How Much Should You Trust Your Power Calculation Results? Power Analysis as an Estimation Problem

Shiyao Liu* Teppei Yamamoto†

July 13, 2020

*Ph.D. Candidate, Department of Political Science, Massachusetts Institute of Technology.

†Associate Professor, Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139. Email: teppei@mit.edu, URL: http://web.mit.edu/teppei/www

Abstract

With the surge of randomized experiments and the introduction of pre-analysis plans, today's political scientists routinely use power analysis when designing their empirical research. An often neglected fact about power analysis in practice, however, is that it requires knowledge about the true values of key parameters, such as the effect size. Since researchers rarely possess definitive knowledge of these parameter values, they often rely on auxiliary information to make their best guesses. For example, survey researchers commonly use pilot studies to explore alternative treatments and question formats, obtaining effect size estimates to be used in power calculations along the way. Field experimentalists often use evidence from similar studies in the past to calculate the minimum required sample size for their proposed experiment. Common across these practices is the hidden assumption that uncertainties about those often empirically obtained parameter values can safely be neglected for the purpose of power calculation.

In this paper, we show that such assumptions are often consequential and sometimes dangerous. We propose a conceptual distinction between two types of power analysis: empirical and non-empirical. We then argue that the former should be viewed as an estimation problem, such that their properties as estimators (e.g., bias, sampling variance) can be formally quantified and investigated. Specifically, we analyze two commonly used variants of empirical power analysis – power estimation and minimum required sample size (MRSS) estimation – asking how reliable these analyses can be under scenarios resembling typical empirical applications in political science. The results of our analytical and simulation-based investigation reveal that these estimators are likely to perform rather poorly in most empirically relevant situations. We offer practical guidelines for empirical researchers on when to (and not to) trust power analysis results.

Key Words: power analysis, sample size, experimental design, research transparency.

1 Introduction

With the surge of randomized experiments and the introduction of pre-analysis plans and research pre-registration, today's political scientists routinely use statistical power analysis. Many researchers, especially those who primarily employ experimental methods, consider power analysis to be an essential part of empirical research. For example, Evidence in Governance and Politics (EGAP), a prominent network of researchers and practitioners engaged in field experiments, recommends power analysis as an "important component of a pre-analysis plan" in their influential Methods Guides (Chen and Grady, 2019). Indeed, EGAP's research registration form asks every registered study whether a power analysis was conducted prior to data collection. It is also common for research grant agencies to either recommend or require power calculation included in study proposals (e.g., National Science Foundation, 2013). In the domain of academic publications, Journal of Experimental Political Science lists statistical power as one of the key criteria reviewers are asked to evaluate "registered reports" submissions on (Journal of Experimental Political Science, nd).

Statistical power analysis, or simply power analysis, refers to various techniques that involve *power* either as an input or an output. Power, or the probability of rejecting the null hypothesis when it is false, is often an important consideration when a researcher designs an empirical study under real-world constraints. For example, a researcher may be constrained by the maximum sample size they can use due to their financial or logistical capacity. In such a scenario, an important pre-study question of interest is whether the conceived study can be expected to achieve the level of statistical power that is high enough to make the study worthwhile. Another common situation is when a researcher seeks to infer how large a sample they will need to achieve the desired power (e.g., 80%), perhaps for the purpose of calculating the budget for a research grant proposal.

In all of the above scenarios, power analysis rests on the key assumption that the researcher has somehow obtained the knowledge of the true values of the study parameters prior to conducting the study. Specifically, power analysis in its simplest form requires two of the three population values as inputs: the standardized effect size (i.e. the raw effect size divided by the standard deviation of the outcome), sample size, and the power itself. Of these, sample size and power are typically either controlled by the researcher as part of the experimental design or externally given based on a budget constraint or by convention. The standardized effect size, however, is a feature of the data generating process itself and therefore almost never known by the researcher. Thus, researchers employing a power analysis must often cope with fundamental uncertainty about the true value of the standardized effect size.

In practice, however, researchers often rely on informal methods to set a value for the standardized effect size in their power analysis, despite the crucial role of the assumption in determining the output. Two

approaches are particularly common in empirical research. First, researchers often employ a pilot study to obtain an estimate of the size of the treatment effect of interest and use that estimate as an input to their power calculation. Second, researchers may look for a previous empirical study testing a similar hypothesis in the literature and use an estimate of the effect size in the study as if it was equivalent to their effect of interest. In both of these approaches, researchers employ existing empirical information about a population parameter (i.e. the standardized effect size) and attempt to make inference about the likely value of a function of the parameter (i.e. power or minimum required sample size). That is, power analysis is an estimation method to solve an empirical problem in these contexts. Despite this, current practice in applied research does not require researchers to formalize the degree of uncertainty in the "estimates" from their power analysis.

In this paper, we propose to call these types of power analyses *empirical power analyses* and distinguish them from the rest of power analysis variants which do not involve the use of empirical information.[1] Specifically, we analyze two types of empirical power analysis techniques: power estimation and minimum required sample size (MRSS) estimation. Viewed as statistical estimation techniques, empirical power analyses can be examined in terms of their statistical properties as estimators, such as bias and sampling uncertainty. We thus investigate the properties of standard power and MRSS estimators both analytically and via Monte Carlo simulations, focusing on the range of the parameter values that we find to correspond well to real-world scenarios in empirical political science based on our survey of the literature. That is, we ask: Can we trust the results of empirical power analyses in typical political science applications? Is the bias in a power or MRSS estimate small enough to be useful given an unbiased estimate of the standardized effect size from a pilot study? How precise are those estimates likely to be when the pilot study contains a typical number of observations?

These questions are crucial to answer for several reasons. First, researchers must often use data from a small pilot study or a loosely related previous research to estimate the true size of their effect of interest. Given the large amount of uncertainty in the effect size estimates, a natural concern is whether the downstream estimate of the power or the MRSS may also be poor. Second, despite the potentially large degree of uncertainty in empirical power analysis results, research practice in empirical political science is increasingly reliant on them. Indeed, researchers employing survey or field experiments routinely use empirical power analysis to make important decisions in the planning stage of their study, including whether to proceed with the study at all. This implies that the misinterpretation of power estimation results could lead to serious inefficiencies, such as missed opportunities and wasted resources. For example, an overestimation

[1]The latter kind of power analysis is therefore not subject to much of our critique in this paper. These non-empirical power analysis can be fruitfully used and recommended in many empirical settings, as we discuss in Section 5.

of the MRSS could discourage a researcher from conducting an actually promising experiment. Conversely, an overestimated power could lead a grant-making agency to funding a project that is in truth bound to fail. Third, even though the stakes are high, the existing literature has not critiqued the practice of power analysis from the perspective of estimation uncertainty in power analysis results.[2]

Overall, our investigation reveals a rather bleak picture of the usefulness of empirical power analysis in political science research. First, we show analytically that both power and MRSS estimates are not unbiased even when an unbiased estimate of the true effect size is available (as it is when, for example, the researcher conducts a pilot study on a random sample from the population of interest). Second, both our survey of the existing methods for bias correction and evidence from our simulation studies indicate that the biases in these estimates are in unknown directions and difficult to correct. Third, our simulation results suggest that estimation uncertainty in power and MRSS estimates is likely to be so large under typical empirical scenarios that the estimates are unlikely to be useful for practical purposes. These results imply that empirical researchers should exercise caution when applying empirical power analysis in their research in the traditional manner. Our advice, instead, is that researchers should primarily use power analysis for non-empirical purposes, such as to derive the required minimum sample size to detect the *desired* effect size based on substantive and/or normative grounds. Should they choose to employ empirical power analyses despite the likely performance problems, researchers should always quantify and report the degree of uncertainty in their power analysis estimates.

The rest of the paper is organized as follows. In Section 2, we set up our notational framework and define key concepts and quantities for our subsequent analysis. Sections 3 and 4 present the results of our analyses of the power and MRSS estimators, respectively, both analytically and via simulations. Section 5 contains our practical recommendations based on these results. Section 6 concludes.

2 Framework: Power Analysis as an Estimation Problem

Consider an empirical setting where the researcher is interested in the average treatment effect (ATE) of a binary treatment $Z \in \{0, 1\}$ on an outcome variable of interest Y. To estimate the ATE, the researcher intends to conduct a classical randomized experiment on a simple random sample of size N_f, randomly assigning n_{f1} subjects to the treatment condition and the remaining $n_{f0} = N_f - n_{f1}$ subjects to the control condition. Prior to conducting the "full" experiment, however, the researcher has data on Z and Y from another randomized experiment using the same treatment and outcome variables, but conducted on a

[2]There exists a growing literature criticizing the use of power analysis on conceptual grounds, notably from Bayesian perspectives (Gelman and Carlin, 2014; Kruschke and Liddell, 2018). Our argument is distinct from this strand of previous research in that we primarily examine power estimates in terms of their frequentist properties, so that the concept of power itself is well-defined and meaningful under our framework.

separate sample of N_p subjects drawn from the same population. We call this separate, pre-study sample a "pilot" experiment, as we intend it as a depiction of a common research practice where researchers conduct a small-scale pilot experiment prior to launching a full study.[3] Specifically, we assume that the researcher observes variables Z and Y on N_p pilot subjects of which n_{p1} were randomly assigned to $Z = 1$ and the remaining $n_{f0} = N_p - n_{p1}$ subjects received $Z = 0$.

A common practice in empirical research resembling this setup is to conduct an empirical power analysis prior to launching the full study, using information from the pilot study. Although tools for more complex power analyses are now available (e.g., Green and MacLeod, 2016; Blair et al., 2019), we focus on the textbook power analysis for the two-sample t-test with the asymptotic normal reference distribution, which is still widely used in practice. That is, suppose the researcher intends to test the null hypothesis of zero ATE using the two-sided t-test of difference in means in their full experiment. Denoting the type-I and type-II errors of the intended test by α and β, respectively, the *(true) power* of the full experiment ψ is given by the following formula:

$$\psi \;=\; 1-\beta \;=\; 1-\Phi\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right)-\frac{\tau}{\sqrt{\frac{S_0^2}{n_{f0}}+\frac{S_1^2}{n_{f1}}}}\right)+\Phi\left(-\Phi^{-1}\left(1-\frac{\alpha}{2}\right)-\frac{\tau}{\sqrt{\frac{S_0^2}{n_{f0}}+\frac{S_1^2}{n_{f1}}}}\right), \qquad (1)$$

where $\Phi(\cdot)$ represent the standard normal cumulative distribution function (CDF), τ the true ATE, $S_1^2 \equiv \mathbb{V}(Y \mid Z = 1)$, and $S_0^2 \equiv \mathbb{V}(Y \mid Z = 0)$.

We now make two simplifying assumptions. First, we assume that the variance of the outcome variable is constant between the treated and control groups, such that $\sigma^2 \equiv S_1^2 = S_0^2$. Second, we assume that the treatment is randomly assigned so as to minimize the sampling variance of the estimated ATE (Neyman, 1923), such that $N_d/2 = n_{d1} = n_{d0}$, $d \in \{f, p\}$. Under these assumptions, it is straightforward to show that the power of the full experiment ψ can be written in terms of only three parameters: the size of the test α, sample size N_f, and the *standardized effect size* τ_{std}, which equals the true ATE scaled to the standard deviation of the outcome variable such that $\tau_{\text{std}} \equiv \tau/\sigma$. Namely, equation (1) can now be written as,

$$\psi \;=\; 1-\Phi\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right)-\frac{\tau_{\text{std}}\sqrt{N_f}}{2}\right)+\Phi\left(-\Phi^{-1}\left(1-\frac{\alpha}{2}\right)-\frac{\tau_{\text{std}}\sqrt{N_f}}{2}\right), \qquad (2)$$

Of the three parameters on the right-hand side of equation (2), two are what we might call *design parameters* which the researcher has control over in designing their study, at least in theory. That is, α can be set to a desired level (e.g. according to the convention of testing significance at $\alpha = .05$) and N_f can be chosen based

[3] Despite the terminology, this setup also encompasses other common scenarios, such as a field experiment where the researcher uses results from a previously published experiment resembling the proposed study to inform their power analysis. In this case, N_f should be interpreted as the effective sample size for a previous study.

on the financial or logistical constraints of the study. However, the remaining parameter, τ_{std}, is *empirical* by nature in the sense that its true value exists in the real world independent of the researcher's decision.

A common approach for calculating power using equation (2) is, thus, to obtain an *estimate* of τ_{std} from the pilot experiment and plug in the estimate to the equation, along with the desired or externally provided, known values of α and N_f. That is, an *empirical power analysis* entails estimating ψ via a plug-in estimator implied by equation (2):

$$\hat{\psi} \;=\; 1 - \Phi\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right) - \frac{\hat{\tau}_{\text{std}}\sqrt{N_f}}{2}\right) + \Phi\left(-\Phi^{-1}\left(1-\frac{\alpha}{2}\right) - \frac{\hat{\tau}_{\text{std}}\sqrt{N_f}}{2}\right), \tag{3}$$

where $\hat{\tau}_{\text{std}} = \hat{\tau}/\hat{\sigma}$, such that

$$\hat{\tau} \;=\; \frac{\sum YZ}{n_{p1}} - \frac{\sum Y(1-Z)}{n_{p0}}, \quad \hat{\sigma} \;=\; \sqrt{\frac{\sum (Y - \sum Y/N_p)^2}{N_p - 1}}, \tag{4}$$

where the summations are over the observations belonging to the pilot sample of size N_p.

Another typical use of the power formula in equation (2) is for calculating the MRSS for the full experiment, or the smallest number of full-experiment observations sufficient to achieve a desired power value ψ. That is, the power of the full experiment will exceed the desired value ψ if and only if:

$$N_f \;\geq\; \frac{4\left[\Phi^{-1}\left(1-\frac{\alpha}{2}\right) - \Phi^{-1}\left(1-\psi\right)\right]^2}{\tau_{\text{std}}^2}, \tag{5}$$

where we ignore the last term on the right-hand side of equation (2), which is negligibly small for most common choices of α.[4] $MRSS$ is then defined as the smallest integer N_f that satisfies the inequality (5). Noting that ψ is also a design parameter that can be set (e.g. at $\psi = .8$ by convention) by the researcher in this context, the *MRSS estimation* consists of the following plug-in estimator of $MRSS$:

$$\hat{MRSS} \;=\; \left\lceil \frac{4\left[\Phi^{-1}\left(1-\frac{\alpha}{2}\right) - \Phi^{-1}\left(1-\psi\right)\right]^2}{\hat{\tau}_{\text{std}}^2} \right\rceil, \tag{6}$$

where $\hat{\tau}_{\text{std}} = \hat{\tau}/\hat{\sigma}$ is given by equation (4).

Our problem, then, is to investigate the statistical properties of the power and MRSS estimators defined by equations (3) and (6), respectively. Before examining each estimator, some general discussion is helpful. First, note that both estimators are nonlinear functions of $\hat{\tau}_{\text{std}}$, which is the ratio of an unbiased estimator

[4] It is easy to see that the term is strictly bounded from above by $\alpha/2$.

$\hat{\tau}$ to a nearly unbiased estimator $\hat{\sigma}$. Specifically, by standard sampling theory, we know that:

$$\mathbb{E}[\hat{\tau}] \;=\; \tau, \quad \mathbb{V}[\hat{\tau}] \;=\; \frac{4\sigma^2}{N_p}, \quad \frac{\hat{\tau}-\tau}{2\sigma/\sqrt{N_p}} \xrightarrow{d} \mathcal{N}(0,1) \text{ as } N_p \to \infty,$$

and

$$\mathbb{E}[\hat{\sigma}^2] \;=\; \sigma^2, \quad \mathbb{V}[\hat{\sigma}^2] \;=\; \frac{1}{N_p}\left(\kappa - \frac{N_p-3}{N_p-1}\sigma^4\right), \quad \frac{\hat{\sigma}^2}{\sigma^2} \overset{\text{approx.}}{\sim} \frac{\chi^2(\nu)}{\nu} \text{ as } N_p \to \infty,$$

where $\kappa = \mathbb{E}[(Y - \mathbb{E}(Y))^4]$ and $\nu = \frac{2\sigma^4}{\mathbb{V}[\hat{\sigma}^2]}$ (O'Neill, 2014). Although these properties imply that both $\hat{\psi}$ and $M\hat{R}SS$ are consistent as N_p grows to infinity, they do not guarantee further desirable properties for the power and MRSS estimators. In particular, since $\hat{\psi}$ and $M\hat{R}SS$ are nonlinear in terms of either $\hat{\tau}$ or $\hat{\sigma}$, these estimates are generally not unbiased. Small-sample biases in these estimators are indeed of great concern since N_p, the sample size of the pilot study, tends to be quite small in many empirical scenarios.

3 Power Estimation

We first investigate the statistical properties of the power estimator defined in equation (3). Although perhaps not as common as MRSS estimation, power calculation is still commonplace in empirical political science, and it is often the way students are first introduced to the idea of power analysis in their first-year quantitative methodology class. A classical application of this occurs when the researcher has no direct control over the sample size for the proposed full experiment (e.g. because of logistical or financial constraints) and wants to decide whether it is worth proceeding with the planned study given the sample size externally fixed by such constraints. For example, in his pre-analysis plan registered on the EGAP repository, Tausanovitch (2015) uses data from a previous pilot study to show that his hypothetical proposed study of 2,000 survey respondents will have 88% chance of detecting the treatment effect that is half as large as the observed effect size in the prior study. An important question, however, is how reliable the reported power of 88% actually is, given that the power is in truth empirically estimated based on data from a previous pilot study. Below, we answer this question in a more general manner via analytical investigations of the properties of the estimator, as well as Monte Carlo simulations.

3.1 Analytical Results

As discussed in Section 2, the crux of the problem is that empirical power calculation is an estimation problem for the estimand that is a nonlinear function of another parameter. While a well-behaved estimator

is available for the latter parameter, that does not imply that a plug-in estimator for the true estimand is also well-behaved due to the nonlinearity. Specifically, recall that our power estimator is given by equation (3). To simplify the problem for the purpose of illustration, we focus on the case of $\alpha = 0.05$ and use the following approximation which ignores the final term[5]:

$$\hat{\psi} \;\simeq\; 1 - \Phi\left(1.96 - \frac{\hat{\tau}_{\text{std}}\sqrt{N_f}}{2}\right). \tag{7}$$

Equation (7) makes it apparent that our power estimate $\hat{\psi}$ is a nonlinear function of our treatment effect estimate $\hat{\tau}_{\text{std}}$ from the pilot study. More specifically, the function approximately takes the form of the normal CDF, which is neither globally convex nor concave. This observation is crucial since either global convexity or concavity would allow us to determine the direction of the bias.[6] Instead, all we can say is that $\hat{\psi}$ can be either upward or downward biased in small samples even when we have an unbiased estimate of $\hat{\tau}_{\text{std}}$ from the pilot study. However, with classical central limit theorem and delta method, as N_p goes to infinity, $\hat{\psi}$ is a consistent estimator for ψ, and the limiting distribution is asymptotically normal. Yet in most pilot studies, N_p is small by definition, and thus asymptotic properties provide no guarantee for the performance of $\hat{\psi}$ with $\hat{\tau}_{\text{std}}$ estimated from pilot data.

Can we make any further progress analytically before we turn to Monte Carlo simulations? As it turns out, with further approximations, we can derive several predictions about the magnitude and the likely directions of the bias under specific scenarios, as we show in Appendix A.1. First, the bias of the power estimator becomes arbitrarily close to zero when the magnitude of the true standardized effect size (τ_{std}) is large, holding other things constant. This is a consequence of the second term of equation (7) becoming close to zero in expectation as τ_{std} increases, making the estimate arbitrarily close to one, to which the true value of the power is also approaching when τ_{std} grows. Second, as the size of the full experiment N_f increases, the bias can become either larger or smaller in either direction, depending on the values of the other parameters. This might seem counterintuitive, since the true power itself approaches one as N_f grows as in the case of τ_{std} growing in size. However, an inspection of equation (7) reveals that the argument of the second term both shifts to the left and becomes more variable as N_f grows, leaving the net effect on $\mathbb{E}[\hat{\psi}]$ ambiguous.

Is it possible to correct the bias with some bias correction methods? We first note that the bias function for $\hat{\psi}$ is non-linear. Traditional simulation-based bias-correction methods such as the bootstrap or jackknife only work well when the bias is a constant or a linear function of the true parameter (Cordeiro and Cribari-Neto, 2014; MacKinnon and Smith Jr, 1998). Nevertheless, we tried these methods in Appendix A.2, and as

[5] As discussed in Section 2, this term is only as large as $\alpha/2$ at most and usually much smaller.

[6] By Jensen's inequality, $\mathbb{E}[g(X)] \leq g(\mathbb{E}[X])$ if $g(X)$ is globally concave, and $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$ if globally convex for a random variable X and a real-valued function $g(\cdot)$.

expected, these methods perform poorly in our setting. Second, MacKinnon and Smith Jr (1998) proposed an analytical bias-correction method when the bias is a non-linear function of the true parameter. However, their proposed method assumes the existence a closed-form analytical bias function, which is not satisfied in our setting (note the normal CDF does not have a closed-form expression). Despite this, we applied a modified version of their proposed method by using a numerically simulated bias function in place of an analytical bias function, only to find that the method failed to produce noticeable improvement. In sum, existing bias-correction methods are unable to correct the bias for $\hat{\psi}$.

Thus, while the results above provide useful insight on the theoretical behavior of the power estimator, their practical values are unfortunately rather limited. Fundamentally, this is because the highly non-linear nature of equation (7) precludes us from making a clear prediction about either the magnitude or the sign of the bias *ex ante*. Therefore, we now turn to an alternative approach: Monte Carlo simulations.

3.2 Simulations

To investigate small-sample properties of the power estimator in equation (3), we simulate repeated sampling from various data generating processes (DGPs) that we consider to match likely empirical scenarios in political science. We vary our DGP with respect to three population parameters that determine the sampling distribution of the power estimator: standardized treatment effect (τ_{std}), sample size of the pilot experiment (N_p), and the intended sample size of the full experiment (N_f). We set the values of these parameters in accordance with the spectrum of most political science applications. Specifically, we use $\tau_{\text{std}} \in [0, 1]$, $N_p \in \{50, 250, 450, 650\}$, and $N_f \in \{100, 500, 900, 1300\}$. We then evaluate the performance of the power estimator with respect to its bias and standard error calculated over $1,000$ Monte Carlo draws.

Simulation Procedure. We conduct the following Monte Carlo experiment for each combination of the τ_{std}, N_p and N_f values:

1. Randomly draw $\frac{N_p}{2}$ realizations of Y for the treatment group from $\mathcal{N}(\tau, 4^2)$, and $\frac{N_p}{2}$ realizations of Y for the control group from $\mathcal{N}(0, 4^2)$ to simulate the pilot study, where $\tau_{\text{std}} = \tau/4$.

2. Calculate the difference-in-means estimate of the treatment effect $\hat{\tau}$.

3. Estimate S_1^2 and S_0^2 with the sample variances of Y in the treatment and control groups, respectively. Denote those by $\widehat{S_1^2}$ and $\widehat{S_0^2}$.

4. Estimate power using a plug-in estimator based on equation (2), setting $\alpha = 0.05$:

$$\widehat{\psi} = 1 - \Phi\left(1.96 - \frac{\hat{\tau}}{\sqrt{\frac{\hat{S}_1^2}{n_{f1}} + \frac{\hat{S}_0^2}{n_{f0}}}}\right) + \Phi\left(-1.96 - \frac{\hat{\tau}}{\sqrt{\frac{\hat{S}_1^2}{n_{f1}} + \frac{\hat{S}_0^2}{n_{f0}}}}\right),$$

where $n_{f0} = n_{f1} = N_f/2$.

5. Evaluate performance of the power estimator by repeating Steps 1 to 4 1,000 times and calculating Monte Carlo estimates of the bias and the standard error.

Results. Figures 1 and 2 present the simulated bias and standard error of the power estimator, respectively, across the values of the pilot sizes, intended full experiment sizes and true standardized treatment effects.

First, for a wide range of parameter values, the bias of the power estimator is substantial in magnitude and can be either positive or negative. Specifically, for a given sample size combination, the bias is positive and largest in magnitude when the true standardized effect size is close to zero. As the effect size increases, the bias promptly decreases and becomes closer to zero, until its sign flips to negative at around $\tau_{\text{std}} = 0.175$. As the standardized effect size further increases, the bias gradually disappears while the true power has maxed out at the theoretical maximum of one. Considering the fact that the range of power is from 0 to 1, and that most researchers want the power of their research to be greater than 0.8, even a bias of 0.1 is quite substantial. And our result indicates that the bias is indeed at least as large as that for a wide range of scenarios.

Second, the power estimator is also highly imprecise for a wide range of the parameter values. For example, with $N_p = 50$ and $N_f = 100$, the standard error exceeds 0.1 for all values of τ_{std} in $[0, 1]$. Given the theoretical range of power and the typical threshold of 0.8, the standard error of 0.1 is substantial. Indeed, we find that the standard error to exceed this value unless the true standard effect size is at least as large as about 0.4, except in unrealistic scenarios where the pilot sample size is much larger than the full experiment size (first column, bottom three plots).

Third, as the analytical results in Appendix A.1 indicate, the bias is generally an increasing function of the ratio of the full experiment sample size to the pilot sample size, holding the standardized treatment effect constant. This result is confirmed by our simulation results. This result is in a sense a contrapositive for our fifth observation below - holding N_f constant, a larger N_p tends to reduce the bias in power estimation.

Fourth, the bias generally becomes negligibly small after the true standardized effect size increases past a threshold of approximately 0.3 to 0.8, depending on the sample size specifications. A good rule of thumb appears that a researcher should not worry about the bias in power estimation should she expect the her true

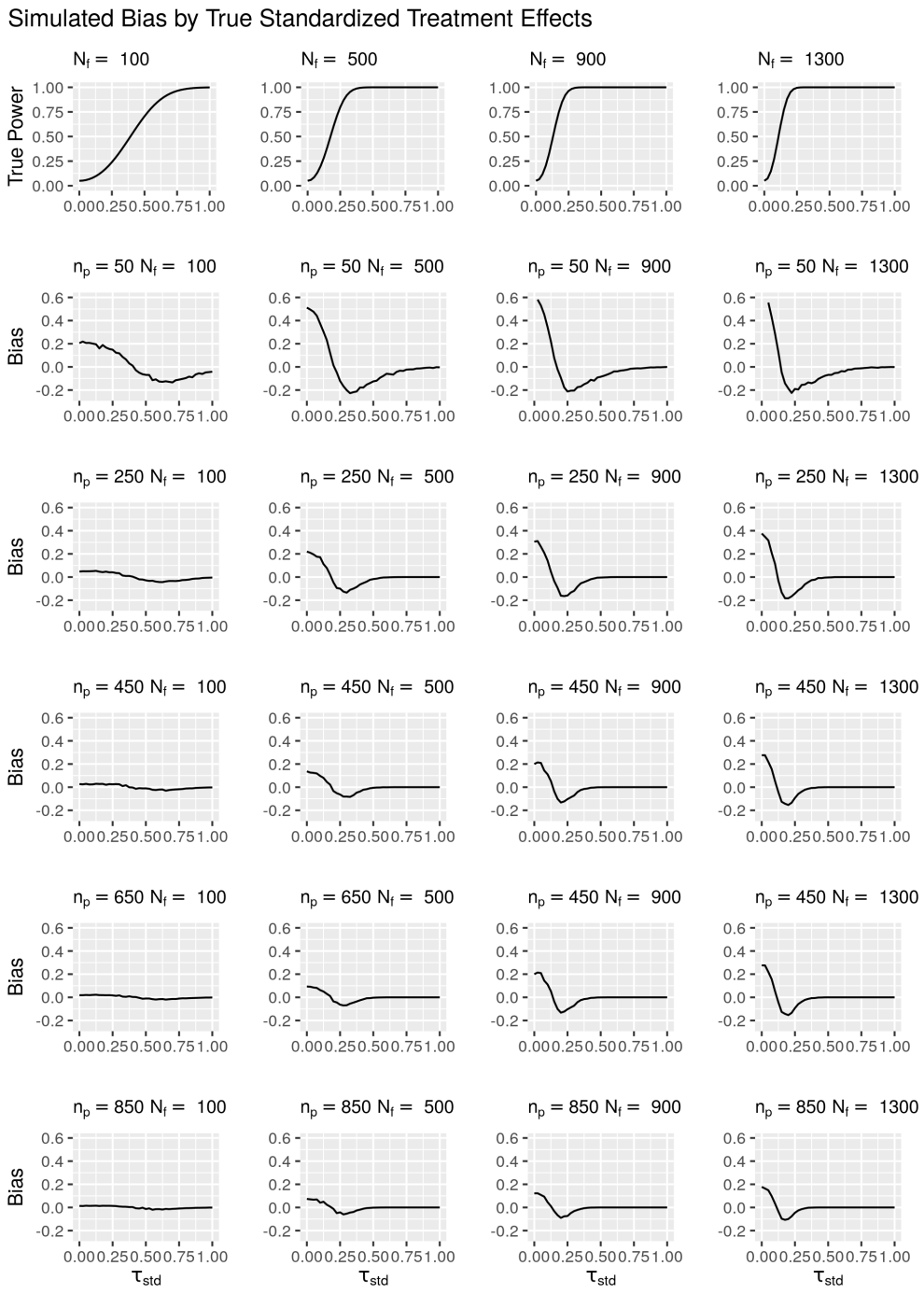

Figure 1: Simulated Bias by True Standardized Treatment Effect. The top row of plots present the true power for each full experiment sample size as a function the standardized effect size. The remaining plots show Monte Carlo estimates of the bias of the power estimator on the vertical axis for a given pilot sample size (row), full experiment sample size (column) and the standardized effect size (horizontal axis in each plot).

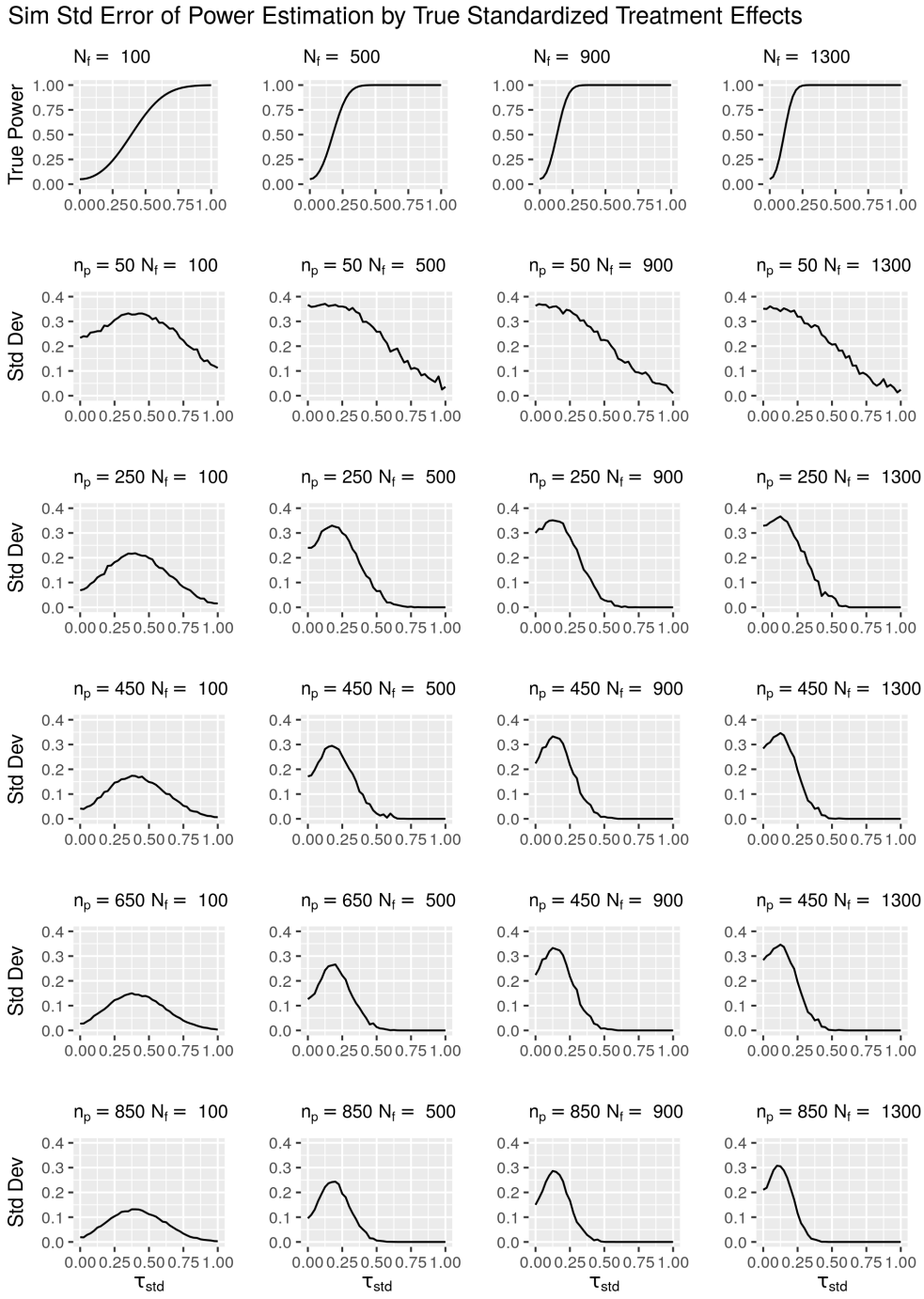

Figure 2: Standard Error by True Standardized Treatment Effect. See legend for Figure 1 for the interpretation of the graph components.

standardized treatment effect to be > 0.5. This is mainly because, with the usual full experiment sample size in political science applications, the power for an experiment with a 0.5 standardized treatment effect is quite close to one.

Fifth, the bias monotonically decreases as the pilot sample size increases. As we know, $\hat{\tau}_{\text{std}}$ is a consistent estimator. With the classical central limit theorem and the delta method, it is straightforward to show that $\hat{\psi}$ is also a consistent estimator for ψ with a rate of convergence at $\sqrt{N_p}$. Moreover, the limiting distribution for $\hat{\psi}$ will be asymptotically normal. Hence, as N_p gets larger, asymptotic properties start to take effect and hence both the bias and the variance reduce.

Finally, as can been seen in the first row of Figure 1, the true power function is convex in one area, but concave in another. Given the estimation uncertainty in $\hat{\tau}_{\text{std}}$, researchers cannot know whether the true standardized treatment effect falls in the convex area or the concave area of the power function unless the pilot sample size is unrealistically large. As we discuss in Appendix A.2, this ambiguity causes severe difficulty for the existing bias correction techniques, along with the other known limitations of these methods.

3.3 Calibrating Simulation Results

A key insight from our simulation study is that the bias and precision of the power estimator crucially depends on the true standardized effect size. A natural question, then, is where on the horizontal axis we are likely to be in the plots in Figures 1 and 2. Should political scientists be concerned about the bias in their power estimates in their typical experiments?

To answer this question, we turned to the reported estimates of the standardized effect sizes in the literature. Specifically, we collected all articles published in four top political science journals[7] in 2015 or later that report ATEs (or similar causal quantities that can be estimated via difference in means) as their main quantities of interest. Our search yielded a total of 164 standardized effect size observations. Although few of these articles directly report the standardized ATE, we were able to calculate the value of τ_{std} (with some assumptions when needed) implied by the uncertainty estimates reported alongside the effect sizes (see Appendix A.4 for more details of the data collection).

Figure 3 presents the distribution of the standardized effect sizes reported in our sample of articles, classified by the type of study. First, it is important to note that our sample is unlikely to be representative of *all* standardized effect sizes in the universe of political science experimental studies, since our sampling frame is published articles in top journals. Because of the well known publication bias and file-drawer effects, these observations are most likely tilted towards the right tail of the true effect size distribution (Schäfer and

[7]The journals are American Journal of Political Science, American Political Science Review, Journal of Politics, and Political Analysis. We chose these because they consistently rank highly on major impact factor rankings and they regularly publish articles containing experimental studies.

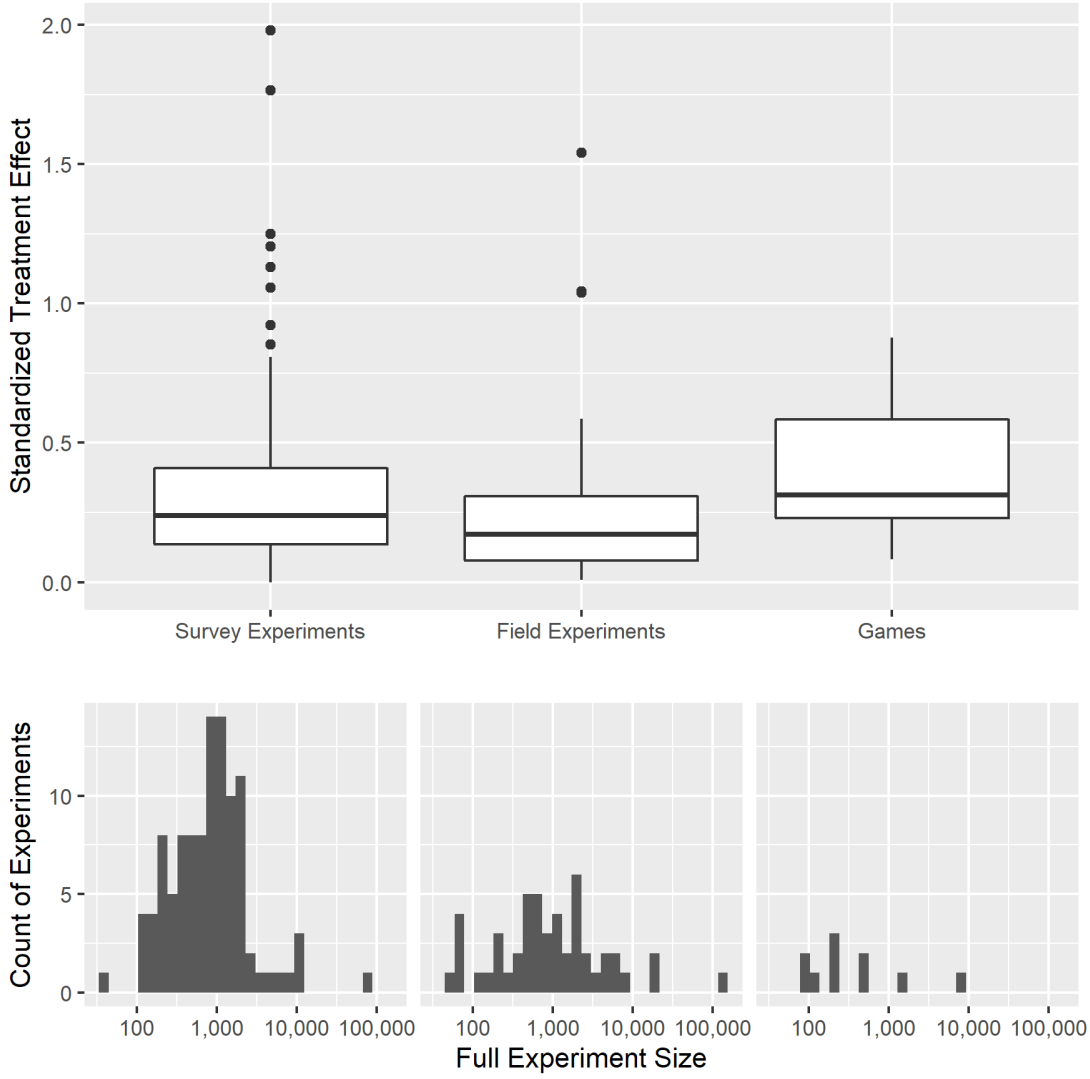

Figure 3: Distribution of Standardized Effect Sizes and Full Experiment Sample Sizes in Top Published Political Science Articles.

Schwarz, 2019). Thus, our reported statistics should be interpreted as overestimates of true standardized effect sizes, possibly by substantial margins.

Strikingly, our estimated distribution of the standardized effect sizes are largely concentrated within the range of 0.1 and 0.4, despite the likely overestimation. The median standardized effect size is approximately 0.25 for survey experiments, and it is slightly smaller for field experiments. The experiments that used economic games either as the treatment or an outcome measure show slightly larger, but not much larger effect sizes.

Comparing these results to the bias and standard error simulations in Figures 1 and 2 reveals a rather bleak picture of the utility of power estimation in political science. That is, the empirical distribution of the standardized effect sizes falls exactly in the region where the bias is highly sensitive to the value of τ_{std} and the standard error is largest. For example, with a typical scenario of the pilot size of 50 and full experiment size of 900 (second row, second column in Figures 1 and 2), the bias is as large as about -0.2 and the standard error is almost as large as 0.35. Indeed, even with an unrealistically large pilot size of 850, one can still expect a bias of 0.1 if the standardized treatment effect is 0.15 and the intended full experiment size is $1,300$ with the standard error of as large as 0.3 (bottom left).

4 Minimum Required Sample Size Estimation

Next, we turn to an alternative form of empirical power analysis: the MRSS estimation. This form of power analysis is probably most commonly used in empirical political science research. For example, in the pre-analysis plan registered on the EGAP repository, Dunham and Lieberman (2013) report that they "estimated the expected effect size via a pilot experiment with a sample of 100" to obtain a range for their likely MRSS (between 342 and 1043 participants) to achieve the power of 0.9, leading to the decision to pursue a sample size of 1,000.[8] We investigate whether this type of empirical power analysis is likely to produce reliable results both analytically and via simulations.

4.1 Analytical Results

We begin by deriving the expectation of the MRSS estimator defined in equation (6) to find its bias. Unfortunately, it turns out that this expectation does not exist, making the bias of $M\hat{R}SS$ undefined. To see

[8] Although this illustrates a typical use of pilot data to inform MRSS estimation, Dunham and Lieberman (2013) employ an ANOVA-based power analysis for their factorial design, as opposed to the t-test analyzed in the current paper. Whether the results in our analysis applies to their study is thus an open question.

this, note that:

$$\mathbb{E}\left[M\hat{R}SS\right] = \mathbb{E}\left\{\frac{4\left[\Phi^{-1}\left(1-\frac{\alpha}{2}\right)-\Phi^{-1}\left(1-\psi\right)\right]^2}{\hat{\tau}^2_{\text{std}}}\right\}$$
$$= 4\left[\Phi^{-1}\left(1-\frac{\alpha}{2}\right)-\Phi^{-1}\left(1-\psi\right)\right]^2 \mathbb{E}\left\{\frac{1}{\hat{\tau}^2_{\text{std}}}\right\}.$$

We now show $\mathbb{E}\left\{\frac{1}{\hat{\tau}^2_{\text{std}}}\right\}$ does not converge as long as the probability density function for $\hat{\tau}_{\text{std}}$, $f_{\hat{\tau}_{\text{std}}}(x)$ is continuous and bounded from above [9]. Letting $\overline{f}$ be the upper bound for $f_{\hat{\tau}_{\text{std}}}(x)$, we have

$$\mathbb{E}\frac{1}{\hat{\tau}^2_{\text{std}}} = \int_{-\infty}^{+\infty}\frac{1}{x^2}f_{\hat{\tau}_{\text{std}}}(x)dx$$
$$= \int_{-\infty}^{-1}\frac{1}{x^2}f_{\hat{\tau}_{\text{std}}}(x)dx + \int_{-1}^{0}\frac{1}{x^2}f_{\hat{\tau}_{\text{std}}}(x)dx + \int_{0}^{1}\frac{1}{x^2}f_{\hat{\tau}_{\text{std}}}(x)dx + \int_{1}^{+\infty}\frac{1}{x^2}f_{\hat{\tau}_{\text{std}}}(x)dx$$

We know $0 \leq f_{\hat{\tau}_p}(x) \leq f_{\hat{\tau}_p}(\tau) = \overline{f}$. Hence, $0 \leq \int_1^{+\infty}\frac{1}{x^2}f_{\hat{\tau}_p}(x)dx \leq \overline{f}$. Similarly, $0 \leq \int_1^{+\infty}\frac{1}{x^2}f_{\hat{\tau}_p}(x)dx \leq \overline{f}$. As a result, for a given N_p, the first term and the last term in the above summation is non-negative and bounded. We now investigate the property for the second term $\int_{-1}^{0}\frac{1}{x^2}f_{\hat{\tau}_p}(x)dx$. First, in the domain of $[-1,0]$, $f_{\hat{\tau}_p}(x)$ is greater than or equal to its minimum within this domain - $\int_{-1}^{0}\frac{1}{x^2}f_{\hat{\tau}_p}(x)dx \geq \min_{x\in[-1,0]} f_{\hat{\tau}_p}(x)\int_{-1}^{0}\frac{1}{x^2}dx = \min_{x\in[-1,0]} f_{\hat{\tau}_p}(x)\int_{0}^{1}\frac{1}{x^2}dx$[10]. Second, similarly, $\int_0^1\frac{1}{x^2}f_{\hat{\tau}_p}(x)dx \geq \min_{x\in[0,1]} f_{\hat{\tau}_p}(x)\int_0^1\frac{1}{x^2}dx$. Hence, the sum of the second and third term will be greater than $2\min_{x\in[-1,1]} f_{\hat{\tau}_p}(x)\int_0^1\frac{1}{x^2}dx$. Yet, $\int_0^1\frac{1}{x^2}dx$ is positive and not upper bounded. As a result, $\mathbb{E}\frac{1}{\hat{\tau}^2_p}$ is a sum of two non-negative terms with an upper bound and another non-negative term without an upper bound, and hence does not converge to a finite value.

The non-existence of $\mathbb{E}[M\hat{R}SS]$ implies existence of wild values in empirical sampling of $M\hat{R}SS$, as we will see later in the simulation. Moreover, a non-existence of $\mathbb{E}[M\hat{R}SS]$ also makes $\mathbb{V}[M\hat{R}SS]$ non-existent becauce the existence of the first moment is a condition for the second moment to be well-defined. In Appendix A.3, we show consistency for $M\hat{R}SS$ as N_p goes to infinity. However, pilot studies are by definition small, and asymptotic results would not help much in such small sample settings.

4.2 Simulations

Since neither the expectation nor the variance of the MRSS estimator exists, standard performance measures for estimators such as the bias and the root mean squared errors are not appropriate. Therefore, we employ an alternative approach to investigating the small sample performance of the MRSS estimator via simulations. Specifically, we simulate $1,000$ realizations of the MRSS estimator from a given DGP and examine their

[9] A probability density function is by definition bounded from below by 0.

[10] Normal pdf is continuous within such domain and thus extreme value theorem holds. The second equality holds because of symmetry of $\frac{1}{x^2}$.

empirical distribution to see how it changes as the pilot sample size and the standardized effect size varies. As before, the values of the parameters are set in accordance with the spectrum of most political science applications. Specifically, we set $\tau_{\text{std}} \in \{0.125, 0.25, 0.5, 1\}$ and $N_p \in \{n | n \in \mathbb{Z}, 10 \leq n \leq 5000\}$.

Simulation Procedure. We follow the following steps for each combination of the τ_{std} and N_p values:

1. Randomly draw $\frac{N_p}{2}$ realizations of Y for the treatment group from $\mathcal{N}(\tau, 4^2)$, and $\frac{N_p}{2}$ realizations of Y for the control group from $\mathcal{N}(0, 4^2)$ to simulate the pilot study, where $\tau_{\text{std}} = \tau/4$.

2. Calculate the difference-in-means estimate of the treatment effect $\hat{\tau}$.

3. Estimate the MRSS using the plug-in estimator in equation (6), setting $\alpha = 0.05$ and $\psi = 0.8$:

$$\hat{MRSS} \;=\; \left\lceil \frac{4\left[1.96 - \Phi^{-1}(0.2)\right]^2}{\hat{\tau}^2_{\text{std}}} \right\rceil.$$

4. Repeat Steps 1 to 3 1,000 times to obtain a simulated sampling distribution of $\hat{MRSS}$.

Results. Figure 4 shows the simulated sampling distributions of the MRSS estimation across four different values of the standardized effect size (top to bottom panels) and different pilot sample sizes (horizontal axis).

The results exhibit a striking degree of sampling variability for the MRSS estimator across a wide range of parameter values. The simulated sampling distribution of the MRSS estimator is extremely right-skewed, with the mean almost always exceeding the 95th percentile across all parameter values (also notice the y-axis in these plots are on the log-10 scale). For example, suppose that a researcher is estimating the true standardized treatment effect of 0.125 (top plot), an empirically likely scenario based on our literature review (Figure 3). With the desired power of $\psi = 0.8$ and testing the null hypothesis at $\alpha = 0.05$, the MRSS would be 2,008 (red horizontal line). However, since the true effect size is unknown to the researcher *a priori*, they decide to run a pilot experiment with the sample size of 100 to obtain an estimate of τ_{std}. Under this scenario, unfortunately, the researcher could obtain a point estimate for $\hat{\tau}_{\text{std}}$ from anywhere between -2.54 and 2.39 out of our 1,000 simulations. Even if we focus on the central 90% of the sampling distribution, the range is still quite wide (-1.17 for the 5th percentile and 2.10 for the 95th percentile). Remarkably, plugging these estimates into the MRSS estimator reveals that the implied range of the MRSS estimator would be between 5 and 27 billion, and the 5th-to-95th percentiles range will be between 12 and 16,807. On average, the estimator vastly overestimates the true MRSS (27.5 million), while the median of the sampling distribution actually underestimates (110).[11]

[11] We note that these numbers should be interpreted only as illustrative of the sampling behavior of the MRSS estimator, as

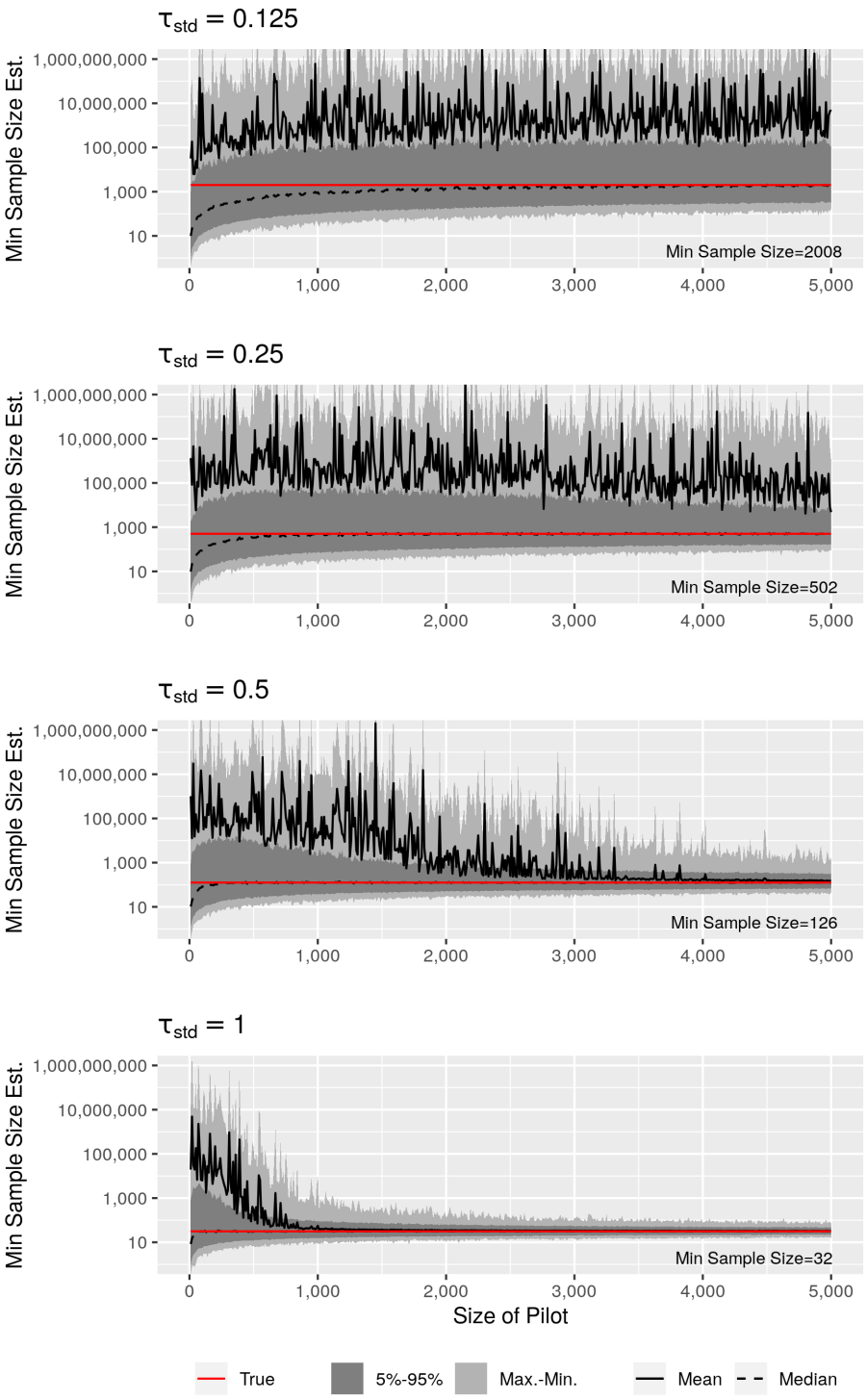

Figure 4: Simulated Sampling Distributions of the MRSS Estimator. Each of the four plots presents characteristics of the simulated sampling distribution of the MRSS estimator (maximum, 95th percentile, mean, median, 5th percentile and minimum) as functions of the pilot study sample size (N_p), with each plot corresponding to an assumed size of the true standardized effect size (τ_{std}). The red horizontal line indicates the true MRSS corresponding to τ_{std} in each plot. Note that the y-axis is on the log-10 scale for interpretability.

Now, suppose instead that the researcher uses the pilot sample of size 5,000 to combat the uncertainty in the MRSS estimate. A sample of this size is rather unrealistic for a typical pilot study in political science field or survey experiments, but conceivable if the estimate instead comes from a previous large-scale study testing the same hypothesis. Unfortunately, the situation remains bleak even under this alternative scenario: the researcher would still get an estimate MRSS from anywhere between 161 and 4.2 billion (313 and 103,864 for the 5th-to-95th percentile range), with the empirical mean and median at 5.1 million and 1,790, respectively.

Finally, consider the most optimistic scenario where the true standardized effect size is 1, an atypically large effect in political science applications based on our literature review (Figure 3). With a large pilot study on a sample of size 1,000, the researcher would now get a MRSS estimate from the more reasonable range of 9 to $13,250$, with the 5th-to-95th percentile range of 16 and 92. The empirical mean and median of the sampling distribution have also been stabilized at 56 and 32, respectively. Although these results might look more promising, one must realize that the true MRSS is as small as 32 at these parameter values. What this implies is that, when the researcher conducts a pilot experiment of size 1,000 in this scenario, the estimated ATE from the pilot study itself is almost always highly statistically significant. An MRSS estimation for a separate "full" experiment thus becomes rather pointless under such a scenario.

In sum, our simulation study reveals that the MRSS estimation based on an empirical estimate of the standardized effect size is likely to be of limited use in typical political science research. Estimates are unusefully variable under most scenarios relevant for political scientists, and when the estimation is likely to yield a reliable result, the treatment effect estimate from the pilot study is already highly statistically significant almost for sure.

4.3 RMSS Estimation in Practice

In practice, researchers tend to estimate the MRSS only when their pilot study turns out to be "promising." That is, researchers often use a pilot study to explore the range of possible design choices, such as the content of the treatment, and then decide on a study specification to be used in the full experiment. Empirical power analysis is frequently conducted in such a context, where the researcher picks out a treatment-outcome combination that yielded an effect estimate that seemingly indicates the existence of an effect but does not quite reach the conventional level of statistical significance.

The research practice of this kind clearly deviates from the setup of the simulations we employed in Section 4.2, which assumes that MRSS estimation is conducted no matter what the pilot treatment effect estimate might turn out to be. Instead, the distribution of the MRSS estimates *that are actually conducted* is

opposed to estimators of the corresponding order statistics or moments of the true sampling distribution of the MRSS estimator. In fact, they are likely to be vastly different (except for the median) in another run of the same simulation procedure, though the general pattern of variability will be the same.

likely to be conditional on the pilot effect estimate attaining a certain range of statistical significance. Under this alternative, more realistic sampling regime, the distribution of the MRSS estimates will be much less variable, since the extreme draws of the estimates shown in Figure 4 correspond to the pilot effect estimates that are close to zero.

Does this imply that our pessimistic conclusion from the simulation analysis is irrelevant in practice? Unfortunately, the answer is no. Indeed, a closer look at the problem reveals that the practice of conditioning empirical power analysis on the event of a promising pilot result is flawed and should be abandoned. Here, we discuss two reasons why this practice is problematic, using our framework of analyzing power analysis as a statistical inference problem.

First, one may prematurely give up on an experiment that would otherwise be both statistically and substantively significant by giving up the full experiment after seeing a large p-value in the pilot study (see also Kraemer et al., 2006). Consider the case where the standardized treatment effect equals 0.25 as an example. The sampling distribution for the standardized treatment effect estimated in a pilot study with the sample size of 100 has the mean of 0.25 and the variance of 1/25. Now, assume that the researcher will only proceed with the research when the t-statistic for the pilot effect estimate falls within a "promising but not quite there" range, say between 1 and 2. Using the standard normal distribution as an approximate reference distribution for the t-statistic, it turns out there is only 34% chance that the researcher would actually conduct the power analysis (and thus potentially proceed to the full experiment), despite the fact that the standardized effect size is 0.25 in truth.[12] In other words, the probability that the researcher would prematurely give up the research which would otherwise be promising is 50%[13].

Second, more fundamentally, it turns out that basing one's decision on whether to proceed with the full experiment on the p-value for the treatment effect estimated from the pilot study makes the pilot study completely irrelevant. We show analytically here that there is a one-to-one mapping from the p-value (or equivalently the t-statistic) to the MRSS. In other words, the p-value and the pilot sample size alone fully determine the MRSS, regardless of whatever values are observed for any of the actual variables in the experiment. Hence, should one base her decision on the p-value from the pilot study, given the size of the pilot, the range of the required minimum sample size for the research would always be the same. This means that one can estimate the MRSS from a hypothetical pilot study of sample size N_p without actually ever conducting it.

Specifically, since we know that $\mathbb{E}[\hat{\tau}_{\text{std}}] = \tau_{\text{std}}$ and $\mathbb{V}[\hat{\tau}_{\text{std}}] = 4/N_p$, the t-statistic for the pilot effect

[12] $1 \leq \frac{\hat{\tau}}{\frac{1}{5}} \leq 2 \Rightarrow \frac{1}{5} \leq \hat{\tau} \leq \frac{2}{5} \Rightarrow \Pr\left(\frac{1}{5} \leq \hat{\tau} \leq \frac{2}{5}\right) = \Phi\left(\frac{\frac{2}{5}-\frac{1}{5}}{\frac{1}{5}}\right) - \Phi\left(\frac{\frac{1}{5}-\frac{1}{5}}{\frac{1}{5}}\right) = 0.34.$

[13] We deduct the probability of the researcher obtaining a statistically significant result from pilot: $1 - 0.34 - \Pr\left(\hat{\tau}\ significant\right) = 1 - 0.34 - \Pr\left(\frac{\hat{\tau}}{\frac{1}{5}} > 2\right) = 1 - 0.34 - \left(1 - \Phi\left(\frac{\frac{2}{5}-\frac{1}{5}}{\frac{1}{5}}\right)\right) = 0.50$

estimate would be $\frac{\hat{\tau}_{\text{std}}}{2} \cdot \sqrt{N_p}$. Now, suppose the researcher adopts the decision rule of proceeding only when the t-statistic is between 1 and 2. Using the standard normal reference distribution, the estimator for the standardized treatment effect should fall between $\frac{2}{\sqrt{N_p}}$ and $\frac{4}{\sqrt{N_p}}$ to satisfy this rule. We plug in this range into the equation for the MRSS to obtain:

$$\frac{\left[\Phi^{-1}\left(1-\frac{\alpha}{2}\right)-\Phi^{-1}\left(1-\psi\right)\right]^2 N_p}{16} \leq N_f \leq \frac{\left[\Phi^{-1}\left(1-\frac{\alpha}{2}\right)-\Phi^{-1}\left(1-\psi\right)\right]^2 N_p}{4}.$$

Noting all parameters to determine the range for the minimum required sample size are known before a pilot is conducted, the pilot study would not provide any additional information to the determination for the range of the minimum required sample size.

In sum, the practice of conditioning whether to conduct a MRSS estimation on the statistical significance of the pilot test should be abandoned. Ironically, it is this very practice that might have given empirical researchers an illusion of usefulness for MRSS estimation: by estimating MRSS only when they see a promising effect estimate in the pilot study, researchers have effectively been pre-fetching MRSS estimates that are confined within the range excluding extreme values.

5 Practical Recommendations

Our analysis of two frequently used techniques for empirical power analysis – power estimation and MRSS estimation – has revealed their serious limitations in the context of applied research in political science. Here, we offer practical guidelines for empirical researchers who seek to employ power analysis in their research.

First, empirical power analysis is generally *not* recommended in most scenarios encountered in political science research, such as field experiments and online survey experiments. Given the range of standardized effect sizes we found in the experimental studies recently published in top political science journals (Figure 3), neither power estimation nor MRSS estimation is likely to produce reliable results. The problem is particularly severe when researchers use ATE and outcome variance estimates from a small pilot study as inputs to power calculation software, since the large uncertainty in those estimates are further amplified by the nonlinear transformation involved in the power or MRSS estimator. Should researchers still decide to proceed with an empirical power analysis, they should always calculate how much estimation uncertainty their power or MRSS estimate contains and report a formal measure of estimation uncertainty (e.g. standard error) along with the point estimate. That is, treat empirical power analysis just like an estimation problem and apply the same reporting standard.

Our critique, however, does not apply to *non-empirical* power analysis techniques which do not involve

an empirical estimation of the standardized effect size. As discussed earlier, power analysis can be non-empirical if all of the parameters are unambiguously specified either by an external constraint or based on a normative criterion. For example, minimum detectable effect size (MDES) calculation is a common method for calculating the smallest standardized treatment effect that can be found statistically significant with a specified probability (ψ), given a significance level (α) and sample size (N_f) to be used for the test. Since these parameters are not empirically estimated and are free of statistical uncertainty, one can safely calculate MDES and compare it against a pre-set threshold value that the researcher considers substantively large enough for the study to be worthwhile. A close variant of this approach is MRSS calculation using a normatively provided effect size. For example, suppose that a grant-making agency or an implementing partner organization proposes a target effect size as a condition of their funding or partnership. The effect size in such scenario is a non-empirical value externally provided without statistical uncertainty, and the MRSS formula can be used for calculating the sample size that corresponds to the specified parameter values. The same logic applies to power calculation with an externally specified effect size. That said, once should also keep in mind that the formulas used for non-empirical power and MRSS calculation are still the same formula, which is of the functional form that causes much of the variability problems we have discussed for empirical power/MRSS estimation. This implies that the results of non-empirical power/MRSS calculation will also be sensitive to even small changes in the pre-specified effect sizes.

As we discussed in Section 4.3, a particularly problematic form of empirical power analysis is to use pilot data to obtain an initial estimate of the target treatment effect, and proceed to a formal power analysis only when the estimate turns out to be "promising" (i.e., moderately statistically significant). Conditional on the level of statistical significance (i.e., p-value or t-statistic) and the pilot sample size, the actual pilot data contain *no information* for power calculation. In other words, if the researcher decides (prior to collecting any data) that they will propose a full study if the pilot study of size N_p produces a test result with a p-value of certain range, then the researcher need not collect any pilot data to calculate the MRSS range for the full experiment.

Given our findings, are pilot studies (or any preexisting study testing the same hypothesis) useful at all for power analysis? One remaining way pilot data can empirically inform power analysis is to transform the standard effect size into the substantively meaningful scale of an actual outcome variable. The MDES calculation, which we recommended as an example of non-empirical power analysis, outputs the minimum detectable *standardized* effect size, i.e., the effect size in standard deviations of the outcome variable. To evaluate whether the result satisfies a normative threshold, one must be able to interpret the MDES output substantively in the original scale of the outcome variable of interest. This, in turn, requires an empirical estimate of the standard deviation of the outcome in most scenarios. As it turns out, estimating the original

effect size from a standard deviation estimate and a fixed standardized effect size is far safer than the power or MRSS estimation. That is, a typical sample size for a pilot experiment turns out to be large enough for a sufficiently precise estimate of the outcome standard deviation, which translates into a reasonably reliable estimate of the raw effect size. Thus, pilot experiments are still an important component of empirical political science research, especially given their other various roles not directly related to power analysis (Leon et al., 2011).

Finally, it is important to reiterate that the above list of recommendations is an addition to the growing body of critiques of power analysis (e.g., Rothman and Greenland, 2018; Gelman and Carlin, 2014). Other scholars have criticized various research practices surrounding power analysis, including caution against post-hoc power calculation using the observed effect size (Gelman, 2019) and against proceeding to a full experiment only when the pilot-based MRSS is considered feasible (Albers and Lakens, 2018). We advise researchers to pay attention to these existing recommendations as well as ours.

6 Conclusion

Power analysis plays an increasingly prominent role in today's political science. Researchers routinely conduct power calculations in the design stage of their empirical research and include their outputs as part of pre-analysis plans or grant proposals. However, not all kinds of power analyses are created equal. In this paper, we introduced a conceptual distinction between empirical power analysis and non-empirical power analysis, defining the former as any form of power analysis that involves empirical data as inputs to the calculation. We then proposed an analytical framework which views empirical power analysis as a statistical estimation problem for the purpose of systematically investigating its reliability. We applied this framework to the two most common forms of empirical power analysis – power estimation and MRSS estimation – to analyze their properties as estimators, with particular focus on the range of parameter values political scientists are likely to encounter in their empirical research.

The results of our theoretical and simulation-based analysis revealed a rather bleak picture of the utility of empirical power analysis in political science research. We found that the power estimates are strongly biased in unpredictable directions and sensitive to small changes in the input effect size estimates when the study parameters such as the true standardized effect size and the pilot sample size take on values typically expected in political science research. We also showed that the MRSS estimator has infinite expectation and variance over repeated samples, making it extremely variable when the pilot sample size and the true standardized effect size are within reasonable ranges. Moreover, we identified an important fallacy in the current practice of empirical power analysis based on pilot data, which involves the estimation of MRSS

conditional on a pre-specified level of statistical significance. Based on these findings, we offered a set of practical recommendations for applied researchers, generally advising against empirical power analysis as it is practiced currently.

Our analysis leaves a number of questions for future investigations. First, we have focused on power analysis for the simplest form of experimental design – the classical randomized experiment of a binary treatment on a simple random sample of subjects – to derive all our results. Although this is the workhorse method behind most of the ubiquitous power calculators and therefore likely to cover much of the actual empirical research, power analysis for other types of study designs of course does exist. Given the move towards more complex experimental designs and the use of pretreatment covariates for improving efficiency, it is important for future work to analyze non-traditional forms of power analysis, some of which might be simulation-based. Second, the poor performance of existing bias correction methods when applied to the power estimation suggests the need for the development of alternative statistical methods or possible design-based solutions to aid empirical power analysis. Finally, the current paper brackets the more fundamental criticism against the concept of power analysis itself, which ties into a broader argument criticizing null hypothesis significance testing in applied research. Indeed, our findings should be understood in the context of the ongoing discipline-wide debate about what constitutes good empirical science.

Appendix

A.1 Properties of the Plug-in Power Estimator

To derive properties of the power estimator in equation (7), we begin by deriving the exact mean and variance of $\hat{\tau}_{\text{std}}$, our estimate of the standardized effect size from the pilot data. We begin by noting that

$$\mathbb{E}[\hat{\tau}_p] \;=\; \tau, \quad \mathbb{V}(\hat{\tau}_p) = \frac{4\sigma^2}{N_p}.$$

Since $\hat{\tau} \perp\!\!\!\perp \hat{\sigma}^2$ (CITE), we have

$$\begin{aligned}\mathbb{E}\hat{\tau}_{\text{std}} = \mathbb{E}\frac{\hat{\tau}}{\sqrt{\hat{\sigma^2}}} &= \mathbb{E}\hat{\tau}\mathbb{E}\frac{1}{\sqrt{\hat{\sigma^2}}} \\ &= \tau\mathbb{E}\frac{1}{\sqrt{\hat{\sigma^2}}}\end{aligned}$$

At this point we need some distributional assumptions to proceed with the derivation without invoking asymptotic results, which we want to avoid since N_p is small in many applied settings. Assume Y is normally distributed. Then, $\frac{(N_p-1)}{\sigma^2}\hat{\sigma^2} \sim \chi^2_{N_p-1}$, and therefore $\sqrt{\frac{N_p-1}{\sigma^2}} \cdot \sqrt{\hat{\sigma^2}} \sim \chi_{N_p-1}$. Indeed, $\sqrt{\frac{\sigma^2}{N_p-1}} \cdot \frac{1}{\sqrt{\hat{\sigma^2}}} \sim Inv - \chi_{N_p-1}$.

Following Lee (2012), when $N_p > 5$,

$$\mathbb{E}\sqrt{\frac{\sigma^2}{N_p-1}} \cdot \frac{1}{\sqrt{\hat{\sigma^2}}} \approx \sqrt{\frac{1}{N_p-\frac{5}{2}}}$$

Hence,

$$\begin{aligned}\mathbb{E}\frac{\hat{\tau}}{\sqrt{\hat{\sigma^2}}} &\approx \frac{\tau\sqrt{\frac{N_p-1}{\sigma^2}}}{\sqrt{N_p-\frac{5}{2}}} \\ &= \frac{\tau}{\sigma}\sqrt{1+\frac{3}{2N_p-5}} \qquad (8)\end{aligned}$$

Thus, $\hat{\tau}_{\text{std}}$ is downward biased when $N_p < 4$ and upward biased when $N_p > 4$. However, the bias is negligibly small even for a moderately sized pilot experiment.

Next, we consider the variance of $\hat{\tau}_{\text{std}}$. Note that

$$\begin{aligned}
\mathbb{E}\left[\frac{\hat{\tau}}{\sqrt{\hat{\sigma}^2}} - \mathbb{E}\frac{\hat{\tau}}{\sqrt{\hat{\sigma}^2}}\right]^2 &= \mathbb{E}\left[\frac{\hat{\tau}^2}{\hat{\sigma}^2} - \mathbb{E}^2\frac{\hat{\tau}}{\sqrt{\hat{\sigma}^2}}\right] \\
&= \mathbb{E}\frac{\hat{\tau}^2}{\hat{\sigma}^2} - \left(\mathbb{E}\frac{\hat{\tau}}{\sqrt{\hat{\sigma}^2}}\right)^2 \\
&= \mathbb{E}\hat{\tau}^2\mathbb{E}\frac{1}{\hat{\sigma}^2} - \left(\mathbb{E}\frac{\hat{\tau}}{\sqrt{\hat{\sigma}^2}}\right)^2 \\
&= \mathbb{E}\hat{\tau}^2\mathbb{E}\frac{1}{\hat{\sigma}^2} - \frac{\tau^2}{\sigma^2}\left(1 + \frac{3}{2N_p - 5}\right)
\end{aligned}$$

As we know $\frac{(N_p-1)}{\sigma^2}\hat{\sigma^2} \sim \chi^2_{N_p-1}$, we have $\frac{1}{(N_p-1)} \cdot \frac{1}{\hat{\sigma^2}}\sigma^2 \sim Inv - \chi^2_{n-1}$. Hence,

$$\mathbb{E}\frac{1}{\hat{\sigma^2}} = \frac{N_p - 1}{N_p - 3} \cdot \frac{1}{\sigma^2}.$$

Further,

$$\begin{aligned}
\mathbb{E}\hat{\tau}^2 &= var(\hat{\tau}) + \tau^2 \\
&= \frac{4\sigma^2}{N_p} + \tau^2
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{V}\hat{\tau}_{\text{std}} &= \left(\frac{4\sigma^2}{N_p} + \tau^2\right)\left(\frac{N_p - 1}{N_p - 3} \cdot \frac{1}{\sigma^2}\right) - \frac{\tau^2}{\sigma^2}\left(1 + \frac{3}{2N_p - 5}\right) \\
&= \frac{4(N_p - 1)}{n(N_p - 3)} + \frac{\tau^2}{\sigma^2}\frac{N_p - 1}{(N_p - 3)(2N_p - 5)}.
\end{aligned} \tag{9}$$

Thus, $\mathbb{E}[\hat{\tau}_{\text{std}}]$ and $\mathbb{V}[\hat{\tau}_{\text{std}}]$ are given by equations (8) and (9), respectively. While we could in theory continue the derivation with these results, we instead opt to ignore the sampling variability of $\hat{\sigma}^2$ for simplicity, effectively assuming $\hat{\tau}_{\text{std}} = \hat{\tau}/\sigma$. This can be justified for two reasons. First, as noted above, $\mathbb{E}[\hat{\tau}_{\text{std}}]$ is approximately unbiased unless N_p is very small. Second, $\mathbb{V}(\hat{\tau}_{\text{std}})$ can be shown to be greater than $\mathbb{V}(\hat{\tau}/\sigma)$ as long as $N_p > 4$, since $\mathbb{V}\frac{\hat{\tau}}{\sigma} = \frac{4}{N_p}$ and

$$\mathbb{V}\hat{\tau}_{\text{std}} - \mathbb{V}\frac{\hat{\tau}}{\sigma} = \frac{4}{N_p} \cdot \left(\frac{2}{N_p - 3}\right) + \tau^2_{\text{std}}\frac{N_p - 1}{(N_p - 3)(2N_p - 5)} > 0$$

when $N_p > 4$.

Now, define the random variable $x = 1.96 - \frac{\hat{\tau}_{\text{std}}\sqrt{N_f}}{2}$ and scalar $x_{true} = 1.96 - \frac{\tau_{\text{std}}\sqrt{N_f}}{2}$, we get $\hat{\psi} = 1 - \Phi(x)$,

where

$$x \sim \mathcal{N}\left(1.96 - \frac{\tau_{\text{std}}}{2}\sqrt{N_f}, \frac{N_f}{N_p}\right)$$

To evaluate $\mathbb{E}\hat{\psi}$, we would need to evaluate $\mathbb{E}\Phi(x)$. Since $\Phi(\cdot)$ cannot be expressed in closed form, we apply the following approximation (CITE):

$$\Phi(z) \approx \begin{cases} 0 & z \leq -2.6 \\ 0.01 & -2.6 < z \leq -2.2 \\ 0.5 - \frac{-4.4z - z^2}{10} & -2.2 < z \leq 0 \\ \frac{4.4z - z^2}{10} + 0.5 & 0 < z \leq 2.2 \\ 0.99 & 2.2 < z \leq 2.6 \\ 1 & z > 2.6 \end{cases}$$

Hence,

$$\begin{aligned} \mathbb{E}\Phi\left(x\right) = 0.01 \times \int_{-2.6}^{-2.2} p(x)dx + \int_{-2.2}^{0} \left[0.5 - \frac{-4.4x - x}{10}\right] p(x)dx \\ + \int_{0}^{2.2} \left[\frac{4.4x - x^2}{10} + 0.5\right] p(x)dx + 0.99 \times \int_{2.2}^{2.6} p(x)dx + \int_{2.6}^{+\infty} p(x)dx \end{aligned} \tag{10}$$

where $p(x)$ is the probability density function for random variable x.

The approximate form of $\mathbb{E}[\hat{\psi}]$ allows us to investigate the likely behavior of the bias of the estimator in several scenarios:

1. If the density function of x, $p(x)$ is mostly to the left of -2.2, $\mathbb{E}\Phi\left(x\right) \approx 0.01$, and $\Phi\left(x_{true}\right) \approx 0.01$, and the bias would be small.

2. If the density function of x, $p(x)$ is mostly concentrated between-2.2 and 0, then

$$\mathbb{E}\Phi\left(x\right) \approx 0.5 - \frac{-\mathbb{E}4.4x - \mathbb{E}x^2}{10}$$

while

$$\Phi\left(x_{true}\right) \approx 0.5 - \frac{-4.4x_{true} - x_{true}^2}{10}$$

the bias of the power would be

$$\begin{aligned}[1-\mathbb{E}\Phi(x)]-[1-\Phi(x_{true})] &= \frac{x_{true}^2-\mathbb{E}x^2}{10}\\ &= -\frac{var(x)}{10}\\ &= -\frac{N_f}{10\centerdot N_p}\end{aligned}$$

3. If the density function of x, $p(x)$ is mostly concentrated between 0 and 2.2, then

$$\mathbb{E}\Phi(x) \approx \frac{4.4\mathbb{E}x-\mathbb{E}x^2}{10}+0.5$$

while

$$\Phi(x) \approx \frac{4.4x_{true}-x_{true}^2}{10}+0.5$$

and the bias of the power would be

$$\begin{aligned}[1-\mathbb{E}\Phi(x)]-[1-\Phi(x)] &= \frac{\mathbb{E}x^2-x_{true}^2}{10}\\ &= \frac{var(x)}{10}\\ &= \frac{N_f}{10\centerdot N_p}\end{aligned}$$

4. If the density function of x, $p(x)$ is mostly to the right of 2.2,$\mathbb{E}\Phi\left(\widehat{X_1}\right) \approx 0.99$, and $\Phi(x_{true}) \approx 0.99$, and the bias would be small.

These observations allow us to make the following conclusions about the bias of $\hat{\psi}$:

- If the intended size for the full experiment is quite large, $x_{true} = 1.96 - \frac{\tau_{\text{std}}}{2}\sqrt{N_f}$ would be quite small. Hence, $p(x)$ would be mostly to the left of -2.2, the bias for power estimation would be small.
- If the true standardized treatment effect is quite large, $x_{true} = 1.96 - \frac{\tau_{\text{std}}}{2}\sqrt{N_f}$ would be quite small as well. $p(x)$ would be mostly to the left of -2.2, the bias for power estimation would be small.
- If the intended size for the full experiment is not large enough to push $p(x)$ to the left of -2.2, the larger the intended full experiment is, the larger the bias, because the absolute value for the bias would be in proportion to ratio of the full experiment size and the pilot size, $\frac{N_f}{N_p}$.
- Given the size of the pilot and the intended full experiment, the direction of the bias could be easily flipped even though there is only a small difference in the true standardized treatment effect. This is

because the direction of the bias all depends whether $p(x)$ is more heavily distributed on the negative part or on the positive part.

A.2 Bias Correction Methods

To fully illustrate the point that the bias for Equation (7) cannot be corrected due to our ignorance of the local convexity and local concavity in the neighborhood of the equation around the true τ, we try several traditional bias correction methods on power estimation. Our simulation results show that none of these bias correction methods work in this setting.

Figure A.1 compares the bias of bias-corrected estimators with the bias of the naive estimator when true τ changes from 0 to 8. In this simulation, we specify $\sigma = 4$, so equivalently the standardized treatment effects range from 0 to 2. The pilot size is set at $N_p = 100$, and the full size is set at $N_f = 800$, a very common scenario for political scientists. We repeat the sampling process by $1,000$ times to obtain the simulated bias.

In Figure A.1, the solid black curve shows the simulated bias for the naive difference-in-means estimator. Consistent with our simulations in Figure 1, the bias looks like a check sign. We over-estimate the power when the treatment effect is smaller, but under-estimate the power when the treatment effect gets larger. Yet, the bias approaches zero when the treatment effect is really large.

The red dashed line and greed dotted line record the bias for two versions of estimators obtained from bootstrapping bias correction method (Tibshirani and Efron, 1993). Despite a slight improvement on the cases when $\tau < 1$, the bias is essentially identical, if not slightly larger, when $\tau > 1$. As a result, these bootstrapping methods we have tried fail to universally reduce the bias. Indeed, their relative performance compared with the naive estimator depends on specific values of true τ, and sample sizes N_p, N_f.

The blue dashed line, the purple solid line and the orange dashed line show the bias for a jackknife bias corrected estimator, and two versions of double bootstrap bias correction estimators (Tibshirani and Efron, 1993). Similar to the case for the bootstrap bias correction method, none of these estimators show a significant improvement in bias reduction.

The light blue dashed line, on the other hand, is an oracle power estimator where we assume the researchers know the true treatment effect, but need to estimate the standard deviation from the pilot study. Thus, the researcher plugs the true treatment effect and the estimated standard deviation into Equation (7) to obtain her power estimation. This setting is not realistic, but it illustrates the validity of our simplifying assumption in the previous section of appendix where we assume away the sampling error for variance estimation and only focus on the sampling error for treatment effect.

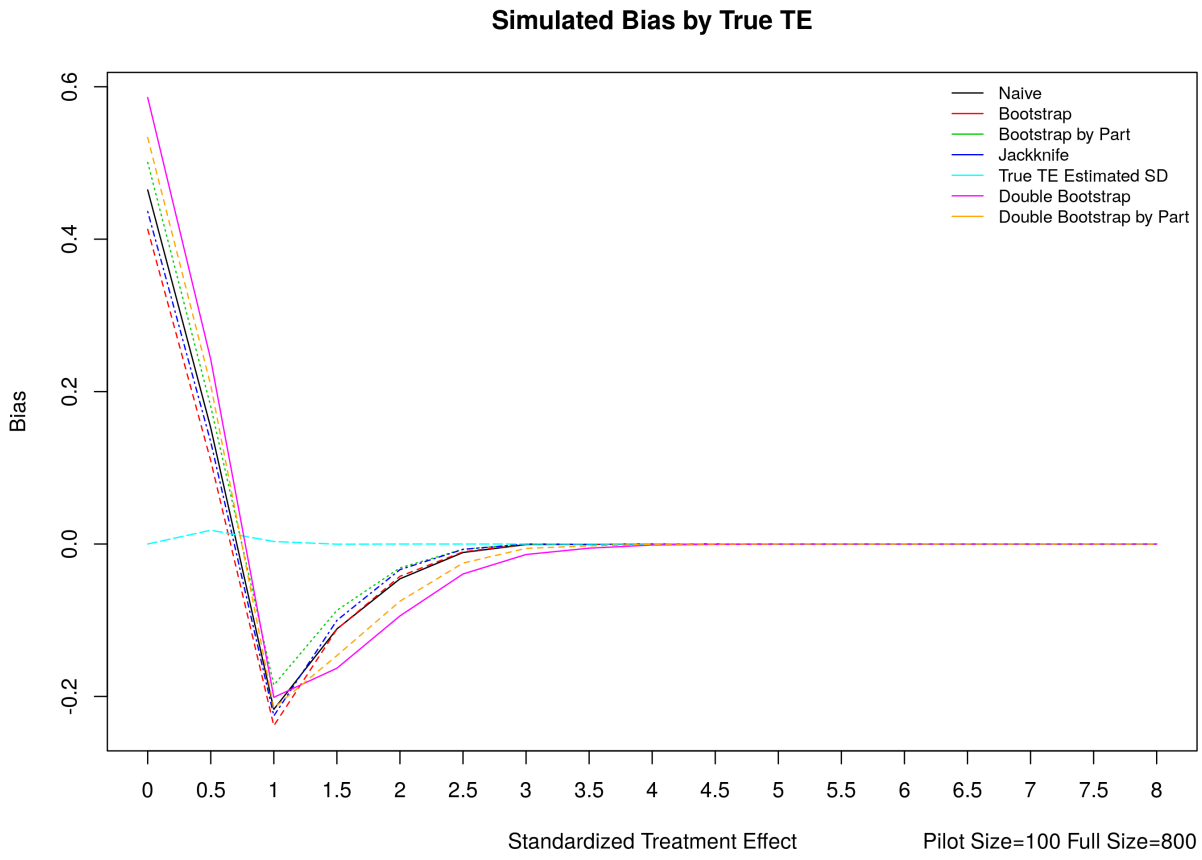

Figure A.1: Comparison of the Bias Correction Techniques Applied to the Power Estimator. The results show that none of the existing techniques appreciably improves estimates over the naive uncorrected estimator. (The light blue dashed line represents the unfeasible "oracle" estimator where the true effect size (but not the standard deviation) is known to the researcher.)

A.3 Consistency for $M\hat{R}SS$

We claim the sequence $M\hat{R}SS_{N_p} = \frac{4\left[\Phi^{-1}\left(1-\frac{\alpha}{2}\right)-\Phi^{-1}(1-\psi)\right]^2}{\hat{\tau}^2_{\text{std}}}$ converges in probability towards $N = \frac{4\left[\Phi^{-1}\left(1-\frac{\alpha}{2}\right)-\Phi^{-1}(1-\psi)\right]^2}{\tau^2_{\text{std}}}$. To show that, we need to find N_{upper} such that for any $\varepsilon > 0$, $\delta > 0$, we have $\mathbb{P}\left(\left|M\hat{R}SS_{N_p} - N\right| \leq \varepsilon\right) < \delta$, for all $N_p \geq N_{upper}$.

With the classical central limit theorem, we have $\hat{\tau}_{\text{std}} \xrightarrow{d} \mathcal{N}\left(\tau, \frac{4}{N_p}\right)$. Thus, for any u and $\frac{1}{5}\delta$, there exists N_k such that for all $N_p \geq N_k$,

$$\mathbb{P}\left(\hat{\tau}_{\text{std}} \leq u\right) \in \left[\Phi\left(\frac{u-\tau}{2/\sqrt{N_p}}\right) - \frac{1}{5}\delta, \Phi\left(\frac{u-\tau}{2/\sqrt{N_p}}\right) + \frac{1}{5}\delta\right]$$

remembering $\mathbb{P}\left(\hat{\tau}_{\text{std}} \leq u\right) = \mathbb{P}\left(\frac{\hat{\tau}_{\text{std}}-\tau}{2/\sqrt{N_p}} \leq \frac{u-\tau}{2/\sqrt{N_p}}\right)$, where $\Phi(.)$ is the standard normal cdf.

Let $C = 4\left[\Phi^{-1}\left(1-\frac{\alpha}{2}\right) - \Phi^{-1}\left(1-\psi\right)\right]^2$, so when $N_p \geq \left(\frac{\Phi^{-1}\left(\frac{1}{5}\delta\right)}{\frac{\tau_{\text{std}}}{2}\left(\sqrt{\frac{C}{C+\varepsilon\tau^2_{\text{std}}}}-1\right)}\right)^2$ and $N_p \geq N_k$

$$\begin{aligned}
\mathbb{P}\left(\frac{C}{\hat{\tau}^2_{\text{std}}} \geq \frac{C}{\tau^2_{\text{std}}} + \varepsilon\right) &= \mathbb{P}\left(\hat{\tau}^2_{\text{std}} \leq \frac{C}{\frac{C}{\tau^2}+\varepsilon}\right) = \mathbb{P}\left(\hat{\tau}_{\text{std}} \leq \sqrt{\frac{C}{\frac{C}{\tau^2}+\varepsilon}}\right) \\
&\leq \Phi\left(\frac{\sqrt{\frac{C}{\frac{C}{\tau^2}+\varepsilon}} - \tau_{\text{std}}}{2/\sqrt{N_p}}\right) + \frac{1}{5}\delta \\
&\leq \Phi\left(\sqrt{N_p}\frac{\tau_{\text{std}}}{2}\left(\sqrt{\frac{C}{C+\varepsilon\tau^2_{\text{std}}}} - 1\right)\right) + \frac{1}{5}\delta \\
&\leq \frac{1}{5}\delta + \frac{1}{5}\delta = \frac{2}{5}\delta
\end{aligned}$$

Similarly, when $N_p \geq \left(\frac{\Phi^{-1}\left(1-\frac{1}{5}\delta\right)}{\frac{\tau_{\text{std}}}{2}\left(\sqrt{\frac{C}{C-\varepsilon\tau^2_{\text{std}}}}-1\right)}\right)^2$ and $N_p \geq N_k$, we have

$$\begin{aligned}
\mathbb{P}\left(\frac{C}{\hat{\tau}^2_{\text{std}}} \leq \frac{C}{\tau^2_{\text{std}}} - \varepsilon\right) &= \mathbb{P}\left(\hat{\tau}^2_{\text{std}} \geq \frac{C}{\frac{C}{\tau^2}-\varepsilon}\right) = \mathbb{P}\left(\hat{\tau}_{\text{std}} \geq \sqrt{\frac{C}{\frac{C}{\tau^2}-\varepsilon}}\right) \\
&\leq 1 - \Phi\left(\frac{\sqrt{\frac{C}{\frac{C}{\tau^2}-\varepsilon}} - \tau_{\text{std}}}{2/\sqrt{N_p}}\right) + \frac{1}{5}\delta \\
&\leq 1 - \Phi\left(\sqrt{N_p}\frac{\tau_{\text{std}}}{2}\left(\sqrt{\frac{C}{C-\varepsilon\tau^2_{\text{std}}}} - 1\right)\right) + \frac{1}{5}\delta \\
&\leq \frac{1}{5}\delta + \frac{1}{5}\delta = \frac{2}{5}\delta
\end{aligned}$$

Hence, for any ε and δ, there exists $N_{upper} = max\left\{\left(\frac{\Phi^{-1}\left(\frac{1}{3}\delta\right)}{\frac{\tau_{\text{std}}}{2}\left(\sqrt{\frac{C}{C+\varepsilon\tau_{\text{std}}^2}}-1\right)}\right)^2, \left(\frac{\Phi^{-1}\left(1-\frac{1}{3}\delta\right)}{\frac{\tau_{\text{std}}}{2}\left(\sqrt{\frac{C}{C-\varepsilon\tau_{\text{std}}^2}}-1\right)}\right)^2, N_k\right\}$ such that when $N_p \geq N_{upper}$, $\mathbb{P}\left(\left|M\hat{R}SS_{N_p} - N\right| \leq \varepsilon\right) = \mathbb{P}\left(\frac{C}{\hat{\tau}_{\text{std}}^2} \geq \frac{C}{\tau_{\text{std}}^2} + \varepsilon\right) + \mathbb{P}\left(\frac{C}{\hat{\tau}_{\text{std}}^2} \leq \frac{C}{\tau_{\text{std}}^2} - \varepsilon\right) \leq \frac{4}{5}\delta < \delta$.

A.4 Details on Data Collected from Journals

We collected all publications that involve the reporting of at least a result on an experiment on American Journal of Political Science, American Political Science Review, Journal of Politics and Political Analysis since 2015. The challenge was that it was generally not conventional for most researchers to report standardized treatment effects. Instead, researchers almost always reported either a t-statistic or a standard error for their treatment effects. In addition, researchers reported the sample size of their experiments. Hence, we recovered the estimated standardized treatment effect with the following formula. With equal sample size for the treated group and the control group, remembering $\hat{\tau} \xrightarrow{d} \mathcal{N}\left(\tau_{std}, \frac{4\sigma^2}{N}\right)$ and thus $\widehat{\mathbb{V}(\hat{\tau})}$ as a consistent estimator for $\frac{4\sigma^2}{N}$, we can recover $\hat{\tau}_{\text{std}}$ by

$$\hat{\tau}_{\text{std}} = \frac{2\hat{\tau}}{\sqrt{\widehat{\mathbb{V}(\hat{\tau})} \times N_f}}$$

with the definition of a t-statistic $t_{\hat{\tau}} = \frac{\hat{\tau}}{\sqrt{\widehat{\mathbb{V}(\hat{\tau})}}}$, $\hat{\tau}_{\text{std}}$ can also be recovered by

$$\hat{\tau}_{\text{std}} = \frac{2t_{\hat{\tau}}}{\sqrt{N_f}}$$

We identified 128 publications across these four journals that involved at least one experiment. For each experiment, we identified its main causal quantity via the following procedure:

1. If the experiment reports a causal quantity in the main text, we consider this causal quantity as its main causal quantity for the experiment.

2. If the experiment does not report a causal quantity in the main text, but report a causal quantity in tables or figures, we consider this causal quantity as its main causal quantity for the experiment.

3. If the experiment reports a causal quantity neither in the main text nor in a table or figure, but report a causal quantity in the appendix, we consider this causal quantity as its main causal quantity for the experiment.

Each experiment could contain multiple "main" causal quantities according to the criteria above. To reduce duplicates, we used the following rules to select one causal quantity into our collection, and discard the

others:

1. If there is only one causal quantity tied to the substantive research hypothesis, we select this causal quantity into our collection.

2. If there are multiple causal quantities tied to the substantive research hypothesis, we select the causal quantity estimated with the simplest model.

3. If there are multiple causal quantities estimated via equivalently simple modes, from a conservative perspective, we select the causal quantity with the largest (standardized) size.

Our resulting dataset contain 164 effect size observations that are either average treatment effects (ATE), or similar causal quantities that can be estimated via difference in means.

References

Albers, C. and D. Lakens (2018, January). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology 74*, 187–195. 5

Blair, G., J. Cooper, A. Coppock, and M. Humphreys (2019). Declaring and diagnosing research designs. *American Political Science Review 113*, 838–859. 2

Chen, L. and C. Grady (2019). 10 things to know about pre-analysis plans. `https://egap.org/resource/10-things-to-know-about-pre-analysis-plans/`. Accessed 13-July-2020. 1

Cordeiro, G. M. and F. Cribari-Neto (2014). *An introduction to Bartlett correction and bias reduction.* Springer. 3.1

Dunham, Y. and E. Lieberman (2013, Nov). Social identity and social risk: Experimental research on race, stigma, and hiv/aids in the united states. 4, 8

Gelman, A. (2019). Don't calculate post-hoc power using observed estimate of effect size. *Annals of Surgery 269*(1), e9–e10. 5

Gelman, A. and J. Carlin (2014). Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science 9*(6), 641–651. 2, 5

Green, P. and C. J. MacLeod (2016). Simr: an r package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution 7*(4), 493–498. 2

Journal of Experimental Political Science (n.d.). FAQ for registered reports. `https://www.cambridge.org/core/journals/journal-of-experimental-political-science/information/faqs-for-registered-reports`. Accessed 13-July-2020. 1

Kraemer, H. C., J. Mintz, A. Noda, J. Tinklenberg, and J. Yesavage (2006). Caution regarding the use of pilot studies to guide power calculations for study proposals. *Archives Of General Psychiatry 63*(5), 484–489. 4.3

Kruschke, J. K. and T. Liddell (2018). The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychonomic Bulletin and Review 25*, 178–206. 2

Lee, P. M. (2012). Bayesian statistics: an introduction. 4th. A.1

Leon, A. C., L. L. Davis, and H. C. Kraemer (2011). The role and interpretation of pilot studies in clinical research. *Journal of psychiatric research 45*(5), 626–629. 5

MacKinnon, J. G. and A. A. Smith Jr (1998). Approximate bias correction in econometrics. *Journal of Econometrics 85*(2), 205–230. 3.1

National Science Foundation (2013). Common guidelines for education research and development. `https://www.nsf.gov/pubs/2013/nsf13126/nsf13126.pdf`. Accessed 13-July-2020. 1

Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, section 9. (translated in 1990). *Statistical Science 5*, 465–480. 2

O'Neill, B. (2014, 10). Some useful moment results in sampling problems. *The American Statistician 68*, 282–296. 2

Rothman, K. J. and S. Greenland (2018). Planning study size based on precision rather than power. *Epidemiology 29*(5), 599–603. 5

Schäfer, T. and M. A. Schwarz (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology 10*, 813. 3.3

Tausanovitch, C. (2015, June). Why do voters support partisan candidates? 3

Tibshirani, R. J. and B. Efron (1993). An introduction to the bootstrap. *Monographs on statistics and applied probability 57*, 1–436. A.2