Power Rules*

Practical Statistical Power Calculations

Carlisle Rainey†

May 29, 2024

Draft *very much* in progress.
Comments welcomed! (crainey@fsu.edu)

Get latest version from GitHub.

Abstract

Recent work emphasizes the importance of statistical power and shows that power in the social sciences tends to be extremely low. In this paper, I offer simple rules that make statistical power more approachable for substantive researchers. The rules describe how researchers can compute power using (1) features of a reference population, (2) an existing study with a similar design and outcome, and/or (3) a pilot study. In the case of balanced, between-subjects designs (perhaps controlling for pre-treatment variables), these rules are sufficient for a complete and compelling power analysis for treatment effects and interactions using only paper-and-pencil. For more complex designs, these rules can provide a useful ballpark prediction before turning to specialized software or complex simulations. Most importantly, these rules help researchers develop a sharp intuition about statistical power. For example, it can be helpful for readers and researchers to know that experiments have 80% power to detect effects that are 2.5 times larger than the standard error and how to easily form a conservative prediction of the standard error using pilot data. These rules lower the barrier to entry for researchers new to thinking carefully about statistical power and help researchers design powerful, informative experiments.

*This essay flows from many discussions with undergraduate students, PhD students, and colleagues. In these discussions, we have thought about power as a minor extension of estimation and inference rather than power as a distinct topic. In this essay, I try to articulate that perspective and help researchers extend their intuitions and expertise with estimation and inference to statistical power. Marco Aviña, Scott Clifford, and Chuck Smith provided valuable comments on earlier drafts.

†Carlisle Rainey is Associate Professor of Political Science, Florida State University, 540 Bellamy, Tallahassee, FL, 32306. (crainey@fsu.edu).

Introduction

Experiments continue to grow in political science (Druckman and Green 2021), and designing an experiment requires us to either select an appropriate sample size or evaluate whether a given sample size is sufficient for our goals.[1] Statistical power, or the chance of rejecting the null hypothesis for a given treatment effect, is a common, powerful way to select or evaluate a sample size. The consequences of low power are increasingly understood: low power can lead to wildly overestimated treatment effects (Gelman and Carlin 2014) and ambiguous results that are difficult to publish (e.g., Alrababa'h et al. 2023). Nonetheless, Arel-Bundock et al. (2022) argue that only about 10% statistical tests in political science have at least 80% power. And political science is hardly unusual (Cohen 1962; Button et al. 2013; Ioannidis, Stanley, and Doucouliagos 2017; Stanley, Carter, and Doucouliagos 2018; Yang et al. 2023; Stommes, Aronow, and Sävje 2023). Underscoring their important findings, Arel-Bundock et al. (2022) emphasize: "our research community must address the problems of low power and selection on significance with institutional, methodological, and theoretical remedies."

One apparent problem is that while political scientists understand estimation and inference for experimental data very well (e.g., least squares regression with robust standard errors), political scientists are less comfortable thinking about statistical power prior to data collection. This apparent discomfort is not without reason. Power analysis requires specialized software or tedious simulation. Most specialized software, such as G*Power (Faul et al. 2007, 2009), is

[1]For example, the *Journal of Politics* guidelines for registered reports notes that "a detailed justification of the planned sample size is essential" (accessed May 2, 2024, persistent link: https://archive.md/g7YQ3). The *Journal of Experimental Political Science* requires authors to "explain how the sample size was determined and note statistical power" (accessed May 2, 2024, persistent link: https://archive.md/7nLKg). Time-sharing Experiments for the Social Sciences (TESS) notes that power analysis is "encouraged" (accessed May 11, 2024, persistent link: https://archive.md/XcOUY). The Civic Health and Institutions Project, a 50 States Survey, (CHIP50) notes that power analysis is "strongly encouraged" (accessed May 11, 2024, persistent link: https://archive.md/SLpgW). The National Institutes of Health (NIH) provides several examples to "demonstrate rigor" and all examples include a discussion of statistical power (accessed May 11, 2024, persistent link: https://archive.md/Jh9Lp). The popular OSF pre-registration template asks users to provide an explicit sample size rationale, which could include a power analysis (Bowman et al. 2020).

designed primarily for psychologists. The language (e.g., Cohen's d) and style of analysis (e.g., ANCOVA) can make the software awkward for political scientists (though of course not insurmountable).[2] Most importantly, political scientists tend to focus on treatment effects on the scale of the outcome (e.g., King, Tomz, and Wittenberg 2000), while psychologists are more comfortable with standardized effect sizes.[3] And while psychologists rely on a range of procedures to test hypotheses, political scientists typically use least squares regression with robust standard errors. In political science, recent conceptual work and statistical software (Blair et al. 2019; Blair, Coppock, and Humphreys 2023) make complex simulations to comprehensively evaluate research designs more accessible, but simulation to obtain statistical power remains challenging compared to analyzing experimental data. Political scientists lack a gateway between the simple, robust tools of inference using experimental data and the complex, tedious tools that allow complete and careful evaluations of research designs. This paper can help create the link.

In this paper, I offer several simple rules for assessing statistical power. These rules do not require special software or simulation and are motivated from statistical analyses common in political science. I tightly connect the rules for statistical power to the logic of estimation and inference for experimental data in political science (least squares regression, robust standard errors, and confidence intervals). There is tremendous value in simple, conceptual rules of thumb—the kind of rules that allow us to predict the statistical power of an experiment with paper, a pencil, and (perhaps also) a pocket calculator.[4] While specialized software

[2]For example,the power resources that the *Journal of Politics* provides include links to several resources for planning sample size in experiments, but many of their resources are directed primarily at psychologists.

[3]Lakens (2022) offers a careful and thorough discussion of sample size justifications for an audience of psychologists, but (appropriately) uses terms common in psychology. For example, Lakens centers much of his discussion around Cohen's d, standardized effect size used often in psychology but less often in political science.

[4]My own view is that paper-and-pencil calculations are valuable, for two reasons. First, predicting statistical power involves a fair amount of guesswork. Specialized software and simulations bring an air of exactness that does not seem to match the spirit of the task. With paper-and-pencil calculations, it seems easier to remember that task is to make an informed guess. Second, paper-and-pencil work constantly reminds us how inputs relate to outputs. Since we are *designing* experiments and have freedom to make different

and complex simulations are helpful (and essential in some cases), simple rules have several benefits: (1) simple rules can make statistical power accessible, rather than seem like a mystical quantity; (2) simple rules allow us to reason fluently about design choices; (3) simple rules can reduce the chance of making big errors by giving us a ballpark starting point even if we ultimately rely on a more sophisticated approach; and (4), in many cases, simple rules are sufficient to justify a sample size and simulations and specialized software are unnecessary. In this paper, I walk through several rules to predict the statistical power of an experiment. For simple randomization schemes, these rules are sufficient to demonstrate adequate power for treatment effects and interactions. I explain the intuition of each rule, describe how we can use the rule to plan or evaluate a study, and give examples. The rules are often sufficient to predict the power for studies of treatment effects and interactions using features of a reference population, existing studies, and/or pilot data.

Review: The Estimation and Inference Framework

Throughout this discussion, I assume that we use ordinary least squares and robust standard errors to estimate the treatment effect. I assume size-0.05, one-sided tests of directional hypotheses evaluated in practice using 90% confidence intervals. (A later section describes how researchers can modify the rules for 95% confidence intervals, if they prefer.) To make the discussion more natural, I borrow the language of survey experiments and describe the experimental units as "respondents," but the logic generalizes to other types of units (e.g., villages).

Formally, I assume potential outcomes $\mathcal{Y}_i(1)$ if assigned to treatment and $\mathcal{Y}_i(0)$ if assigned to control for respondent i of N. An experiment might use more than two conditions, but for the sake of the discussion below, I focus on the statistical power of a comparison of just *two*

design choices, it can be helpful to be constantly reminded of the relationship between inputs and outputs.

of these conditions, which I call treatment and control.[5] For these two conditions, we want to make inferences about the average treatment effect $\tau = \frac{1}{N}\sum_{i=1}^{N}\mathcal{Y}_i(1) - \frac{1}{N}\sum_{i=1}^{N}\mathcal{Y}_i(0)$ (which I refer to as a "treatment effect" throughout).

To make inferences about τ, we randomly assign n_1 respondents to treatment and n_0 respondents to control, where $N = n_1 + n_0$.[6] We create an indictor variable D_i that equals 1 if the respondent is assigned to treatment and 0 if the respondent is assigned to control. Then we create the observed outcome $Y_i = \mathcal{Y}_i(D_i)$ by assigning Y_i the potential outcome under treatment if $D_i = 1$ and the potential outcome under control if $D_i = 0$. Then we estimate the treatment effect using $\widehat{\tau} = \frac{1}{n_1}\sum_{i=1}^{N} D_i \cdot Y_i - \frac{1}{n_0}\sum_{i=1}^{N}(1 - D_i) \cdot Y_i$, which we compute in practice using least squares estimates of the model $Y_i = \alpha + \tau \cdot D_i + \epsilon_i$. In some cases, we might increase the precision of the estimates by including variables measured pre-treatment as control variables in the regression model.

To estimate the sampling variance of $\widehat{\tau}$, we use Neyman's (1990) conservative variance estimator $\widehat{Var}(\widehat{\tau}) = \frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}$, where s_1^2 and s_0^2 represent the sample variance in the treatment and the control group, respectively. Then we can estimate the standard error of $\widehat{\tau}$ with $\widehat{SE}_{\widehat{\tau}} = \sqrt{\widehat{Var}(\widehat{\tau})}$.

In practice, we use a HC2 heteroskedasticity-robust variance estimator (or a closely related variant), which Samii and Aronow (2012) show is equivalent to Neyman's estimator. For testing, Li and Ding (2017) provide several additional (seemingly innocuous) regularity conditions such that a finite-sample version of the central limit theorem holds and $\left[\widehat{\tau} - \Phi_{std}^{-1}\left(1 - \frac{\alpha}{2}\right) \cdot \widehat{SE}_{\widehat{\tau}}, \widehat{\tau} + \Phi_{std}^{-1}\left(1 - \frac{\alpha}{2}\right) \cdot \widehat{SE}_{\widehat{\tau}}\right]$ is an asymptotically conservative $\left(1 - \frac{\alpha}{2}\right) \cdot 100\%$ confidence interval. We can use the 90% confidence interval $= \left[\widehat{\tau} - 1.64 \cdot \widehat{SE}_{\widehat{\tau}}, \widehat{\tau} + 1.64 \cdot \widehat{SE}_{\widehat{\tau}}\right]$

[5] In many experiments, there will be multiple comparisons of interest. In this cases, we need to compute the statistical power for each test separately. As the *Journal of Politics* guidelines note: "power must be calculated for each test individually" (accessed May 3, 2024; persistent link: https://archive.md/g7YQ3).

[6] Note that N represents the *total number of the respondents* in the comparison, while n_i represents the number of respondents *in condition* i. Below, I assume equal numbers of respondents in each condition, so I let n represent the number of respondents per condition, which is a convenient quantity when thinking about experimental design and statistical power.

to conduct size-0.05 tests of directional hypotheses by rejecting a null hypotheses if and only if *all* values in the 90% confidence interval are consistent with the research hypothesis (Rainey 2014; McCaskey and Rainey 2015).

Rules 1 and 2: What is Statistical Power?

In a designed experiment, we usually have a theoretically motivated research hypothesis that we would like to test. This research hypothesis implies a null hypothesis that we hope to reject. For simplicity, I assume throughout much of this paper that we hypothesize a *positive* treatment effect so that $H_R : \tau > 0$. The extension to a negative treatment effect is obvious.

Imagine that we repeatedly re-randomize to treatment and control and compute $\widehat{\tau}$ for each repetition. Then $\widehat{\tau}$ will vary across these repeated experiments. We refer to the distribution of $\widehat{\tau}$ across these hypothetical repeated experiments as the *sampling distribution* of $\widehat{\tau}$. This sampling distribution is typically centered over the treatment effect (i.e., $\widehat{\tau}$ is unbiased or nearly so). Importantly, the standard deviation of the sampling distribution of $\widehat{\tau}$ is called the *standard error* and $\widehat{SE}_{\widehat{\tau}}$ is usually a precise estimator of the standard error.

If the research hypothesis $H_R : \tau > 0$ is correct, then the sampling distribution of $\widehat{\tau}$ lies mostly to the right of zero. However, we will only claim that $\tau > 0$ if the entire 90% confidence interval is larger than zero, or, equivalently, if $\widehat{\tau} - 1.64 \cdot \widehat{SE}_{\widehat{\tau}} > 0$. This means that we must design our experiment so that the sampling distribution falls relatively far from zero.

We say that the statistical *power* of an experiment is the chance of rejecting the null hypothesis *for a particular treatment effect* τ. For example, we say that an experiment has "80% power to detect a treatment effect of ___" if $\widehat{\tau} - 1.64 \cdot \widehat{SE}_{\widehat{\tau}} > 0$ in 80% of repeated experiments. Similarly, we say that the experiment has "95% power to detect a treatment effect of ___" if the $\widehat{\tau} - 1.64\widehat{SE}_{\widehat{\tau}} > 0$ in 95% of repeated experiments. By convention, 80% power is the minimal standard for "adequate" statistical power (Cohen 1988), but others

recommend higher power, such as 95%.[7]

First, how can we compute statistical power? It turns out that statistical power is determined by the treatment effect and the standard error of the estimate. To compute the statistical power of an experiment, we must make assumptions or "informed guesses" about certain quantities. I denote these assumed values with tilde. For example, I denote the assumed treatment effect as $\widetilde{\tau}$.[8] This just means "the treatment effect we are assuming for the sake of computing statistical power."

How can we compute the statistical power of an experiment? An intuitive approach uses the construction of the 90% confidence interval. Recall that we will reject the null hypothesis if the lower-bound of the 90% confidence interval falls above zero. The arms of the 90% confidence interval are about 1.64 standard errors wide.[9] This means that we will reject the null hypothesis if the estimate of the treatment effect is larger than 1.64 standard errors. Thus, the power of the study is the percent of the sampling distribution that is larger than 1.64 standard errors. Recall that the sampling distribution is approximately normally distributed and centered over the treatment effect τ with standard deviation equal to the standard error of the estimate of τ. Thus, computing the power is simply computing the percent of the sampling distribution above $1.64 \cdot SE$.

Rule 1 (Power; Most Intuitively): Power equals $1 - \Phi(1.64 \cdot SE; \mu = \tau, \sigma = SE)$, where $\Phi(z; \mu, \sigma)$ is the normal CDF, SE is the standard error of the estimated treatment

[7]For example, the *Journal of Politics* "a power of at least 0.90, ideally 0.95, to detect anticipated/relevant effect sizes" for registered reports (accessed May 3, 2024; persistent link: https://archive.md/g7YQ3). *Nature Human Behavior* requires that "the a priori power must be 0.95 or higher for all proposed hypothesis tests" for registered reports (accessed May 13, 2024; persistent link: https://archive.md/xpGdj).

[8]We have three τs floating around: the *actual* treatment effect τ, the treatment effect *assumed for the sake of computing power* $\widetilde{\tau}$, and the *estimated* treatment effect $\widehat{\tau}$. I read the tilde notation as "assumed to compute power." For example, I read $\widetilde{\tau}$ as "the treatment effect assumed to compute power." The tilde notation also means "Wait!—think about this carefully, it is an assumption requiring judgment."

[9]I write "about" 1.64 standard errors because in practice we use the *estimated* standard error, not the actual standard error. However, the estimated standard error is sufficiently precise to treat the two interchangeably in this context.

effect, and τ is the treatment effect.

See Gerber and Green (2012, 92) for a similar presentation.

Dividing through by the standard error and then subtracting $\frac{\tau}{SE}$, we see that power also equals $1 - \Phi\left(1.64 - \frac{\tau}{SE}; \mu = 0, \sigma = 1\right)$; the values of $\Phi\left(z; \mu = 0, \sigma = 1\right)$ are provided in common z tables, where $z = 1.64 - \frac{\tau}{SE}$. This allows us to compute power using the *standard* normal CDF.

Rule 2 (Power; From a z Table): Power equals $1 - \Phi_{std}\left(1.64 - \frac{\tau}{SE}\right)$, where $\Phi_{std}(z)$ is the standard normal CDF (as found in a standard z table), SE is the standard error of the estimated treatment effect, and τ is the treatment effect.

Rule 2 allows us to connect statistical power to the familiar z table in the appendices of many statistics textbooks or the `pnorm()` function in R. But, more importantly, it drives home an important intuition—power is determined by the key ratio $\frac{\tau}{SE}$. To estimate power, we must make an informed assumption about the effect of interest and a good prediction of standard error of the estimate. Further, when we consider changes to the experimental design and the consequences for statistical power, it can be helpful to think about the numerator or the denominator of the key ratio $\frac{\tau}{SE}$: How do you make the effect as large as possible? And how do you make the estimate as precise as possible?

What Treatment Effect Should I Assume?

To compute statistical power, we must make an informed assumption about the treatment effect. Statements about statistical power always have the form: "the experiment has ___% power *to detect a treatment effect of* ___." But obviously, we do not know the treatment effect with any precision, else we would not need to run the experiment. Thus, we must assume. Lakens (2022, 6) writes: "The goal of an a-priori power analysis is to achieve

sufficient power, given a specific *assumption* of the effect size a researcher wants to detect" [emphasis mine].

For any given experiment, there are three treatment effects worth thinking about: (1) the "best guess" of the effect, (2) the smallest plausible effect, and (3) the smallest substantively meaningful effect. *By default, we should focus on the statistical power for the smallest substantively meaningful effect.*[10] However, when the smallest plausible effect is much larger than the smallest effect of substantive interest, we might compute power for the smallest plausible effect.[11] Importantly, we must use our judgment: judgment about the substantive importance of effects, judgment about the empirical plausibility of various effects, and judgment about the relevance of each to the argument. As Kirk (1996, 755) notes, "researchers have an obligation to make this kind of judgment."[12]

As a crude approximation, we might feel tempted to use a rule of thumb such as declaring a 0.15 SD increase in the outcome is a "small" effect (e.g., Lovakov and Agadullina 2021). While rules of thumb can be useful as a starting point, they are only a first approximation—rules of thumb about "small" effects do not account for the question under study, so we must supplement these rules of thumb with additional arguments. Lakens, Scheel, and Isager (2018, see esp. pp. 261-263) and Lakens (2022, see esp. pp. 10-13) offer helpful further discussion of how we might choose an treatment effect of interest.

Importantly, we should *not* use pilot data to estimate the treatment effect for the purpose

[10]Rainey (2014) refers this this threshold as the "smallest substantively meaningful effect." Others refers to the same concept as the "smallest effect size of interest" (SESOI) (Lakens, Scheel, and Isager 2018), "minimum effect of interest" (MEI), and "minimum meaningful effect" (MME).

[11]If the researchers are confident that a treatment effect is much larger than the smallest substantively meaningful effect, then they might decide it is wasteful to power their experiment for a treatment effect that is much smaller than the smallest plausible effect. This strikes me as a reasonable argument.

[12]In full, Kirk (1996, 755) writes:

> Researchers have an obligation to make this kind of judgment. No one is in a better position than the researcher who collected and analyzed the data to decide whether or not the results are trivial. It is a curious anomaly that researchers are trusted to make a variety of complex decisions in the design and execution of an experiment, but in the name of objectivity, they are not expected or even encouraged to decide whether data are practically significant.

of a power analysis (Leon, Davis, and Kraemer 2011; Albers and Lakens 2018). A pilot data set is too small to estimate the treatment effect with sufficient accuracy—the estimate from the pilot data might even have the wrong sign! (Though see Perugini, Gallucci, and Costantini (2014) for a conserative approach and Lakens (2014) on sequential analysis.)

Making a thoughtful assumption about the treatment effect is perhaps the most challenging component of a power analysis. We are doing a study to learn about a treatment effect, so it feels uncomfortable to make declarations about the effect beforehand. It feels similarly awkward to divide effects into important and unimportant bins. After all, "important" is ill-defined and varies continuously. It is not all obvious how to make this judgment. To make this less awkward, I can say this: it is important to make a thoughtful assumption about the treatment effect, but simply making an *explicit* assumption moves us most of the way toward a useful power analysis. When we declare our assumption—regardless of whether we get it exactly right—we will have useful power analysis. We can say that "our experiment has ___% power to detect a treatment effect of ___."

Rule 3: From the SD to the SE

In practice, we can use the standard deviation of the outcome in the population under study to predict the standard error of the estimated treatment effect. To motivate this rule, recall the regression model, $y_i = \alpha + \tau D_i + \epsilon_i$ that allows us to estimate the treatment effect τ via least squares. By assuming $\epsilon_i \sim N(0, \sigma^2)$ we can obtain the classical standard error estimate $SE_{\widehat{\tau}}^{classic} = \sqrt{\frac{\sigma^2}{N \cdot \overline{D} \cdot (1-\overline{D})}}$, where $\overline{D}$ represents the fraction assigned to the treatment group (or the average of the D_i) and N represents the total sample size in the two conditions. Assuming balanced assignment to treatment and control where $\overline{D} = \frac{1}{2}$ and $n_1 = n_0 = n$, we have $SE_{\widehat{\tau}}^{classic} = \sqrt{\frac{\sigma^2}{2 \cdot n \cdot \frac{1}{4}}} = \frac{2 \cdot \sigma}{\sqrt{2 \cdot n}}$. This is a helpful result. We can think of σ as the standard deviation of the outcome within each experimental condition. Let $\widetilde{SD(Y)}$ denote

an assumption or "guesstimate" about the value of σ. If we can make a good guess $\widetilde{SD(Y)}$ about the value of σ, then we can translate that into a good guess about the value of $SE_{\widehat{\tau}}$ in our study.

To obtain a suitable value $\widetilde{SD(Y)}$ to plug in for σ, we can use the standard deviation of a measure of the outcome (or a similar outcome) in a reference population. For a survey experiment, perhaps a similar survey question has been asked in the American National Election Study or the Cooperative Election Study. We could also reference a previous experiment using the same outcome. Plugging the standard deviation into the equation for the standard error, we can predict that the standard error in our study will be about $\frac{2 \cdot \widetilde{SD(Y)}}{\sqrt{2 \cdot n}}$. Importantly, n represents the sample size *per condition* (not the total sample size).

Rule 3 (SD to SE): We can use features of a reference population to predict the standard error of the estimated treatment effect in a planned study. The standard error will be about $\frac{2 \cdot \widetilde{SD(Y)}}{\sqrt{2 \cdot n}}$, where n is the sample size per condition and $\widetilde{SD(Y)}$ is the standard deviation of the outcome in a reference population.

Importantly, this prediction relies on the assumption that $\epsilon_i \sim N\left(0, \sigma^2\right)$, which is not guaranteed by the design. In fact, one can easily construct creative examples where classical and robust standard errors diverge substantially. However, the classical standard errors are close enough to motivate Rule 3. Indeed, Samii and Aronow (2012, 370) write that "design-based estimators that exploit the randomization distribution while eschewing regression assumptions may not be as different from classical regression estimators as may seem at first glance." In particular, Samii and Aronow (2012, 370) show that design-based and classical standard errors are equivalent for balanced designs and constant treatment effects. Thus, while we should prefer robust standard errors in practice, classical standard errors allow a simple, closed-form approximation that help us understand the power of our hypothesis tests using simple, helpful rules.

Rule 4: Adjusting for Adjustment

We can sometimes substantially decrease the standard error by adjustment for covariates measured pre-treatment. Clifford, Sheagley, and Piston (2021) suggest a promising strategy of measuring the outcome of interest both before and after the treatment and controlling for the pre-treatment measure of the outcome in the regression model. Using the classical framework from above, Cox and McCullagh (1982, 547) show that regression adjustment using k pre-treatment covariates changes the standard error by a factor of about $\sqrt{(1-\rho^2)\cdot\left(1+\frac{k}{(N-3)+k}\right)}$, where ρ^2 is the population R^2 for a regression predicting the outcome using the pre-treatment controls within each condition (see also Boldt, Lin, and Rainey 2024). In the case of a single pre-treatment covariate, ρ is the correlation between the outcome and the pre-treatment covariate within each condition. For many designs, N is large relative to k, so that $\frac{k}{(N-3)+k} \approx 1$ and the factor becomes about $\sqrt{(1-\rho^2)}$ (Bloom 1995; Meyvis and Van Osselaer 2017). We can use this factor to translate the standard error without adjustment into the standard error with adjustment using $\sqrt{(1-\rho^2)}\cdot SE_{\widehat{\tau}}^{noadj.} = SE_{\widehat{\tau}}^{adj.}$. Alternatively, we can say that adjustment shrinks the standard error (e.g., from Rule 3) by about $\left[1-\sqrt{(1-\rho^2)}\right]\cdot 100\%$. We can use this relationship to compute statistical power by plugging in an assumed value for ρ^2, which I denote as $\widetilde{R^2}$. As before, we can use a reference population to obtain a suitable value of $\widetilde{R^2}$ to adjust the predicted standard error, such as using the R^2 of a regression of the outcome on the control variables in a similar population.

Rule 4 (Adjustment): We can use features of a reference population to predict how much regression adjustment using pre-treatment control variables will shrink the standard error of an unadjusted estimate in a planned study. Regression adjustment will shrink the standard error of the unadjusted estimate by about $\left[1-\sqrt{\left(1-\widetilde{R^2}\right)}\right]\cdot 100\%$, where $\widetilde{R^2}$ is the R^2 of a regression of the outcome on the control variables in a reference population.

Rule 5: From the SE to the MDE

It turns out that the effect we can detect with 80% or 95% power is determined by the standard error. Bloom (1995) refers to this critical treatment effect as the "minimum detectable effect" (MDE).[13] The arms of the 90% confidence interval are 1.64 standard errors wide. Thus, the power of the study is the percent of the sampling distribution that is larger than 1.64 standard errors. To find the minimum detectable effect, we need to solve for the treatment effect that positions 80% of the sampling distribution above 1.64. If we plug 0.80 into the inverse of the standard normal CDF, we obtain $\Phi_{std}^{-1}(0.8) = 0.84$.[14] To position 80% of the sampling distribution above 1.64 standard errors, the sampling distribution must be centered $(1.64 + 0.84) = 2.48$ standard errors above zero, which we can safely treat as 2.5. For 95% power, we can compute $(1.64 + \Phi_{std}^{-1}(0.95)) = (1.64 + 1.64) = 3.3$ standard errors (see Bloom 1995 for further discussion). Figure 1 shows the logic of this relationship graphically.

Rule 5 (Bloom's Rule; SE to MDE): An experiment has 80% power to detect a treatment effect that is 2.5 times the standard error and 95% power to detect a treatment effect that is 3.3 times the standard error.

This rule is also helpful for readers! Even if the authors of a study do not discuss their statistical power, you can quickly compute the minimum detectable effect with 80% power by multiplying the standard error times 2.5 and the minimum detectable effect with 95% power by multiplying the standard error times 3.3. However, see Hoenig and Heisey (2001) on the potential dangers of using statistical power as a data analysis tool.[15]

[13] This is distinct from the smallest estimate of the treatment effect that *can be* statistically significant, which is 1.64 standard errors.

[14] In R, we can use `qnorm(0.80)`.

[15] Hoenig and Heisey (2001, 5) write: "power calculations tell us how well we might be able to characterize nature in the future given a particular state and statistical study design, but they cannot use information in the data to tell us about the likely states of nature."

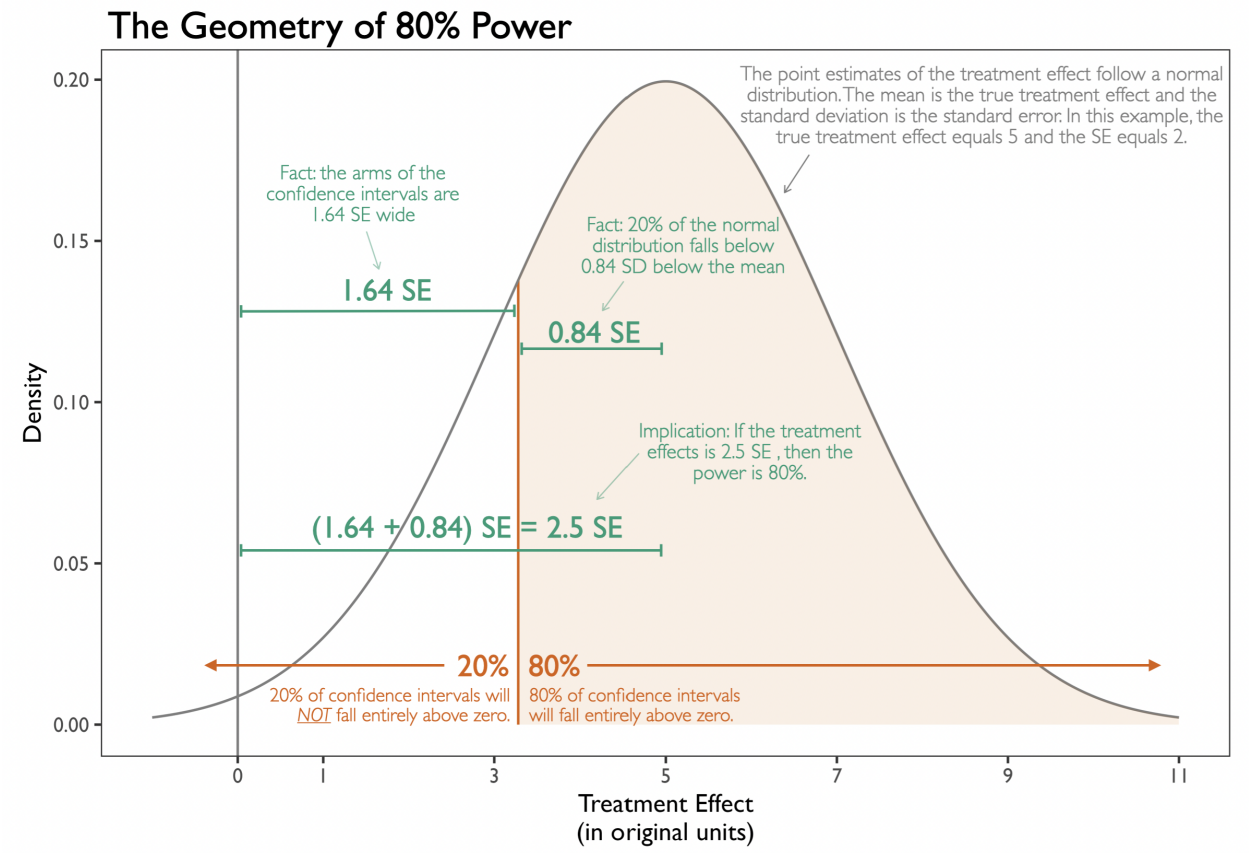

Figure 1: This figure shows why experiments have 80% power to detect effects that are 2.5 standard errors. First, the arms of the 90% confidence interval are 1.64 standard errors wide, so the power of the study is the percent of the sampling distribution that is larger than 1.64 standard errors. If we plug 0.80 into the inverse of the normal CDF, we obtain 0.84. To position 80% of the sampling distribution above 1.64 standard errors, the sampling distribution must be centered (1.64 + 0.84) standard errors = 2.48 standard errors above zero, which we can safely treat as 2.5.

Example: Evaluating Power for Fixed Sample Size

In many research contexts, we do not choose the sample size. For example, it is common to include survey experiments on modules of the Cooperative Election Study (CES), which include 1,000 respondents. Or we might have the opportunity to include an experiment on a survey designed for another purpose. Or we might want to conduct an experiment on the entire population of interest. In this context, we simply need to know whether our already-determined sample size provides sufficient statistical power. Rules 3, 4, and 5 allow us to work in this context (as well as Rules 7 and 9 below).

As an example, imagine we are planning to replicate Ahler and Sood's (2018) finding that correcting respondents' misperceptions of their out-party reduces affective polarization. If we are planning to replicate the finding on a CES module, then we will have 1,000 respondents total or 500 respondents in the treatment and control conditions. Do 1,000 respondents provide sufficient statistical power?

The original experiment shows that correcting respondents' misperceptions of their out-party reduces affective polarization. In their experiment, the treatment has two steps. First, Ahler and Sood (1) ask Republicans to report their perceptions of the percent of Democrats with certain demographic attributes and (2) ask Democrats to report their perceptions of the percent of Republicans with certain demographic attributes. After asking respondents to report their perceptions of the out-party, Ahler and Sood provide respondents with the correct information. Compared to a control group that was neither asked their perceptions nor given the correct information, the treatment group evaluated supporters of the out-party more favorably on a 101-point feeling thermometer scale. They estimate a treatment effect of 6.4 points on the 101-point scale with a 95% confidence interval of [3, 10]. As part of a much larger study, Broockman, Kalla, and Westwood (2022) closely replicate Ahler and Sood's result and estimate the treatment effect is 3.9 with a 90% confidence interval of [1.1,

6.6].[16] In both cases, the 90% confidence interval includes only positive effects, so the authors reject the null hypothesis that the treatment effect is less than or equal to zero and conclude that the treatment has the hypothesized positive effect on the feeling thermometer toward supporters of the out-party.

How could we use features of a reference population along with Rules 3, 4, and 5 to evaluate our study?[17] Our goal is to use Rule 3 to predict the standard error, use Rule 4 to account for any control variables we plan to use, and use Rule 5 to find the minimum detectable effect with 80% and 95% power. If that minimum detectable effect is sufficiently small, then we will say that our study has adequate statistical power.

Without Control Variables: Rules 3 and 5

Using Rule 3, the standard error will be about $\frac{2 \cdot \widetilde{SD(Y)}}{\sqrt{2 \cdot n}}$. In this context, $n = 500$ because we have two experimental conditions and $N = 1,000$ in a CES module. However, we must plug a suitable value for $\widetilde{SD(Y)}$ into the equation. For this application, and many others, we can think of $\widetilde{SD(Y)}$ as the standard deviation of the outcome in the control group (or treatment group if that is easier).

In practice, we can look to standard deviations in reference populations "like" our experimental sample.[18] For example, the 2020 American National Elections Study (ANES) asks a *similar* survey question. The ANES version asks respondents to report their feelings toward the Democratic [Republican] party, while Ahler and Sood ask respondents to report

[16] Broockman, Kalla, and Westwood (2022) report a slightly different analysis in the main text of their paper. Most notably, they use the difference between evaluations of the "people who are" the in-party and "people who are" the out-party. Using their replication data, I run an analysis that treats evaluations of "people who are" the out-party as the outcome (instead of the difference). This latter, alternative analysis is similar to Ahler and Sood's approach, so I focus on it here.

[17] In planning a replication of Ahler and Sood's result, we would ideally translate the standard errors from the original study (as well as Broockman, Kalla, and Westwood's (2022) replication) into a standard error and minimum detectable effect for our study—see Rule 7 below. But, for the sake of this example, suppose this valuable information is not available. Also, triangulating from several sources is helpful.

[18] In a pinch, we can use the crude rule of thumb that the range of a variable tends to be about four times the standard deviation (Wan et al. 2014), so $101/4 = 25.25$, though this rule should be used only as a first approximation, if used at all.

their feelings toward *supporters of* the Democratic [Republican] party. The standard deviation of the responses to the similar ANES question is 20.8.

Using Rule 3, the standard error in our replication will be about $\frac{2 \cdot \widetilde{SD(Y)}}{\sqrt{2 \cdot n}} = \frac{2 \cdot 20.8}{\sqrt{2 \cdot 500}} = 1.32$. Using Rule 5, a standard error of 1.32 means that we will have 80% power to detect a treatment effect of $2.5 \cdot SE = 2.5 \cdot 1.32 = 3.30$ points on the 101-point scale and have 95% power to detect a treatment effect of $3.3 \cdot SE = 3.3 \cdot 1.32 = 4.36$ points. As substantive experts, if we determine that these effects are acceptable (e.g., they correspond to the smallest substantively meaningful effect), then the experiment has adequate statistical power.

With Control Variables: Rules 3, 4, and 5

To shrink the standard error (and the minimum detectable effect along with it), we can consider the adjustment strategy of Broockman, Kalla, and Westwood (2022). They control for a seven-party party identification scale and partisan strength. Rule 4 states that regression adjustment will shrink the unadjusted standard error by about $\left[1 - \sqrt{1 - \widetilde{R^2}}\right] \cdot 100\%$, but we need to choose a suitable value of $\widetilde{R^2}$ to plug into the equation.

In the 2020 ANES, the seven-point party identification scale and partisan strength scale have R^2 of 5% for the similar feeling thermometer toward the Democratic and Republican parties (rather than "supporters of" those parties). For $\widetilde{R^2} = 5\%$, this adjustment strategy will shrink the standard error by about $\left[1 - \sqrt{1 - \widetilde{R^2}}\right] = \left[1 - \sqrt{1 - 0.05}\right] = 2.5\%$. This strategy is perhaps worth pursuing, but will not change the standard error dramatically.[19]

Adding the adjustment to our calculation, our standard error will be about $\frac{2 \cdot \widetilde{SD(Y)} \cdot \sqrt{1 - \widetilde{R^2}}}{\sqrt{2 \cdot n}} = \frac{2 \cdot 20.8 \cdot \sqrt{1 - 0.05}}{\sqrt{2 \cdot 500}} = 1.28$. Using Rule 5, a standard error of 1.28 means that we will have 80% power to detect a treatment effect of $2.5 \cdot SE = 2.5 \cdot 1.28 = 3.20$ points on the 101-point scale

[19] It seems reasonable, too, that these two variables will explain less variation in feelings toward "supporters of" the Democratic and Republican parties, so this R^2 of 5% might be too high. Indeed, re-analysis of Broockman, Kalla, and Westwood's data shows that these two control variables produce an R^2 of 0.9% in the control group and 1.8% in the treatment group. An R^2 of 1.4% (which is the average of the two) will shrink the standard error by about 0.7%. This difference is negligible compared to a strategy without control variables.

and have 95% power to detect a treatment effect of $3.3 \cdot SE = 3.3 \cdot 1.28 = 4.22$ points. The minimum detectable effects using Broockman, Kalla, and Westwood's adjustment strategy are practically the same as those without any adjustment.

However, Clifford, Sheagley, and Piston (2021) offer a potentially promising approach here. They suggest measuring the outcome both before and after the treatment and then controlling for the pre-treatment measure. This pre-treatment measure should be *strongly* related to the post-treatment outcome. If we let $\widetilde{R^2} = 40\%$—though we should do additional work to confirm this working assumption[20]—then the standard error will be about $\frac{2 \cdot 20.8 \cdot \sqrt{1-0.40}}{\sqrt{2 \cdot 500}} = 1.02$. For a standard error of 1.02, we will have 80% power to detect a treatment effect of $2.5 \cdot SE = 2.5 \cdot 1.02 = 2.55$ points on the 101-point scale and 95% power to detect a treatment effect of $3.3 \cdot SE = 3.3 \cdot 1.02 = 3.37$ points. This adjustment strategy might make a difference in whether we consider our 1,000-respondent study adequately powered!

Rule 6: From the SD to the Sample Size

Rule 5 is helpful—it gives us the minimum detectable effect for a given study. But it does not produce the sample size that we need to obtain 80% or 95% power. To find the sample size that produces our desired minimum detectable effect (i.e., usually the smallest substantively meaningful effect), we can combine Rules 3, 4, and 5. Suppose we want our experiment to have 80% power to detect the effect $\widetilde{\tau}$. Then we need $\widetilde{\tau} = \frac{SE}{2.5} = \frac{\frac{\widetilde{SD(Y)}}{\sqrt{2 \cdot n}}}{2.5}$. Solving for n, we find that we will need $2 \cdot \left(\frac{2.5 \cdot \widetilde{SD(Y)}}{\widetilde{\tau}} \right)^2$ respondents per condition to obtain 80% power. If we are adjusting for control variables, we need $2 \cdot \left(\frac{2.5 \cdot \widetilde{SD(Y)} \cdot \sqrt{\left(1-\widetilde{R^2}\right)}}{\widetilde{\tau}} \right)^2$ respondents per condition for 80% power. For 95% power to detect the effect $\widetilde{\tau}$, the factor changes from 2.5 to 3.3 and

[20] In a small pilot study, Culter, Pietryka, and Rainey (2024) validate a strategy of measuring the feelings toward supporters of the out-party before and after the treatment. In their small pilot with 250 respondents, they find that the pre-treatment measure has an R^2 of about 73%. This suggests that the pre-post strategy will shrink the standard error by about 48%.

we need $2 \cdot \left(\frac{3.3 \cdot \widetilde{SD(Y)}}{\widetilde{\tau}} \right)^2$ or $2 \cdot \left(\frac{3.3 \cdot \widetilde{SD(Y)} \cdot \sqrt{\left(1 - \widetilde{R^2}\right)}}{\widetilde{\tau}} \right)^2$ respondents per condition, respectively.[21]

Rule 6 (SD to Sample Size): We can use features of a reference population to compute the sample size we will need in a planned study. For 80% power to detect the treatment effect $\widetilde{\tau}$, the sample size per condition will need to be about $2 \cdot \left(\frac{2.5 \cdot \widetilde{SD(Y)}}{\widetilde{\tau}} \right)^2$ (or $2 \cdot \left(\frac{2.5 \cdot \widetilde{SD(Y)} \cdot \sqrt{\left(1 - \widetilde{R^2}\right)}}{\widetilde{\tau}} \right)^2$ if including control variables), where $\widetilde{SD(Y)}$ is the standard deviation of the outcome in a reference population and $\widetilde{R^2}$ is the R^2 of a regression of the outcome on the control variables in a reference population. For 95% power, the factor changes from 2.5 to 3.3.

Rule 7: From an Existing Study to the Standard Error

The square root law tells us that the standard error depends on the sample size by a factor of $\frac{1}{\sqrt{n}}$. For example, the classical standard errors that I use to motivate these rules have the form $SE^{classic}_{\widehat{\tau}} = \frac{2 \cdot \sigma}{\sqrt{2 \cdot n}} = \frac{2 \cdot \sigma}{\sqrt{2}} \cdot \frac{1}{\sqrt{n}}$. This implies a useful relationship between two studies A and B that vary in their sample size.[22] The ratio of two standard errors $SE^A_{\widehat{\tau}}$ and $SE^B_{\widehat{\tau}}$ with sample sizes n^A and n^B per condition, respectively, has the form $\frac{SE^A_{\widehat{\tau}}}{SE^B_{\widehat{\tau}}} = \frac{\frac{2 \cdot \sigma}{\sqrt{2 \cdot n^A}}}{\frac{2 \cdot \sigma}{\sqrt{2 \cdot n^B}}}$. The two $\frac{2 \cdot \sigma}{\sqrt{2}}$ cancel and we have $\frac{SE^A_{\widehat{\tau}}}{SE^B_{\widehat{\tau}}} = \sqrt{\frac{n^B}{n^A}}$. Solving for $SE^B_{\widehat{\tau}}$, we have $SE^B_{\widehat{\tau}} = \sqrt{\frac{n^A}{n^B}} \cdot SE^A_{\widehat{\tau}}$. Thus, if we know the sample size and standard error of an existing study, then we can use the existing study to predict the standard error in our planned study: our standard error will be about $\sqrt{\frac{n^{existing}}{n^{planned}}} \cdot SE^{existing}_{\widehat{\tau}}$. Importantly, the existing study should have a similar population,

[21] This leads to another useful rule of thumb, though this one is rough. Suppose that we are interested in effects that are 10% of a standard deviation. (Lovakov and Agadullina (2021) find that about one in five studies have effects smaller than 11% of a standard deviation.) Then the term $\left(\frac{\widetilde{SD(Y)}}{\widetilde{\tau}} \right)^2$ equals 100, so we can simply multiply the factors $2 \cdot 2.5^2 = 12.5$ and $2 \cdot 3.3^2 = 21.8$ by 100 to find the required sample sizes of 1,250 and 2,180 per condition to detect these "small" effects with 80% and 95% power.

[22] Or less exactly, but more usefully, two *similar* studies that vary in their sample size.

design, and outcome measure. The sample size, treatment, and treatment effect can differ; otherwise the studies should be similar.

Rule 7 (Existing Study to SE): We can use an existing, similar study to predict the standard error of the estimated treatment effect in a planned study. The standard error will be about $\sqrt{\frac{n^{existing}}{n^{planned}}} \cdot SE_{\widehat{\tau}}^{existing}$, where $n^{existing}$ is the number of respondents per condition in the existing study, $SE_{\widehat{\tau}}^{existing}$ is the estimated standard error in the existing study, and $n^{planned}$ is the number of respondents per condition in the planned study.

Rule 8: From an Existing Study to the Sample Size

We can also use an existing study to compute the sample size we need to obtain a desired power level. From Rule 7, we know that $SE_{\widehat{\tau}}^{planned} = \sqrt{\frac{n^{existing}}{n^{planned}}} \cdot \widehat{SE}_{\widehat{\tau}}^{planned}$. If we want 80% power to detect the effect $\widetilde{\tau}$ in our planned study, then we need $SE_{\widehat{\tau}}^{planned} = \frac{\widetilde{\tau}}{2.5}$ (see Rule 5). Plugging this into Rule 7 and solving for $n^{planned}$, we will need about $n^{existing} \cdot \left(\frac{2.5}{\widetilde{\tau}} \cdot \widehat{SE}_{\widehat{\tau}}^{existing}\right)^2$ respondents per condition. For 95% power, we will need about $n^{exisiting} \cdot \left(\frac{3.3}{\widetilde{\tau}} \cdot \widehat{SE}_{\widehat{\tau}}^{existing}\right)^2$ respondents per condition.

Rule 8 (Existing Study to Sample Size): We can use an existing, similar study to compute the sample size we will need in a planned study. For 80% power to detect the treatment effect $\widetilde{\tau}$, we will need about $n^{existing} \cdot \left[\frac{2.5}{\widetilde{\tau}} \cdot \widehat{SE}_{\widehat{\tau}}^{existing}\right]^2$ respondents per condition, where $n^{existing}$ is the number of respondents per condition in the existing study and $SE_{\widehat{\tau}}^{existing}$ is the estimated standard error in the existing study For 95% power, the factor changes from 2.5 to 3.3.

Rule 9: From Pilot Data to the SE

It is common to run a small pilot study prior to the experiment for reasons unrelated to statistical power.[23] For example, we might like to learn whether respondents can accurately remember certain details of a vignette. If we have pilot data available, we can also use these pilot data to predict the standard error in the full study. Suppose, for example, that we have tentatively planned to run a full study with 1,000 respondents. We might run a pilot study on 100 respondents to make sure that no issues arise. We can perform the planned analysis on the pilot data and use Rule 7 to predict the standard errors (and minimum detectable effect) in the full study.[24] For example, if the small pilot has a standard error of 2.0, then a good guess of the standard error in the full study is $\sqrt{\frac{100}{1{,}000}} \cdot 2.0 = 0.32 \cdot 2.0 = 0.63$.

However, translating the standard error from a small pilot study is meaningfully different from translating the standard error from another full study. Recall that the standard error estimate from the pilot data is only an *estimate.* And while the standard error estimate is usually accurate, pilot studies are typically small enough to worry about the noise in the estimate. With a noisy estimate of the standard error, we might happen to substantially under-estimate the standard error (and thus over-estimate power). To protect against running an under-powered study, I recommend predicting the standard error *conservatively* from pilot data.

To understand how severely you might under-estimate (or over-estimate) the standard error using Rule 7, recall the that standard error of a sample standard deviation is about $SE\,(\text{sample SD}) = \text{sample SD} \cdot \sqrt{\frac{1}{4 \cdot n}}$ so that the factor $\sqrt{\frac{1}{4 \cdot n}} \cdot 100\%$ gives us a typical error in the estimate of the standard error as a percentage. For $n = 50$ respondents per condition,

[23] Importantly, one should *not* use pilot data to estimate the treatment effect $\widehat{\tau}$. Any pilot study will not be sufficiently precise to provide a useful estimate of the treatment effect Leon, Davis, and Kraemer (2011); Albers and Lakens (2018).

[24] We should not pay attention to the point estimates in the pilot data; the point estimates are too noisy to be useful for any purpose.

the standard error estimate is typically off by about $\sqrt{\frac{1}{4 \cdot n}} \cdot 100\% = \sqrt{\frac{1}{4 \cdot 50}} \cdot 100\% = 7\%$ (either too high or too low), with errors larger than about 14% being unusual. To protect against an under-powered study, we can increase the predicted standard error by a factor of $2 \cdot \sqrt{\frac{1}{4 \cdot n}} + 1 = \sqrt{\frac{1}{n}} + 1$, which roughly corresponds to the upper bound of a 95% confidence interval for the standard error. Plugging this conservative standard error into Rule 7 above, we obtain $\sqrt{\frac{n^{pilot}}{n^{planned}}} \left[\left(\sqrt{\frac{1}{n^{pilot}}} + 1 \right) \cdot \widehat{SE}_{\widehat{\tau}}^{pilot} \right]$ as a *conservative* estimate of the standard error.

Rule 9 (Pilot Data to SE): We can use pilot data to predict the standard error of the estimated treatment effect in a planned study. Conservatively, the standard error will be about $\sqrt{\frac{n^{pilot}}{n^{planned}}} \left[\left(\sqrt{\frac{1}{n^{pilot}}} + 1 \right) \cdot \widehat{SE}_{\widehat{\tau}}^{pilot} \right]$, where n^{pilot} is the number of respondents per condition in the pilot data, $SE_{\widehat{\tau}}^{pilot}$ is the estimated standard error using the pilot data, and $n^{planned}$ is the number of respondents per condition in the planned study.

Rule 10: From Pilot Data to the Sample Size

Perhaps most importantly, we can use the logic of Rule 8 to make adjustments to the sample size of the planned study. Suppose that we want to design an experiment with 80% power to detect the effect $\widetilde{\tau}$. By Rule 5, we need the standard error to be $\frac{\widetilde{\tau}}{2.5}$. Using Rule 9, we can conservatively predict that the standard error will be $\sqrt{\frac{n^{pilot}}{n^{planned}}} \left[\left(\sqrt{\frac{1}{n^{pilot}}} + 1 \right) \cdot \widehat{SE}_{\widehat{\tau}}^{pilot} \right]$. Setting this equal to $\frac{\widetilde{\tau}}{2.5}$ and solving for $n^{planned}$, we can conservatively predict that $n^{pilot} \cdot \left[\frac{2.5}{\widetilde{\tau}} \cdot \left(\sqrt{\frac{1}{n^{pilot}}} + 1 \right) \cdot \widehat{SE}_{\widehat{\tau}}^{pilot} \right]^2$ respondents per condition will give us 80% power to detect the effect $\widetilde{\tau}$.

Rule 10 (Pilot Data to Sample Size): We can use pilot data to conservatively predict the sample size we will need in a planned study. For 80% power to detect the treatment effect $\widetilde{\tau}$, we will (conservatively) need about $n^{pilot} \cdot \left[\frac{2.5}{\widetilde{\tau}} \cdot \left(\sqrt{\frac{1}{n^{pilot}}} + 1 \right) \cdot \widehat{SE}_{\widehat{\tau}}^{pilot} \right]^2$ respondents per condition, where n^{pilot} is the number of respondents per condition in the pilot data

and $SE^{pilot}_{\widehat{\tau}}$ is the estimated standard error using the pilot data. For 95% power, the factor changes from 2.5 to 3.3.

More Examples: Finding the Sample Size

In the examples above, we imagined a research context in which the sample size was not under our control. But any many contexts, we can control (or must choose) the sample size. Continuing our example of Ahler and Sood's (2018) experiment from above, we can use features of a reference population, similar existing studies, and pilot data to predict the required sample size. For the Ahler and Sood example, we have all three available. For simplicity, suppose that we would like 95% power to detect a treatment effect of 3 points on the 101-point scale.[25]

Features of a Reference Population

Above, we use the ANES to motivate setting $\widetilde{SD(Y)}$ to 20.8 and $\widetilde{R^2}$ to 5%. Rule 6 tells us that, without control variables, we will need about $2 \cdot \left(\frac{3.3 \cdot \widetilde{SD(Y)}}{\widetilde{\tau}}\right)^2 = 2 \cdot \left(\frac{3.3 \cdot 20.8}{3}\right)^2 = 1,050$ respondents per condition for 95% power. With control variables that predict the outcome with an R^2 of 5%, this becomes $2 \cdot \left(\frac{3.3 \cdot \widetilde{SD(Y)} \cdot \sqrt{\left(1-\widetilde{R^2}\right)}}{\widetilde{\tau}}\right)^2 = 2 \cdot \left(\frac{3.3 \cdot 20.8 \cdot \sqrt{(1-0.05)}}{3}\right)^2 = 995.$

An Existing Study: Ahler and Sood (2018)

Ahler and Sood (2018) have about 268 respondents per condition with a standard error of about 1.8 on the 101-point scale. If we mimic their design, then we can use Rule 8 to predict that we will need $n^{existing} \cdot \left[\frac{3.3}{\widetilde{\tau}} \cdot \widehat{SE}^{existing}_{\widehat{\tau}}\right]^2 = 268 \cdot \left[\frac{3.3}{3} \cdot 1.8\right]^2 = 1,051$ respondents per condition for 95% power to detect a treatment effect of 3 points on the 101-point scale.

[25] A 3-point effect on the 101-point scale is the lower bound of Ahler and Sood's 95% confidence interval.

An Existing Study: Broockman, Kalla, and Westwood (2022)

Broockman, Kalla, and Westwood (2022) have about 502 respondents per condition with a standard error of about 1.67 on the 101-point scale. If we mimic their design, then we can use Rule 8 to predict that we will need $n^{existing} \cdot \left[\frac{3.3}{\widetilde{\tau}} \cdot \widehat{SE}_{\widehat{\tau}}^{existing}\right]^2 = 502 \cdot \left[\frac{3.3}{3} \cdot 1.7\right]^2 = 1,694$ respondents per condition for 95% power to detect a treatment effect of 3 points on the 101-point scale.[26]

A Pilot Study: Culter, Pietryka, and Rainey (2024)

Culter, Pietryka, and Rainey (2024) conduct a pilot study with about 85 respondents per condition. Their study uses the pre-post strategy suggested by Clifford, Sheagley, and Piston (2021) and they include the pre-treatment measure as a control variable. They analyze the small pilot data set as they plan to analyze the full data set and find a standard error of 2.13. Using Rule 10, they will need (conservatively) about $n^{pilot} \cdot \left[\frac{3.3}{\widetilde{\tau}} \cdot \left(\sqrt{\frac{1}{n^{pilot}}} + 1\right) \cdot \widehat{SE}_{\widehat{\tau}}^{pilot}\right]^2 = 85 \cdot \left[\frac{3.3}{3} \cdot \left(\sqrt{\frac{1}{85}} + 1\right) \cdot 2.13\right]^2 = 573$ respondents per condition for 95% power to detect a treatment effect of 3 points on the 101-point scale. Cutler, Pietryka, and Rainey's approach requires many fewer observation because their pre-treatment measure is highly predictive of the outcome.

Rule 11: From 80% Power to 95% Power

In thinking about the tradeoff between 80% and 95% power it can be helpful to have a rule connecting sample size requirements. Recall from Rule 7 that $\frac{SE_{\widehat{\tau}}^A}{SE_{\widehat{\tau}}^B} = \sqrt{\frac{n^B}{n^A}}$. For an assumed treatment effect $\widetilde{\tau}$, the standard errors that yield 80% and 95% power are given by $\frac{\widetilde{\tau}}{2.5}$ and $\frac{\widetilde{\tau}}{3.3}$,

[26] In Broockman, Kalla, and Westwood's data, the standard deviation of the outcome is about 27, which is much larger than the standard deviation of the similar measure in the ANES (about 21) and the standard deviation of Ahler and Sood's measure (about 22). This highlights the importance of triangulating power calculations using multiple sources to motivate the assumptions.

respectively. Then we have $\frac{SE^A_{\widehat{\tau}}}{SE^B_{\widehat{\tau}}} = \frac{\frac{\tilde{\tau}}{2.5}}{\frac{\tilde{\tau}}{3.3}} = \frac{3.3}{2.5} = \sqrt{\frac{n^B}{n^A}}$. Then we can see that $\frac{n^B}{n^A} = \left(\frac{3.3}{2.5}\right)^2 = 1.74$. This means that to increase the power from 80% to 95%, we need increase our sample size by about 74%. With less rounding error, this becomes 75%.

Rule 11 (80% to 95% Power): Increasing the sample size by 75% will increase the power from 80% to about 95%.

How Are Interactions Different?

Up to this point in the paper, I have focused on a hypothesis about the treatment effect (i.e., "average treatment effect" or ATE). But what if we have a hypothesis about an *interaction* (i.e., a difference in treatment effects across two scenarios)? The general logic of the rules extends to the interaction, with two minor modifications.

Assume a 2×2 factorial design with two treatments. Then we have potential outcomes $\mathcal{Y}_i(1,1)$ if assigned to both treatments, $\mathcal{Y}_i(1,0)$ if assigned to the first treatment, $\mathcal{Y}_i(0,1)$ if assigned to the second treatment, and $\mathcal{Y}_i(0,0)$ if assigned neither treatment for respondent i of N. Then we can define the interaction as $\delta = \left[\frac{1}{N}\sum_{i=1}^{N}\mathcal{Y}_i(1,1) - \frac{1}{N}\sum_{i=1}^{N}\mathcal{Y}_i(0,1)\right] - \left[\frac{1}{N}\sum_{i=1}^{N}\mathcal{Y}_i(1,0) - \frac{1}{N}\sum_{i=1}^{N}\mathcal{Y}_i(0,0)\right]$.

Imagine we assign n respondents to each of the four conditions. We create two indicator variables D_1 and D_2 that equal 1 if the respondent is assigned to the first and second treatment, respectively, and 0 otherwise. Then we create the observed outcome $Y = \mathcal{Y}(D_1, D_2)$. Then we estimate the interaction in practice using least squares estimates of the model $Y = \beta_0 + \beta_1 \cdot D_1 + \beta_2 \cdot D_2 + \delta \cdot D_1 \cdot D_2 + \epsilon$. The parameter δ is the interaction. Inference about this parameter is the same as described above (least squares regression and robust standard errors; inferences about δ using the 90% confidence interval). By assuming $\epsilon_i \sim N(0, \sigma^2)$, the classical variance for the least squares estimates of the parameter vector $\beta = [\beta_0, \beta_1, \beta_2, \delta]'$ is $Var\left(\widehat{\beta}\right) = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\sigma^2$ for the design matrix $\mathbf{X} = [1, D_1, D_2, D_1 \cdot D_2]$. The experimental

design requires n repetitions of each of the four unique rows of $\mathbf{X}$, which allows us to simplify $\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ and find that the bottom right element of $\left(\mathbf{X}'\mathbf{X}\right)^{-1}\sigma^2$—the variance of $\widehat{\delta}$—equals $\frac{4}{n}$. This gives us $SE_{\widehat{\delta}}^{classic} = \sqrt{\frac{4\cdot\sigma^2}{n}} = \frac{2\cdot\sigma}{\sqrt{n}}$, where n represents the number of respondents assigned to each condition.[27]

The standard error for an interaction has a different relationship to the standard deviation, so two of the rules above change for interactions.

- **A Modification to Rules 3 and 6.** Rules 3 and 6 help us use features of a reference population to predict the standard error for the estimate of a treatment effect and approximate the required sample size, respectively. When estimating an interaction rather than a treatment effect, we need to adjust these approximations.
  - First, Rule 3 requires adjustment. As explained above, the standard error for the estimated interaction is about $\frac{2\cdot\widetilde{SD(Y)}}{\sqrt{n}}$ rather than $\frac{2\cdot\widetilde{SD(Y)}}{\sqrt{2\cdot n}}$ for the treatment effect.
  - Second, Rule 6 requires adjustment. Suppose we want our experiment to have 80% power to detect the interaction $\widetilde{\delta}$. Then we need $\widetilde{\delta} = 2.5 \cdot SE = 2.5 \cdot \frac{2\cdot\widetilde{SD(Y)}}{\sqrt{n}}$. Solving for n, we find that $n = 4 \cdot \left(\frac{2.5\cdot\widetilde{SD(Y)}}{\widetilde{\delta}}\right)^2$. If we are adjusting for control variables, we have $4 \cdot \left(\frac{2.5\cdot\widetilde{SD(Y)}\cdot\sqrt{\left(1-\widetilde{R^2}\right)}}{\widetilde{\delta}}\right)^2$. Notice that the first factor changes from 2 in Rule 3 to 4 for interaction.
- **A Modification to Rules 9 and 10.** Second, Rules 9 and 10 require a small adjustment to account for the change from $N = 2 \cdot n$ in the treatment effect context to $N = 4 \cdot n$ in the interaction context—the standard error will be more accurate with a larger total sample size. The adjustment factor changes from $\left(\sqrt{\frac{1}{n^{pilot}}} + 1\right)$ in the

[27] Alternatively, we can think of the estimate of the interaction effect as the difference in differences of sample means $\widehat{\delta} = \left[\overline{Y}_{11} - \overline{Y}_{01}\right] - \left[\overline{Y}_{10} - \overline{Y}_{00}\right]$, where $\overline{Y}_{11}$ represents the sample average for respondents where $D_1 = 1$ and $D_2 = 1$ and so on. Assuming independence of the sample means, the variance of these differences is the sum of the variances of each term, so that $Var(\widehat{\delta}) = \frac{\sigma^2}{n} + \frac{\sigma^2}{n} + \frac{\sigma^2}{n} + \frac{\sigma^2}{n} = \frac{4\cdot\sigma^2}{n}$. This also give us $SE_{\widehat{\delta}}^{classic} = \sqrt{\frac{4\cdot\sigma^2}{n}} = \frac{2\cdot\sigma}{\sqrt{n}}$. This motivation that Gelman uses in his widely-shared blog post (accessed May 3, 2024; persistent link: https://archive.md/FOTo7).

treatment effect context to $\overbrace{\sqrt{\frac{1}{2 \cdot n^{pilot}}}}^{\text{modified}} + 1$ in the interaction context.

Comparing the standard error for the interaction to the standard error for the treatment effect leads to three important observations. First, for the treatment effect, we had $SE_{\widehat{\tau}}^{classic} = \sqrt{\frac{\sigma^2}{2 \cdot n \cdot \frac{1}{4}}} = \frac{2 \cdot \sigma}{\sqrt{2 \cdot n}} = \frac{1}{\sqrt{2}} \cdot \frac{2 \cdot \sigma}{\sqrt{n}}$; thus the standard error for the interaction is $\sqrt{2}$ times larger (about 41% larger) than the standard error for the treatment effect. In order to make the standard errors the same, we need to double the sample size *in each condition*. Second, the 2×2 factorial design has twice as many conditions and thus requires twice the sample size in total. Third, the interaction is likely much smaller than the treatment effect. In a widely shared blog post, Andrew Gelman famously suggests that "you need 16 times the sample size to estimate an interaction than to estimate a main effect" (accessed May 3, 2024; persistent link: https://archive.md/FOTo7). While this rule is too crude to apply generally—it makes strong assumptions about the treatment effect of interest—it does starkly highlight the demands of estimating an interaction relative to the treatment effect. It can help our intuition to partition Gelman's 16x claim into its constituent parts: you need twice the number of respondents per condition for a similarly precise estimate, twice the number of conditions, and four times the number of respondents when the parameter is half the size.

Example: Interactions

Robbins et al. (2024) study how respondents evaluate in-party legislators when they criticize an out-party president following a covert operation. They hypothesize that Republicans, for example, will report higher approval of Republican legislators that are more critical of the Democratic president, but that this effect will be much larger when the operation fails than when the operation succeeds. To test their hypothesis, they design a 2×2 factorial experiment. In each condition, a vignette describes (1) a covert operation that succeeds or

fails and (2) legislators that aggressively criticize or ignore the operation.

Informed by Myrick (2020), the authors expected that criticism would increase approval by about 0.33 points on a seven-point scale when the operation succeeded, but by about 1.00 points when the operation failed. This implies an interaction of about 0.67 points.

Myrick (2020) conducted a similar experiment and measured her outcomes similarly. After analyzing her data closely, the authors expected their seven-point approval outcome would have a standard deviation of about 2.0. Using Rule 3 (and modifying for interaction), the authors would need about $4 \cdot \left(\frac{2.5 \cdot 2.0}{0.67}\right)^2 = 223$ respondents per condition (or 892 respondents in total) for 80% power to detect an interaction of 0.67. For 95% power, the authors would need about $4 \cdot \left(\frac{3.3 \cdot 2.0}{0.67}\right)^2 = 389$ respondents per condition (or 1,556 respondents in total).

The authors also completed a pilot study with 75 respondents per condition. They conducted their planned analysis on the pilot data and estimated a standard error of 0.40 for the interaction. Using Rule 10, conservatively, the authors would need about $n^{pilot} \cdot \left[\frac{2.5}{\widetilde{\tau}} \cdot \left(\sqrt{\frac{1}{2 \cdot n^{pilot}}} + 1\right) \cdot \widehat{SE}_{\widehat{\tau}}^{pilot}\right]^2 = 75 \cdot \left[\frac{2.5}{0.67} \cdot \left(\sqrt{\frac{1}{2 \cdot 75}} + 1\right) \cdot 0.40\right]^2 = 195$ respondents per condition (or 780 respondents total) to have 80% power to detect an interaction of 0.67. For 95% power, they would need about $75 \cdot \left[\frac{3.3}{0.67} \cdot \left(\sqrt{\frac{1}{2 \cdot 75}} + 1\right) \cdot 0.40\right]^2 = 340$ respondents per condition (or 1,360 respondents total) to have 95% power.[28]

How to Use the Rules

Table 1 shows how you can use the rules to establish that your planned experiment is well-powered. Broadly, there are three strategies, or ways to arrange our thinking:

1. For a given sample size and statistical power, show that the minimum detectable effect

[28] Robbins et al. ultimately used 375 respondents per condition (or 1,500 respondents total). Using Rule 9, the authors would conservatively predict the standard error would be about $\sqrt{\frac{n^{pilot}}{n^{planned}}} \cdot \widehat{SE}_{\widehat{\delta}}^{pilot} = \sqrt{\frac{75}{375}} \cdot 0.40 = 0.19$. In the full, 1,500-respondent study, the standard error was 0.18, which suggests a statistical power of about $1 - \Phi_{std}\left(1.64 - \frac{\widetilde{\tau}}{SE}\right) = 1 - \Phi_{std}\left(1.64 - \frac{0.67}{0.18}\right) = 98\%$ to detect an interaction of 0.67.

corresponds to a "small" effect, either the smallest substantively meaningful effect or the smallest plausible effect given the available evidence.

2. For given sample size and effect of interest, show that the statistical power of the experiment is greater than a chosen threshold (e.g., 80% or 95%).
3. For a given statistical power and effect of interest, show that the sample size meets (or exceeds) the required sample size.

However, we should not draw strong distinctions between these strategies—the strategies differ only in the order in which the sample size, effect of interest, and statistical power are chosen. In the end, it simply matters that the three components combine to form a well-powered experiment.

To make each of these arguments, we can gather data from three potential sources, ideally relying on multiple.

1. *Features of a reference population.* We can find studies that measure our outcome (and perhaps control variables) on a similar population. We can use the standard deviation and R^2 in this reference population to predict the standard deviation and R^2 in our planned study.
2. *Existing similar study.* We can find an existing similar study. We can adjust the standard error estimates in the existing study to predict the standard error in our planned study. Importantly, the existing study should have a similar population, design, and outcome measure. The sample size, treatment, and treatment effect can differ; otherwise the studies should be similar.
3. *Pilot data.* We can use pilot data. We can adjust the standard error estimates from the pilot data to predict the standard error in our planned study. Because the standard error estimates in the pilot data will be quite noisy, we can make a conservative adjustment.

Table 1 describes how we might use these various strategies.

Table 1: This table shows how we can use the rules to answer questions about the experiment using features of a reference population, existing studies, and pilot data.

Goal	**Source**	**Estimation Strategy**	**Required Information**	**Rule Sequence**
I know my sample size and the statistical power I want. What is my minimum detectable effect?	features of reference population	no control variables	$SD(Y)$ in reference population	Rule 3 → Rule 5
		control variables	$SD(Y)$ and R^2 of control variables in reference population	Rule 3 → Rule 4 → Rule 5
	existing similar study	—	estimated standard error from existing study	Rule 7 → Rule 5
	pilot data	—	estimated standard error from pilot data	Rule 9 → Rule 5
I know my sample size and the effect of interest. What is my statistical power?	features of reference population	no control variables	$SD(Y)$ in reference population	Rule 3 → Rule 2
		control variables	$SD(Y)$ and R^2 of control variables in reference population	Rule 3 → Rule 4 → Rule 2
	existing similar study	—	estimated standard error from existing study	Rule 7 → Rule 2
	pilot data	—	estimated standard error from pilot data	Rule 9 → Rule 2
I know the statistical power I want and the effect of interest. What sample size do I need?	features of reference population	no control variables	$SD(Y)$ in reference population	Rule 6
		control variables	$SD(Y)$ and R^2 of control variables in reference population	Rule 6
	existing similar study	—	estimated standard error from existing study	Rule 8
	pilot data	—	estimated standard error from pilot data	Rule 10

Extensions

What if I want power other than 80% or 95%?

Throughout the paper, I discuss rules to obtain the minimum detectable effect and required sample size for 80% or 95% power. The rules for 80% power have a factor of 2.5 that can be changed to 3.3 for 95% power. These factors are the minimum detectable effects, in standard errors (i.e., see Rule 5). What about other power levels? Table 2 provides the minimum detectable effect, in standard errors, for a variety of power levels.[29] We can use these factors rather 2.5 for 80% power (or 3.3 for 95% power) in the relevant rules. Table 2 also provides the percent change in sample size to change the power from 80% to various other levels (i.e., see Rule 11).

Table 2: This table shows the minimum detectable effect, in standard errors, for a variety of power levels. We can use these values to find the minimum detectable effect for power levels other than 80% and 95% (i.e., as in Rule 5). We can also use these factors rather 2.5 for 80% power (or 3.3 for 95% power) in the relevant rules. For convenience, this table also shows the percent change in sample size required to change the statistical power from 80% to the various levels.

Power	**Minimum Detectable Effect (in Standard Errors)**	**Percent Change in Sample Size from 80% Power**
20%	0.8	-90%
40%	1.4	-69%
60%	1.9	-42%
80%	**2.5**	**0%**
90%	2.9	39%
95%	**3.3**	**75%**
98%	3.7	121%
99%	4.0	155%
99.9%	4.7	263%

[29] I round these minimum detectable effects to the nearest tenth. For greater precision, we can use `pnorm(0.95) + pnorm(0.80)`, where `0.80` corresponds to 80% power and can be changed to the desired power level.

What if I want to use a 95% confidence interval?

Checking that a 90% confidence interval contains only values consistent with the research hypothesis corresponds to a size-0.05 test. However, some researchers prefer to use a 95% confidence interval, which corresponds to size-0.025 test for directional claims. Table 3 provides the minimum detectable effects for researchers using a 95% confidence interval.[30] We can use these factors rather 2.5 for 80% power in the relevant rules.

Table 3: This table is an alternate version of Table 2 for researchers who wish to use a 95% confidence interval rather than a 90% confidence interval. The table shows the minimum detectable effect, in standard errors, for a variety of power levels. Researchers can use these values to find the minimum detectable effect for power levels other than 80% and 95% (i.e., as in Rule 5. Researchers can also use these factors rather 2.5 for 80% power (or 3.3 for 95% power) in the relevant rules. For convenience, this table also shows the percent change in sample size required to change the statistical power from 80% to the various levels.

Power	**Minimum Detectable Effect (in Standard Errors)**	**Percent Change in Sample Size from 80% Power**
20%	1.1	-84%
40%	1.7	-63%
60%	2.2	-38%
80%	2.8	0%
90%	3.2	34%
95%	3.6	66%
98%	4.0	105%
99%	4.3	134%
99.9%	5.0	225%

To achieve the same statistical power with a 95% confidence interval (which corresponds to a size-0.025 test) compared to a 90% confidence interval (which corresponds to a size-0.05 test), researchers need to increase their sample size by about 27% for 80% power and 20% for 95% power.

For example, I noted above that Ahler and Sood (2018) have about 268 respondents per

[30] I round these minimum detectable effects to the nearest tenth. For greater precision, we could use `pnorm(0.975) + pnorm(0.80)`, where `0.80` corresponds to 80% power and can be changed to the desired power level.

condition with a standard error of about 1.8 on the 101-point scale. Suppose researchers want 80% power to detect a treatment effect of 3 points on the 101-point scale with a 95% confidence interval. Using Rule 8 modified for a 95% confidence interval, researchers would need about $n^{existing} \cdot \left[\frac{2.8}{\widehat{\tau}} \cdot \widehat{SE}_{\widehat{\tau}}^{existing}\right]^2 = 268 \cdot \left[\frac{2.8}{3} \cdot 1.8\right]^2 = 756$ respondents per condition (note the 2.8 in place of the 2.5; using a 90% confidence interval would require about 555 respondents per condition).

Related Work

Cohen (1988) provides the foundational introduction to power analysis, and Cohen (1990) offers an engaging personal and historical perspective. DeGroot and Schervish (2010, ch. 9) and Casella and Berger (2002, ch. 8) provide a through technical discussion. Greenland et al. (2016) offers an accessible and intuitive discussion organized around potential *mis*understandings of the concepts. Bloom (1995) offers a particularly practical and intuitive approach using the minimum detectable effect. Cohen (1992), Lenth (2001), and Meyvis and Van Osselaer (2017) also offer practical guides. Lakens (2022) offers a broader perspective on sample size justification beyond statistical power. Blair et al. (2019) and Blair, Coppock, and Humphreys (2023) offer a useful conceptual framework and software for comprehensive evaluation of research designs.

Conclusion

To address the problem of low power in political science research (Arel-Bundock et al. 2022) and increasing requirements that articles, registered reports, and grant applications include a power analysis, I offer several useful rules that help political scientists reason about statistical power. For simple randomized experiments to estimate treatment effects and/or interactions,

these rules might be sufficient to justify a sample size. For more complex designs, these rules can serve as a starting point for more complex simulations.

Perhaps most importantly, these rules help develop intuitions about statistical power. For example, these rules can help us quickly appreciate the tradeoffs between spreading survey respondents across eight conditions rather than four, or between using fewer respondents through a more expensive provider or more respondents through less expensive provider.

In this paper, I focused mostly on the choice of sample size, but there are more (and sometimes better!) ways to increase statistical power. We must get the ratio of the assumed treatment effect and the standard error above 2.5 and ideally above 3.3. Rule 5 makes the task clear—we need a treatment with oomph to make the numerator large and strategies like regression adjustment, precise measurement instruments, attentive respondents, and large samples to make the standard error small.[31] And keep Rule 5 top of mind. I hope these rules help us think carefully about these efforts.

[31]While framed as defending null results, Kane (2024) describes several strategies to increase treatment effects and decrease the standard error.

References

Ahler, Douglas J., and Gaurav Sood. 2018. "The Parties in Our Heads: Misperceptions about Party Composition and Their Consequences." *The Journal of Politics* 80(3):964–981.

Albers, Casper, and Daniël Lakens. 2018. "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.

Alrababa'h, Ala' et al. 2023. "Learning from Null Effects: A Bottom-Up Approach." *Political Analysis* 31(3):448–456.

Arel-Bundock, Vincent, Ryan C Briggs, Hristos Doucouliagos, Marco M Aviña, and T.D. Stanley. 2022. "Quantitative Political Science Research is Greatly Underpowered.".

Blair, Graeme, Alexander Coppock, and Macartan Humphreys. 2023. *Research Design in the Social Sciences: Declaration, Diagnosis, and Redesign.* Princeton: Princeton University Press.

Blair, Graeme, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. 2019. "Declaring and Diagnosing Research Designs." *American Political Science Review* 113(3):838–859.

Bloom, Howard S. 1995. "Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs." *Evaluation Review* 19(5):547–556.

Boldt, Damian, Winston Lin, and Carlisle Rainey. 2024. "Regression Adjustment in Survey Experiments: A Practical Perspective." Manuscript in preparation.

Bowman, Sara D, Alexander C DeHaven, Timothy M Errington, Tom E Hardwicke, David T Mellor, Brian A Nosek, and Courtney K Soderberg. 2020. "OSF Prereg Template.".

Broockman, David E., Joshua L. Kalla, and Sean J. Westwood. 2022. "Does Affective Polarization Undermine Democratic Norms or Accountability? Maybe Not." *American Journal of Political Science* 67(3):808–828.

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

Casella, George, and Roger L. Berger. 2002. *Statistical Inference.* 2nd ed. Pacific Grove, CA: Duxbury.

Clifford, Scott, Geoffrey Sheagley, and Spencer Piston. 2021. "Increasing Precision without Altering Treatment Effects: Repeated Measures Designs in Survey Experiments." *American Political Science Review* 115(3):1048–1065.

Cohen, Jacob. 1962. "The statistical power of abnormal-social psychological research: A review." *The Journal of Abnormal and Social Psychology* 65(3):145–153.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences.* 2 ed. Hillside, NJ: Lawrence Erlbaum Associates.

Cohen, Jacob. 1990. "Things I have learned (so far)." *American Psychologist* 45:1304–1312.

Cohen, Jacob. 1992. "A power primer." *Psychological Bulletin* 112(1):155–159.

Cox, David R, and Peter McCullagh. 1982. "A biometrics invited paper with discussion. some aspects of analysis of covariance." *Biometrics* pp. 541–561.

Culter, Austin Lloyd, Matthew Pietryka, and Carlisle Rainey. 2024. "Merely Asking: A Replication of Ahler and Sood (2018)." Manuscript in preparation.

DeGroot, Morris H, and Mark J Schervish. 2010. *Probability and Statistics*. 4 ed. Upper Saddle River, NJ: Pearson.

Druckman, James N., and Donald P. Green. 2021. "A New Era of Experimental Political Science." In *Advances in Experimental Political Science*, ed. James N. Druckman, and Donald P. Green. Cambridge: Cambridge University Press pp. 1–16.

Faul, Franz, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. "G* Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences." *Behavior Research Methods* 39(2):175–191.

Faul, Franz, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. "Statistical Power Analyses Using G* Power 3.1: Tests for Correlation and Regression Analyses." *Behavior Research Methods* 41(4):1149–1160.

Gelman, Andrew, and John Carlin. 2014. "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors." *Perspectives on Psychological Science* 9(6):641–651.

Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: W. W. Norton & Company.

Greenland, Sander, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman. 2016. "Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations." *European Journal of Epidemiology* 31(4):337–350.

Hoenig, John M, and Dennis M Heisey. 2001. "The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis." *The American Statistician* 55(1):19–24.

Ioannidis, John P. A., T. D. Stanley, and Hristos Doucouliagos. 2017. "The Power of Bias in Economics Research." *The Economic Journal* 127(605):F236–F265.

Kane, John V. 2024. "More than meets the ITT: A guide for anticipating and investigating nonsignificant results in survey experiments." *Journal of Experimental Political Science* p. 1–16.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44(2):347.

Kirk, Roger E. 1996. "Practical Significance: A Concept Whose Time Has Come." *Educational and Psychological Measurement* 56(5):746–759.

Lakens, Daniël. 2014. "Performing high-powered studies efficiently with sequential analyses." *European Journal of Social Psychology* 44(7):701–710.

Lakens, Daniël. 2022. "Sample Size Justification." *Collabra: Psychology* 8(1).

Lakens, Daniël, Anne M. Scheel, and Peder M. Isager. 2018. "Equivalence Testing for Psychological Research: A Tutorial." *Advances in Methods and Practices in Psychological Science* 1(2):259–269.

Lenth, Russell V. 2001. "Some Practical Guidelines for Effective Sample Size Determination." *The American Statistician* 55(3):187–193.

Leon, Andrew C., Lori L. Davis, and Helena C. Kraemer. 2011. "The role and interpretation of pilot studies in clinical research." *Journal of Psychiatric Research* 45(5):626–629.

Li, Xinran, and Peng Ding. 2017. "General Forms of Finite Population Central Limit Theorems with Applications to Causal Inference." *Journal of the American Statistical Association* 112(520):1759–1769.

Lovakov, Andrey, and Elena R. Agadullina. 2021. "Empirically derived guidelines for effect size interpretation in social psychology." *European Journal of Social Psychology* 51(3):485–504.

McCaskey, Kelly, and Carlisle Rainey. 2015. "Substantive Importance and the Veil of Statistical Significance." *Statistics, Politics and Policy* 6(1–2).

Meyvis, Tom, and Stijn M J Van Osselaer. 2017. "Increasing the Power of Your Study by Increasing the Effect Size." *Journal of Consumer Research* 44(5):1157–1173.

Myrick, Rachel. 2020. "Why So Secretive? Unpacking Public Attitudes toward Secrecy and Success in US Foreign Policy." *The Journal of Politics* 82(3):828–843.

Neyman, Jerzy. 1990. "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles (with Discussion). Section 9 (Translated)." *Statistical Science* 5:465–472.

Perugini, Marco, Marcello Gallucci, and Giulio Costantini. 2014. "Safeguard Power as a Protection Against Imprecise Power Estimates." *Perspectives on Psychological Science* 9(3):319–332.

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

Robbins, Caroline, Alessandro Brunelli, Jose Casto, Ainsley Coty, Andrew Louis, Bryanna Major, Maria A Martinez, Yadianis L Ojeda, Larissa Pontes, Elke Schumacher, Omer Turkomer, Luzmi Valenzuela, Valeria Veras, and Carlisle Rainey. 2024. "Overt Consequences of Covert Actions: Success, Failure, and Voters' Preferences for Legislative Oversight." osf.io/preprints/socarxiv/9p5h8.

Samii, Cyrus, and Peter M. Aronow. 2012. "On equivalencies between design-based and regression-based variance estimators for randomized experiments." *Statistics and Probability Letters* 82(2):365–370.

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

Stommes, Drew, P. M. Aronow, and Fredrik Sävje. 2023. "On the reliability of published findings using the regression discontinuity design in political science." *Research and Politics* 10(2):205316802311664.

Wan, Xiang, Wenqian Wang, Jiming Liu, and Tiejun Tong. 2014. "Estimating the Sample Mean and Standard Deviation from the Sample Size, Median, Range and/or Interquartile Range." *BMC Medical Research Methodology* 14(1).

Yang, Yefeng, Alfredo Sánchez-Tójar, Rose E. O'Dea, Daniel W. A. Noble, Julia Koricheva, Michael D. Jennions, Timothy H. Parker, Malgorzata Lagisz, and Shinichi Nakagawa. 2023. "Publication bias impacts on effect size, statistical power, and magnitude (Type M) and sign (Type S) errors in ecology and evolutionary biology." *BMC Biology* 21(1).

A The Rules

Rule 1 (Power; Most Intuitively): Power equals $1 - \Phi(1.64 \cdot SE; \mu = \tau, \sigma = SE)$, where $\Phi(z; \mu, \sigma)$ is the normal CDF, SE is the standard error of the estimated treatment effect, and τ is the treatment effect.

Rule 2 (Power; From a z Table): Power equals $1 - \Phi_{std}\left(1.64 - \frac{\tau}{SE}\right)$, where $\Phi_{std}(z)$ is the standard normal CDF (as found in a standard z table), SE is the standard error of the estimated treatment effect, and τ is the treatment effect.

Rule 3 (SD to SE): We can use features of a reference population to predict the standard error of the estimated treatment effect in a planned study. The standard error will be about $\frac{2 \cdot \widetilde{SD(Y)}}{\sqrt{2 \cdot n}}$, where n is the sample size per condition and $\widetilde{SD(Y)}$ is the standard deviation of the outcome in a reference population.

Note: Rule 3 must be modified to handle interaction. See the section "How Are Interactions Different?"

Rule 4 (Adjustment): We can use features of a reference population to predict how much regression adjustment using pre-treatment control variables will shrink the standard error of an unadjusted estimate in a planned study. Regression adjustment will shrink the standard error of the unadjusted estimate by about $\left[1 - \sqrt{\left(1 - \widetilde{R^2}\right)}\right] \cdot 100\%$, where $\widetilde{R^2}$ is the R^2 of a regression of the outcome on the control variables in a reference population.

Rule 5 (Bloom's Rule; SE to MDE): An experiment has 80% power to detect a treatment effect that is 2.5 times the standard error and 95% power to detect a treatment effect that is 3.3 times the standard error.

Note: Rule 5 can be modified to compute the minimum detectable effect for power other than 80% and 95%. See Table 2. It can also be modified for researchers using 95% confidence intervals rather than 90% confidence intervals as I recommend. See Table 3.

Rule 6 (SD to Sample Size): We can use features of a reference population to compute the sample size we will need in a planned study. For 80% power to detect the treatment effect $\widetilde{\tau}$, the sample size per condition will need to be about $2 \cdot \left(\frac{2.5 \cdot \widetilde{SD(Y)}}{\widetilde{\tau}}\right)^2$ (or $2 \cdot \left(\frac{2.5 \cdot \widetilde{SD(Y)} \cdot \sqrt{\left(1-\widetilde{R^2}\right)}}{\widetilde{\tau}}\right)^2$ if including control variables), where $\widetilde{SD(Y)}$ is the standard deviation of the outcome in a reference population and $\widetilde{R^2}$ is the R^2 of a regression of the outcome on the control variables in a reference population. For 95% power, the factor changes from 2.5 to 3.3.

Note: Rule 6 must be modified to handle interaction. See the section "How Are Interactions Different?"

Rule 7 (Existing Study to SE): We can use an existing, similar study to predict the standard error of the estimated treatment effect in a planned study. The standard error will be about $\sqrt{\frac{n^{existing}}{n^{planned}}} \cdot SE_{\widehat{\tau}}^{existing}$, where $n^{existing}$ is the number of respondents per condition in the existing study, $SE_{\widehat{\tau}}^{existing}$ is the estimated standard error in the existing study, and $n^{planned}$ is the number of respondents per condition in the planned study.

Rule 8 (Existing Study to Sample Size): We can use an existing, similar study to compute the sample size we will need in a planned study. For 80% power to detect the treatment effect $\widetilde{\tau}$, we will need about $n^{existing} \cdot \left[\frac{2.5}{\widetilde{\tau}} \cdot \widehat{SE}_{\widehat{\tau}}^{existing}\right]^2$ respondents per condition, where $n^{existing}$ is the number of respondents per condition in the existing study and $SE_{\widehat{\tau}}^{existing}$ is the estimated standard error in the existing study For 95% power, the factor changes from 2.5 to 3.3.

Rule 9 (Pilot Data to SE): We can use pilot data to predict the standard error of the estimated treatment effect in a planned study. Conservatively, the standard error will be about $\sqrt{\frac{n^{pilot}}{n^{planned}}}\left[\left(\sqrt{\frac{1}{n^{pilot}}}+1\right)\cdot\widehat{SE}_{\widehat{\tau}}^{pilot}\right]$, where n^{pilot} is the number of respondents per condition in the pilot data, $SE_{\widehat{\tau}}^{pilot}$ is the estimated standard error using the pilot data, and $n^{planned}$ is the number of respondents per condition in the planned study.

Note: Rule 9 must be modified to handle interaction. See the section "How Are Interactions Different?".

Rule 10 (Pilot Data to Sample Size): We can use pilot data to conservatively predict the sample size we will need in a planned study. For 80% power to detect the treatment effect $\widetilde{\tau}$, we will (conservatively) need about $n^{pilot} \cdot \left[\frac{2.5}{\widetilde{\tau}} \cdot \left(\sqrt{\frac{1}{n^{pilot}}}+1\right) \cdot \widehat{SE}_{\widehat{\tau}}^{pilot}\right]^2$ respondents per condition, where n^{pilot} is the number of respondents per condition in the pilot data and $SE_{\widehat{\tau}}^{pilot}$ is the estimated standard error using the pilot data. For 95% power, the factor changes from 2.5 to 3.3.

Note: Rule 10 must be modified to handle interaction. See the section "How Are Interactions Different?".

Rule 11 (80% to 95% Power): Increasing the sample size by 75% will increase the power from 80% to about 95%.