Statistical Science
2020, Vol. 35, No. 2, 254–271
https://doi.org/10.1214/19-STS732

A Generalized Approach to Power Analysis for Local Average Treatment Effects

Kirk Bansak

Abstract. This study introduces a new approach to power analysis in the context of estimating a local average treatment effect (LATE), where the study subjects exhibit noncompliance with treatment assignment. As a result of distributional complications in the LATE context, compared to the simple ATE context, there is currently no standard method of power analysis for the LATE. Moreover, existing methods and commonly used substitutes—which include instrumental variable (IV), intent-to-treat (ITT) and scaled ATE power analyses—require specifying generally unknown variance terms and/or rely upon strong and unrealistic assumptions, thus providing unreliable guidance on the power of tests of the LATE. This study develops a new approach that uses standardized effect sizes to place bounds on the power for the most commonly used estimator of the LATE, the Wald IV estimator, whereby variance terms and distributional parameters need not be specified nor assumed. Instead, in addition to the effect size, sample size and error tolerance parameters, the only other parameter that must be specified by the researcher is the compliance rate. Additional conditions can also be introduced to further narrow the bounds on the power calculation. The result is a generalized approach to power analysis in the LATE context that is simple to implement.

Key words and phrases: Experimental design, local average treatment effects, noncompliance, principal stratification, statistical power.

1. INTRODUCTION

Power analysis has long been recognized as a vital study design tool (Cohen, 1962). Running simple power analyses provides researchers with concrete and reliable information to help determine their budgetary requirements, choose a sample size and form reasonable expectations on the magnitude of treatment effects they will be able to detect. This helps researchers avoid an eventuality in which a study has failed to produce meaningful findings not because there is nothing interesting to find but rather due to insufficient power to overcome fundamentally noisy data. The results of a power analysis can also help researchers avoid running certain studies altogether if the costs are simply too prohibitive in light of the probability of successful detection of a meaningful effect. Yet various fields of research are replete with studies that have failed to report power analyses and implemented drastically under-powered designs (Tversky and Kahneman, 1971, Tsang, Colley and Lynd, 2009). Indeed, many researchers' (and funders') time, energy and money have been put at risk by neglect of power analysis in the early stages of research design. In practice, however, power analyses can often be challenging to properly implement.

Consider the standard experimental setting in which units are assigned to one of two conditions, a treatment condition and a control condition, and the goal is to determine whether the treatment has an effect on some outcome variable of interest. Further, it may also be possible that *uptake* of the treatment is not perfectly determined by *assignment* of the treatment, as some units may not comply with their assignment status. To define the causal effects of interest, this study employs the potential outcomes framework presented by Neyman (1923) and Rubin (1974) and postulates a data-generating distribution on quadruples $(Y_i(0), Y_i(1), D_i(0), D_i(1)) \in \mathbb{R} \times \mathbb{R} \times \{0,1\} \times \{0,1\}$. For any unit i, the $Y_i(d)$, $d \in \{0,1\}$, denote the outcome that unit i would exhibit if it undertook treatment status and took the treatment ($d = 1$) or if it undertook control status and did not take the treatment ($d = 0$). Additionally, the $D_i(z)$, $z \in \{0,1\}$, denote the treatment uptake status that unit i would exhibit

Kirk Bansak is Assistant Professor, Department of Political Science, University of California San Diego, Social Sciences Building 301, La Jolla, California 92093-0521, USA (e-mail: kbansak@ucsd.edu).

if assigned to the treatment condition ($z = 1$) or if assigned to the control condition ($z = 0$). Throughout this study, we suppose we will observe a sample of N independent and identically distributed units of the form $(Y_i, D_i, Z_i) \in \mathbb{R} \times \{0, 1\} \times \{0, 1\}$, where for each unit i the $(Y_i(0), Y_i(1), D_i(0), D_i(1))$ is drawn from the distribution noted above, Z_i is a treatment assignment, $D_i = D_i(Z_i)$ is the realized treatment uptake, and $Y_i = Y_i(D_i)$ is the realized outcome.

In the simple context of full compliance with the treatment assignment (i.e., where all units always take the treatment if assigned to it and do not take the treatment if not assigned to it), then $D_i(0) = 0$, $D_i(1) = 1$, and hence $D_i = Z_i$, for all i. In this setting, researchers are generally interested in the average treatment effect (ATE), which is the difference between the expected outcome units would attain if they took up the treatment and the expected outcome units would attain if they did not take up the treatment. The ATE, denoted here by δ, is defined formally as

$$\delta = E\big[Y_i(1) - Y_i(0)\big].$$

Given full compliance and random assignment of the treatment, consistent and unbiased estimation of δ is straightforward. Even in this simple case, however, a complication for performing a prestudy power analysis for the test of the null hypothesis that $\delta = 0$ is the need, *a priori*, for estimates of or assumptions about the variances of $Y_i(0)$ and $Y_i(1)$ (Bloom, 2006, Duflo, Glennerster and Kremer, 2007). This complication is well known among applied researchers, and fortunately, there also exist fixes for this problem in the full-compliance setting, as will be described later.

Less well known, however, is the extent to which such complications become exacerbated in the context of estimating treatment effects when the study units exhibit noncompliance with treatment assignment. In this case, even if assigned to the treatment, a unit may not necessarily take the treatment, and vice versa. The distribution of $(Y_i(0), Y_i(1), D_i(0), D_i(1))$ now features four sub-types of units or "principal strata" that are defined as a function of the $D_i(z)$: "compliers," defined as the stratum for which $D_i(1) = 1$ and $D_i(0) = 0$; "always-takers," for which $D_i(1) = 1$ and $D_i(0) = 1$; "never-takers," for which $D_i(1) = 0$ and $D_i(0) = 0$; and defiers, for which $D_i(1) = 0$ and $D_i(0) = 1$. In the presence of noncompliance, the ATE generally cannot be identified.[1] However, under a set of assumptions presented by Angrist, Imbens and Rubin (1996), it is possible to identify a "local average treatment effect" (LATE), which is the ATE for the compliers, or those who take the treatment when assigned to the treatment and do not otherwise.[2] The LATE, which will be denoted by τ, is defined formally as

$$\tau = E\big[Y_i(1) - Y_i(0)\big|D_i(1) - D_i(0) = 1\big].$$

This study considers the problem of performing a power analysis in the presence of noncompliance for the test of the null hypothesis that $\tau = 0$. Due to the existence of multiple principal strata, the possibility of distinct marginal distributional behavior across those strata, the focus on local identification of the average treatment effect for the compliers, and the inability to completely differentiate compliers from other principal strata in observed data, the number of distributional parameters that impact the power vastly proliferates in the LATE context. In fact, the power of the test that $\tau = 0$ not only depends upon the rate of compliance, with lower compliance resulting in lower power, but is also impacted by the different conditional means and variances of the outcome across the principal strata as well as the relative sizes of the principal strata across the distribution (i.e., probabilities of belonging to each stratum).

As a result, there is currently no standard method of power analysis in the LATE context. In addition, existing methods require specifying distributional assumptions that are difficult to make and/or come with hidden, implicit assumptions about the various principal strata that are unlikely to reflect the reality of one's applied data. Recognizing the complexity of LATE power analysis, some researchers settle for performing scaled ATE or "intent-to-treat" (ITT) power analyses, discussed later, even when their ultimate estimand of interest is the LATE. This is a precarious practice given that, as will be shown, the results of scaled ATE and ITT power analyses can diverge substantially from the results of a LATE power analysis. This state of affairs is problematic given how common noncompliance is in many research environments, including field experiments, clinical trials and randomized controlled trials (RCTs) using encouragement designs. These types of studies also tend to be among the most expensive, generating strong incentives for well-calibrated power analyses.

This study introduces a new approach to LATE power analysis employing the Wald IV estimator. Specifically, by using a standardized LATE effect size, this study shows how bounds can be placed on the power of the test of the null hypothesis that $\tau = 0$ whereby neither variance components nor patterns of noncompliance and heterogeneity need to be specified. Instead, in addition to the

[1] Noncompliance would not pose a problem for identification of the ATE if units' noncompliance behavior were independent of their potential outcomes. However, in practice this may often be unlikely, as study subjects can be motivated to select into the treatment or control condition based on their expectations of their own potential outcomes under each condition.

[2] The assumptions and identification of the LATE will be discussed more fully in Section 3.

effect size, sample size and error tolerance parameters, the only other parameter that must be specified by the researcher is the compliance rate. In contrast, nine other underlying parameters that affect power need not be specified. This study focuses on the Wald IV estimator because it is the most accessible and commonly used estimator of the LATE among applied researchers. In addition, in contrast to other estimators of the LATE, such as those based on maximum likelihood estimation and Bayesian methods (e.g., Imbens and Rubin, 1997), the Wald IV estimator is nonparametric.

As usual, the effect size and sample size parameters can be isolated in the power analysis introduced in this study, providing "worst-case-scenario" formulas for minimum detectable effect sizes and required sample sizes. Additional assumptions can also be made to further narrow the bounds on the power calculation to avoid overconservatism. The result is a generalized approach to power analysis in the LATE context that is simultaneously conservative, disciplined and simple to implement. As a central reference and summary of the main recommendations, Table 1 provides the conservative formulas for power, minimum detectable effect size and required sample size under a variety of scenarios considered in the study, offering researchers a principled strategy for proceeding with conservative power analyses for the LATE. The approach can also be extended to tests in fuzzy regression discontinuity designs (Hahn, Todd and Van der Klaauw, 2001) that use the instrumental variable (IV) estimator in the discontinuity window around the threshold, as well as quasi-experiments that apply the IV framework to observational data.

To introduce a frame of reference, Section 2 will briefly discuss power analysis in the standard ATE context with full compliance. Section 3 will then introduce the LATE, and proceed to highlight problems with existing power analysis methods and general challenges to analyzing power in the LATE context. Sections 4 and 5 will present the new method of LATE power analysis introduced by this study. Section 6 will summarize the main results and recommendations, as well as provide an illustration of how the method could be used in practice by applying it in the context of the National JTPA Study. Sections 7 and 8 discuss how to extend the framework to allow for covariate adjustment and multivalued treatments. Section 9 concludes.

2. POWER ANALYSIS FOR AVERAGE TREATMENT EFFECTS

Consider the goal of understanding how some intervention (a treatment) impacts an outcome of interest in an experimental setting where we can assume full compliance with treatment assignment. As before, suppose we observe a sample of N independent and identically distributed units of the form $(Y_i, D_i, Z_i) \in \mathbb{R} \times \{0,1\} \times \{0,1\}$, where for the ith unit Z_i is the treatment assignment, and the outcome Y_i and treatment uptake D_i are generated according to the data-generating distribution of potential outcomes noted in the previous section. Given full compliance, $Z_i = D_i$ for all i. Further, let $\widehat{\overline{Y(0)}}$ and $\widehat{\overline{Y(1)}}$ denote the averages of the observed outcomes for the sampled units actually assigned to the control and treatment conditions, respectively, which constitute unbiased and consistent estimates of $E[Y_i(0)]$ and $E[Y_i(1)]$ given random assignment of the treatment and the "stable unit treatment value assumption" (SUTVA) (Rubin, 1978, 1980, 1990). As a result, the difference-in-means estimator, $\hat{\delta} = \widehat{\overline{Y(1)}} - \widehat{\overline{Y(0)}}$, is unbiased and consistent for the average treatment effect (ATE), the true value of which is $\delta = E[Y_i(1) - Y_i(0)]$.

As shown elsewhere (Cohen, 1988, Bloom, 2006, Duflo, Glennerster and Kremer, 2007), given asymptotic normality of the difference-in-means estimator, power analysis for the test of the null hypothesis that $\delta = 0$ with a two-sided alternative then proceeds with the following equation:

$$\Phi\left(-c^* + \frac{\delta}{\sqrt{V_N}}\right) + \Phi\left(-c^* - \frac{\delta}{\sqrt{V_N}}\right) = 1 - \beta,$$

where $\Phi(\bullet)$ denotes the standard normal cumulative distribution function, δ denotes a hypothesized true ATE value, V_N denotes the sampling variance of the estimator $\hat{\delta}$, $1-\beta$ denotes the power to correctly reject the null hypothesis (β denotes the type-II error rate), and c^* denotes the critical value corresponding to the tolerable type-I error rate (α) and hypothesis test type. For the standard two-tailed test of the null hypothesis that $\delta = 0$, $c^* = \Phi^{-1}(1 - \frac{\alpha}{2})$. Conventionally, $V_N \equiv \frac{\text{Var}(Y_i(1))}{N_1} + \frac{\text{Var}(Y_i(0))}{N_0}$, given N units with N_1 assigned to treatment and $N_0 = N - N_1$ assigned to control.[3]

In order to use the power formula above, the analyst must specify V_N, which requires explicitly or implicitly specifying the variances of $Y_i(0)$ and $Y_i(1)$. This requirement presents a possibly serious practical complication. While previous studies and/or existing data can often help to inform these variance specifications, there often do not exist any data that is recent or closely related enough to serve as a useful benchmark, particularly in the case where a researcher is interested in a novel outcome variable or new population of interest. One might devise an idea, based on theoretical expectations, about what a conservative variance might look like. However, in the very plausible case that these expectations are inaccurate, too

[3] Imbens and Rubin (2015) show that this formula for V_N is conservative given complete randomization of N units with a predetermined number N_1 assigned to treatment and $N_0 = N - N_1$ assigned to control.

high a guess will lead to an overpowered study while too low a guess will lead to an unsuccessful study. In both cases, the researcher's resources are at risk of being wasted.

"Effect sizes" have long been established as the standard solution to this problem in the ATE context with full compliance (Cohen, 1988, Bloom, 2006, Duflo, Glennerster and Kremer, 2007). In cases where the variances are unknown and/or absolute effect magnitudes are difficult to interpret, a common recommendation is to employ the effect size $\frac{\delta}{\sigma}$—rather than the absolute effect δ—where σ is the standard deviation of a reference outcome distribution. In other words, effect sizes are measures of treatment effects that are standardized with reference to the distribution of the outcome variable. Most commonly used is $\sigma = \sqrt{E[\text{Var}(Y_i|D_i)]}$, the expected within-group standard deviation of the outcome. By employing effect sizes, the result is that the variance terms in the power formula drop out, thus obviating the inconvenient need to estimate or guess variance values. In addition, there exist general benchmarks for what constitutes a small, medium and large effect size (e.g., Cohen, 1988) and meta-analyses within individual fields of study have enabled researchers to develop discipline-specific guidance on effect size significance (Lipsey, 1990, Chapter 3).

3. POWER IN THE LATE CONTEXT

3.1 Local Average Treatment Effect (LATE)

In many studies, the subjects exhibit noncompliance: some units assigned to the treatment condition do not take the treatment, and/or some units assigned to the control condition do take the treatment. This problem is pervasive across many research settings—including field experiments, clinical trials and RCTs using encouragement designs—as subjects often cannot be forced to take the treatment, and some subjects are able to access the treatment even when not assigned to it (Gerber and Green, 2012). As explained earlier, in the presence of noncompliance, the ATE generally cannot be identified, but it is possible to identify the local average treatment effect (LATE), which is the ATE for the compliers (Angrist, Imbens and Rubin, 1996). In the case of one-sided noncompliance, the LATE is the ATE for the treated (given no always-takers) or the ATE for the untreated (given no never-takers).

In their seminal study applying the potential outcomes framework to the identification and estimation of the LATE, Angrist, Imbens and Rubin (1996) begin by considering N units indexed by i and defining the potential outcomes $D_i(\mathbf{Z})$ and $Y_i(\mathbf{Z}, \mathbf{D})$, where $\mathbf{Z}$ and $\mathbf{D}$ correspond to the N-dimensional treatment assignment and uptake vectors across the units. $D_i(\mathbf{Z}) \in \{0, 1\}$ denotes the treatment uptake that unit i would exhibit given the full treatment assignment vector, and $Y_i(\mathbf{Z}, \mathbf{D}) \in \mathbb{R}$ denotes the outcome that unit i would exhibit given the full treatment assignment and treatment uptake vectors. Note that while this potential outcomes notation differs from that employed earlier in the present study, Angrist, Imbens and Rubin (1996) make a set of assumptions that simplify the potential outcomes to the form employed earlier here. Specifically, Angrist, Imbens and Rubin (1996) introduce the following assumptions:

ASSUMPTION 1 (Stable unit treatment value assumption (SUTVA)). Let $(\mathbf{Z}, \mathbf{D})$ and $(\mathbf{Z}', \mathbf{D}')$ be pairs of treatment assignment and uptake vectors. If $Z_i = Z_i'$, then $D_i(\mathbf{Z}) = D_i(\mathbf{Z}')$. If $Z_i = Z_i'$ and $D_i = D_i'$, then $Y_i(\mathbf{Z}, \mathbf{D}) = Y_i(\mathbf{Z}', \mathbf{D}')$.

ASSUMPTION 2 (Random assignment of the treatment). $P(\mathbf{Z} = \mathbf{a}) = P(\mathbf{Z} = \mathbf{a}')$ for all $\mathbf{a}$ and $\mathbf{a}'$ such that $\iota^T \mathbf{a} = \iota^T \mathbf{a}'$ where ι is the N-dimensional column vector with all elements equal to one.

ASSUMPTION 3 (Exclusion restriction). $\mathbf{Y}(\mathbf{Z}, \mathbf{D}) = \mathbf{Y}(\mathbf{Z}', \mathbf{D})$ for all $\mathbf{Z}$, $\mathbf{Z}'$ and for all $\mathbf{D}$.

ASSUMPTION 4 (Nonzero average causal effect of Z on D). $E[D_i(1) - D_i(0)] \neq 0$.

ASSUMPTION 5 (Monotonicity). $D_i(1) \geq D_i(0)$ for all i.

As shown by Angrist, Imbens and Rubin (1996), given these assumptions, the potential outcomes for Y are reduced to $Y_i(d)$, $d \in \{0, 1\}$, as introduced earlier in this study. The $Y_i(d)$ denote the outcome that unit i would exhibit if unit i assumed treatment status and took the treatment ($d = 1$) or if it assumed control status and did not take the treatment ($d = 0$), irrespective of all other units. Further, the potential outcomes for D are also reduced to the earlier notation, $D_i(z)$, $z \in \{0, 1\}$, which denote the treatment uptake status that unit i would exhibit if unit i was assigned to the treatment condition ($z = 1$) or assigned to the control condition ($z = 0$), irrespective of all other units.

We can thus postulate, as introduced and defined earlier, a data-generating distribution on the quadruples $(Y_i(0), Y_i(1), D_i(0), D_i(1)) \in \mathbb{R} \times \mathbb{R} \times \{0, 1\} \times \{0, 1\}$. Recall that compliers are defined as units for whom $D_i(1) - D_i(0) = 1$. In contrast, always-takers are units for whom $D_i(1) = D_i(0) = 1$, and never-takers are units for whom $D_i(1) = D_i(0) = 0$. Note the existence of defiers, or units for whom $D_i(1) - D_i(0) = -1$ (i.e., units that take the treatment if not assigned to it and do not take the treatment if assigned to it), is ruled out by Assumption 5.

Following Angrist, Imbens and Rubin (1996), the LATE (denoted here by τ) is defined formally as the ATE, or average causal effect of D on Y, for compliers:

$$\tau = E\big[Y_i(1) - Y_i(0)\big|D_i(1) - D_i(0) = 1\big].$$

Under Assumptions 1–5, Angrist, Imbens and Rubin (1996) show that this estimand is equivalent to the ratio between the average causal effect of Z on Y (intent-to-treat effect, or ITT), which will be denoted by γ, and the average causal effect of Z on D (first-stage effect), which will be denoted by π. The first-stage effect is also equivalent to the compliance rate given the assumptions. That is

$$\tau = \frac{\gamma}{\pi},$$
$$\gamma = E\big[Y_i\big(D_i(1)\big) - Y_i\big(D_i(0)\big)\big],$$
$$\pi = E\big[D_i(1) - D_i(0)\big] = P\big(D_i(1) - D_i(0) = 1\big).$$

Now suppose we observe a sample of N independent and identically distributed units of the form $(Y_i, D_i, Z_i) \in \mathbb{R} \times \{0, 1\} \times \{0, 1\}$, where for each unit i the $(Y_i(0), Y_i(1), D_i(0), D_i(1))$ is drawn from the distribution noted above, $D_i = D_i(Z_i)$ given the treatment assignment Z_i, and $Y_i = Y_i(D_i)$. Given the assumptions, the LATE can be estimated consistently by the Wald IV estimator, which will be denoted by $\hat{\tau}$:

$$\hat{\tau} = \frac{\widehat{\mathrm{Cov}}(Y_i, Z_i)}{\widehat{\mathrm{Cov}}(D_i, Z_i)},$$

where $\widehat{\mathrm{Cov}}$ denotes the sample covariance.

In contrast to other estimators of the LATE, such as those based on maximum likelihood estimation and Bayesian methods (e.g., Imbens and Rubin, 1997), the Wald IV estimator is nonparametric and does not require assumptions about the probability distributions underlying the data. The Wald IV estimator is also the most accessible and commonly used estimator of the LATE among applied researchers. The asymptotic variance of the estimator given independent and identically distributed observations, as shown by Imbens and Angrist (1994), is

$$V_N^{\hat{\tau}} = \frac{E[\epsilon_i^2\{Z_i - E[Z_i]\}^2]}{N\mathrm{Cov}^2(D_i, Z_i)},$$

where $\epsilon_i = Y_i - E[Y_i] - \tau(D_i - E[D_i])$.

In the face of noncompliance, researchers often weigh the merits of focusing on the ITT vs. LATE as the ultimate estimand of interest from the perspective of their own research goals and questions (e.g., Imbens, 2014a, Kitagawa, 2014, Swanson and Hernán, 2014, Imbens, 2014b). The ITT measures the average effect of treatment *assignment* in the presence of noncompliance. This is an ideal estimand for researchers wishing to understand the overall system-wide effect of introducing an intervention into the study context. However, the ITT does not capture a causal effect of the treatment itself. In contrast, the LATE measures the average causal effect of the treatment *uptake* for the compliers. While the compliers are a subset of the underlying population, note that they are often the sub-population of interest, as they are precisely the subset of individuals who can actually be induced to take (or not take) the treatment. In contrast, it is often not relevant or useful to understand the effect of a treatment for a subpopulation who will never end up taking the treatment (or who will always take it no matter what).

By measuring a causal effect of the treatment, the LATE thereby allows researchers to understand the efficacy of the treatment itself. This can be a critical task for a number of research goals. First, it allows for more direct scientific investigation of the underlying causal phenomenon. Second, it facilitates efforts to improve the design of the intervention such that it becomes more efficacious at the individual level. Third, it is also key for determining the cost-efficiency of the treatment in many contexts. Given that costly interventions often scale proportionately to the number of applications/dosages actually delivered, rather than simply the number assigned, it is crucial to measure the cost-efficiency of delivered treatment applications/dosages, which the LATE allows for but the ITT does not. In short, for studies focused on understanding the efficacy of treatments and measuring their causal effects, the LATE is often a more interesting, informative, and/or policy-relevant estimand.[4]

3.2 Proliferation of Parameters Affecting Power for the LATE

In general, we may separate power analysis parameters into three groups: error tolerance parameters, investigation parameters and distribution parameters. The error tolerance parameters are α (type-I error tolerance) and β (type-II error tolerance), where β is a parameter only in the case where we are solving for a different parameter rather than calculating the power $(1 - \beta)$. The investigation parameters are sample size and effect magnitude/size. These are the parameters of fundamental interest that motivate the use of a prestudy power analysis. Finally, the distribution parameters are the parameters that characterize the distribution(s) of the population(s) of interest. In contrast to the tolerance parameters, which are selected by convention or on a discretionary basis, and the investigation parameters, which the researcher seeks to learn about in order to make research design decisions, the distribution parameters are matters of inconvenience. While they are (usually) not of strict interest to the researcher, the distribution parameters have a dramatic impact on statistical power, and they must be specified at values that are known or believed to reflect reality in order for a power analysis to be properly calibrated and hence informative.

As described earlier, in the standard ATE context with full compliance using absolute effect magnitudes, the

[4]Recall, also, that in study designs that ensure the absence of always-takers (never-takers) the LATE becomes the ATE for the treated (untreated).

power formula requires specification of the variance of the ATE estimator. This variance depends upon two distribution parameters: the potential outcome variances of both the treatment and control conditions. In addition, by employing standardized effect sizes, these distribution parameters can be dispensed with. In the LATE context, a power formula would also entail specifying the variance of the estimator. In contrast to the ATE context, however, this variance depends upon many more distribution parameters in the LATE context. If we consider the compliance rate π to be an additional investigation parameter in the LATE context, then there are in fact nine distribution parameters that affect the variance of the Wald IV estimator and hence also affect the power.

The reason for this proliferation of parameters has to do with marginal distributional heterogeneity across the principal strata. Specifically, in addition to the investigation parameters, the estimator variance is also affected by: (1) the complier control condition potential outcome mean, $E[Y_i(0)|D_i(1) - D_i(0) = 1]$; (2) the complier control condition potential outcome variance, $\text{Var}[Y_i(0)|D_i(1) - D_i(0) = 1]$; (3) the complier treatment condition potential outcome variance, $\text{Var}[Y_i(1)|D_i(1) - D_i(0) = 1]$; (4) the never-taker control condition potential outcome mean, $E[Y_i(0)|D_i(1) = D_i(0) = 0]$; (5) the never-taker control condition potential outcome variance, $\text{Var}[Y_i(0)|D_i(1) = D_i(0) = 0]$; (6) the always-taker treatment condition potential outcome mean, $E[Y_i(1)|D_i(1) = D_i(0) = 1]$; (7) the always-taker treatment condition potential outcome variance, $\text{Var}[Y_i(1)|D_i(1) = D_i(0) = 1]$; (8) the proportion of never-takers, $P(D_i(1) = D_i(0) = 0)$; and (9) the proportion of always-takers, $P(D_i(1) = D_i(0) = 1)$.[5] These properties are illustrated for the Wald IV estimator in the Supplementary Materials (Bansak, 2020, SM) Appendix B (Tables B1–B2), which presents the results of a series of simulations illustrating the power of the estimator as the marginal distributional characteristics of the principal strata are varied.

3.3 Limitations of Existing Methods for Power Analysis

Given the expectation of noncompliance with treatment assignment, a researcher wishing to perform a power analysis in order to inform the study design (e.g., number of subjects) has a few existing options. However, the unique characteristics of the LATE context, namely the existence of multiple principal strata characterized by marginal distributional heterogeneity, significantly limit the reliability of these existing methods.

The first existing option is to apply a standard power analysis to the ITT. This may seem problematic at face value, of course, since the ITT is a different target estimand than the LATE. Indeed, for researchers who intend to focus on and estimate the LATE, the problem with ITT power analyses is that, for a given data-generating distribution, the power to detect nonzero effects for the ITT difference-in-means estimator will deviate from that of estimators of the LATE, as highlighted in previous work by Jo (2002). This phenomenon is illustrated in the SM Appendix B (Table B3), which presents the results of simulations in which the power for the LATE (Wald IV estimator) and ITT (difference-in-means estimator) change at different rates as the simulation specifications are altered. In fact, as the simulations show, the power for tests of the LATE may be higher or lower than the power for tests of the ITT depending upon the heterogeneity across the principal strata. This may be somewhat surprising because the Wald IV estimator is simply a scaled version of the ITT, where the ITT is divided by the compliance rate. However, the compliance rate is a quantity that must also be estimated, and that estimate is generally correlated with the estimate of the ITT, resulting in the Wald IV estimator having distinct statistical properties from the ITT difference-in-means estimator.

To illustrate the problem more vividly, consider a hypothetical superpopulation whose potential outcomes are depicted in Figure 1 based on a specific data-generating distribution of the quadruples $(Y_i(0), Y_i(1), D_i(0), D_i(1)) \in \mathbb{R} \times \mathbb{R} \times \{0, 1\} \times \{0, 1\}$. Compliers comprise 30% of the superpopulation: $P(D_i(1) - D_i(0) = 1) = 0.3$. Always-takers and never-takers each comprise 35%: $P(D_i(1) = D_i(0) = 1) = 0.35$ and $P(D_i(1) = D_i(0) = 0) = 0.35$. Finally, defiers do not exist: $P(D_i(1) - D_i(0) = -1) = 0$. Note that compliance rates of 0.3 and lower are prevalent in field experiments, encouragement-based RCTs and natural experiments.[6] The LATE has a value of 5 in the superpopulation displayed in Figure 1.[7] In addition, the never-takers have a slightly higher mean potential outcome value under control and always-takers have a slightly

[5] It should be noted that the sum of (8), (9) and π must be one, removing a degree of freedom in the specification of distribution parameters. In addition, it should be noted that the treatment condition potential outcome mean for the compliers is not included since it is simply the sum of the compliers' control condition mean and the LATE. Further, treatment (control) condition parameters are not included for the never-takers (always-takers) because the treatment (control) condition never manifests in the data for the never-takers (always-takers) by definition.

[6] For instance, the estimated compliance rate in a vote canvassing field experiment run by Gerber and Green (2000) was around 0.3, the estimated compliance rate in an influenza vaccination encouragement design evaluated by Hirano et al. (2000) was approximately 0.12, and in a natural experiment evaluated by Angrist (1990) on the effect of Vietnam War veteran status on civilian earnings, the estimated compliance rate (effect of draft eligibility on veteran status) ranged from 0.10 to 0.16 for white American citizens born from 1950–1952.

[7] Specifically, the potential outcomes are normally distributed with a variance of 9 and means of 0 and 5 for the compliers under control and under treatment, respectively.

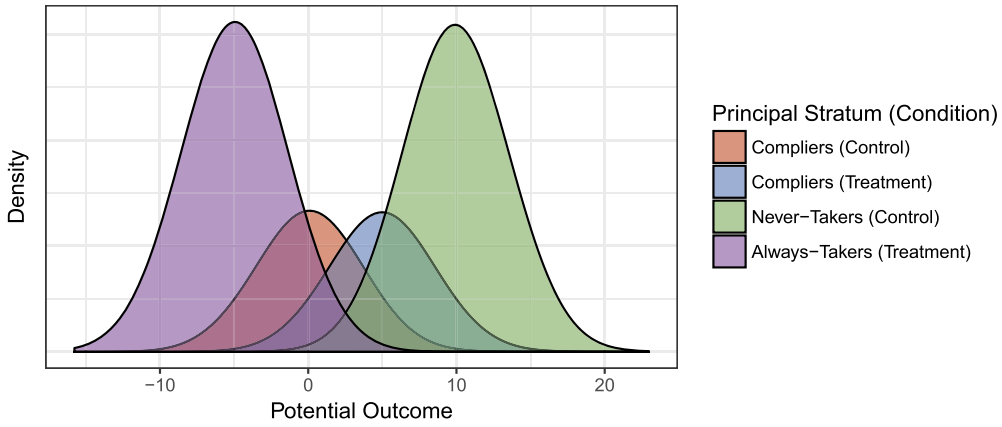

FIG. 1. *Distribution of potential outcomes $Y_i(d)$ in superpopulation.*

lower mean potential outcome value under treatment, consistent with a common scenario in which subjects with initially high outcome levels have no incentive to take the treatment and subjects with low outcome values are particularly motivated to access the treatment regardless of their assignment.[8] Note that only the treatment potential outcomes are relevant for the always-takers, and only the control potential outcomes are relevant for the never-takers. 10,000 samples of size 650 were randomly drawn from this superpopulation, and each unit had a probability of 0.5 of being assigned to the treatment. Whether each unit actually took the treatment and its realized outcome were determined jointly by its treatment assignment and the principal stratum to which it belonged, yielding for each sample 650 independent and identically distributed units of the form $(Y_i, D_i, Z_i) \in \mathbb{R} \times \{0, 1\} \times \{0, 1\}$. Using the observed values for each sample, tests of the hypotheses that the ITT and LATE are zero were performed with the difference-in-means and Wald IV estimators, respectively, thereby allowing for a simulated comparison of the power of each test.[9] The power for the ITT was approximately 0.77, while the power for the LATE was substantially lower at 0.61. While this is only a single hypothetical data-generating distribution, it conveys an important general message: the results of an ITT power analysis can provide extremely inaccurate guidance (e.g., a miscalibrated sample size recommendation) for researchers planning ultimately to focus on and estimate the LATE. As a general rule, the mismatch between the power for the ITT and power for the LATE will be more pronounced given a lower compliance rate and greater distributional heterogeneity across the principal strata.

A second existing option is a scaled ATE power analysis, which is a commonly used approach in which the results of a standard ATE power analysis are scaled by an appropriate function of the compliance rate. Using this approach, Duflo, Glennerster and Kremer (2007) present a formula for computing minimum detectable effects in the presence of noncompliance based on a simple scaling of the standard ATE formula, the result of which follows from the ATE estimator variance being divided by the compliance rate squared. The rationale behind this process is based on the fact that the Wald IV estimator is simply the ratio of the ITT to the compliance rate. However, the scaled ATE power analysis treats the compliance rate as a known value when, as already explained above, the compliance rate must be estimated and that estimate is generally correlated with the ITT estimate. The resulting problem with this approach, which Duflo, Glennerster and Kremer do not make explicit but has been shown elsewhere (Baiocchi, Cheng and Small, 2014), is that a number of strong and unrealistic assumptions are required for this scaling of the standard ATE power analysis to yield the (approximately) correct power for tests using the Wald IV estimator. Specifically, it must be the case that (a) the never-takers have the same mean outcome value as the untreated compliers, (b) the always-takers have the same mean outcome value as the treated compliers, and (c) all groups have the same within-condition outcome variance. If any of those assumptions are violated, the true power of the test of the hypothesis that the LATE equals zero can diverge dramatically from the power implied by this scaled ATE power analysis. SM Appendix B (Table B4) demonstrates this result, illustrating how the scaled ATE power analysis, similar to an ITT power analysis, can provide extremely unreliable guidance on power for the LATE.

Finally, as a third option, power analyses specifically for instrumental-variable (IV) effects have been introduced in the epidemiology literature (Pierce, Ahsan and VanderWeele, 2011, Freeman, Cowling and

[8]Specifically, the potential outcomes for the always-takers (under treatment) and never-takers (under control) are normally distributed with a variance of 9 and means of -5 and 10, respectively.

[9]The simulated power is the proportion of samples for which the test rejects the null hypothesis of no effect. Two-sided hypothesis tests with $\alpha = 0.05$ were used.

Schooling, 2013, Brion, Shakhbazov and Visscher, 2013, Wang et al., 2018). In particular, Freeman, Cowling and Schooling (2013), Brion, Shakhbazov and Visscher (2013) and Wang et al. (2018) all introduce power formulas for IV effects. However, there are two major limitations to the approaches taken by these studies. First, they require specifying a number of variance components, about which a researcher may not have good preexisting knowledge or priors. Second, they proceed from a classic IV perspective and hence neglect the extent to which these variance components depend upon the distinct distributional behavior of the principal strata.

For instance, the formulas presented by Freeman, Cowling and Schooling and Brion, Shakhbazov and Visscher both require specifying $\text{Var}(D_i)$.[10] This presents a challenge given noncompliance with the treatment assignment, as $\text{Var}(D_i)$ is a function of both the first stage effect π (which is also the compliance rate) and the proportion of always-takers versus never-takers. In other words, to choose an informative value of $\text{Var}(D_i)$, one must specify not only a hypothetical compliance rate but also the precise pattern of noncompliance. In addition, Freeman, Cowling and Schooling also require specifying $\text{Var}(Y_i|D_i)$, while Brion, Shakhbazov and Visscher require specifying the biased asymptotic value of the least squares estimator of the effect of D on Y. Finally, the approach taken by Wang et al. requires specifying the ITT, standard deviation of the potential outcome under control, standard deviation of the error from regressing the treatment on the instrument, and the correlation between the potential outcome under control and the error from regressing the treatment on the instrument. In many study contexts in the social sciences, medicine, public health, program evaluation and other fields, the researcher will lack solid estimates or priors on one or more of these parameters. In such contexts, the formulas offered by Freeman, Cowling and Schooling, Brion, Shakhbazov and Visscher and Wang et al. cannot be used reliably.

3.4 General Complications for Designing LATE Power Analyses

As highlighted above, there are significant limitations and liabilities associated with existing methods of power analysis in the presence of noncompliance. As a result, this is an area in the applied methodological literature that requires new approaches and solutions. Yet there are notable impediments to developing flexible and reliable approaches to power analysis in the LATE context.

A first complication that has been recognized for some time relates to the local identification of the LATE (Jo, 2002). As already described, the possibility of heterogeneous potential outcome distributions across the three principal strata (compliers, never-takers and always-takers) combined with the possibility of different patterns of noncompliance leads to the proliferation of parameters that affect power in the LATE context. Because these parameters jointly factor into the variance of LATE estimators, specifying a hypothetical value for the estimator variance to enable a power analysis involves making explicit or implicit assumptions about all of these parameters.

This first problem leads to a second complication in terms of being able to specify standardized effect sizes in such a way that the variance components of the power analysis drop out. Whereas in the ATE context given perfect compliance the fix is fairly simple and hence enables the analyst to minimize the number of assumptions that must be made, such a fix has been elusive in the LATE context given imperfect compliance. The result is that existing power analyses in the LATE context, whether analytic or simulation-based, have inconveniently required (explicit or implicit) distributional assumptions that may not match the reality of the data that will eventually be collected.

4. INTRODUCING A GENERALIZED APPROACH TO POWER ANALYSIS FOR THE LATE

The remainder of this study introduces a new method of LATE power analysis that addresses the problems described above and provides a more reliable tool than existing methods. The innovation and contribution of this new method is in showing how, by employing effect sizes, bounds can be placed on the power formula whereby neither variance components nor patterns of heterogeneity and noncompliance need to be specified. Instead, in addition to the effect size, sample size and error tolerance parameters, the only other parameter that must be specified by the researcher is the compliance rate. In other words, only tolerance and investigation parameters must be specified; the analyst need not specify nor even make assumptions about the estimator variance or any of the underlying distribution parameters.

4.1 Deriving a Modified Power Formula

As all of the results that follow pertain to estimating the LATE, the standard LATE assumptions (1–5) apply. As before, consider a sample of N independent and identically distributed units of the form $(Y_i, D_i, Z_i) \in \mathbb{R} \times \{0, 1\} \times \{0, 1\}$, where the LATE will be estimated using the Wald IV estimator, $\hat{\tau} = \frac{\widehat{\text{Cov}}(Y_i, Z_i)}{\widehat{\text{Cov}}(D_i, Z_i)}$. Similar to the ATE context, the results also invoke the asymptotic normality of the estimator. Hence, the power formula for the test of

[10] $\text{Var}(D_i)$ enters both formulas directly as well as through the need to specify ρ_{DZ} (the correlation between D_i and Z_i), which can only be mapped from a hypothetical first-stage effect π by specifying both $\text{Var}(D_i)$ and $\text{Var}(Z_i)$.

the null hypothesis that $\tau = 0$ begins as

$$\Phi\left(-c^* + \frac{\tau}{\sqrt{V_N^{\hat{\tau}}}}\right) + \Phi\left(-c^* - \frac{\tau}{\sqrt{V_N^{\hat{\tau}}}}\right) = 1 - \beta.$$

Further assume that assignment to the treatment is randomized with equal probability of being assigned to the treatment and control conditions. (Note that this equal assignment probability assumption will be relaxed later.)

ASSUMPTION 6 (Equal assignment probability). $P(Z_i = 1) = 0.5$ for all units $i = 1, 2, \ldots, N$,

The LATE power analysis introduced in this study proceeds by defining an effect size of interest. Following conventional practice using effect sizes in ATE power analyses, the effect size is defined in standardized terms with reference to the expected within-assignment-group standard deviation of the outcome:

DEFINITION 1. Define the effect size of interest as

$$\kappa = \frac{\tau}{\sqrt{E[\mathrm{Var}(Y_i|Z_i)]}}.$$

As will be discussed later, defining the effect size in this manner with reference to treatment assignment groups is appealing for a number of reasons. In particular, it provides a structure for determining reasonable effect sizes in advance of a study. In addition, it leads the resulting LATE power analysis derived below to nest the standard ATE power analysis as a special case with full compliance. More discussion is provided in a later section.

By focusing on the effect size, κ takes the place of τ as one of the three investigation parameters, along with π and N. In order to derive a LATE power formula that does not require specifying distribution parameters, or terms that depend on them, $\frac{\tau}{\sqrt{V_N^{\hat{\tau}}}}$ must be expressed exclusively in terms of investigation parameters. Recall, however, the complexity of the estimator variance: $V_N^{\hat{\tau}} = \frac{E[\epsilon_i^2\{Z_i - E[Z_i]\}^2]}{N\mathrm{Cov}^2(D_i, Z_i)}$ where $\epsilon_i = Y_i - E[Y_i] - \tau(D_i - E[D_i])$. Further consider that given imperfect compliance in the LATE context, and hence selection into (or out of) the treatment for some subjects, ϵ is not an intrinsically meaningful disturbance. In particular, ϵ is not orthogonal to D and hence does not have a conditional expectation of 0; by extension, $E[\epsilon_i^2]$ is not a substantively meaningful term.

As a result of the distributional complexities in the LATE context, it is not possible to derive a point calculation for the power of the Wald IV estimator without specifying its variance or the underlying distribution parameters. However, given Assumption 6 (equal assignment probability) and Definition 1, a set of tight bounds can be derived for the power of the Wald IV estimator. For notational convenience and without loss of generality, assume that $\kappa > 0$ (and hence also $\tau > 0$) is being investigated. Specifically, the following bounds can be put on $\frac{\tau}{\sqrt{V_N^{\hat{\tau}}}}$:

PROPOSITION 1. *Given Assumptions 1–6,*

$$\frac{0.5\kappa\pi\sqrt{N}}{\sqrt{1 + \kappa^2 E[\nu_i^2] + 2\kappa\sqrt{E[\nu_i^2]}}} \leq \frac{\tau}{\sqrt{V_N^{\hat{\tau}}}} \leq \frac{0.5\kappa\pi\sqrt{N}}{\sqrt{1 + \kappa^2 E[\nu_i^2] - 2\kappa\sqrt{E[\nu_i^2]}}},$$

where $\nu_i = D_i - E[D_i] - \pi(Z_i - E[Z_i])$.[11]

Of particular interest for study design purposes is the lower bound, which can provide the basis for a lower (and hence conservative) bound for the power. Notably, this reexpression leaves only one remaining term that is not an investigation parameter, $E[\nu_i^2]$. However, since D is binary, a practical and conservative reexpression of $E[\nu_i^2]$ can be undertaken by setting $E[\nu_i^2]$ to its largest possible value as a function of π. The result is a final lower (conservative) bound on $\frac{\tau}{\sqrt{V_N^{\hat{\tau}}}}$:

PROPOSITION 2. *Given Assumptions 1–6:*

$$\frac{0.5\kappa\pi\sqrt{N}}{1 + \kappa\sqrt{(0.5 - \frac{\pi}{2})(0.5 + \frac{\pi}{2})}} \leq \frac{\tau}{\sqrt{V_N^{\hat{\tau}}}}.$$

As can be seen, this final lower bound contains only the three investigation parameters: κ, π and N. Furthermore, this lower bound is tight—it cannot be raised without making additional assumptions—thus providing a tight lower bound for the power. In addition, setting $E[\nu_i^2]$ to its largest possible value also results in an approximate (slightly low) upper bound, $\frac{0.5\kappa\pi\sqrt{N}}{|1 - \kappa\sqrt{(0.5 - \frac{\pi}{2})(0.5 + \frac{\pi}{2})}|}$, though this quantity is of less practical value for study design than the conservative lower bound.

The bound in Proposition 2 can be plugged into the power formula $\Phi(-c^* + \tau/\sqrt{V_N^{\hat{\tau}}}) + \Phi(-c^* - \tau/\sqrt{V_N^{\hat{\tau}}})$ to produce a tight lower bound on the power

$$\Phi\left(-c^* + \frac{0.5\kappa\pi\sqrt{N}}{1 + \kappa\sqrt{(0.5 - \frac{\pi}{2})(0.5 + \frac{\pi}{2})}}\right) + \Phi\left(-c^* - \frac{0.5\kappa\pi\sqrt{N}}{1 + \kappa\sqrt{(0.5 - \frac{\pi}{2})(0.5 + \frac{\pi}{2})}}\right) \leq 1 - \beta.$$

Values of the investigation parameters κ, π and N—as well as a type-I error tolerance α to calculate c^*—can then be selected in order to calculate the lower bound on the power, $1 - \beta$, of the test of the null hypothesis that

[11] Note that, without loss of generality, it is assumed here that $\kappa > 0$ (and hence also $\tau > 0$) is being investigated. If κ and τ were negative, the inequalities would be reversed.

the LATE is zero. Importantly, there do not exist any variance terms in the modified power formula. As a result, it provides a bound that captures any distributional patterns among all three principal strata. In other words, the result is a generalized power analysis for the Wald IV estimator of the LATE that is free of additional distributional assumptions and does not require specification of the estimator variance or its underlying distribution parameters.

In addition, as with standard power analyses, the modified power formula can be rearranged to solve for the investigation parameters. Instead of calculating power based on a specific effect size (κ), compliance rate (π) and sample size (N), it can be more useful to select a desired power level and solve for one of the other parameters by fixing the rest. Of particular interest in this case should be κ, solving for which will yield the minimum detectable effect size (MDES),[12] and N, solving for which will yield the required sample size.

4.2 Solving for the Minimum Detectable Effect Size

Again without loss of generality, assume that $\kappa > 0$ is being investigated. Then, for reasonably high levels of power—which includes conventional power levels, such as 0.8—the second term in the power formula is negligible, simplifying the formula to

$$\Phi\left(-c^* + \frac{\tau}{\sqrt{V_N^{\hat{\tau}}}}\right) = 1 - \beta.$$

Recalling that $c^* = \Phi^{-1}(1 - \frac{\alpha}{2})$ for a two-sided test, this can then be reexpressed as

$$\frac{\tau}{\sqrt{V_N^{\hat{\tau}}}} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) + \Phi^{-1}(1 - \beta).$$

Let $M = \Phi^{-1}(1 - \frac{\alpha}{2}) + \Phi^{-1}(1 - \beta)$, which is called the "multiplier." M can then be plugged in for $\frac{\tau}{\sqrt{V_N^{\hat{\tau}}}}$ in the bounds presented above. This allows κ to then be isolated such that the MDES can be computed as a function of the other parameters. This results in a tight upper bound on the MDES, corresponding to the tight lower bound on the power, that can then be used as a conservative value for study design purposes:

$$\kappa \leq \frac{2M}{\pi\sqrt{N} - 2M\sqrt{(0.5 - \frac{\pi}{2})(0.5 + \frac{\pi}{2})}}.$$

As before, an approximate lower bound on the MDES can also be expressed in the same manner, though the MDES lower bound is of less practical value than the MDES upper bound for study design.[13]

[12]This term is borrowed from Bloom (2006), who used it in the ATE context as an extension of minimum detectable effects measured in absolute terms (Bloom, 1995).

[13]The MDES approximate lower bound is $2M/\left(\pi\sqrt{N} + 2M\sqrt{(0.5 - \frac{\pi}{2})(0.5 + \frac{\pi}{2})}\right) \leq \kappa$.

4.3 Solving for the Sample Size

Instead of isolating κ as above, N can be isolated in order to solve for the required sample size. Continuing to assume without loss of generality that $\kappa > 0$ is being investigated, the following is the tight upper bound on the required sample size corresponding to the tight lower bound on the power

$$N \leq \frac{4M^2\left(1 + \kappa\sqrt{(0.5 - \frac{\pi}{2})(0.5 + \frac{\pi}{2})}\right)^2}{\kappa^2\pi^2}.$$

This upper bound can then be used as a conservative value for study design purposes. Once again, an approximate lower bound on the required sample size can be similarly expressed, with the same caveat that such a quantity holds less practical value than the upper bound.[14]

4.4 Narrowing the Bounds

By providing a strict lower bound on the power, the method presented above offers a disciplined and reliable means of performing a conservative power analysis for the LATE. However, there is a tradeoff between conservatism and efficiency. If the lower bound is too conservative, it will lead to underestimation of the power and hence overestimation of the MDES and sample size required. This could then result in sub-optimal outcomes, such as a study being over-funded to achieve the conservative sample size or perhaps not funded at all if the sample size requirements exceed the financial resources available. As a result, it would be useful to narrow the bounds on the power formula where possible.

Continuing to assume without loss of generality that $\kappa > 0$, it can be shown that the lower bound on the power formula can be substantially raised when $\bar{Y}_{NT} \leq \bar{Y}_C \leq \bar{Y}_{AT}$, where $\bar{Y}_C$, $\bar{Y}_{NT}$ and $\bar{Y}_{AT}$ denote the expected realized outcome value for compliers, never-takers and always-takers (e.g., $\bar{Y}_C = E[Y_i|Complier] = E[Y_i|D_i(1) - D_i(0) = 1]$).

ASSUMPTION 7 (Ordered means). $\bar{Y}_{NT} \leq \bar{Y}_C \leq \bar{Y}_{AT}$ where $\bar{Y}_C$, $\bar{Y}_{NT}$ and $\bar{Y}_{AT}$ denote the expected realized outcome value for compliers, never-takers and always-takers.

In the case of one-sided noncompliance, Assumption 7 (ordered means) can be simplified to $\bar{Y}_{NT} \leq \bar{Y}_C$ or $\bar{Y}_C \leq \bar{Y}_{AT}$, depending upon the direction of noncompliance. It should also be noted that, in the case where noncompliance is almost one-sided (i.e. very few always-takers or very few never-takers), the sparse principal stratum will have only a negligible impact on estimation. Thus,

[14]The required sample size approximate lower bound is $\frac{4M^2\left(1 - \kappa\sqrt{(0.5 - \frac{\pi}{2})(0.5 + \frac{\pi}{2})}\right)^2}{\kappa^2\pi^2} \leq N$.

as a practical matter, Assumption 7 can be simplified to $\bar{Y}_{NT} \leq \bar{Y}_C$ or $\bar{Y}_C \leq \bar{Y}_{AT}$ as long as the sparse principal stratum is deemed sufficiently small.

As a result, if the researcher is comfortable making Assumption 7, then the lower bound of the power formula can be raised by using the following:

PROPOSITION 3. *Given Assumptions 1–7,*

$$\frac{0.5\kappa\pi\sqrt{N}}{\sqrt{1+\kappa^2(0.5-\frac{\pi}{2})(0.5+\frac{\pi}{2})}} \leq \frac{\tau}{\sqrt{V_N^{\hat{\tau}}}}.$$

Plugging this into the power formula yields:

$$\Phi\left(-c^* + \frac{0.5\kappa\pi\sqrt{N}}{\sqrt{1+\kappa^2(0.5-\frac{\pi}{2})(0.5+\frac{\pi}{2})}}\right) + \Phi\left(-c^* - \frac{0.5\kappa\pi\sqrt{N}}{\sqrt{1+\kappa^2(0.5-\frac{\pi}{2})(0.5+\frac{\pi}{2})}}\right) \leq 1-\beta.$$

By the same process described earlier, it is possible to solve for κ and N to derive new (lowered) upper bounds on the MDES and required sample size:

$$\kappa^* \leq \frac{2M}{\sqrt{N\pi^2 - 4M^2(0.5-\frac{\pi}{2})(0.5+\frac{\pi}{2})}},$$

$$N^* \leq \frac{4M^2(1+\kappa^2(0.5-\frac{\pi}{2})(0.5+\frac{\pi}{2}))}{\kappa^2\pi^2},$$

where again $M = \Phi^{-1}(1-\frac{\alpha}{2}) + \Phi^{-1}(1-\beta)$.

When would Assumption 7 (ordered means) be reasonable? Roughly speaking, there are two factors to consider when assessing the plausibility of this assumption. The first relates to effect heterogeneity. Specifically, it should be the case that always-takers (never-takers) select into (out of) the treatment because treatment uptake for them is associated with effects that are larger (smaller) than the average treatment effect for the compliers, or at least similarly sized. For instance, in the case of a positive and beneficial treatment, we must expect the noncomplying study subjects to be sufficiently rational that they are selecting into (out of) the treatment in anticipation of a particularly good (bad) effect on their outcome. Alternatively, selection into and out of the treatment could also be made for arbitrary reasons that are uncorrelated with individual effects. The second factor relates to baseline outcome levels in the absence of the treatment. Specifically, we must expect that always-takers (never-takers) do not have baseline outcome levels that are particularly low (high) compared to that of the compliers.

Precisely when the assumption should be expected to hold, in light of the two factors described above, will certainly be context dependent. However, specific research design steps can be taken to increase its plausibility. First, studies can often be designed so as to exclude one type of noncompliance. Indeed, many experiments are designed to prevent those not assigned to the treatment from accessing it and thus ensuring the absence of always-takers. By achieving one-sided noncompliance, the analyst need only consider two principal strata, rather than three. The well-known National JTPA Study is one such example (Bloom et al., 1997). In this experimental study, subjects were randomly assigned such that they were either given an offer to enroll in a job training program (assigned to treatment) or excluded from participating in the training for an 18-month period (assigned to control). However, many subjects given access to the job training program decided not to receive the training, resulting in a large chunk of never-takers. While there were some enterprising individuals who gained access to the job training in spite of not being assigned to it, their numbers were so small that there was virtually one-sided noncompliance.

Another research design step that can be taken is to impose reasonable restrictions on the study population of interest to ensure more similar baseline outcome levels across the principal strata. Again, the JTPA experiment is illustrative, as eligibility to participate in the experiment was restricted to those with economic disadvantages and barriers to employment. Had such restrictions not been made, the study may have included employed and/or higher income professionals who likely would have opted out of the job training program regardless of treatment assignment, boosting the baseline economic outcome levels of the never-takers.

In fact, as shown in the SM Appendix C, the final results from the JTPA experiment were consistent with the ordered means assumption in terms of an outcome variable that measured the participants' earnings in the 30-month period following their random assignment. Appendix C also presents the results of two other studies that were consistent with the ordered means assumption. One is a vote-canvassing field experiment (Gerber and Green, 2000). The other is a fuzzy regression discontinuity design on the effect of naturalization on political integration (Hainmueller, Hangartner and Pietrantuono, 2016). That the ordered means assumption was met in all three of these studies, which involved distinct study designs and research topics, demonstrates the plausibility of this assumption in various research domains.

In contrast, however, another common scenario in field experiments and encouragement-based RCTs is where subjects with initially high outcome levels have no incentive to take the treatment and subjects with low outcome values are particularly motivated to access the treatment regardless of their assignment. This pattern can lead to an ordering of the principal strata means that is the opposite of the ordered means assumption, as illustrated earlier in Figure 1. If the researcher believes such a scenario to be

possible, the ordered means assumption should not be applied, and the conservative lower bound on the power as reflected by Proposition 2 should be used.

4.5 Discussion on Effect Sizes

As explained earlier, the effect size of interest in this study (κ) is defined similarly to the way effect sizes are conventionally defined in ATE power analyses, with reference to the expected standard deviation of the outcome within treatment assignment groups. The difference, of course, is that full compliance is assumed in the ATE case, and hence treatment assignment is equivalent to treatment uptake. Nonetheless, the treatment assignment groups remain conceptually and practically useful reference groups in the LATE case for several reasons.

First, the treatment assignment groups are a mathematically natural reference group, allowing standard ATE power analysis results to nest as a special case within the LATE power formula presented in this study. Consider that in the special case of full compliance ($\pi = 1$), the LATE becomes the ATE. Further, given $\pi = 1$, the LATE power bounds presented in this study, as laid out in Proposition 2, are simplified to a single value:

$$1 - \beta = \Phi\left(-c^* + \frac{\kappa\sqrt{N}}{2}\right) + \Phi\left(-c^* - \frac{\kappa\sqrt{N}}{2}\right).$$

Solving for the minimum detectable effect size and required sample size yields $\kappa = \frac{2M}{\sqrt{N}}$ and $N = \frac{4M^2}{\kappa^2}$, where again $M = \Phi^{-1}(1 - \frac{\alpha}{2}) + \Phi^{-1}(1 - \beta)$. These results are identical to the conventional ATE power analysis results given asymptotic normality of the estimator and equal probability of assignment to treatment and control (Cohen, 1988, Bloom, 2006).

Second, the treatment assignment groups contain a natural reference point for defining a standardized effect size. In particular, the distribution of the outcome under assignment to control represents a natural state of the world in the absence of intervention, and hence $\text{Var}(Y_i|Z_i = 0)$ is one of the few baseline values that can often be reliably measured or estimated in advance of a study by analyzing data on the baseline population.[15] Further, while $\text{Var}(Y_i|Z_i = 1)$ cannot be measured in advance, it may be reasonable to assume it is relatively close in value to $\text{Var}(Y_i|Z_i = 0)$. In such cases, $E[\text{Var}(Y_i|Z_i)] \approx \text{Var}(Y_i|Z_i = 0)$, and hence the effect size of interest is defined (approximately) with reference to a naturally occurring distribution that is measurable prior to study implementation.

[15] In contrast, noncompliance leads to nonrandomization of D, which means $\text{Var}(Y_i|D_i = 0)$ will not be accurately reflected by prestudy estimates.

Third, because $\text{Var}(Y_i|Z_i = 0)$ may be measurable or estimable in advance, this allows for approximate mapping of effect sizes to absolute effects in the power formula. As long as the researcher is comfortable assuming that $\text{Var}(Y_i|Z_i = 1)$ will not diverge substantially from $\text{Var}(Y_i|Z_i = 0)$, then the researcher may estimate $\hat{\omega}_0 = \sqrt{\widehat{\text{Var}(Y_i|Z_i = 0)}}$, use that estimate as an approximate value for $\sqrt{E[\text{Var}(Y_i|Z_i)]}$, and hence replace κ with $\frac{\tau}{\hat{\omega}_0}$. The results presented above could then be modified to solve for a minimum detectable absolute effect (i.e., solve for τ itself) or solve for the required sample size in terms of τ.

Irrespective of the availability of reliable estimates for $\hat{\omega}_0$, researchers may also determine a target MDES by surveying previous studies and meta-analyses within their own fields of study (e.g., Lipsey, 1990, Chapter 3). In his seminal presentation of the topic, Cohen (1988) offered the conventional benchmarks in the social and behavioral sciences of 0.2, 0.5 and 0.8 as small, medium and large effect sizes, respectively. These general conventions may be useful as rough guidance. However, what is considered a small or large effect size inevitably varies across disciplines and research topics. Accordingly, it is advisable for the researcher to more carefully characterize the effect size scale within the research context at hand, in consultation with relevant data from previous studies and meta-analyses, as is the case for any power analysis irrespective of study design and compliance levels.

4.6 Comparing the Bounds to Simulations

To further validate the LATE power bounds derived in this study, Figure 2 compares the bounds to simulated power curves, where power is plotted as a function of κ. As in the simulation presented earlier, the simulations presented here also each specify a data-generating distribution of the quadruples $(Y_i(0), Y_i(1), D_i(0), D_i(1)) \in \mathbb{R} \times \mathbb{R} \times \{0, 1\} \times \{0, 1\}$, randomly draw from that distribution and randomize the treatment assignment variable, generating samples of independent and identically distributed units of the form $(Y_i, D_i, Z_i) \in \mathbb{R} \times \{0, 1\} \times \{0, 1\}$.

The solid black lines denote the analytic upper and lower bounds of the power, while the dashed black line denotes the alternative lower bound under Assumption 7 (ordered means). The colored lines denote the power curves that were simulated by specifying the full set of investigation and distribution parameters. For all of the curves (analytic and simulated), the following parameters are fixed: $\pi = 0.5$, $N = 1500$, $\alpha = 0.05$. In addition, for the simulated power curves, the following seven of the nine distribution parameters are fixed: $E[Y_i(0)|Complier] = 0$, $\text{Var}(Y_i(0)|Complier) = 64$, $\text{Var}(Y_i(1)|Complier) = 64$, $\text{Var}(Y_i(0)|NeverTaker) = 144$, $\text{Var}(Y_i(1)|AlwaysTaker) = 16$, $P(NeverTaker) =$

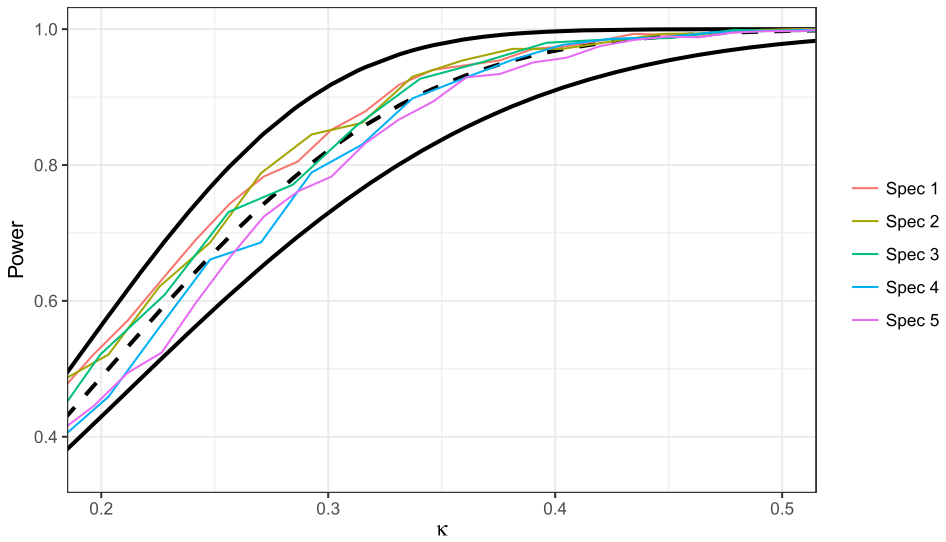

FIG. 2. *Simulated power vs. analytic bounds.*

0.25, and $P(AlwaysTaker) = 0.25$. In contrast, the final two distribution parameters, $E[Y_i(0)|NeverTaker]$ and $E[Y_i(1)|AlwaysTaker]$, vary across the five different simulation specifications shown in different colors. The five sets of values of $E[Y_i(0)|NeverTaker]$ and $E[Y_i(1)|AlwaysTaker]$, starting with the first specification, are as follows: $(-20, 20)$, $(-10, 10)$, $(-3, 3)$, $(10, -10)$ and $(20, -20)$.[16] This ensures that the simulation includes specifications that both do and do not meet Assumption 7 (ordered means), and hence allows for detailed evaluation of the bounds.

Figure 2 provides a simple demonstration of the performance of the power bounds presented in this study. As can be seen, the simulated curves fall within the analytic bounds denoted by the solid black lines. Furthermore, the alternative lower bound (the dashed black line) also bounds the appropriate simulated curves. For specifications 1 and 2, Assumption 7 (ordered means) is met by design at all values of κ, and hence the alternative lower bound applies. Accordingly, the curves for these specifications lie above the alternative lower bound. In contrast, the ordered means assumption is violated by specifications 4 and 5. Thus, it is no surprise that the curves for these specifications lie below the alternative lower bound. Additional graphical illustrations of the relationships between power and the investigation parameters are provided in the SM Appendix D.

5. RELAXING THE EQUAL ASSIGNMENT PROBABILITY ASSUMPTION

In some situations, the researcher may have reason put an unequal probability on assignment to the treatment and control conditions. For instance, treatment assignment/encouragement may be costly. For such cases, it will be useful to relax Assumption 6.

5.1 Results with $P(Z_i = 1) = p_z$

In the IV-LATE literature, the simplifying assumption of homoskedasticity that $E[\epsilon_i^2|Z_i] = E[\epsilon_i^2]$ is often made. While there may be few cases in which this assumption is likely to hold exactly, it is often sufficiently reasonable such that it does not substantially affect statistical inference. The assumption that $E[\epsilon_i^2|Z_i] = E[\epsilon_i^2]$ can be useful here.

ASSUMPTION 8 (Homoskedasticity). $E[\epsilon_i^2|Z_i] = E[\epsilon_i^2]$.

However, because the simplifying assumption that $E[\epsilon_i^2|Z_i] = E[\epsilon_i^2]$ is not a conservative one, it is useful to induce conservatism elsewhere. In order to do this, we can consider the limiting value of $E[\nu_i^2]$ at 0.25.

Continuing to assume without loss of generality that $\kappa > 0$ (and hence also $\tau > 0$) is under investigation, the following lower bound on $\frac{\tau}{\sqrt{V_N^{\hat{\tau}}}}$ then follows without making Assumption 6 (equal assignment probability):

PROPOSITION 4. *Given Assumptions 1–5 and 8, and any value* $p_z = P(Z_i = 1)$:

$$\frac{\kappa\pi\sqrt{p_z(1-p_z)N}}{1+0.5\kappa} \leq \frac{\tau}{\sqrt{V_N^{\hat{\tau}}}}.$$

As before, to derive the lower bound on the power, the bound above can simply be plugged into the power formula $\Phi(-c^* + \tau/\sqrt{V_N^{\hat{\tau}}}) + \Phi(-c^* - \tau/\sqrt{V_N^{\hat{\tau}}})$. Again, κ and N can be isolated such that the MDES and required sample size can be computed as a function of the other

[16] In these simulations, the underlying super populations were generated with normally distributed potential outcomes with means and variances according to the specifications described here.

parameters. The resulting conservative upper bounds on the MDES and required sample size are as follows:

$$\kappa \leq \frac{2M}{2\pi\sqrt{Np_z(1-p_z)} - M},$$
$$N \leq \frac{M^2(1+0.5\kappa)^2}{p_z(1-p_z)\kappa^2\pi^2},$$

where again $M = \Phi^{-1}(1-\frac{\alpha}{2}) + \Phi^{-1}(1-\beta)$.

5.2 Narrowing the Bounds While Relaxing the Equal Assignment Probability Assumption

As before, the lower bound of the power can be increased under Assumption 7 (ordered means). The result is the following alternative lower bound.

PROPOSITION 5. *Given Assumptions 1–5 and 7–8, and any value* $p_z = P(Z_i = 1)$:

$$\frac{\kappa\pi\sqrt{p_z(1-p_z)N}}{\sqrt{1+0.25\kappa^2}} \leq \frac{\tau}{\sqrt{V_N^{\hat{\tau}}}}.$$

Solving for κ and N to derive alternative (lowered) MDES and required sample size upper bounds leads to the following:

$$\kappa^* \leq \frac{2M}{\sqrt{4\pi^2 Np_z(1-p_z) - M^2}},$$
$$N^* \leq \frac{M^2(1+0.25\kappa^2)}{p_z(1-p_z)\kappa^2\pi^2},$$

where again $M = \Phi^{-1}(1-\frac{\alpha}{2}) + \Phi^{-1}(1-\beta)$.

Appendix E in the SM presents results comparing simulated power curves to the analytic bounds given $P(Z_i = 1) = 0.25$, similar to the results shown earlier in Figure 2. Appendix E demonstrates that the analytic bounds derived for the general case where $P(Z_i = 1) = p_z$ perform as well as the bounds derived for the special case of $P(Z_i = 1) = 0.5$.

6. OVERVIEW AND THE METHOD IN CONTEXT

Table 1 presents a summary of the main results for the LATE power analysis introduced in this study, providing the recommended formulas under the various scenarios considered. The formulas presume the use of the Wald IV estimator to test the null hypothesis that the LATE equals 0 with a two-sided alternative. Recall that the formulas were derived, without loss of generality, under the assumption that $\kappa > 0$.[17] The formulas in Table 1 provide the conservative values for each quantity of interest depending upon whether the probability of treatment assignment is equal or unequal and whether the ordered means assumption is met or not. This includes conservative values for the minimum detectable effect size ($\tilde{\kappa}$), required sample size ($\tilde{N}$), and the power ($\widetilde{1-\beta}$), all computed as a function of the other parameters, with $M = \Phi^{-1}(1-\frac{\alpha}{2}) + \Phi^{-1}(1-\beta)$ and $c^* = \Phi^{-1}(1-\frac{\alpha}{2})$.

In sum, to perform a conservative power analysis for the LATE, researchers should first identify which of the four cells in Table 1 best characterizes their particular study context. They can then compute their quantity of interest (e.g., required sample size) based on hypothetical values of the other parameters using the formulas provided in the table.

If uncertain whether or not the ordered means assumption is likely to be met, it is recommended that researchers operate as if the assumption is *not* met so as to err on the side of conservatism. Refer to the earlier discussion on the factors to consider when assessing the plausibility of the ordered means assumption. Also recall that in the case where noncompliance is one-sided or almost one-sided (i.e., no/few always-takers or no/few never-takers), the ordered means assumption can be simplified to $\bar{Y}_{NT} \leq \bar{Y}_C$ or $\bar{Y}_C \leq \bar{Y}_{AT}$ as long as the sparse principal stratum is deemed sufficiently small. In addition, researchers should refer to the earlier discussion on how effect sizes may be mapped to absolute effects. Finally, researchers should also note that corner cases exist whereby negative or nonreal numbers may be computed for $\tilde{\kappa}$, which as a practical matter correspond to prohibitively large effect sizes. Hence, if a negative or nonreal number is computed for $\tilde{\kappa}$, researchers should conclude that it will be impossible to detect the effect in the scenario under consideration.

To illustrate how the method of LATE power analysis presented in this study could be used, the method is applied to the context of the National JTPA Study (Bloom et al., 1997). As described earlier, subjects were randomly assigned such that they were either allowed to enroll in a job training program (assigned to treatment) or excluded from the training for an 18-month period (assigned to control). However, many subjects exhibited noncompliance: many assigned to the treatment decided not to enroll in the training program, while a few assigned to the control gained access to the job training program. The outcome of interest here is the subjects' earnings in the 30-month period following their random assignment.

For the purposes of this illustration, two different values of π will be employed. The first is its estimated value as observed in the JTPA data,[18] which is 0.63. This is, of

[17] If the effect is expected to have a negative value, researchers can simply treat κ as the absolute value of the effect size and continue using the same formulas.

[18] The dataset used here is the tabulation of the JTPA study data by Abadie, Angrist and Imbens (2002). The data correspond to adult participants in the JTPA experiment for whom 30-month earnings were measured.

TABLE 1
Summary of results: conservative formulas for LATE power analysis

	$P(Z=1)=0.5$ Equal assignment probability	$P(Z=1)\neq 0.5$ Unequal assignment probability
$\neg(\bar{Y}_{NT}\leq\bar{Y}_C\leq\bar{Y}_{AT})$ Ordered means not met	$\tilde{\kappa}:=\frac{2M}{\pi\sqrt{N}-2M\sqrt{(0.5-\frac{\pi}{2})(0.5+\frac{\pi}{2})}}$	$\tilde{\kappa}:=\frac{2M}{2\pi\sqrt{Np_z(1-p_z)}-M}$
	$\tilde{N}:=\frac{4M^2\left(1+\kappa\sqrt{(0.5-\frac{\pi}{2})(0.5+\frac{\pi}{2})}\right)^2}{\kappa^2\pi^2}$	$\tilde{N}:=\frac{M^2(1+0.5\kappa)^2}{p_z(1-p_z)\kappa^2\pi^2}$
	$\widetilde{1-\beta}:=\Phi\left(-c^*+\frac{0.5\kappa\pi\sqrt{N}}{1+\kappa\sqrt{(0.5-\frac{\pi}{2})(0.5+\frac{\pi}{2})}}\right)+\Phi\left(-c^*-\frac{0.5\kappa\pi\sqrt{N}}{1+\kappa\sqrt{(0.5-\frac{\pi}{2})(0.5+\frac{\pi}{2})}}\right)$	$\widetilde{1-\beta}:=\Phi\left(-c^*+\frac{\kappa\pi\sqrt{p_z(1-p_z)N}}{1+0.5\kappa}\right)+\Phi\left(-c^*-\frac{\kappa\pi\sqrt{p_z(1-p_z)N}}{1+0.5\kappa}\right)$
$\bar{Y}_{NT}\leq\bar{Y}_C\leq\bar{Y}_{AT}$ Ordered means met	$\tilde{\kappa}:=\frac{2M}{\sqrt{N\pi^2-4M^2(0.5-\frac{\pi}{2})(0.5+\frac{\pi}{2})}}$	$\tilde{\kappa}:=\frac{2M}{\sqrt{4\pi^2Np_z(1-p_z)-M^2}}$
	$\tilde{N}:=\frac{4M^2\left(1+\kappa^2(0.5-\frac{\pi}{2})(0.5+\frac{\pi}{2})\right)}{\kappa^2\pi^2}$	$\tilde{N}:=\frac{M^2(1+0.25\kappa^2)}{p_z(1-p_z)\kappa^2\pi^2}$
	$\widetilde{1-\beta}:=\Phi\left(-c^*+\frac{0.5\kappa\pi\sqrt{N}}{\sqrt{1+\kappa^2(0.5-\frac{\pi}{2})(0.5+\frac{\pi}{2})}}\right)+\Phi\left(-c^*-\frac{0.5\kappa\pi\sqrt{N}}{\sqrt{1+\kappa^2(0.5-\frac{\pi}{2})(0.5+\frac{\pi}{2})}}\right)$	$\widetilde{1-\beta}:=\Phi\left(-c^*+\frac{\kappa\pi\sqrt{p_z(1-p_z)N}}{\sqrt{1+0.25\kappa^2}}\right)+\Phi\left(-c^*-\frac{\kappa\pi\sqrt{p_z(1-p_z)N}}{\sqrt{1+0.25\kappa^2}}\right)$

course, not necessarily something the researcher would know precisely in advance, but it provides a useful point of reference. The second value for π will be 0.4, which we may view as a researcher's conservative guess prior to the actual study. We will fix p_Z at its observed value of 0.67, since this is a value over which the researcher has control, and hence the formulas in the far right column of Table 1 are applicable. We set α and β at their conventional levels of 0.05 and 0.2, respectively. We can then specify a range of effect sizes (κ's) to determine the conservative sample size required ($\tilde{N}$ in Table 1) under these specifications. Furthermore, because we know in retrospect the pooled within-assignment-group variance of the outcome (earnings), we can map the κ values to absolute effect values (τ's). Note also that in the JTPA experiment this value is virtually identical to $\widehat{\text{Var}}(Y_i|Z_i=0)$, which could have been estimated in advance of the study via a baseline survey given that assignment to control represents a natural state of the world in the absence of intervention.[19] As a result, the κ values could have been mapped to τ values even in the absence of retrospective data, to the benefit of implementing the power analysis.

The results given $\pi=0.63$ are shown in Table 2, with the conservative recommendation for the required sample size, $\tilde{N}$, provided with and without making Assumption 7 (ordered means). For instance, given a desired effect size of 0.1, the conservative sample size recommendation would be approximately 10,000 observations to achieve a level of power of 0.8 to reject the null hypothesis that $\tau=0$ without making Assumption 7, while it would be approximately 9000 observations given Assumption 7. The actual LATE effect size estimate in the pooled adult

TABLE 2
LATE Power Analysis, Given $\pi=0.63$ *and* $p_Z=0.67$

κ	τ	Recommended $\tilde{N}$ without ordered means assumption	Recommended $\tilde{N}$ with ordered means assumption
0.05	837.94	37,588	35,799
0.10	1675.89	9861	8966
0.15	2513.83	4594	3998
0.20	3351.78	2706	2258
0.25	4189.72	1811	1453
0.30	5027.66	1314	1016
0.35	5865.61	1008	752
0.40	6703.55	805	581
0.45	7541.50	663	464
0.50	8379.44	559	380

[19] $\sqrt{\widehat{E[\text{Var}(Y_i|Z_i)]}}=16{,}759$, while $\sqrt{\widehat{\text{Var}(Y_i|Z_i=0)}}=16{,}180$, a difference of about 3%.

TABLE 3
LATE Power Analysis, Given $\pi = 0.4$ *and* $p_Z = 0.67$

κ	τ	Recommended $\tilde{N}$ without ordered means assumption	Recommended $\tilde{N}$ with ordered means assumption
0.05	837.94	93,241	88,804
0.10	1675.89	24,461	22,242
0.15	2513.83	11,395	9916
0.20	3351.78	6712	5602
0.25	4189.72	4493	3605
0.30	5027.66	3260	2521
0.35	5865.61	2501	1867
0.40	6703.55	1997	1442
0.45	7541.50	1644	1151
0.50	8379.44	1387	943

sample was 0.11.[20] Thus, it is no surprise that given the actual sample size of 11,204 adult participants and the fact that the ordered means assumption was ultimately met in this study, the LATE estimate in the study is statistically significant ($p < 0.001$). While 0.11 would generally be considered a relatively small effect size—according to the rough guidance presented by Cohen (1988), 0.2, 0.5 and 0.8 are benchmarks for small, medium and large effect sizes in the social and behavioral sciences—the JTPA study was of sufficiently large scale to detect this effect in the pooled adult sample.

Table 3 displays the results given $\pi = 0.4$. As shown, an increase in the amount of noncompliance leads to a disproportionately large increase in the sample size requirements. While noncompliance is assumed to increase by a factor of about 1.6, the required sample size given any particular κ increases by a factor of about 2.5. As these results show, had the compliance rate π actually been 0.4, it is likely that the JTPA study would have failed to find a statistically significant effect of the training program on earnings, even in the pooled adult sample. The method of LATE power analysis presented in this study is designed to alert researchers to such possibilities of under-powered designs before studies are launched without requiring researchers to make the collection of strong assumptions involved in other approaches to LATE power analysis.

7. POWER WITH COVARIATES

The standard LATE assumptions establish the consistency of the Wald IV estimator without covariate adjustment, but covariates can still be used to improve the precision of the estimates. As a result, researchers sometimes employ covariate adjustment in order to attain a more powerful LATE estimator. A common approach is to use linear two-stage least squares (2SLS), which is equivalent to modeling and estimating linear first-stage and intent-to-treat relationships (Angrist and Pischke, 2009, pp. 120-122):

$$D_i = \mathbf{W}_i \boldsymbol{\eta} + \pi Z_i + \nu_i^*, \tag{1}$$

$$Y_i = \mathbf{W}_i \boldsymbol{\xi} + \gamma Z_i + \zeta_i^*, \tag{2}$$

where $\mathbf{W}_i$ corresponds to a set of covariates, as well as an intercept. Provided that the covariates contained in $\mathbf{W}_i$ are pretreatment-assignment covariates—that is, they are independent of Z_i and hence do not result in biased estimates of π and γ—then the LATE can be estimated consistently by $\frac{\hat{\gamma}}{\hat{\pi}}$, where $\hat{\gamma}$ and $\hat{\pi}$ are the linear least squares estimators. In addition, if $\mathbf{W}$ helps to explain variation in D and/or Y that is left unexplained by Z, then the covariate adjustment can also decrease the variance of $\frac{\hat{\gamma}}{\hat{\pi}}$. As a result, linear 2SLS with covariate adjustment has the potential to offer a more powerful estimator of the LATE, and the method presented in this study can be extended to incorporate these gains.

DEFINITION 2. Define the following:

$$R_{DW}^2 = \frac{\sigma^2 - \sigma^{*2}}{\sigma^2} \quad \text{and} \quad R_{YW}^2 = \frac{\omega^2 - \omega^{*2}}{\omega^2},$$

where $\sigma^2 = E[\nu_i^2]$ as defined in the proof of Proposition 1, $\omega^2 = E[\zeta_i^2]$ as defined in the proof of Proposition 1, $\sigma^{*2} = E[\nu_i^{*2}]$ from equation (1), and $\omega^{*2} = E[\zeta_i^{*2}]$ from equation (2).

R_{DW}^2 measures the proportion of variation in D left unexplained by Z that is explained by the covariates contained in $\mathbf{W}$, while R_{YW}^2 measures the proportion of variation in Y left unexplained by Z that is explained by the covariates contained in $\mathbf{W}$. Given Definition 2, covariate adjustment in the 2SLS framework can be employed to yield the following bounds for use in the power formula (continuing to assume that $\kappa > 0$ and $\tau > 0$ are under investigation):

$$\begin{aligned}
&(0.5\kappa\pi\sqrt{N}) \\
&\quad /\Big(\Big((1 - R_{YW}^2) + \kappa^2(1 - R_{DW}^2)E[\nu_i^2] \\
&\quad + 2\kappa\sqrt{(1 - R_{YW}^2)(1 - R_{DW}^2)E[\nu_i^2]}\Big)^{1/2}\Big) \\
&\le \frac{\tau}{\sqrt{\widehat{V_N^{2SLS}}}} \\
&\le (0.5\kappa\pi\sqrt{N}) \\
&\quad /\Big(\Big((1 - R_{YW}^2) + \kappa^2(1 - R_{DW}^2)E[\nu_i^2] \\
&\quad - 2\kappa\sqrt{(1 - R_{YW}^2)(1 - R_{DW}^2)E[\nu_i^2]}\Big)^{1/2}\Big).
\end{aligned}$$

[20] The estimate of the LATE of the training program on earnings is \$1849. This divided by the expected within-assignment-group standard deviation of earnings in the sample, 16,759, yields an effect size of 0.11.

As previously, this formula can be modified to both relax Assumption 6 (equal assignment probability) and employ Assumption 7 (ordered means), and $E[\nu_i^2]$ replaced with either $(0.5 - \frac{\pi}{2})(0.5 + \frac{\pi}{2})$ or 0.25 depending on whether Assumption 6 is made. More detail is provided in the SM Appendix F. It must be emphasized that the results described in this section only apply given the standard LATE assumptions (1–5) as well as independence between Z and $\mathbf{W}$. In other words, the assumptions necessary for the consistency of the estimator must be met without covariate adjustment, the purpose of covariate adjustment must simply be to decrease the variance of the estimator, and the covariates must not be affected by Z.

8. POWER WITH VARIABLE TREATMENTS

In cases where the endogenous treatment is no longer binary but rather has variable intensity (e.g. drug dosage, years of schooling), Angrist and Imbens (1995) have shown that the Wald IV estimator can still be used under Assumptions 1–5. In this case, however, the Wald IV estimator is consistent for a new estimand they call the average causal response (ACR), which is "a weighted average of causal responses to a unit change in treatment, for those whose treatment status is affected by the instrument" (p. 435). In other words, like with the LATE, the estimand only pertains to those subjects for whom the instrument has a nonzero effect on treatment uptake/dosage, but the ACR is a weighted average rather than a simple average of the individual-level causal effects of the treatment on the outcome.[21]

In spite of the modified estimand, the general properties of the Wald IV estimator, including its variance, remain the same. Furthermore, the assumption of a binary treatment is not critical in the derivation of the power formulas introduced in this study. The binary treatment assumption was employed in determining values for $E[\nu_i^2]$, but a linear rescaling of a multivalued treatment to the interval $[0, 1]$ would mean the conservative value of $E[\nu_i^2] = 0.25$ would remain valid. As arbitrary linear transformations of variables do not affect statistical power, the method of power analysis presented in this study can also be applied to variable treatments.[22] Yet the researcher must keep in mind that given a variable treatment, the estimand that is identified is the ACR rather than the LATE, and π can no longer be interpreted simply as the compliance rate.

[21] See Angrist and Imbens (1995) for more details on the weighting formula.

[22] Intuitively, this rescaling would not affect the power, even though it would mean a rescaling of $E[\nu_i^2]$, because it would result in a commensurate rescaling of π.

9. CONCLUSION

This study proposed a new approach to power analysis in the LATE context that makes three important contributions. First, in contrast to previous approaches, it does not involve distributional assumptions about the various principal strata. Second, and most importantly, it provides a tight lower bound on the power while removing the need to specify or make assumptions about variance components or distributional heterogeneity across the principal strata. Third, it shows how additional assumptions can be made to raise the lower bound to better balance conservatism with efficiency.

By providing bounds on the power that are free of distributional assumptions, this study introduces a reliable and disciplined way of computing power conservatively without the inefficiencies of other approaches (e.g., setting arbitrarily high variances) that can lead to excessively conservative calculations. The result is a generalized approach to power analysis in the LATE context that is simultaneously conservative, disciplined and simple to implement.

ACKNOWLEDGMENTS

The author thanks Avidit Acharya, Matthew Blackwell, Justin Grimmer, Jens Hainmueller, Andy Hall, Dominik Hangartner, Kosuke Imai, Guido Imbens, Mike Tomz, Stefan Wager, Teppei Yamamoto and Xiang Zhou for helpful comments.

SUPPLEMENTARY MATERIAL

Supplement to "A Generalized Approach to Power Analysis for Local Average Treatment Effects" (DOI: 10.1214/19-STS732SUPP; .pdf). The supplementary materials contain the proofs of all propositions along with additional results referenced in the main text.

REFERENCES

ABADIE, A., ANGRIST, J. and IMBENS, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica* **70** 91–117. MR1926256 https://doi.org/10.1111/1468-0262.00270

ANGRIST, J. D. (1990). Lifetime earnings and the Vietnam era draft lottery: Evidence from Social Security administrative records. *Am. Econ. Rev.* **80** 313–336.

ANGRIST, J. D. and IMBENS, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *J. Amer. Statist. Assoc.* **90** 431–442. MR1340501

ANGRIST, J. D., IMBENS, G. W. and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **91** 444–455.

ANGRIST, J. D. and PISCHKE, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton Univ. Press, Princeton, NJ.

BAIOCCHI, M., CHENG, J. and SMALL, D. S. (2014). Instrumental variable methods for causal inference. *Stat. Med.* **33** 2297–2340. MR3257582 https://doi.org/10.1002/sim.6128

BANSAK, K. (2020). Supplement to "A Generalized Approach to Power Analysis for Local Average Treatment Effects." https://doi.org/10.1214/19-STS732SUPP.

BLOOM, H. S. (1995). Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation Review* **19** 547–556.

BLOOM, H. S. (2006). The core analytics of randomized experiments for social research. Technical report, MDRC.

BLOOM, H. S., ORR, L. L., BELL, S. H., CAVE, G., DOOLITTLE, F., LIN, W. and BOS, J. M. (1997). The benefits and costs of JTPA Title II-A programs: Key findings from the National Job Training Partnership Act Study. *J. Hum. Resour.* **32** 549–576.

BRION, M.-J. A., SHAKHBAZOV, K. and VISSCHER, P. M. (2013). Calculating statistical power in Mendelian randomization studies. *Int. J. Epidemiol.* **42** 1497–1501.

COHEN, J. (1962). The statistical power of abnormal-social psychological research. *J. Abnorm. Soc. Psychol.* **65** 145–153.

COHEN, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ.

DUFLO, E., GLENNERSTER, R. and KREMER, M. (2007). Using randomization in development economics research: A toolkit. Technical report, Centre for Economic Policy Research, London.

FREEMAN, G., COWLING, B. J. and SCHOOLING, C. M. (2013). Power and sample size calculations for Mendelian randomization studies using one genetic instrument. *Int. J. Epidemiol.* **42** 1157–1163.

GERBER, A. S. and GREEN, D. P. (2000). The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment. *Am. Polit. Sci. Rev.* **94** 653–663.

GERBER, A. S. and GREEN, D. P. (2012). *Field Experiments*: *Design, Analysis, and Interpretation*. W. W. Norton & Company, New York.

HAHN, J., TODD, P. and VAN DER KLAAUW, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* **69** 201–209.

HAINMUELLER, J., HANGARTNER, D. and PIETRANTUONO, G. (2016). Naturalization fosters the long-term political integration of immigrants. *Proc. Natl. Acad. Sci. USA* **112** 12651–12656.

HIRANO, K., IMBENS, G. W., RUBIN, D. B. and ZHOU, X.-H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics* **1** 69–88.

IMBENS, G. W. (2014a). Instrumental variables: An econometrician's perspective. *Statist. Sci.* **29** 323–358. MR3264545 https://doi.org/10.1214/14-STS480

IMBENS, G. (2014b). Rejoinder: "Instrumental variables: An econometrician's perspective" [MR3264546; MR3264547; MR3264548; MR3264549; MR3264545]. *Statist. Sci.* **29** 375–379. MR3264550 https://doi.org/10.1214/14-STS496

IMBENS, G. W. and ANGRIST, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62** 467–475.

IMBENS, G. W. and RUBIN, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *Ann. Statist.* **25** 305–327. MR1429927 https://doi.org/10.1214/aos/1034276631

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*: *An Introduction*. Cambridge Univ. Press, New York. MR3309951 https://doi.org/10.1017/CBO9781139025751

JO, B. (2002). Statistical power in randomized intervention studies with noncompliance. *Psychol. Methods* **7** 178–193.

KITAGAWA, T. (2014). Instrumental variables before and LATEr [discussion of MR3264545]. *Statist. Sci.* **29** 359–362. MR3264546 https://doi.org/10.1214/14-STS494

LIPSEY, M. W. (1990). *Design Sensitivity*: *Statistical Power for Experimental Research*. Sage Publications, Newbury Park, CA.

NEYMAN, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych* **10** 1–51.

PIERCE, B. L., AHSAN, H. and VANDERWEELE, T. J. (2011). Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. *Int. J. Epidemiol.* **40** 740–752.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688.

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58. MR0472152

RUBIN, D. B. (1980). Comment: Randomization analysis of experimental data: The Fisher randomization test. *J. Amer. Statist. Assoc.* **75** 591–593.

RUBIN, D. B. (1990). Comment on J. Neyman and causal inference in experiments and observational studies: "On the application of probability theory to agricultural experiments. Essay on principles. Section 9" [Ann. Agric. Sci. **10** (1923), 1–51]. *Statist. Sci.* **5** 472–480. MR1092987

SWANSON, S. A. and HERNÁN, M. A. (2014). Think globally, act globally: An epidemiologist's perspective on instrumental variable estimation [discussion of MR3264545]. *Statist. Sci.* **29** 371–374. MR3264549 https://doi.org/10.1214/14-STS491

TSANG, R., COLLEY, L. and LYND, L. D. (2009). Inadequate statistical power to detect clinically significant differences in adverse event rates in randomized controlled trials. *J. Clin. Epidemiol.* **62** 609–616.

TVERSKY, A. and KAHNEMAN, D. (1971). The belief in the law of small numbers. *Psychol. Bull.* **76** 105–110.

WANG, X., JIANG, Y., ZHANG, N. R. and SMALL, D. S. (2018). Sensitivity analysis and power for instrumental variable studies. *Biometrics* **74** 1150–1160. MR3908133