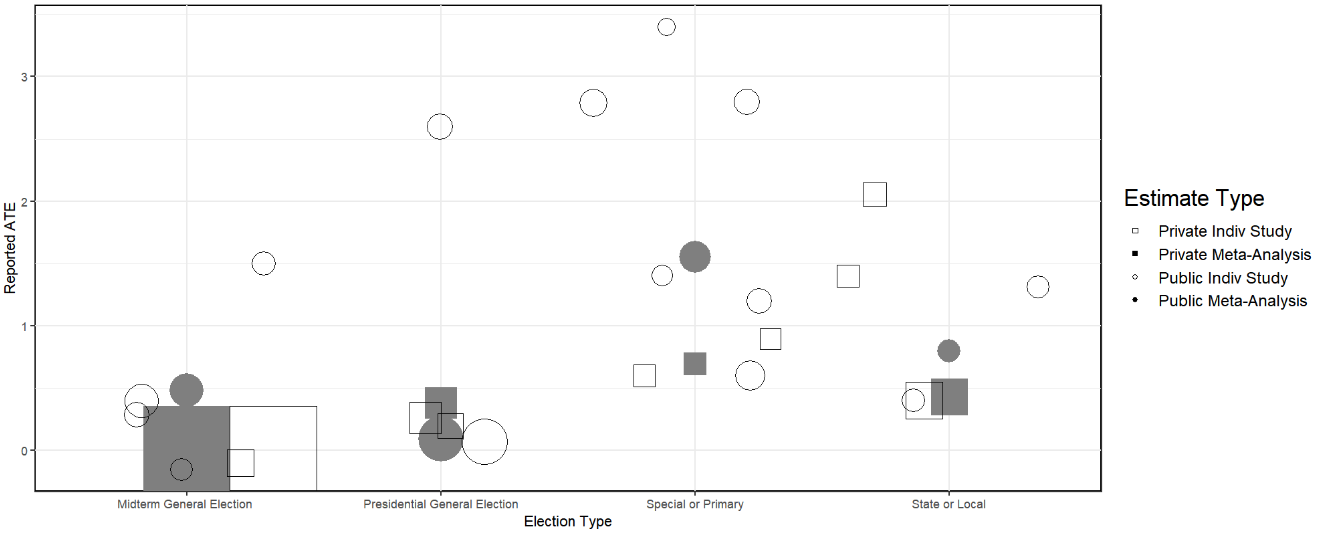


Figure 4: Treatment Effects of Postcards-to-Voters RCT Studies for Which Sufficient Information Is Available to Find Standard Errors

Results are presented in Figure 4, which depicts the treatment effects and reporting status of this collection of studies, subdivided by the election context in which the campaign took place. These estimates are scaled by their precisions (the inverse of their squared standard errors) such that more precisely estimated effects are larger in the graph. For each type of election, we also summarize the public and non-public estimates using meta-analysis.

Of the two negative results, one was reported, and one was not; of twenty-one positive results, thirteen were reported and eight were not. This suggests a $\gamma_{>0} = \frac{13}{21}$ and $\gamma_{\leq 0} = \frac{1}{2}$. The ratio of $\gamma_{>0}$ to $\gamma_{\leq 0}$ is then $\frac{\frac{13}{21}}{\frac{1}{2}} = \frac{26}{21} = 1.24$. This ratio implies that reporting bias may exist insofar as positive results were more likely to be made publicly available. However, this bias appears to be relatively minor.

We note too that the case of postcards-to-voters research provides a clear example of how the decision to report is continuous rather than binary on two important axes: the degree of information shared and the degree of dissemination. On the first dimension, the degree of data transparency ranges from a mention of a treatment effect in a newspaper article or blogpost[4] to the publication of a full report comparable in its depth to an academic study (albeit without replication code or the underlying raw data). In terms of degrees of dissemination, some studies

[4] See here for an example: https://demcastusa.com/2021/05/12/yes-postcards-to-voters-do-boost-interest-and-turnout/

were shared in email exchanges with one of the authors of this chapter; others were sent to email lists of practitioners and some academics; and others were discussed in newspapers and academic conferences. Both dimensions of reporting are important, researchers conducting meta-analyses can only use a study if they are somehow aware of its findings;[5] studies that are technically public but invisible through usual academic search channels cannot contribute to the accumulation of knowledge except to the subset of scholars aware of their existence. Even studies publicized in the websites of individual proprietary organizations may be difficult to locate without prior information about which organizations conduct a particular type of research.

In aggregate, both proprietary and non-proprietary researchers are broadly in accord that sending postcards to voters has a small but significant positive effect on turnout. The precision weighted average from all twenty-three studies puts this effect between 0.4 and 0.9 percentage points. This effect is somewhat larger than the 0.4 percentage point effect of conventional nonpartisan GOTV mailers found in a prior meta-analysis and many times larger than the near-zero effects of advocacy campaigns (Green and Gerber, 2019).

There are two clear links between this example and the theoretical section. First is the value of large-N studies and the presence of extrinsic information. One study reports results from a nationwide study with $N = 7.7$ million, a treatment effect of 0.07 percentage points and a standard error of 0.03. This study is helpful to us because it is sufficiently large (and precisely estimated) to help us rule out both a null effect and a large mobilization effect. Per the logic in our theoretical section, this result is sufficiently large to mitigate selective reporting risk.

What makes this study less useful is how it differs subtly from much of the other research we aggregate. This massive study occurred during the 2020 election; a great challenge of meta-analysis is that it assumes comparable studies are conducted among a single super-population. Even for GOTV meta-analyses, where this assumption is more plausible than in most meta-analyses on other subjects, the circumstances of each type of election are different, a phenomenon already observed in existing work on campaign effects (Kalla and Broockman, 2018). The 2020 election particularly generated several low treatment effects, possibly due to the high overall turnout and

[5] Meta-analysis also requires that researchers have sufficient information to estimate the precision of the study, a challenge that forced us to visually estimate magnitude of error bars for studies where basic information (e.g., the size of treatment and control groups) was not reported.

the unique circumstances of the COVID epidemic. We also observe our highest estimated ATEs in studies conducted during special elections or during primaries when the absence of distinguishing partisan cues leaves voters less certain of their choices.

Second is the importance of extrinsic information. We cannot say whether we have an exhaustive list of the organizations that could have conducted a postcards-to-voters GOTV experiment, but because we have both positive and non-positive ATE estimates, we can make guesses about industry-wide selective reporting norms.

Conclusion

In the preceding sections, we have outlined the threat that selective reporting in proprietary research poses to meta-analysis and situated this challenge in the context of similar problems of nonstatistical uncertainty. We demonstrated than in some cases the threat of selective reporting is minimized if a finite number of organizations could have carried out a treatment and can be reasonably assumed to honestly indicate the existence of such studies if asked. Having access to previously undisclosed research likewise allows us to derive estimates of reporting probability, which in turn allows us to estimate the bias created by selectively reporting. We note that large-N experiments mitigate the problem of selective reporting but that there are circumstances—notably when the number of studies is small and we cannot estimate reporting probabilities — where we may be better off ignoring selectively reported studies altogether.

Before outlining our proposed solutions, we note that when the stakes — and resulting legal liability — are high, government has historically stepped in to eliminate the threat of selective reporting. Skepticism about the credibility of research findings generated by private organizations has long guided the regulation of biomedical research, where the outcomes of randomized trials can have enormous humanitarian and financial implications. Although pharmaceutical firms make the case for the efficacy of interventions they hope to market, final assessment by regulatory bodies is based on randomized trials conducted by independent researchers. Even academic researchers who propose, test, and develop health interventions play no direct role in the trials that determine regulatory approval.

Social science is far from this stringent level of regulatory review, and indeed many of the scientific failings of social science resemble the deficiencies of biomedical research in less tightly regulated domains, such as research appearing in academic journals. For example, the problems of replication that seem rife among prominent publications in social psychology (van Aert et al., 2019) seem endemic also to published cancer research (Mullard, 2021; Rodgers and Collings, 2021). Begley and Ellis (2012) successfully replicated just six out of fifty-three cancer studies, the lowest rate in comparable therapeutic fields. Concerns about potential biases stemming from conflicts of interest have long informed biomedical checklists for meta-analysts (Duval and Tweedie, 2000). The greater stakes in lives and dollars of shoddy biomedical research motivate these practices, but the same broad incentive structure persists for lower stakes research, posing a grave threat to meta-analysis.

We note too that in more directly political applications—namely polling in elections—informed observers recognize that selectively released research carried out by interested parties is inherently unreliable.[6] Reporters discount polls publicized by campaigns, and the government disallows drugs whose efficacy is proved only by private research. But social science researchers still include privately conducted, selectively reported studies in meta-analyses.

As we have already discussed, the norms used to limit academic file drawer problems are not readily applicable to many proprietary contexts. As a purely practical matter, proprietary research concerns data that are often valuable to organizations in their own right (because they could prove valuable to competitors or reveal information about organizational practices). Academic research transparency norms are difficult to apply in these cases. Further—as we discussed in our introduction—the lack of centralized research norms and platforms make academic pre-registration reforms infeasible, as these norms are loosely enforced by journals and the academic community.

Data transparency issues also raise the specter of p-hacking, which may interact with selective reporting to further exacerbate the problem. In the broadest sense p-hacking refers to statistical and reporting practices that enable researchers to change their results so that they attain statistical significance (see, e.g., Brodeur et al. 2020; Humphreys et al. 2013). In theory this could include continuing to collect data until results are significant. However, we note that p-hacking in the

[6] https://fivethirtyeight.com/features/most-primary-polls-so-far-are-internal-polls-thats-a-problem/

context of RCTs presents a different challenge. Because data collection is usually concluded before results are known, a more limited form of p-hacking is possible where researchers selectively report models with statistically significant results. By including different control variables or using different operationalizations of the outcome variables (for example selecting one survey question of out of several that could reasonably represent the same underlying phenomenon), experimental researchers can manipulate what readers ultimately see.

Results that do not on their face support the efficacy of a particular treatment could be p-hacked, or selectively reported to avoid negative publicity for a proprietary organization. In academic research p-hacking is constrained, albeit imperfectly, by a combination of pre-analysis plans—which typically list a specific model or set of models the researchers commit to running—and data transparency. If full data and replication files are available, it is feasible for peer review or post-publication critics to identify selective model reporting. But in the absence of such transparency, the potential for selective reporting introduces an extra layer of nonstatistical uncertainty. An experimental difference-in-means estimator is unbiased in expectation, but not when estimates are reported based on whether they seem congenial.

We note that p-hacking adds a layer of complexity to selective reporting insofar as proprietary researchers could selectively report certain model specifications in place of concealing a study altogether. The possibility that a subset of publicized results could be biased in this way further complicates meta-analysis.[7] Our postcards meta-analysis included a number of studies that lacked sufficient data-transparency to shed light on whether p-hacking took place. In several cases, the opacity with which the results were presented required us to calculate standard errors by physically measuring the number of pixels on a graphed confidence interval. When authors disclose few statistical details, p-hacking is difficult to detect.[8]

[7] Thankfully, in large-Nexperimental research p-hacked results are less likely to produce dramatically different treatment effects compared to observational studies manipulated in the same manner. This is because covariates are unlikely to be substantially imbalanced between experimental groups in large sample sizes.

[8] In the absence of data transparency practices, pre-analysis plans may be the only way to meaningfully constrain p-hacking (unless proprietary researchers are willing to produce specification curves as in Simonsohn et al. 2020). Any proprietary study at risk of selective reporting is likely to provide greater scope for p-hacking. Consider an experiment testing the effectiveness of a GOTV treatment that produces a null result but one that under a handful of model specifications is positive. Having filed a pre-analysis plan and knowing they will have to share data upon publication, academic researchers can either attempt to publish the null or keep it in the file drawer. A proprietary researcher on the other hand could publicize the finding as a null, p-hack to a positive result, or conceal it altogether.

So what is the solution to the broader selective reporting problem and to p-hacking in this context? We think there are two complementary approaches. First, if donors, investors, and private sector actors share scientific goals of acquiring reliable knowledge, the problem will be self-correcting because proprietary organizations will know their findings will be devalued unless they solve their selective reporting problem. To the extent funders seek to accurately assess intervention effectiveness, their incentives are aligned with these practices. Proprietary organizations could maintain their own public pre-registration websites, but this internal archive would not solve the transparency problem unless there is monitoring by outsiders.

A more robust approach would consist of seeking to bind proprietary organizations or their researcher employees by disclosure norms. Interest groups or charitable foundations could require pre-registration for organizations carrying out experiments to be considered for funding opportunities or political contracts or other perks.

Perhaps less effectively, journals that publish or discuss proprietary findings could prominently note whether the implementing organization has a pre-registration policy or whether researchers can produce a complete list of studies carried out by the organization that investigate the same treatment or class of treatments. As we discussed, even knowing whether a non-reported experiment exists informs our meta-analysis. This labeling could serve to discount the value of non-preregistered experiments to non-academic audiences. Journals could also encourage more extensive sharing of proprietary data in anonymized or altered form such that as much transparency as possible is preserved and the tracks of p-hackers are easier to follow.

One interesting and, to our knowledge, unanswered question is whether researchers already intuitively downplay results generated by proprietary sources. Using social science prediction markets such as the Social Science Prediction Platform,[9] one could use a between-subjects design to test whether the same research findings are accorded different probative value depending on whether they are described as coming from public or proprietary research groups. In the same vein, it would be instructive to know whether any public versus proprietary distinction in the weight researchers place on a given finding persists even when studies are said to otherwise meet standards of research transparency such as sharing replication files. This question relates to our

[9] https://socialscienceprediction.org/

strategies for mitigating selective reporting bias; any of the interventions we described require key stakeholders to recognize the costs of selective reporting and ways it distorts efforts to determine the effectiveness of important real-world treatments across fields.

We conclude by noting some qualitative risk factors for selective reporting. We have already discussed the fundamental difference in incentives between proprietary and academic researchers when it comes to following transparency norms. However, some circumstances exacerbate selective reporting risks. First, some organizations or firms stand or fall based on the success of a given treatment or service. Second, if those who oversee a study have built reputations as proponents of a particular intervention, they may have incentives to limit distribution of results that call its effectiveness into question. Finally, organizations may have formal policies preventing researchers from sharing any findings without post-study approval from the implementing organization. Whatever the specific circumstances, proprietary experimental research has the potential to dramatically expand researchers' ability to answer important questions in political science, but only if we contend seriously with the problem of selective reporting.

References

Andrews, I. and M. Kasy (2019). Identification of and correction for publication bias. *American Economic Review 109*(8), 2766–2794. 4

Begley, C. G. and L. M. Ellis (2012). Drug development: Raise standards for preclinical cancer research. *Nature 483*(7391), 531–533. 18

Bond, R. M., C. J. Fariss, J. J. Jones, A. D. I. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler (2012). A 61-million-person experiment in social influence and political mobilization. *Nature 489*(7415), 295–298. 1, 7, 8

Brodeur, A., N. Cook, and A. Heyes (2020). Methods matter: *p*-hacking and publication bias in causal analysis in economics. *American Economic Review 110*(11), 3634–3660. 18

Cools, S., H. Finseraas, and O. Rogeberg (2021). Local immigration and support for anti-immigration parties: A meta-analysis. *American Journal of Political Science 65*(4), 988–1006. 5

Duval, S. and R. Tweedie (2000). A nonparametric 'trim and fill' method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association 95*(449), 89–98. 4, 5, 18

Franco, A., N. Malhotra, and G. Simonovits (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science 345*(6203), 1502–1505. 2, 4

Gerber, A. and N. Malhotra (2008). 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. 2

Gerber, A. S., D. P. Green, and D. Nickerson (2001). Testing for publication bias in political science. *Political Analysis 9*(4), 385–392. 2, 4

Green, D. P. and A. S. Gerber (2016). Voter mobilization, experimentation, and translational social science. *Perspectives on Politics 14*(3), 738–749. 3

Green, D. P. and A. S. Gerber (2019). *Get Out the Vote: How to Increase Voter Turnout* (4th ed.). Washington, D. C.: Brookings Institution Press. 16

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics 9*(1), 61–85. 5

Hedges, L. V. and I. Olkin (1985). *Statistical Methods for Meta-Analysis*. New York, NY: Academic Press. 5

Humphreys, M., R. S. de la Sierra, and P. Van der Windt (2013). Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis 21*(1), 1–20. 18

Jones, J. J., R. M. Bond, E. Bakshy, D. Eckles, and J. H. Fowler (2017). Social influence and political mobilization: Further evidence from a randomized experiment in the 2012 U.S. presidential election. *PLoS ONE 12*(4), e0173851. 8, 9

Kalla, J. and D. E. Broockman (2018). The minimal persuasive effects of campaign contact in general elections: Evidence from 49 field experiments. *The American Political Science Review 112*(1), 148–166. 16

McCrary, J., G. Christensen, and D. Fanelli (2016). Conservative tests under satisficing models of publication bias. *PLoS ONE 11*(2), e0149590. 5

Mullard, A. (2021). Half of top cancer studies fail high-profile reproducibility effort. *Nature 600*(7889), 368–369. 18

Paluck, E. L., R. Porat, C. S. Clark, and D. P. Green (2021). Prejudice reduction: Progress and challenges. *Annual Review of Psychology 72*, 533–560. 4

Rodgers, P. and A. Collings (2021). What have we learned? *eLife 10*, e75830. 18

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin 86*(3), 638–641. 2

Simonsohn, U., J. P. Simmons, and L. D. Nelson (2020). Specification curve analysis. *Nature Human Behaviour 4*(11), 1208–1214. 19

Thompson, S., U. Ekelund, S. Jebb, A. K. Lindroos, A. Mander, S. Sharp, R. Turner, and D. Wilks (2011). A proposed method of bias adjustment for meta-analyses of published observational studies. *International Journal of Epidemiology 40*(3), 765–777. 5

van Aert, R. C. M., J. M. Wicherts, and M. A. L. M. van Assen (2019). Publication bias examined in meta-analyses from psychology and medicine: A meta-meta-analysis. *PLoS ONE 14*(4), e0215052. 18