北京大学中国经济研究中心
China Center for Economic Research
讨论稿系列
Working Paper Series

E2026002 2026-01-23

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

Shiyao Liu Teppei Yamamoto

Abstract

Political scientists routinely use power analysis when designing their empirical research. However, it is often neglected that power analysis relies on untested assumptions about the true values of key parameters, such as the effect size. Researchers commonly use auxiliary empirical information to make guesses about those parameters, such as results from a pilot study or a similar experiment reported in the literature. In this paper, we show that such practice is problematic due to neglected uncertainties in the empirically obtained parameter values. We propose a conceptual distinction between empirical and non-empirical power analyses and analyze the former as an estimation problem, investigating their statistical properties both analytically and via simulations. Our results indicate that estimators for power and minimum required sample size tend to perform poorly under scenarios resembling typical political science applications. We offer practical guidelines for empirical researchers on when to (and not to) trust power analysis results.

Keywords : power analysis, sample size, experimental design, research transparency.

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

Shiyao Liu* Teppei Yamamoto†‡

January 20, 2026

*Assistant Professor, China Center for Economic Research, Institute of South-South Cooperation and Development, National School of Development, Peking University, Beijing 100871, China. Email: shiyaoliu@nsd.pku.edu.cn

†Professor, Faculty of Political Science and Economics, Waseda University, Tokyo, Japan. Email: tyamam@waseda.jp, URL: https://teppeiyamamoto.github.io/

‡We thank Tara Slough, Matthew Simonson, Clayton Webb, and seminar participants at MIT Gov/Lab, New York University in Abu Dhabi, Peking University, and the University of Hong Kong for their helpful comments and suggestions. We are also grateful to Wenxin Lin, Yilin Lu and Tomoya Sasaki for excellent research assistance. An earlier version of this article was presented at the 2020 Annual Meeting of the Society for Political Methodology, 2024 Asian PolMeth and the 2024 Annual Meeting of the American Political Science Association.

Abstract

Political scientists routinely use power analysis when designing their empirical research. However, it is often neglected that power analysis relies on untested assumptions about the true values of key parameters, such as the effect size. Researchers commonly use auxiliary empirical information to make guesses about those parameters, such as results from a pilot study or a similar experiment reported in the literature. In this paper, we show that such practice is problematic due to neglected uncertainties in the empirically obtained parameter values. We propose a conceptual distinction between empirical and non-empirical power analyses and analyze the former as an estimation problem, investigating their statistical properties both analytically and via simulations. Our results indicate that estimators for power and minimum required sample size tend to perform poorly under scenarios resembling typical political science applications. We offer practical guidelines for empirical researchers on when to (and not to) trust power analysis results.

Keywords: 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 employing experimental methods, consider power analysis 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" (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 calculations to be 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).

Power analysis refers to various statistical 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 a level of statistical power that is sufficiently high to render 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 statistics textbooks, power is described as a quantity that is calculated given the true values of parameters for a hypothesis test. In practice, however, power analysis often rests on empirical information. 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), the sample size, and the power itself. While the latter two parameters typically come from external constraints, such as research budget or convention, the standardized effect size is a feature of the data-generating process itself and, therefore, is almost never known by the researcher. Thus, researchers employing power analysis often use some empirical information to cope with the fundamental uncertainty about the standardized effect size.

More specifically, two approaches are particularly common in empirical research. First, researchers often employ a pilot study to obtain an estimate of the treatment effect and use that estimate as an input to their power calculation. Second, researchers may look for a previous empirical study testing a similar hypothesis and use an estimate of the effect size in the study as if it were equivalent to their effect of interest. Both of these approaches use existing empirical information about a population parameter (i.e., the standardized effect size) to make inferences 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 used to solve empirical problems 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.

To illustrate the current practice, we survey the political science pre-registrations created in the Open Science Foundation (OSF) registry in 2024 with some discussion of sample size rationale.[1] Table 1 summarizes the result. Of the 580 pre-registered studies, 84 (or 14.5%) fall under our "empirical power analysis" category, explicitly stating that they refer to either a pilot or a previous study to determine their hypothesized treatment effect. This is the second largest category among the pre-registrations that cite any reason, only next to the "cost/resource constraints" category which cite some resource constraints (166 studies, or

[1]We collected all pre-registration entries from that year that had a unique identifier in the metadata under the "sample size rationale" label. From the total of 1,393 entries, we exclude 742 entries that are either empty or for qualitative studies and 71 entries that refers to attached supplemental files. All classifications are hand-coded.

Type	Count	Proportion
No reference of inputs	246	42.4%
Cost/resource constraints	166	28.6%
Empirical power analysis	84	14.5%
Sample size of pervious studies	35	6.0%
Part of larger studies	31	5.3%
Universe	18	3.1%
Total	580	100.0%

Table 1: Review of Sample Size Rationale Entry in Metadata of OSF in 2024

28.6%). It is also worth noting that some entries in the latter category specifically mention the lack of prior studies as part of their rationale. Of the remaining pre-registrations, the vast majority (246) include no reference to how they chose their hypothesized treatment effects. We suspect that many of these come from informal beliefs formed on related studies that have previously been conducted. In sum, the practice of referring to a pilot or a previous study to get an "estimate" for the hypothesized treatment effect appears to be quite prevalent.

In this paper, we propose to call these types of power analyses *empirical power analyses* and distinguish them from the variants that do not use empirical information[2]. 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 parameter values that we find to correspond well with 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 often need

[2]The latter is therefore not subject to much of our critique in this paper. See our discussion in Section 5.

to use data from a small pilot study or a loosely related previous study. Given the large amount of uncertainty in the estimates from such studies, 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 misinterpreting power estimation results could lead to serious inefficiencies, such as missed opportunities and wasted resources. For example, an overestimation of the MRSS could discourage a researcher from conducting an experiment that is actually promising. 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 power analysis for this estimation uncertainty[3].

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 biased even when an unbiased estimate of the true effect size is available (as it may be 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 are 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. 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

[3]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.

desired effect size based on substantive 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 a setting where the researcher studies the average treatment effect (ATE) of binary treatment $Z \in \{0, 1\}$ on outcome Y. They plan a full randomized experiment with sample size N_f on a simple random sample from the population: n_{f1} subjects randomly assigned to the treatment ($Z = 1$), $n_{f0} = N_f - n_{f1}$ in control ($Z = 0$). Beforehand, they have data from a "pilot" study[4] – another randomized experiment with the same treatment and outcome variable, but on a separate random sample of N_p subjects from the same population with n_{p1} subjects randomly assigned $Z = 1$ and the remaining $n_{f0} = N_p - n_{p1}$ assigned $Z = 0$.

Researchers often use pilot data for empirical power analysis before the full study. Despite the availability of complex tools (e.g., Green and MacLeod, 2016; Blair et al., 2019), we focus on the widely used textbook two-sample t-test power analysis with the asymptotic normal reference distribution. Suppose they test against the zero Average Treatment Effect (ATE) via a two-sided t-test in the full experiment, letting α and β be the probabilities of type-I and type-II errors, respectively, the *(true) power* of the full experiment ψ is:

[4]Despite the terminology, the setup encompasses a scenario where researchers use results from previously published experiments resembling the proposed study to conduct power analysis, where N_p is interpreted as the effective sample size from previous studies.

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

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

We make two simplifying assumptions: (1) the outcome variance is constant across the treated and the control; (2) the treatment is randomly assigned to minimize the sampling variance of the estimated ATE (Neyman, 1923), which indicates the equal treatment allocation across the treatment group and the control group. The assumptions are not overly restrictive: first, they are of practical relevance, as most researchers use constant variance in their study designs. Second, Appendix A.6 shows that power estimation bias stems mainly from the imprecise τ (true effect) estimation, not that of σ, so differing group standard deviations still produce similar biases.

Assumption 1. *(constant outcome variance)* $\sigma^2 \equiv S_1^2 = S_0^2$

Assumption 2. *(equal treatment allocation)* $N_d/2 = n_{d1} = n_{d0}$, $d \in \{f, p\}$

We also clarify that the assumptions that we are *NOT* yet making are: (1) the parametric assumption for the underlying distribution of the outcomes, and (2) whether the outcomes are subject to an identical and independent distribution, as our results are based on the sampling theory (O'Neill, 2014). Thus, the analytical results derived in this paper, unless otherwise stated, shall apply to all cases regardless of the underlying distribution of the outcome.

Under these assumptions, the power of the full experiment ψ depends only on three parameters: the size of the test α, intended full sample size N_f, and the *standardized effect size* τ_{std}, defined as the true ATE scaled to the standard deviation of the outcome, $\tau_{\text{std}} \equiv \tau/\sigma$. Equation (1) simplifies to

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

Of the three parameters in equation (2), two are *design parameters* the researchers in theory have control of: α, conventionally set at the level of $\alpha = 0.05$, and N_f, chosen under the cost or logistical constraints. The third parameter, τ_{std}, is *empirical*, whose true value exists independently of the research design.

A common way to calculate power via equation (2) is to estimate $\hat{\tau}_{\text{std}}$ from a pilot experiment, then plug this estimate, plus known α and N_f, into the equation. Thus, *empirical power analysis* uses the following plug-in estimator:

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

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

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

where all summations and terms (n_{p1}, n_{p0}, N_p) refer to the pilot sample.

Another use of equation (2) is calculating the full experiment's MRSS – the smallest sample size for the desired power ψ. The full experiment meets ψ iff:

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

ignoring the negligible[5] last term in equation (2).

$MRSS$ is the smallest integer N_f satisfying this inequality (5). Since ψ is a researcher-set

[5]The term is strictly bounded from above by $\alpha/2$.

design parameter, conventionally at $\psi = .8$, the *MRSS estimation* adopts a plug-in estimator:

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

with $\hat{\tau}_{\text{std}} = \hat{\tau}/\hat{\sigma}$ from equation (4).

Before we examine the statistical properties of the power and MRSS estimators (equations (3) and (6)), some general discussion is helpful. Noting that both estimators are nonlinear functions of $\hat{\tau}_{\text{std}}$, the ratio of an unbiased estimator $\hat{\tau}$ to a nearly unbiased estimator $\hat{\sigma}$, by standard sampling theory (see Appendix A.2 for more details),

$$\mathbb{E}[\hat{\tau}] \;=\; \tau, \quad \mathbb{V}[\hat{\tau}] \;=\; \frac{4\sigma^2}{N_p}, \quad \frac{\hat{\tau}-\tau}{2\sigma/\sqrt{N_p}} \xrightarrow{d} \mathcal{N}(0,1) \text{ as } N_p \to \infty, \tag{7}$$

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).

While these properties establish the consistency of both $\hat{\psi}$ and $\widehat{MRSS}$ as $N_p \to \infty$, they do not ensure additional desirable characteristics. Notably, since $\hat{\psi}$ and $\widehat{MRSS}$ are nonlinear functions of $\hat{\tau}$ or $\hat{\sigma}$, these estimates are generally biased – a direct consequence of Jensen's inequality. Such small-sample biases are particularly problematic, as pilot study sample sizes (N_p) tend to be relatively small in most empirical settings.

3 Power Estimation

We first examine the power estimator in equation (3). Though less common than MRSS estimation, power calculation is standard in empirical political science and foundational in

methodology courses, where students first encounter these concepts. Researchers adopt this estimator when constrained by fixed full-experiment sample sizes, and need to assess the viability of a study.

For example, 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 pilot study.

We naturally ask, how reliable the reported power of 88% actually is, given that the power is 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, empirical power calculation is to estimate the output of a nonlinear function of other parameters. While the estimators of the latter parameters may behave well, plug-in estimates for the target may not, due to the non-linear property of the function. Focusing on $\alpha = 0.05$ and approximating by ignoring the negligible final term[6],

$$\hat{\psi} \;\simeq\; 1 - \Phi\left(1.96 - \frac{\hat{\tau}_{\text{std}}\sqrt{N_f}}{2}\right). \tag{8}$$

Equation (8) shows that $\hat{\psi}$ is a nonlinear function of the pilot-derived $\hat{\tau}_{\text{std}}$, following a standard normal CDF. The standard normal CDF is neither globally convex nor concave, which prevents us from determining the bias direction via Jensen's inequality[7]: $\hat{\psi}$ may over- or under-estimate its true value, even with the unbiased $\hat{\tau}_{\text{std}}$. While $\hat{\psi}$ is indeed consistent and asymptotically normal as $N_p \to \infty$, pilot sizes are typically small, leaving asymptotic properties less relevant in this setting.

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

[7] 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 $g(\cdot)$.

Proposition 1. *Under Assumptions 1 and 2,* $\mathbb{E}[\hat{\psi}] \neq \psi$*, i.e.* $\hat{\psi}$ *is a biased estimator for* ψ*.*

Proof. See Appendix A.1. □

Further Approximation. Before Monte Carlo simulations, further approximations in Appendix A.2 yield key predictions about bias of the power estimator. The parametric assumption of the outcome Y subject to a normal distribution is introduced to proceed with the derivation without invoking asymptotic results, since N_p is small in most applied settings.

First, bias nears zero as the true τ_{std} grows, holding other parameters fixed: this is because the negative term in equation (8) approaches zero in expectation, so $\hat{\psi}$ and true power both grow towards 1. Second, a larger N_f can increase or decrease bias, and the direction depends on other parameters: this is because the argument of the negative term in equation (8) shifts left and grows more variable, which leaves the net change of $\mathbb{E}[\hat{\psi}]$ ambigious.

Bias-Correction. We attempt to correct the bias of $\hat{\psi}$ in Appendix A.3, but standard bias-correction methods fail in our setting, for two reasons: First, $\hat{\psi}$ has a nonlinear bias. Traditional methods like bootstrap or jackknife work poorly here as they assume a constant or linear bias (Cordeiro and Cribari-Neto, 2014). As expected, they fail in our setting. Second, the non-linear method proposed by MacKinnon and Smith Jr (1998) requires a closed-form analytical bias function, which we lack because the normal CDF has no closed form. A modified version with a simulated bias function, as presented in Appendix A.3, also failed to improve the results. In short, existing methods cannot correct the bias of $\hat{\psi}$.

While the above results offer insights into the power estimator's theoretical behavior, their practical value is limited. Fundamentally, the non-linearity of equation (8) prevents *ex ante* predictions of bias magnitude or direction. We thus turn to Monte Carlo simulations. To conduct Monte Carlo simulations, unlike the analytical results above, we have to introduce parametric data-generating processes (DGPs) for practical reasons, yet we save discussions on the choice of various DGPs in Appendix A.6.

3.2 Simulations

To study the small-sample properties of the power estimator as in equation (3), we simulate repeated sampling from various data-generating processes (DGPs) that match empirical political science scenarios. We vary three key parameters: standardized effect $\tau_{\text{std}} \in [0, 1]$, pilot size $N_p \in \{50, 250, 450, 650\}$, and full-experiment size $N_f \in \{100, 500, 900, 1300\}$ – all aligned with common political science applications. We assess the estimator's bias and standard error via 1,000 Monte Carlo draws.

Simulation Procedure For each τ_{std}, N_p, N_f combination, we run this Monte Carlo experiment:

1. Simulate the pilot: Draw $\frac{N_p}{2}$ treatment-group Y values from $\mathcal{N}(\tau, 4^2)$, and $\frac{N_p}{2}$ control-group values from $\mathcal{N}(0, 4^2)$, where $\tau_{\text{std}} = \tau/4$.

2. Compute the difference-in-means treatment effect estimator $\hat{\tau}$.

3. Estimate treatment/control sample variances of the outcome: $\widehat{S_1^2}$ and $\widehat{S_0^2}$.

4. Calculate power via a plug-in estimator based on equation (2), $\alpha = 0.05$, $n_{f0} = n_{f1} = N_f/2$:

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

5. Repeating Steps 1 to 4 for 1,000 times to estimate bias and 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, across many parameter values, the power estimator's bias is substantial and can be positive or negative. For fixed sample sizes, e.g. the subplot of the second row and third

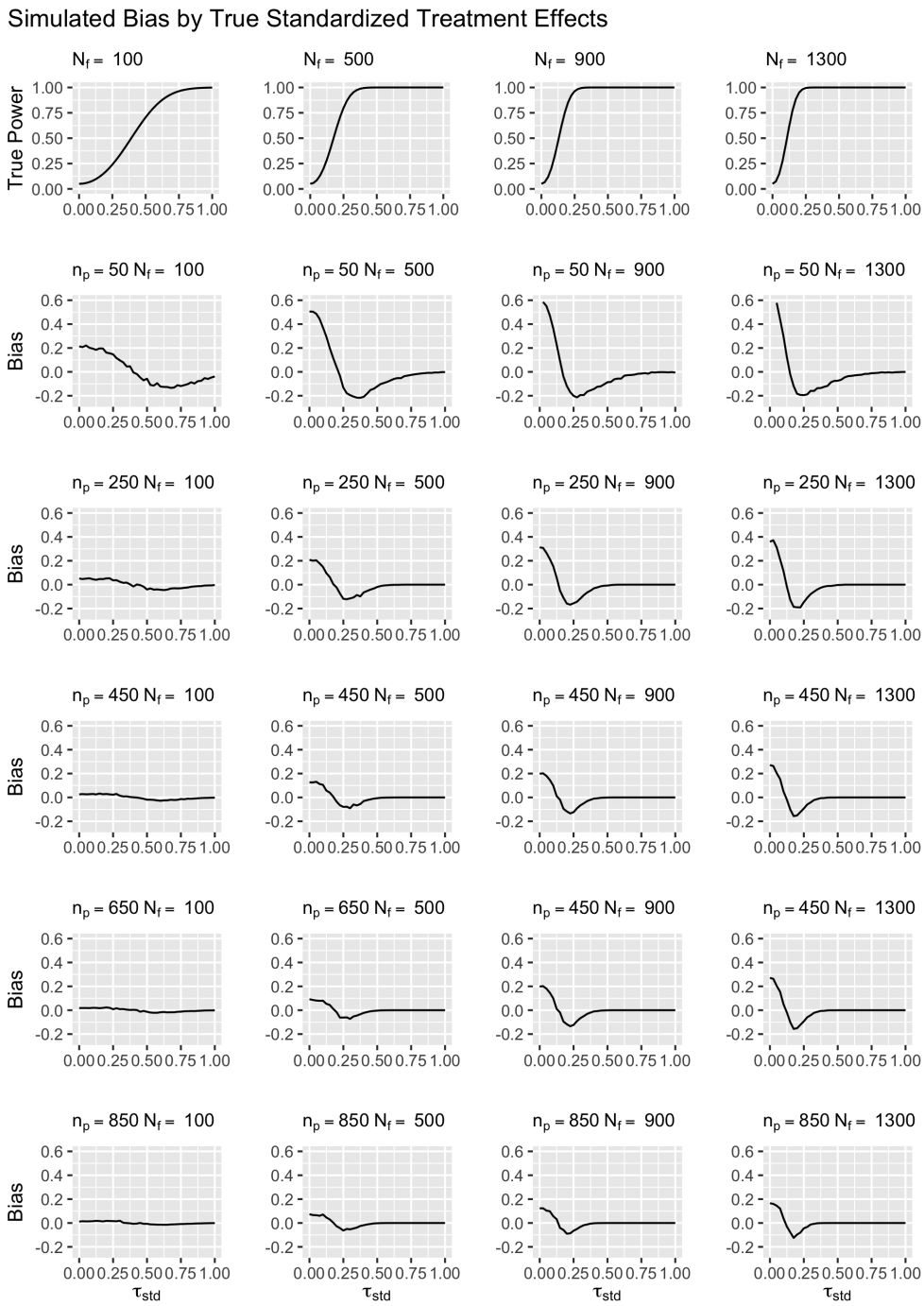

Figure 1: Simulated Bias by True Standardized Treatment Effect. The top row of plots presents the true power for each full experiment sample size as a function of the standardized effect size. The remaining plots show Monte Carlo estimates of the bias of the power estimator on the vertical axis for a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

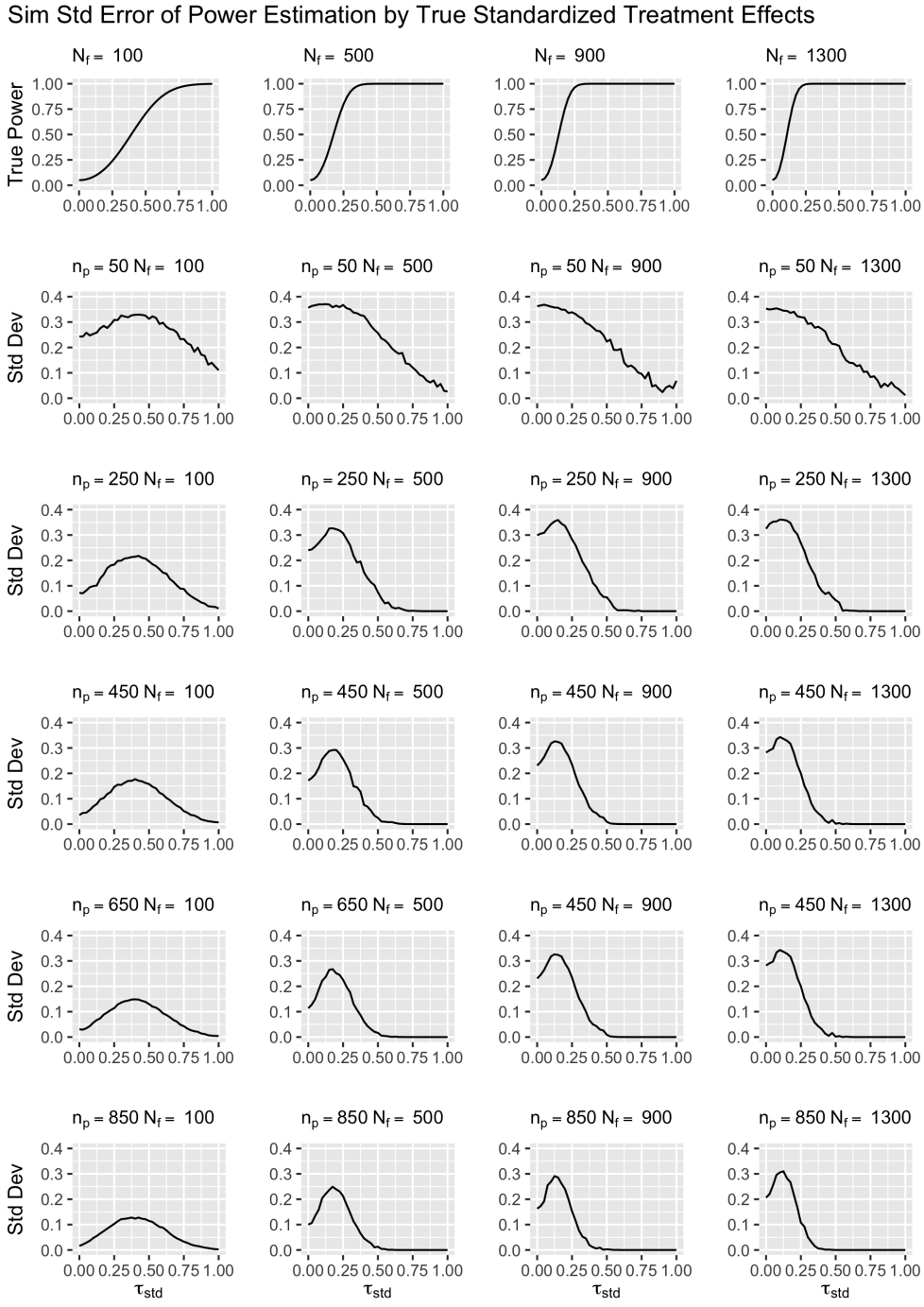

Figure 2: Standard Error by True Standardized Treatment Effect. See legend for Figure 1 for the interpretation of the graph components.

column in Figure 1, bias is positive and largest when τ_{std} is close to zero. The bias decreases to zero, flips negative at around $\tau_{\text{std}} = 0.175$, then fades as true power hits 1. Since power ranges between 0 and 1, and most design aims for a power greater than 0.8, a bias of 0.1 is significant – our results show a bias greater than 0.1 in a wide range of scenarios.

Second, the power estimator is also highly imprecise across various parameter values. For example, with $N_p = 50$ and $N_f = 100$, in Figure 2, the standard error exceeds 0.1 for all $\tau_{\text{std}} \in [0, 1]$. In other sub-plots, the standard error stays below 0.1 only if τ_{std} is greater than approximately 0.4, except in unrealistic cases where pilot size far exceeds full experiment size (first column, bottom three subplots).

Third, viewing Figure 1 horizontally, as Appendix A.2 shows analytically, bias generally increases with the N_f/N_p ratio, holding τ_{std} constant. We confirm this with the simulation. This is the contrapositive of a later observation: with fixed N_f, larger N_p tends to reduce the bias of power estimation.

Fourth, bias becomes negligible once true τ_{std} exceeds 0.3 to 0.8, varying by sample sizes. A helpful rule of thumb is that when the expected $\tau_{\text{std}} > 0.5$, there is no need to worry about the bias. A typical political science N_f makes power for $\tau_{\text{std}} = 0.5$ close to 1.

Fifth, bias decreases monotonically as pilot size N_p increases. Since $\hat{\tau}_{\text{std}}$ is consistent, $\hat{\psi}$ is also consistent for ψ with a convergence rate of $\sqrt{N_p}$ and asymptotically normal. Larger N_p activates these asymptotic properties, reducing bias and variance until they vanish.

Finally, the first row of Figure 1 shows that the true power function is convex in some areas and concave in others. Due to the estimation uncertainty of $\hat{\tau}_{\text{std}}$, researchers cannot tell if the true $\hat{\tau}_{\text{std}}$ falls in the convex or concave area, unless the pilot size is unrealistically large. As Appendix A.3 discusses, this ambiguity, plus other method limitations, hinders existing bias correction approaches.

3.3 Calibrating Simulation Results

A key insight from our simulations is that the power estimator's bias and precision depend crucially on the true standardized effect size. Naturally, we ask: where do we fall on the horizontal axes of Figures 1 and 2? Should political scientists worry about the bias of the power estimation in a typical experiment?

To answer, we rely on reported standardized effect sizes in the literature. We collected articles from 4 top political science journals[8] published between 2015 and 2024, focusing on those with ATEs (or similar quantities) reported. This yielded 410 standardized effect size observations[9].

Figure 3 shows the distribution of standardized effect sizes from our sample, classified by study type. Before the interpretation, we note that our sample from top journal articles is unrepresentative of *all* political science experimental effect sizes: publication bias and file-drawer effects (Schäfer and Schwarz, 2019) skew observations toward the right tail, i.e., larger effects. Thus, our reported statistics overestimate true standardized effect sizes, possibly by significant margins.

Strikingly, our estimated standardized effect sizes are mostly concentrated between 0.1 and 0.4, even with the likely overestimation. The median is about 0.18 for survey experiments, slightly smaller for field experiments, and slightly (but not much) larger for experiments using economic games as treatment or outcome.

Comparing these results to our bias and standard error simulations in Figures 1 and 2 paints a bleak picture of power estimation in political science. Empirical standardized effect sizes between 0.1 and 0.4 fall exactly where bias is most sensitive to τ_{std} and where the standard error is the largest. For example, with a typical $N_p = 50$ and $N_f = 900$ and $\tau_{\text{std}} = 0.25$, the bias reaches about -0.2, and the standard error is approximately 0.35. Even

[8] American Journal of Political Science, American Political Science Review, Journal of Politics, Political Analysis, chosen for high impact and frequent experimental studies.

[9] Although few articles directly report standardized ATE, we calculated τ_{std} (with assumptions if needed) using reported uncertainty estimates (see Appendix A.7 for details).

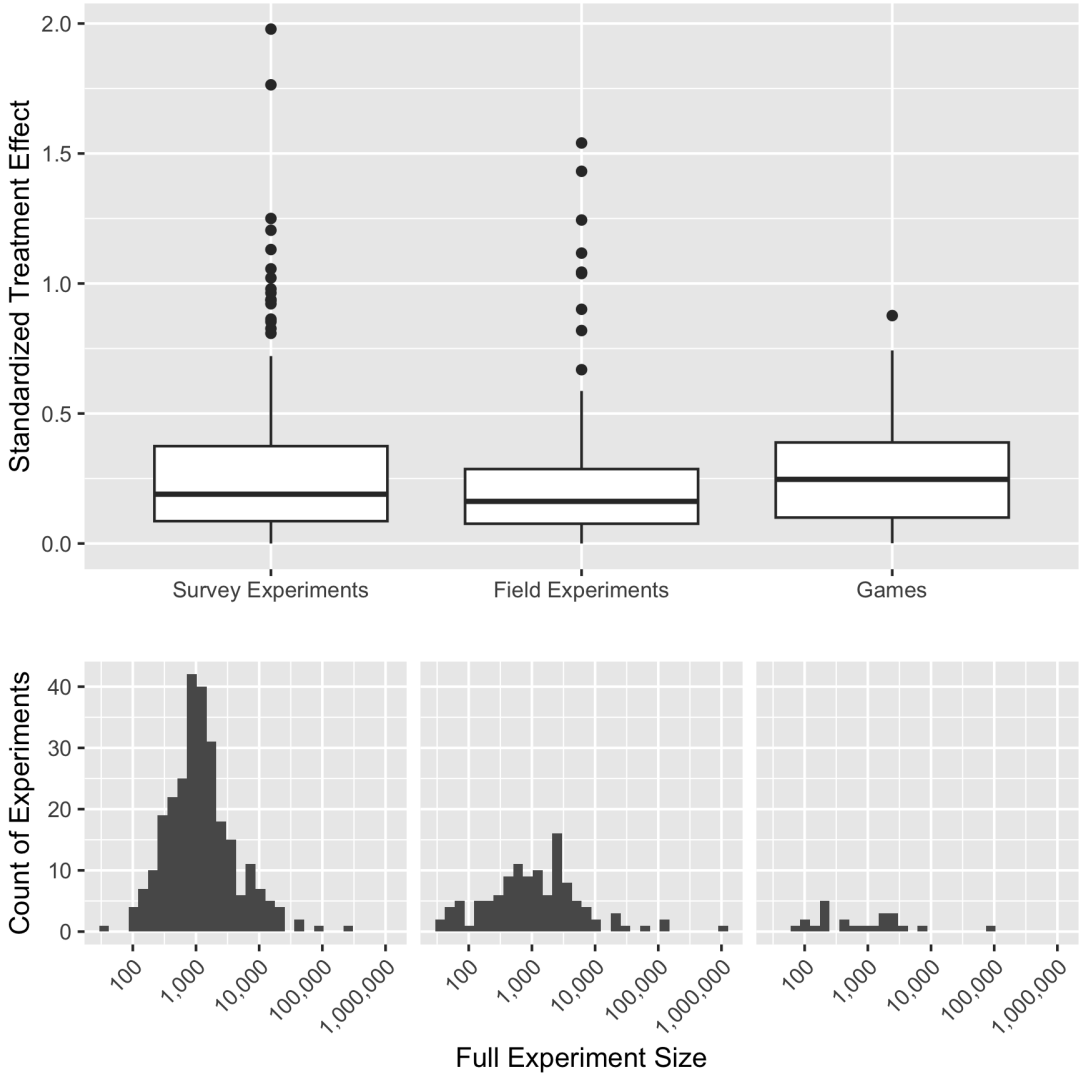

Figure 3: Distribution of Standardized Effect Sizes and Full Experiment Sample Sizes in Top Published Political Science Articles (2015-2024)

with an unrealistically large $N_p = 850$, bias remains 0.1 and the standard error hits 0.3 when $\tau_{\text{std}} = 0.15$ and $N_f = 1300$.

4 Minimum Required Sample Size Estimation

Next, we turn to MRSS estimation – likely the most common form of empirical power analysis in political science. For example, Dunham and Lieberman (2013)'s EGAP-registered pre-analysis plan[10] reports using a 100-participant pilot to estimate expected effect size, deriving an MRSS range of 342 to 1,043 for a 90% power, and choosing N=1,000 for the final study. We investigate the reliability of this method using both analytical and simulation approaches.

4.1 Analytical Results

We start by deriving the expectation of the MRSS estimator (equation (6)) to find its bias. However, this expectation does not exist, leaving the bias of $\widehat{MRSS}$ undefined. Observe:

$$\begin{aligned}\mathbb{E}\left[\widehat{MRSS}\right] &= \mathbb{E}\left\{\frac{4\left[\Phi^{-1}\left(1-\frac{\alpha}{2}\right)-\Phi^{-1}\left(1-\psi\right)\right]^2}{\hat{\tau}^2_{\text{std}}}\right\} \\ &= 4\left[\Phi^{-1}\left(1-\frac{\alpha}{2}\right)-\Phi^{-1}\left(1-\psi\right)\right]^2 \mathbb{E}\left\{\frac{1}{\hat{\tau}^2_{\text{std}}}\right\}.\end{aligned}$$

In Appendix A.4, we show $\mathbb{E}\left\{\frac{1}{\hat{\tau}^2_{\text{std}}}\right\}$ diverges under the regularity assumption 3. Thus, $\mathbb{E}\widehat{MRSS}$ diverges under this condition as well.

Assumption 3. *(continuous and bounded density for $\hat{\tau}_{std}$) $f_{\hat{\tau}_{std}}(x)$, the probability density function for $\hat{\tau}_{std}$, is continuous and bounded from both below and above.*

Proposition 2. *Under Assumptions 1, 2 and 3, $\mathbb{E}\left[\widehat{MRSS}\right]$ does not exist.*

Proof. See Appendix A.4. □

[10]Notably, they used ANOVA-based power analysis for a factorial design, unlike our t-test, so our results may not directly apply.

The non-existence of $\mathbb{E}\left[\widehat{MRSS}\right]$ means empirical samples of $\mathbb{E}\left[\widehat{MRSS}\right]$ may have wild values, as simulations later show. It also makes $\mathbb{V}[\widehat{MRSS}]$ non-existent, since the first moment is required for the second moment to be defined. Although Appendix A.5 shows the consistency of $\widehat{MRSS}$ for large N_p, the pilot sizes are by definition small, so asymptotic results offer little help.

Again, we clarify here that the analytical results should apply regardless of the parametric distribution of the outcome Y, as our analytical results are based on the sampling theory. Admittedly, the following Monte Carlo simulations require specified parametric DGPs, our main simulation results are robust under different DGPs as in Appendix A.6.

4.2 Simulations

Since MRSS has no defined expectation or variance, standard performance measures, such as bias or root mean squared errors, do not apply. We instead investigate its small-sample performance via simulations of 1,000 MRSS realizations per DGP to examine how the empirical distribution changes with different pilot sizes N_p and τ_{std}. Parameters are set to match political science applications: $\tau_{\text{std}} \in \{0.125, 0.25, 0.5, 1\}$, $N_p \in \{10 \leq n \leq 5000, n \in \mathbb{Z}\}$.

Simulation Procedure For each τ_{std}, N_p combination, we run this Monte Carlo experiment:

1. Simulate the pilot: Draw $\frac{N_p}{2}$ treatment-group Y values from $\mathcal{N}(\tau, 4^2)$, and $\frac{N_p}{2}$ control-group values from $\mathcal{N}(0, 4^2)$, where $\tau_{\text{std}} = \tau/4$.

2. Compute the difference-in-means treatment effect estimator $\hat{\tau}$.

3. Calculate MRSS via a plug-in estimator based on equation (6), $\alpha = 0.05$, $\psi = 0.8$:

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

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

Results Figure 4 shows the simulated sampling distributions of the MRSS estimation across four different values of the standardized effect size (from the top to the bottom panels) and different pilot sample sizes (along the horizontal axis). The y-axis is log 10 scaled.

MRSS shows striking sampling variability across a wide range of parameters. Its simulated distribution is highly right-skewed, where the empirical mean nearly always exceeds the 95th percentile. Take $\tau_{\text{std}} = 0.125$, an empirically likely scenario in Figure 3 as an example, with $\psi = 0.8$, and $\alpha = 0.05$, its true MRSS is 2,008. With a pilot $N_p = 100$, $\hat{\tau}_{\text{std}}$ ranges from -2.42 to 2.52 out of the 1,000 simulations. The central 90% spans -1.16 (5^{th} percentile) to 1.42 (95^{th} percentile). This makes MRSS range between 5 and 561 million (5^{th} to 95^{th} percentile: 12 - 7,109). The mean of MRSS estimator (604 thousand) wildly overestimates the true MRSS, while the median (112) underestimates it[11]

Suppose a researcher uses a 5,000-participant pilot to reduce MRSS uncertainty – unrealistically large for typical political science field/survey pilots, but possible from a prior large study. Even so, the situation remains bleak: estimated MRSS ranges between 154 and 332 million (5^{th} to 95^{th} percentiles: 311 – 170,683), with a mean of 727 thousand and a median of 1,861.

Finally, take the most optimistic scenario: true $\tau_{\text{std}} = 1$, atypically large in political science as shown in Figure 3. With a 1,000-participant pilot, MRSS estimates range between 9 and 555 (5^{th} to 95^{th} percentiles: 16 – 97), with stabilized mean/median at 41/31. While this seems promising, true MRSS here is just 32 – and the pilot's estimated ATE is almost always highly significant, making MRSS for a separate full experiment arguably pointless.

Our simulations show MRSS estimation, based on empirical standardized effect size estimates, has limited use in typical political science applications. Estimates are unhelpfully

[11]These numbers only illustrate the MRSS estimator's sampling behavior – not estimating the true distribution's order statistics or moments. They will likely differ by a large amount (except the median) in another simulation run, but the overall variability pattern will remain the same.

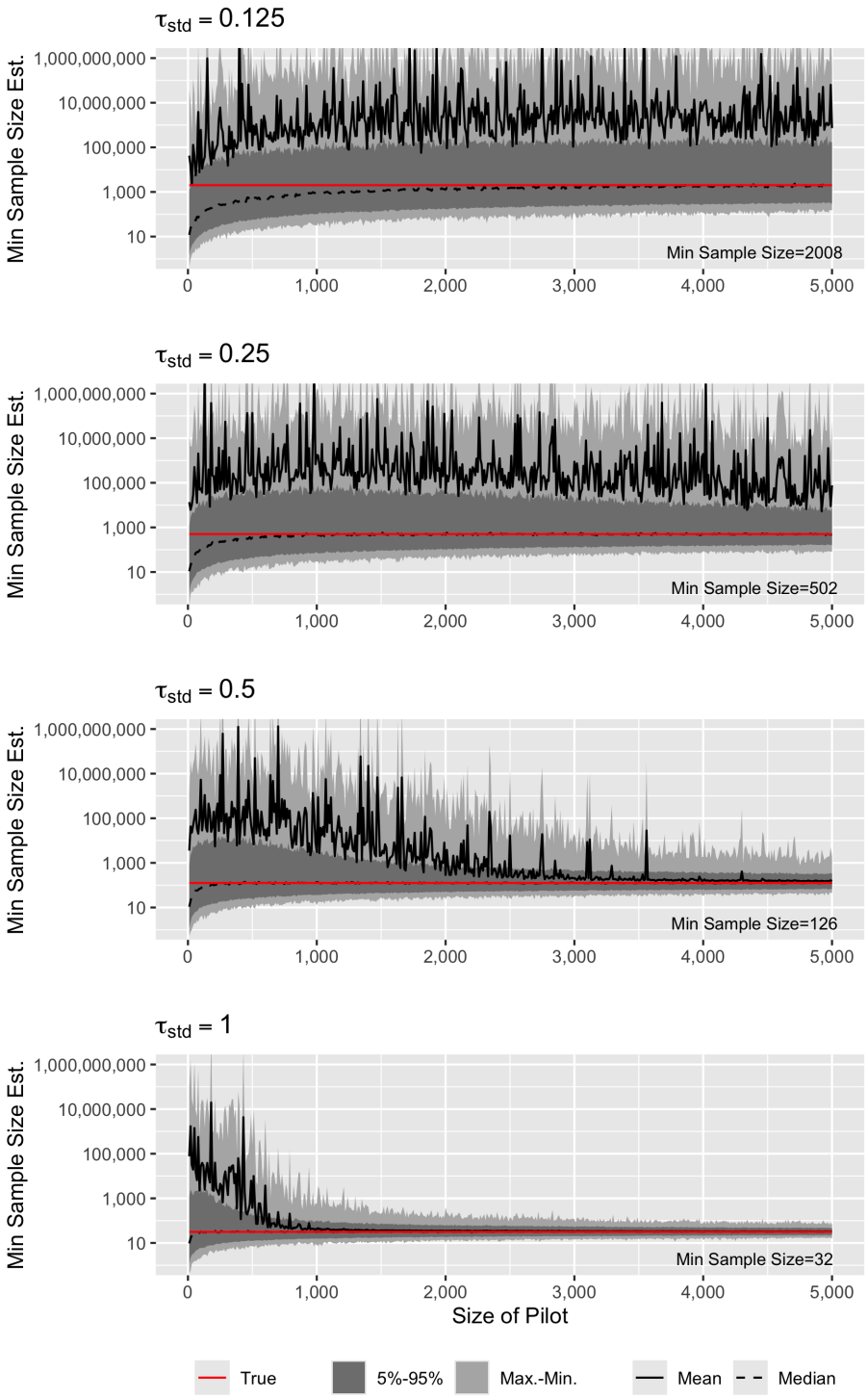

Figure 4: Simulated Sampling Distributions of the MRSS Estimator. Each of the four plots presents characteristics of the simulated sampling distribution of the MRSS estimator (maximum, 95th percentile, mean, median, 5th percentile and minimum) as functions of the pilot study sample size (N_p), with each 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.

variable in most relevant scenarios, and when estimation is reliable, the pilot's treatment effect estimate is already almost certainly highly significant.

4.3 MRSS Estimation in Practice

In practice, researchers tend to estimate MRSS only when pilots are "promising." Pilots often help explore design choices (e.g., treatment content) to finalize full experiment specifications. Empirical power analysis typically occurs when researchers select treatment-outcome combinations with effects that seem to indicate the existence of an impact, yet do not meet conventional statistical significance.

This research practice deviates from our Section 4.2 simulations, which assume an MRSS estimation regardless of the estimated pilot effect. Instead, under this practice, the distribution of *actually conducted* MRSS estimates is likely conditional on the estimated pilot effect falling within a specific range of statistical significance. Under this more realistic regime, MRSS variability is much lower: extreme MRSS draws in Figure 4 correspond to those close-to-zero estimated pilot effects, and are not part of the *actually conducted* MRSS estimates.

However, this does not make our pessimistic simulation conclusion irrelevant. In fact, conditioning empirical power analysis on a promising pilot result is flawed and should be abandoned. We discuss two reasons for this below, using our framework that treats power analysis as an estimation problem. Before we proceed, we clarify that the two reasons below are generally applicable regardless of the underlying distribution of the outcome Y, as the results are based on sampling theories.

First, researchers may prematurely abandon statistically and substantively significant experiments by dropping the full study after a pilot with a large p-value (see also Kraemer et al., 2006). Take $\tau_{\text{std}} = 0.25$ as an example: a 100-participant pilot has $\hat{\tau}_{\text{std}}$ with mean 0.25 and variance 0.04 per equation (7). Suppose the researcher only proceeds if the pilot t-stat is "promising but not significant," i.e., between 1 and 2. Using a standard normal

approximiation for the t-statistic, there's only a 34% chance[12] that they will conduct a power analysis, and potentially run the full experiment, even though the true τ is 0.25. 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, basing full experiment decisions on pilot treatment effect p-values makes the pilot data irrelevant. We show analytically that there is a one-to-one mapping between pilot p-values (or t-statistics) and MRSS. In other words, the decision rule, using pilot p-value and sample size alone, fully determines MRSS, ignoring actual experimental data obtained from the pilot. Thus, given the pilot size, one can estimate MRSS from a hypothetical N_p pilot without conducting the real one.

Specifically, since $\mathbb{E}[\hat{\tau}_{\text{std}}] = \tau_{\text{std}}$ and $\mathbb{V}[\hat{\tau}_{\text{std}}] = 4/N_p$, the pilot t-statistic against the zero effect null is $\frac{\hat{\tau}_{\text{std}}}{2} \cdot \sqrt{N_p}$. Suppose the researcher proceeds only if this t-statistic is between 1 and 2. Using the standard normal reference distribution, $\hat{\tau}_{\text{std}}$ must fall between $\frac{2}{\sqrt{N_p}}$ and $\frac{4}{\sqrt{N_p}}$. Plugging this into equation (6) gives:

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

All parameters for this N_f range, α, ψ, N_p, are known pre-pilot. Thus, the pilot provides no additional information to determine the minimum required sample size range.

In sum, the practice of conditioning MRSS estimation on the statistical significance of a pilot study is not advised. Ironically, this practice itself gave researchers the illusion that MRSS is useful: by only estimating MRSS when pilots show promising effects, they can pre-fetch estimates that are less extreme.

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

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

5 Practical Recommendations

Our analysis of two common empirical power analysis techniques – power estimation and MRSS estimation – highlights their serious limitations for political science applications. Below, we offer practical guidelines for empirical researchers using power analysis.

First, empirical power analysis is generally *not* recommended for most political science scenarios – such as field or online survey experiments. Given the range of standardized effect sizes in recently published top-journal experimental studies (Figure 3), neither power nor MRSS estimation will likely produce reliable results. The problem is exacerbated when using ATE/outcome variance estimates from small pilots, as their large uncertainty is amplified by the nonlinear transformations in power or MRSS estimators. If researchers still proceed with empirical power analysis, they must calculate the estimation uncertainty in their power/MRSS estimate and report a standard error alongside the point estimate, treating it like any other estimation problem and adopting the same reporting standards.

Our critique does not apply to *non-empirical* power analysis – techniques that avoid empirical estimation of the standardized effect size. Power analysis is non-empirical when all parameters are clearly set by external constraints or normative criteria, and thus no statistical uncertainty is involved. For example, the minimum detectable effect size (MDES) calculation[14] identifies the smallest standardized effect detectable with specified power (ψ), significance level (α), and full sample size (N_f). Researchers can then compare MDES to a pre-set threshold for substantive importance. A related approach is the MRSS calculation using a normatively defined effect size, e.g., a grant agency's required target without statistical uncertainty. That said, we caution that non-empirical power and MRSS use the same formulas as empirical methods – meaning results remain sensitive to even small changes in pre-specified effect sizes.

[14]The minimum detectable effect size is shown to be $\tau_{std} \geq \frac{2\left[\Phi^{-1}\left(1-\frac{\alpha}{2}\right)-\Phi^{-1}(1-\psi)\right]}{\sqrt{N_f}}$ or $\tau \geq \frac{2\left[\Phi^{-1}\left(1-\frac{\alpha}{2}\right)-\Phi^{-1}(1-\psi)\right]}{\sqrt{N_f}}\sigma$.

As noted in Section 4.3, a particularly problematic empirical power analysis practice is using pilot data to estimate the target treatment effect, then conducting a formal power analysis only if the estimate is "promising" (moderately statistically significant). Given the pilot's statistical significance (p-value/t-statistic) and sample size, the actual pilot data contains *no additional information* for power calculation. In short, if researchers pre-decide, before collecting data, to propose a full study only if a pilot of size N_p has a p-value in a particular range, they do not need to collect any pilot data to determine the full experiment's MRSS range.

Given our findings, are pilots, or preexisting same-hypothesis studies, useful 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. MDES, our recommended non-empirical power tool, produces standardized effect sizes, minimum detectable effects in outcome standard deviations. To judge if this meets a normative threshold, researchers need to interpret MDES in the outcome's original scale, which usually requires an empirical estimate of the outcome's standard deviation. Estimating raw effect sizes from the estimated standard deviation and a standardized effect is far safer than power/MRSS estimation. Typical pilot sizes are large enough for precise outcome standard deviation estimates, yielding reliable raw effect estimates. Thus, pilots still matter in empirical political science, in addition to their other roles unrelated to power analysis (Leon et al., 2011).

Finally, our recommendations add to the growing body of power analysis critiques (e.g., Rothman and Greenland, 2018; Gelman and Carlin, 2014). Other scholars have criticized practices like post-hoc power calculation using observed effect sizes (Gelman, 2019) and only proceeding with full experiments if pilot-based MRSS seems feasible (Albers and Lakens, 2018). We advise researchers to follow both these existing recommendations and ours.

6 Conclusion

Power analysis is increasingly prominent in political science – researchers routinely use it in study design and include the outputs in pre-analysis plans or grant proposals. But not all power analyses are equal. We introduce a conceptual distinction between empirical power analysis, which uses empirical data as inputs, and non-empirical power analysis, which does not. We then propose an analytical framework that treats empirical power analysis as a statistical estimation problem to systematically investigate its reliability, without the necessity to assume the underlying distribution of the outcome. We apply this framework to the two most common empirical forms, the power estimation and the MRSS estimation, to analyze their properties as estimators, with a focus on parameter ranges that political scientists may likely encounter.

Our theoretical and simulation analyses reveal that empirical power analysis has poor utility in political science. Power estimates are strongly biased with an unpredictable direction and sensitive to small input changes, when the key parameters of the true standardized effect and the pilot size are within the range for typical political science applications. MRSS also has an infinite expectation and variance, leading to extreme variability for reasonable pilot sizes and true effects. We further identify a fallacy in current pilot-based empirical power analysis: estimating MRSS conditional on pre-specified statistical significance. From these findings, we offer practical recommendations and generally advise against empirical power analysis as it is currently practiced.

Our analysis leaves several questions for future research. First, we focus on power analysis for simple randomized experiments with a binary treatment and under simple random sampling. It is the basis for most power calculations, covering much of the empirical work, but other designs also exist. With trends toward complex experiments and using pre-treatment covariates to boost efficiency, future work may analyze non-traditional power analysis, such as simulation-based ones. Second, existing bias correction methods perform poorly for power estimation, and this suggests a demand for alternative statistical or design-based

bias-correction solutions for 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.

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.

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

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.

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

Cox, D. R. and N. Reid (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society: Series B (Methodological) 49*(1), 1–18.

Dunham, Y. and E. Lieberman (2013, Nov). Social identity and social risk: Experimental research on race, stigma, and hiv/aids in the united states. `https://osf.io/6tx42`. Accessed 29-August-2025.

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

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.

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.

Huang, A. and P. J. Rathouz (2017). Orthogonality of the mean and error distribution in generalized linear models. *Communications in Statistics-Theory and Methods 46*(7), 3290–3296.

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.

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.

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.

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

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.

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

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

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.

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

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

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.

Shah, A. K. (1985). A simpler approximation for areas under the standard normal curve. *The American Statistician 39*(1), 80–80.

Tausanovitch, C. (2015, June). Why do voters support partisan candidates? `https://osf.io/nz4cf`. Accessed 29-August-2025.

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

Appendix

A.1 Proof of Proposition 1

Note $\hat{\psi} = 1 - \Phi\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right) - \frac{\hat{\tau}_{\text{std}}\sqrt{N_f}}{2}\right) + \Phi\left(-\Phi^{-1}\left(1-\frac{\alpha}{2}\right) - \frac{\hat{\tau}_{\text{std}}\sqrt{N_f}}{2}\right)$, we work on the second term $\Phi\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right) - \frac{\hat{\tau}_{\text{std}}\sqrt{N_f}}{2}\right)$ and the third term follows.

We expand $\Phi\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right) - \frac{\hat{\tau}_{\text{std}}\sqrt{N_f}}{2}\right)$ into a Taylor series around the true τ_{std},

$$\begin{aligned}\Phi\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right) - \frac{\hat{\tau}_{\text{std}}\sqrt{N_f}}{2}\right) \approx & \Phi\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right) - \frac{\tau_{std}\sqrt{N_f}}{2}\right) \\ & - \frac{1}{\left(\frac{2}{\sqrt{N_f}}\right)}\phi\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right) - \frac{\tau_{std}\sqrt{N_f}}{2}\right)(\hat{\tau}_{\text{std}} - \tau_{std}) \\ & + \frac{1}{2\left(\frac{2}{\sqrt{N_f}}\right)^2}\phi'\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right) - \frac{\tau_{std}\sqrt{N_f}}{2}\right)(\hat{\tau}_{\text{std}} - \tau_{std})^2\end{aligned}$$

where $\Phi(.)$ is the standard normal CDF, $\phi(.)$ the standard normal probability distribution function (PDF), and $\phi'(.)$ the first derivative of the standard normal PDF.

We apply the expectation on both sides and get

$$\begin{aligned}\mathbb{E}\left\{\Phi\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right) - \frac{\hat{\tau}_{\text{std}}\sqrt{N_f}}{2}\right)\right\} \approx & \Phi\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right) - \frac{\tau_{std}\sqrt{N_f}}{2}\right) \\ & - \frac{1}{\left(\frac{2}{\sqrt{N_f}}\right)}\phi\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right) - \frac{\tau_{std}\sqrt{N_f}}{2}\right)\mathbb{E}\left(\hat{\tau}_{\text{std}} - \tau_{std}\right) \\ & + \frac{1}{2\left(\frac{2}{\sqrt{N_f}}\right)^2}\phi'\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right) - \frac{\tau_{std}\sqrt{N_f}}{2}\right)\mathbb{E}\left(\hat{\tau}_{\text{std}} - \tau_{std}\right)^2\end{aligned}$$

Note the remainder term on the right hand side does not equal zero, because: (1) $\mathbb{E}[\hat{\tau}_{\text{std}}] \neq$

τ_{std}, and (2) the second-order term equals $\frac{\mathbb{V}(\hat{\tau}_{\text{std}})}{2\left(\frac{2}{\sqrt{N_f}}\right)^2}\phi'\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right)-\frac{\tau_{std}\sqrt{N_f}}{2}\right)$, which is indeed $\frac{1}{2}\cdot\frac{N_f}{N_p}\phi'\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right)-\frac{\tau_{std}\sqrt{N_f}}{2}\right)$.

A.2 Properties of the Plug-in Power Estimator

To derive properties of the power estimator in equation (8), 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 make two simplifying assumptions: (1) the outcome variance is constant across the treated and the control, so $\sigma^2 \equiv S_1^2 = S_0^2$; (2) the treatment is randomly assigned to minimize the sampling variance of the estimated ATE (Neyman, 1923), so $N_d/2 = n_{d1} = n_{d0}$, $d \in \{f, p\}$.

We also clarify that the assumptions that we are *NOT* making are: (1) the parametric assumption for the underlying distribution of the outcomes in the population, and (2) whether the outcomes are subject to an identical and independent distribution, as our results are based on the sampling theory (O'Neill, 2014). Thus, the analytical results derived in this paper shall apply to all cases regardless of the underlying distribution of the outcome.

Thus, utilizing the fact that all observations in the treated group are drawn randomly from the population of the potential outcomes under treatment, where the population distribution has mean τ and variance σ^2, the sample mean of the treated group, which itself is a random variable, has a mean of τ, and a variance of $\frac{\sigma^2}{\frac{1}{2}N_p} = \frac{2\sigma^2}{N_p}$, under the standard random sampling theory (O'Neill, 2014). For a similar reason, the sample mean of the control group has a mean of 0, and a variance of $\frac{2\sigma^2}{N_p}$.

Recalling that $\hat{\tau}$ is the difference in the estimated sample means, we have

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

Since $\hat{\tau} \perp\!\!\!\perp \hat{\sigma}^2$ (Cox and Reid, 1987; Huang and Rathouz, 2017), we have

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

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.

Assumption A1. *(outcome normality) The outcome variable 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,

$$\begin{aligned}\mathbb{E}\left[\frac{\hat{\tau}}{\sqrt{\hat{\sigma^2}}}\right] &\approx \frac{\tau\sqrt{\frac{N_p-1}{\sigma^2}}}{\sqrt{N_p-\frac{5}{2}}} \\ &= \frac{\tau}{\sigma}\sqrt{1+\frac{3}{2N_p-5}}\end{aligned} \tag{9}$$

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

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

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,

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

Further,

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

Thus,

$$\begin{aligned}
\mathbb{V}\left[\hat{\tau}_{\text{std}}\right] &= \left(\frac{4\sigma^2}{N_p} + \tau^2\right)\left(\frac{N_p - 1}{N_p - 3} \cdot \frac{1}{\sigma^2}\right) - \frac{\tau^2}{\sigma^2}\left(1 + \frac{3}{2N_p - 5}\right) \\
&= \frac{4(N_p - 1)}{n(N_p - 3)} + \frac{\tau^2}{\sigma^2}\frac{N_p - 1}{(N_p - 3)(2N_p - 5)}.
\end{aligned} \tag{10}$$

Thus, $\mathbb{E}[\hat{\tau}_{\text{std}}]$ and $\mathbb{V}[\hat{\tau}_{\text{std}}]$ are given by equations (9) and (10), 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 $\mathbb{V}\left[\frac{\hat{\tau}}{\sigma}\right] = \frac{4}{N_p}$

and

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

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 $\mathbb{E}\left[\hat{\psi}\right]$, we would need to evaluate $\mathbb{E}\left[\Phi(x)\right]$. Since $\Phi(\cdot)$ cannot be expressed in closed form, we apply the following approximation (Shah, 1985):

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

Hence, for a random variable x,

$$\begin{aligned} \mathbb{E}\left[\Phi\left(x\right)\right] \approx 0.01 \times \int_{-2.6}^{-2.2} p(x)dx + \int_{-2.2}^{0} \left[0.5 - \frac{-4.4x - x}{10}\right] p(x)dx \\ + \int_{0}^{2.2} \left[\frac{4.4x - x^2}{10} + 0.5\right] p(x)dx + 0.99 \times \int_{2.2}^{2.6} p(x)dx + \int_{2.6}^{+\infty} p(x)dx \end{aligned} \tag{11}$$

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, $\mathbb{E}\Phi(x) \approx 0.01$, and $\Phi(x_{true}) \approx 0.01$, and the bias would be small.

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

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

while

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

the bias of the power would approximately 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 approximately 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,$\mathbb{E}\left[\Phi\left(\widehat{X_1}\right)\right] \approx 0.99$, and $\Phi\left(x_{true}\right) \approx 0.99$, 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. Thus, $p(x)$ 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 where $p(x)$ is more heavily distributed, whether on the negative part or on the positive part.

A.3 Bias Correction Methods

To fully illustrate the point that the bias for Equation (8) 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 have tried several conventional 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 depicts 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 sufficiently 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, the 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 reducing the bias.

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 (8) to obtain her power estimation. This setting is not realistic, but it illustrates the validity of our simplifying assumption in the main text where we assume for homogenous variance for the outcome of the treated units and the control units, as well as in the previous section of appendix where we assume away

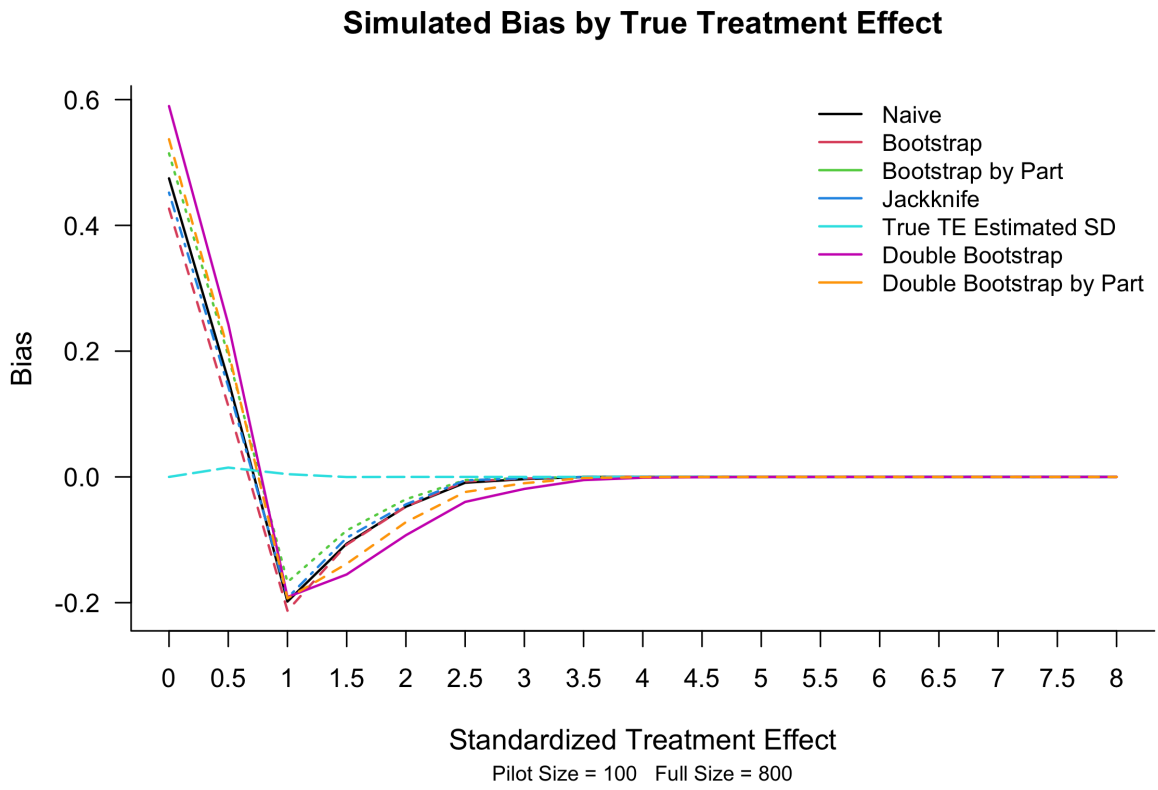

Figure A.1: Comparison of the Bias Correction Techniques Applied to the Power Estimator. The results show that none of the existing techniques appreciably improves estimates over the naive uncorrected estimator. (The light blue dashed line represents the unfeasible "oracle" estimator where the true effect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

the sampling error for variance estimation and only focus on the sampling error for treatment effect.

A.4 Divergence of $\mathbb{E}\left[\widehat{MRSS}\right]$

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

$\widehat{MRSS}$ undefined. To see this, note that:

$$\begin{aligned}\mathbb{E}\left[\widehat{MRSS}\right] &= \mathbb{E}\left\{\frac{4\left[\Phi^{-1}\left(1-\frac{\alpha}{2}\right)-\Phi^{-1}\left(1-\psi\right)\right]^2}{\hat{\tau}^2_{\text{std}}}\right\} \\ &= 4\left[\Phi^{-1}\left(1-\frac{\alpha}{2}\right)-\Phi^{-1}\left(1-\psi\right)\right]^2 \mathbb{E}\left\{\frac{1}{\hat{\tau}^2_{\text{std}}}\right\}.\end{aligned}$$

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[15]. Letting $\overline{f}$ be the upper bound for $f_{\hat{\tau}_{\text{std}}}(x)$, we have

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

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, i.e. $\int_{-1}^{0}\frac{1}{x^2}f_{\hat{\tau}_p}(x)dx \geq \min_{x\in[-1,0]} f_{\hat{\tau}_p}(x)\int_{-1}^{0}\frac{1}{x^2}dx = \min_{x\in[-1,0]} f_{\hat{\tau}_p}(x)\int_{0}^{1}\frac{1}{x^2}dx$[16]. 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 $2\times\min_{x\in[-1,1]} f_{\hat{\tau}_p}(x)\int_0^1\frac{1}{x^2}dx$. Yet, $\int_0^1\frac{1}{x^2}dx$ is positive and not upper bounded. As a result, $\mathbb{E}\left\{\frac{1}{\hat{\tau}^2_p}\right\}$ 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.

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

[16]The standard normal probability distribution function is continuous within such domain and thus extreme value theorem holds. The second equality holds because of symmetry of $\frac{1}{x^2}$.

A.5 Consistency for $\widehat{MRSS}$

We claim that the sequence $\widehat{MRSS}_{N_p} = \frac{4\left[\Phi^{-1}\left(1-\frac{\alpha}{2}\right)-\Phi^{-1}(1-\psi)\right]^2}{\hat{\tau}^2_{\text{std}}}$ 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 $\mathbb{P}\left(\left|\widehat{MRSS}_{N_p} - N\right| \leq \varepsilon\right) < \delta$, 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 $\Phi(\centerdot)$ is the standard normal cumulative distribution function.

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

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

Thus, 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 $N_p \geq N_{upper}$, $\mathbb{P}\left(\left|\widehat{MRSS}_{N_p} - N\right| \leq \varepsilon\right) = \mathbb{P}\left(\frac{C}{\hat{\tau}_{\text{std}}^2} \geq \frac{C}{\tau_{\text{std}}^2} + \varepsilon\right) + \mathbb{P}\left(\frac{C}{\hat{\tau}_{\text{std}}^2} \leq \frac{C}{\tau_{\text{std}}^2} - \varepsilon\right) \leq \frac{4}{5}\delta < \delta$.

A.6 Decomposition of the Standardized Effect $\hat{\tau}_{std}$

To further study whether it is $\hat{\tau}$, the estimated treatment effect, or $\hat{\sigma}$, the estimated standard deviation of the outcome, that leads to the bias in the power estimation. We further conduct the following two simulation exercises:

Simulation Procedure with True σ We conduct the following Monte Carlo experiment for each combination of the τ_{std}, N_p and N_f values, same as those in Section 3:

1. Randomly draw $\frac{N_p}{2}$ realizations of Y for the treatment group such that $Y_1 = \mu_1 + \varepsilon_1$, and $\frac{N_p}{2}$ realizations of $Y_0 = \mu_c + \varepsilon_c$ for the control group, where $\varepsilon_1 \sim \frac{4}{\sqrt{3}}t(3)$ and $\varepsilon_c \sim \frac{4}{\sqrt{3}}t(3)$. Such set-up indicates $S_1^2 \equiv \mathbb{V}(Y \mid Z=1) = S_0^2 \equiv \mathbb{V}(Y \mid Z=0) = 4^2$, and the true average treatment effect being $\mu_1 - \mu_c$.

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

3. Assume that a researcher knows the true variances of Y in the treatment and control groups, S_1^2 and S_0^2, respectively. Denote those by $\widehat{S_1^2} = \widehat{S_0^2} = 4^2$.

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 for 1,000 times and calculating Monte Carlo estimates of the bias and the standard error.

In Figure A.2, we replicate Figure 1 with the same parameter space of $\hat{\tau}_{std}$, N_p and N_f, but a different data generation process whose outcome variable features a fatter tail and a different simulation procedure that assumes the knowledge of the true σ. The simulations look very similar to those in Figure 1. This indicates, first, our observations for the bias in power estimation is robust to a different outcome generation process (normal DGP in Figure 1 vs student-t DGP in Figure A.2). Second, the knowledge of the true standard deviation of the outcome variable does not alleviate the bias in power estimation.

Simulation Procedure with True τ We conduct the following Monte Carlo experiment for each combination of the τ_{std}, N_p and N_f values, same as those in Section 3:

1. Randomly draw $\frac{N_p}{2}$ realizations of Y for the treatment group such that $Y_1 = \mu_1 + \varepsilon_1$, and $\frac{N_p}{2}$ realizations of $Y_0 = \mu_c + \varepsilon_c$ for the control group, where $\varepsilon_1 \sim \frac{4}{\sqrt{3}}t(3)$ and $\varepsilon_c \sim \frac{4}{\sqrt{3}}t(3)$. Such set-up indicates $S_1^2 \equiv \mathbb{V}(Y \mid Z = 1) = S_0^2 \equiv \mathbb{V}(Y \mid Z = 0) = 4^2$, and the true average treatment effect being $\mu_1 - \mu_c$.

2. Assume the researcher knows the true μ_1 and μ_c. Let $\hat{\tau} = \tau = \mu_1 - \mu_c$.

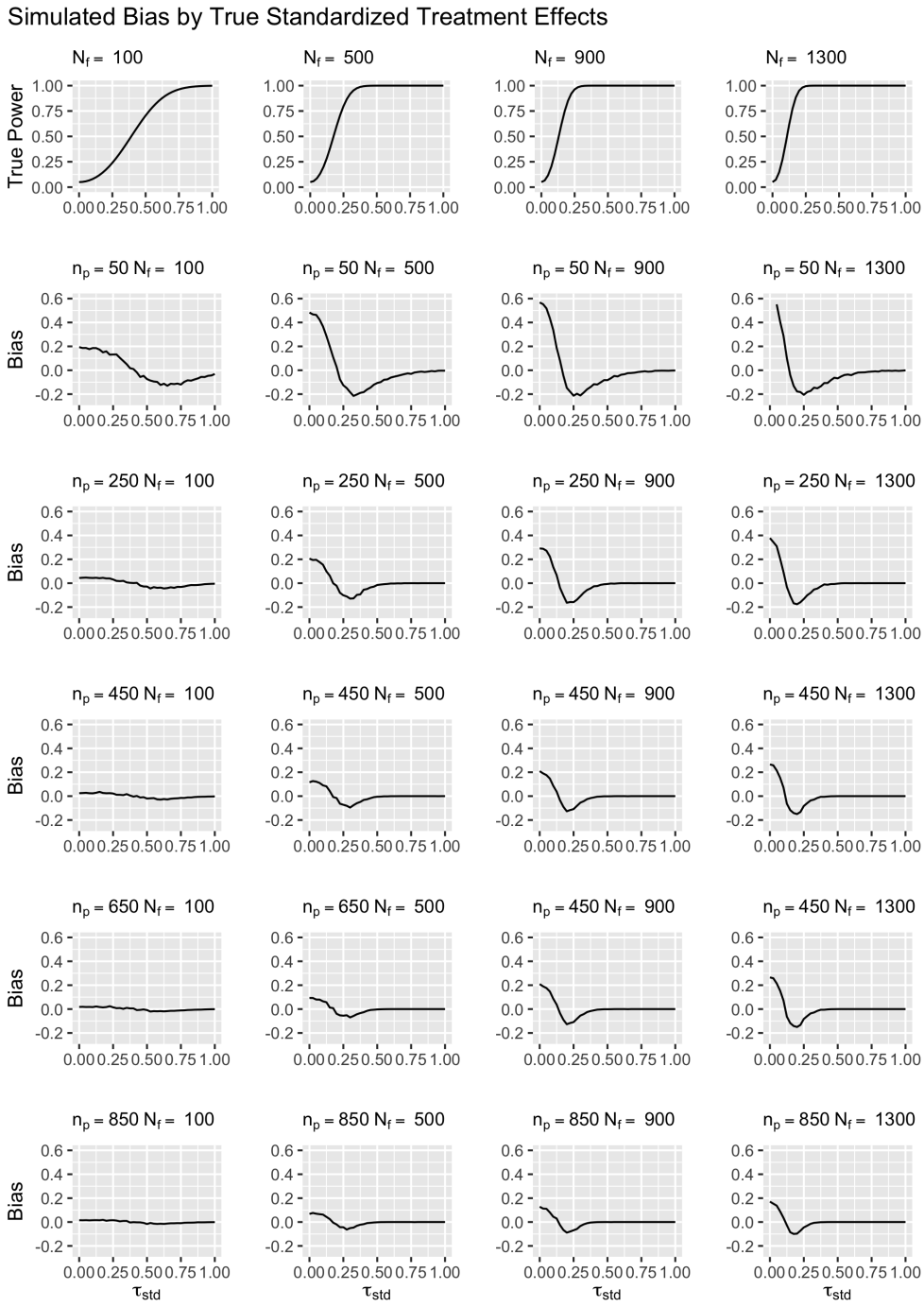

Figure A.2: Simulated Bias by True Standardized Treatment Effect with Knowledge of True σ. The top row of plots present the true power for each full experiment sample size as a function the standardized effect size. The remaining plots show Monte Carlo estimates of the bias of the power estimator on 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). The outcome is generated according to a scaled student-t distribution with a degree of freedom at 3.

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 for 1,000 times and calculating Monte Carlo estimates of the bias and the standard error.

In Figure A.3, we replicate Figure 1 with the same parameter space of $\hat{\tau}_{std}$, N_p and N_f, but a different data generation process whose outcome variable features a fatter tail and a different simulation procedure which assumes the knowledge of true τ. The simulation results indicate that the bias of power estimation is much smaller compared with that in Figure 1 and Figure A.2. This indicates the bias in the power estimation is mainly driven by the imprecise estimation of τ, the true treatment effect, rather than that of σ, the standard deviation of the outcome.

A.7 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 between 2015 and 2024. 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

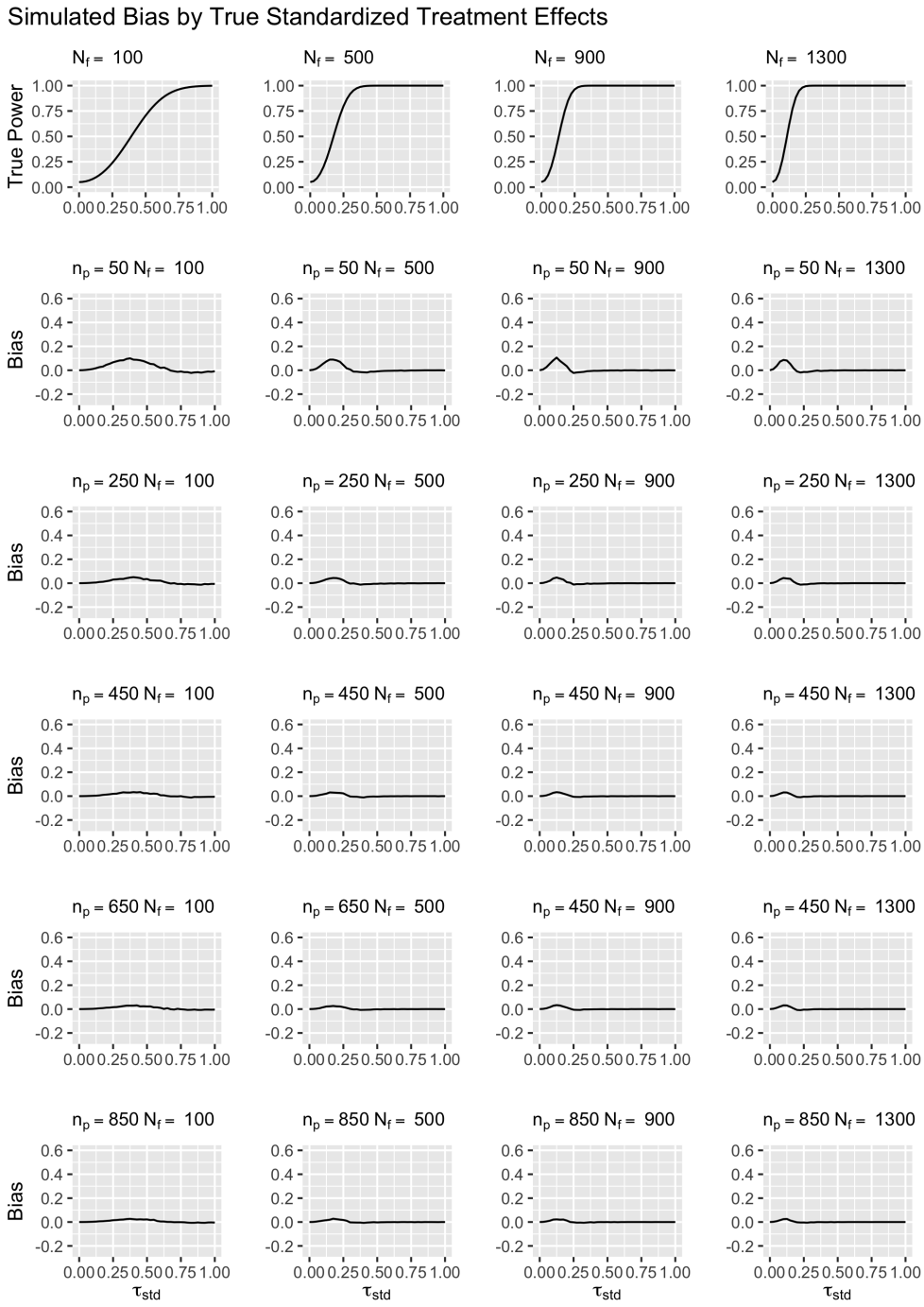

Figure A.3: Simulated Bias by True Standardized Treatment Effect with Knowledge of True τ. The top row of plots present the true power for each full experiment sample size as a function the standardized effect size. The remaining plots show Monte Carlo estimates of the bias of the power estimator on 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). The outcome is generated according to a scaled student-t distribution with a degree of freedom at 3.

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 305 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.

4. We exclude the quantity (and the publication) if the main text and the appendix of the paper does not report at least a conventional numeric standard error or a t-statistic (e.g. when the author adopts permutation tests, or have just reported the results in a figure but not numbers in the appendix) – this means we cannot infer the standardized treatment effects without looking into replication files.

Our resulting dataset contain 410 effect size observations that are either average treatment effects (ATE), or similar causal quantities that can be estimated via difference in means. This number is larger than the number of publications identified because some publications contain multiple studies (experiments).