Beyond generalization of the ATE: Designing randomized trials to understand treatment effect heterogeneity

Elizabeth Tipton

Northwestern University

April 2020

Abstract:

Researchers conducting randomized trials have increasingly shifted focus from the average treatment effect to understanding moderators of this effect. Current methods for exploring moderation focus on model selection and hypothesis tests. At the same time, recent developments in the design of randomized trials have argued for the need for population-based recruitment in order to generalize well. In this paper, we show that a different population-based recruitment strategy can be implemented to increase the precision of estimates of treatment effect moderators, and we explore the trade-offs between optimal designs for the average treatment effect and moderator effects.

In education and social welfare, randomized trials are increasingly used to evaluate the effects of interventions; for example, in the U.S., over the past 15 years, over 300 field experiments have been funded by the Institute of Education Sciences alone (Chhin, Taylor, & Wei, 2018). These experiments are typically designed to answer questions regarding the average treatment effect (ATE) of interventions, and these ATEs are then provided to practitioners making curricular and policy decisions. As this policy-making goal has become more prominent, so too have concerns with the generalizability of results from these experiments to policy relevant populations, given that the recruitment of sites into these studies is based nearly entirely on convenience (Olsen, Orr, Bell, & Stuart, 2013). Recent studies indicate that the samples included in these experiments are far from representative of clearly defined, policy-relevant target populations (e.g., Fellers, 2017; Stuart, Olsen, Bell, & Orr, 2012, Tipton et al 2016). Furthermore, studies suggest that the ATE estimated in these samples can exhibit large biases when estimating the ATE in target populations (e.g., Stuart, Olsen, Bell, and Orr, 2012).

Over the past decade, statisticians have begun to address sample selection bias using a range of strategies. One approach involves combining data from an already-conducted experiment with representative data from a target population using estimators that, under key assumptions, provide unbiased estimates of the population ATE. These include propensity score reweighting estimators (Stuart, Cole, Bradshaw, & Leaf, 2011; Tipton, 2013) as well as regression, machine-learning, and bounding approaches (Chan, 2017; Kern, Stuart, Hill, & Green, 2016). When there is under-coverage – whereby certain subgroups are observed in the population but not in the sample – these estimators can require severe extrapolations (Tipton, 2013).

Given these concerns, another area of statistics focuses instead on improving generalizations through improved sample selection (Olsen, Bell, & Nichols, 2018; Tipton, 2014b; Tipton et al., 2014; White, Rowan, Hansen, & Lycurgus, 2019). Tipton and colleagues provide a "bias-robust" approach in which a target population is divided into strata based upon a large set of potential site-level moderators. Dimension reduction methods (e.g., *k*-means) are used to reduce the number of strata, and the total sample required is allocated proportionally to these strata. Tipton and Olsen (2018) note that within-stratum recruitment can proceed either probabilistically or based upon convenience. While probabilistic selection is statistically ideal, it is not typically feasible in practice, since recruiters operate under resource constraints and response-rates are often low (for an exception, see Yeager et al., 2019). They show, however, that the use of stratification alone increases the similarity between the sample and target population on the set of moderators used in the stratification design; thus, under the assumption that these are *all* of the site-level moderators ('sampling ignorability'), the sample ATE will be an unbiased estimate of the population ATE.

Importantly, this sample selection approach requires that researchers *know* which variables moderate the effect of the intervention. Unfortunately, to date little is collectively known regarding sources of treatment effect variation across studies. Certainly, one reason for this dearth of information could be that tests of site-level moderators can have less power than that of ATEs (Dong, Kelcey, & Spybrook, 2018; Spybrook, Kelcey, & Dong, 2016). But power is a function not only of statistical precision, but also of the size of the interaction between the treatment and moderator itself – e.g., the difference in average treatment effects across subgroups – which is not necessarily small (e.g., Yeager et al., 2019) . Furthermore, as we will show in this paper, the size of this interaction is affected *both* by the degree of treatment effect heterogeneity

in the population (which affects the parameter) *and* the degree of heterogeneity in the moderator in the sample (which affects the estimator). Thus, we argue that statistical power for moderators is not immutable, but can be greatly increased with careful attention to sample selection and study design.

In this paper, we provide a framework for guiding site selection in field experiments so as to precisely estimate and test hypotheses regarding *both* the ATE and site-level moderators of intervention effects in a target population. To do so, we begin by focusing on statistical power for moderators, showing how sample heterogeneity can affect both the parameter and estimator. We then review the literature on optimal experimental design in multi-factor studies, linking these results to the ideal designs for testing effect size moderation in field experiments. Since this approach, while optimal for estimation of moderators, is typically not optimal for estimation of the population ATE, we propose an approach that offers a compromise between these designs. In order to illustrate how this could be used in actual studies, we provide an example based upon a previous evaluation of *Success for All*, an elementary school reading curriculum. The paper concludes with a discussion of both the benefits and limitations of the approach developed here.

Effect size difference and standardization

We begin by defining an experiment in which persons $i = 1, \ldots, n_j$ are nested in sites $j = 1, \ldots, J$, letting T_j indicate if a site is randomly assigned to the treatment condition (versus a control condition). Let $Y_{ij}(0)$ and $Y_{ij}(1)$ be the potential outcomes for person *i* in site *j* and assume for now there is a single site-level moderator, X_j (with $\mathrm{E}(X_j) = 0$). We can thus define the potential outcomes under each condition as,

$$Y_{ij}(T_j = 0) = \gamma_{00} + \gamma_{10} X_j \quad (1)$$

$$Y_{ij}(T_j = 1) = \gamma_{01} + \gamma_{11} X_j$$

and the effect of the treatment on person *i* in site *j* as,

$$\Delta_{ij} = Y_{ij}(1) - Y_{ij}(0) = (\gamma_{01} - \gamma_{00}) + (\gamma_{11} - \gamma_{10})X_j = \gamma_1 + \gamma_3 X_j. \quad (2)$$

Here the ATE of the intervention is $\gamma_1 = \gamma_{01} - \gamma_{00}$, which occurs when $E(X_j) = 0$, and the effect of the site-level moderator on the intervention effect is $\gamma_3 = \gamma_{11} - \gamma_{10}$. Note that Δ_{ij} is never observed, since in an experiment, persons (in randomized block designs) or sites (in cluster randomized designs) are randomized to treatment. However, as a result of random assignment, the moderator X_j is independent of treatment T_j. Furthermore, the potential outcomes ($Y_{ij}(0)$, $Y_{ij}(1)$) are independent of all other confounders conditional on random assignment T_j and the moderator X_j.

In cluster randomized trials – our focus in this paper – these parameters γ_1 and γ_3 can be estimated using a linear model with a treatment interaction,

$$Y_{ij} = \gamma_0 + \gamma_1 T_j + \gamma_2 X_j + \gamma_3 T_j * X_j + r_j + e_{ij} \quad (3)$$

where Y_{ij} is the observed outcome, T_j is –1/2 if assigned to the control condition and +1/2 if assigned to the treatment condition, $e_{ij} \sim N(0, \sigma^2)$ and $r_j \sim N(0, \tau^2)$. This is the general framework used in most cluster-randomized trials for both estimating and testing a treatment interaction.

Importantly, if we assume that the sample is selected in order to represent the population (probabilistically or otherwise, with assumptions), as a result of random assignment, γ_1 is the average *causal* effect of the intervention in the population. Furthermore, γ_3 is the difference in causal effects of the intervention for every 1-unit change in X_j in the population. However, since X_j is not randomly assigned, the treatment by moderator effect γ_3 should not in general be interpreted causally. This does not render the moderator effect meaningless, however, as it provides important descriptive information on the stability of the causal effect in the population

(for a longer discussion of causality and moderators, see Hong, 2015; VanderWeele, 2015). For example, whether causal or not, it is helpful to know if the effect of an intervention is larger in the schools with more (versus fewer) resources or in those serving primarily minority (versus majority) students.

In order to make results of a trial more comparable across outcomes, the ATE is typically reported in standard deviation units. That is, the ATE of focus is the standardized parameter

$$\delta_m = \frac{\gamma_1}{\sqrt{(\sigma^2 + \tau^2)}}, \tag{4}$$

which is interpreted in relation to the population standard deviation of the outcome Y. By standardizing the effect size for the ATE, it is easier to interpret the relative size of a treatment impact compared to other treatments or population characteristics (e.g., Hill, Bloom, Black, & Lipsey, 2008). Dong and colleagues (2018) show that we can similarly standardize the moderator by treatment interaction into the *standardized effect size difference* (SESD),

$$\delta_{ds} = \gamma_3 \sqrt{\frac{\sigma_x^2}{\sigma^2 + \tau^2}}, \tag{5}$$

where σ_x is the standard deviation of X. Now δ_{ds} can be interpreted as the effect of a 1 standard deviation change in X on the standardized treatment effect.

Here an important question is which standard deviation σ_x should be used for standardization. To date, in the literature on power analysis in randomized trials, the *sample* standard deviation is used. However, we argue that – following the literature on generalization – the *population* standard deviation is more appropriate. This is particularly important when power analyses are to be conducted *a priori*, since population standard deviations are available for many populations and site-level moderators in education and social welfare (e.g., the *Common Core of Data*, NCES). Additionally, this is important when SESDs are compared across studies.

Standardizing in relation to the population allows the size of the interaction in the population to be separated clearly from estimation of this interaction.

This distinction is particularly important for categorical moderators. At the site level, these often include dichotomized versions of underlying continuous variables (e.g., low-SES, high-minority schools). There are two issues to keep in mind here. First, as Gelman (2008) discusses, the dichotomized difference δ_{dc} is actually equivalent to a *two* standard deviation[1] change in a continuous moderator δ_{ds}. Second, *how* an underlying continuous moderator is divided into categories also affects the size of the effect size difference δ_{dc}. To see how, imagine the simple case of urbanicity. Here the contrast could be between "urban" and "rural" (excluding the intermediary "suburb" and "town") or between "urban + suburb" and "town + rural" (as occurs when researchers collapse across categories). If treatment impacts are a function of population density (an underlying continuous variable), then clearly the first operationalization based on the extremes will lead to a larger δ_{dc} than the latter. As with continuous moderators, interpretation is thus improved when the levels of the categorical variable are labelled in relation to the population, not the sample.

Throughout the remainder of this paper, for sake of generality, we will assume that continuous covariates are standardized in relation to the *population* standard deviation, which without loss of generality, we will set to $\sigma_x = 1$. We will assume, however, that the variation in the *sample*, S_x, may be larger or smaller than this population standard deviation. Since we

[1] To see why, note that if there are two categories with the proportion Q in the smallest group, the standard deviation of this variable is $\sqrt{Q(1-Q)}$. Thus, when these groups are equally allocated (in the population), this standard deviation is 0.50, half that of a standardized continuous moderator. Gelman argues for *doubling* the coefficient δ_{ds} of the continuous moderators so as to be on the same scale as the categorical moderators; conversely, one could *half* the coefficient δ_{dc} of categorical moderators to make the scales equivalent (which is how we will proceed in this paper).

assume that categorical moderators can be related to these continuous moderators via standardization as well (in relation to the population), we will focus throughout on the general case of continuous moderators.

Benchmarking the SESD

Standardizing in relation to the population is helpful, but without benchmarks akin to those found for the ATE, determining if an effect is "meaningful" is difficult. This is particularly important when determining an appropriate SESD to use in power analyses. Here we propose an approach for benchmarking that relates the SESD to the ATE and the distribution of underlying treatment impacts.

Assume that there is a single standard normal covariate $X \sim N(0,1)$ in the population, that the ATE is δ_m, and δ_{ds} is the SESD. Then we have

$$\bar{\delta}_i = E(\delta_i|X) = \delta_m + \delta_{ds}X \sim N(\delta_m, \delta_{ds}) \quad (6)$$

and if the ATE $\delta_m > 0$ then the treatment impact is *positive* for $100(1-\alpha)\%$ of the population when

$$\delta_{ds} < \frac{\delta_m}{z_\alpha} \quad (7)$$

where z_α is a critical value from the standard normal for a given α–level. For example, if we desire an intervention in which the treatment effect is positive not only on average, but for at least 84% of the population as well (i.e., $z_\alpha = 1$ for $\alpha = 0.16$), then we are interested in cases in which $\delta_{ds} < \delta_m$. A more relaxed criteria may be that we want to ensure that the effect is positive for 75% of the population, in which case $z_\alpha \approx 2/3$ and thus we need $\delta_{ds} < 1.5\delta_m$. On the other extreme, we may want to ensure that a treatment is not only effective on average, but also for *nearly all* sites in the population; in this case, we might focus on cases in which the effect is positive for 95% of the sites, in which case $z_\alpha = 1.64$ and $\delta_{ds} < \delta_m/1.64$. We are not arguing

that this is the definitive approach to interpreting the SESD, as certainly empirical information from existing studies would be more useful. However, the benefit of this approach is that it allows us to anchor our interpretation of the effect of a moderator in relation to the ATE itself.

Sample variation and the MDESD for moderators

Our focus is now on estimating and testing hypotheses regarding γ_3 in its population standardized form δ_{ds}. General properties of these estimators can be found in Raudenbush and Bryk (2002) with applications to cluster-randomized trials and methods for power analysis found in Dong, Kelcey, and Spybrook (2018) and Spybrook, Kelcey, and Dong (2016). These power analysis methods are integrated into free software (see Dong, Kelcey, Spybrook, & Maynard, 2007) and researchers are encouraged to explore power for tests of moderation when designing trials (e.g., Institute of Education Sciences, 2018).

The easiest framework for exploring power is via the minimum detectable effect size difference (MDESD). For the model given in Equation (3), the population MDESD can be written (Dong et al., 2018),

$$MDESD_p(|\delta_{ds}|) = M_v \sqrt{\frac{\left(1 - R^2_{|X}\right)\rho n + (1-\rho)}{P(1-P)\sigma_x^2 \nu n}} \tag{8}$$

where $M_v = t_{\alpha/2} + t_{1-\beta}$ is for two-tailed tests, with J sites, n units within each site, degrees of freedom $\nu = J - 4$, and in which P is the proportion of the sample in treatment, $\rho = \tau^2/(\sigma^2 + \tau^2)$ is the intra-class correlation, $R^2_{|X} = 1 - \frac{\tau^2_{|X}}{\tau^2}$ is the proportion of between-school variation explained by the covariate X and interaction $X * T$, and $\sigma_x^2 = 1$ is variation in the covariate X in the population.

When the sample is randomly selected from the population (the case that Dong and colleagues focus on), the variation in X in the sample is likely to be close to that in the population (i.e., $S_x \approx \sigma_x$). More generally, however, we can define the *sample* MDESD as,

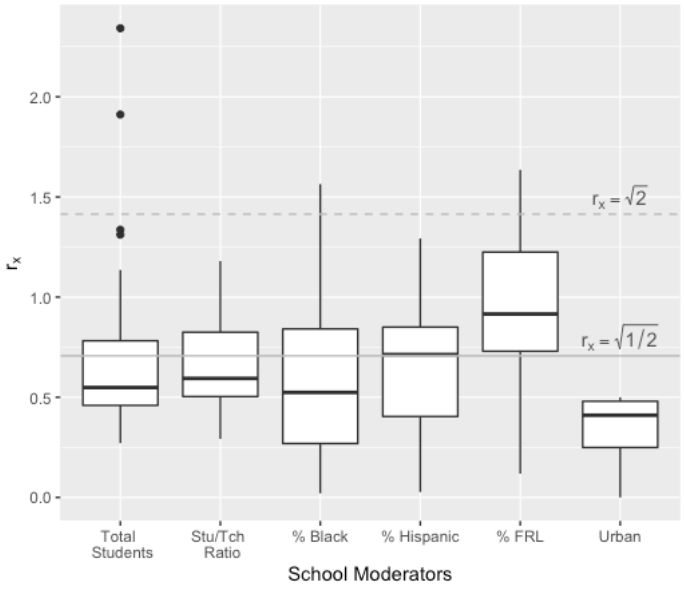
$$MDESD(|\delta_{ds}|) = \frac{MDESD_p(|\delta_{ds}|)}{r_x} \quad (9)$$

where $r_x = S_x/\sigma_x$ is the ratio of the sample standard deviation of X to the population standard deviation. The above relationship makes clear how sample variation can affect the MDESD. When the sample is more heterogenous than the population ($r_x > 1$) the MDESD *decreases* (and thus power increases), while when the sample is more homogenous ($r_x < 1$) the MDESD *increases* (and power decreases). Thus, moving from a very homogenous sample (e.g., $r_x = \sqrt{1/2}$) to a very heterogenous sample (e.g., $r_x = \sqrt{2}$) decreases the MDESD by *half* (and thus, equivalently, increases power). More importantly, the effect of r_x^2 on the MDESD is proportional to the effect of the degrees of freedom ($\nu = J - 4$). Thus, a representative sample (e.g., $r_x = 1$) with $J = 20$ (and $\nu = 16$) sites has an equivalent MDESD to a homogenous sample (e.g., $r_x = \sqrt{1/2}$) with $J = (20 - 4) * 2 + 4 = 36$ and $\nu = 32$) sites.

An important question is the extent to which this actually matters in practice. To explore this, we turn to a review of 34 cluster-randomized trials conducted between 2011 – 2015 that were funded by the Institute of Education Sciences (Spybrook, Wang, & Tipton, 2019). For each study, the ratio of the sample to population standard deviation was calculated for each of six site-level moderators. In Figure 1, results are aggregated across the 34 studies for each moderator. In the vast majority of studies, most samples are more homogeneous than the population (i.e., $r_x = \sqrt{.5}$, the bottom dashed line). Overall this means that the samples of convenience that are

standard in field experiments in education not only make generalizing to a population ATE difficult, but they also jeopardize our ability to explore treatment effect heterogeneity clearly.

Figure 1. Ratio of sample to population SDs (r_x) in 34 RCTs



Designing a field experiment to test moderators

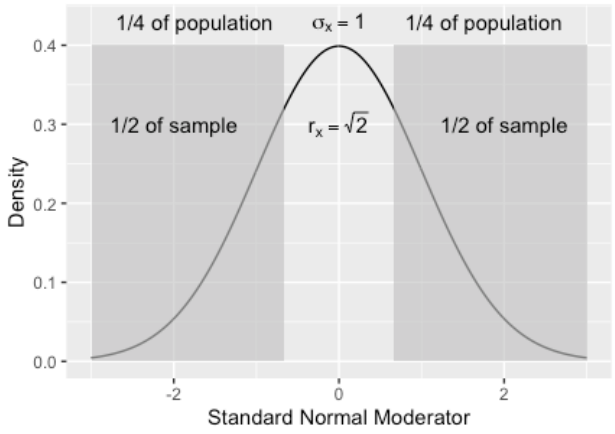
In this section, we develop a strategy to increase the heterogeneity of moderators in samples used in field experiments. We assume that researchers have available to them data on the target population, including information on moderators. In education research in the U.S., the *Common Core of Data* (NCES) provides many demographic moderators for most public schools annually. In social welfare, Tipton and Peck (2017) show that the *American Community Survey* and other Bureau of Labor Statistics datasets can be useful for these purposes.

Begin by assuming that we have a single continuous moderator of interest, X_j. Our focus is on estimation of the treatment by covariate interaction, γ_3. Recall that $\gamma_3 = \gamma_{11} - \gamma_{10}$, the difference in relationships between X_j and the outcomes Y_{ij} under treatment and control respectively. Since the variance of this interaction is an additive function of a difference (i.e.,

$V(\hat{\gamma}_3) = V(\hat{\gamma}_{11}) + V(\hat{\gamma}_{10})$), a design that minimizes variance for estimation of γ_{11} and γ_{10} will also be optimal for an estimate of the difference, γ_3.

While to date, there has been no discussion of optimal designs for sample selection for treatment effect moderators (γ_3) , there is a long history of optimal designs for estimation of the relationships between *factors* and outcomes (e.g., γ_{10}) in laboratory experiments. If we assume that the covariate X_j is normally distributed, Feldt (1961) shows an *extreme groups approach* (*EGA*) – in which each of two dichotomized groups of X_j contain between 22-25% of the population – is optimal. In Figure 2, we illustrate this, assuming that the two extreme groups are defined to include 25% of the population[2] (i.e., $|X| > 0.68$). Note that this results in a sample in which $r_x = \sqrt{2}$.

Figure 2. Optimal allocation of sample for a standard normal moderator



In practice, interventions are likely to be moderated by more than a single covariate, with the simplest case involving two continuous moderators. We can now extend Equation (1) to

[2] Importantly, selecting the sample from the extremes of the distribution of *X* does not mean that the covariate will be included in the model in its dichotomized form. To the contrary, the model has both greater power and interpretability when *X* is included as a continuous covariate (Cohen, 1983; Preacher, Rucker, MacCallum, & Nicewander, 2005).

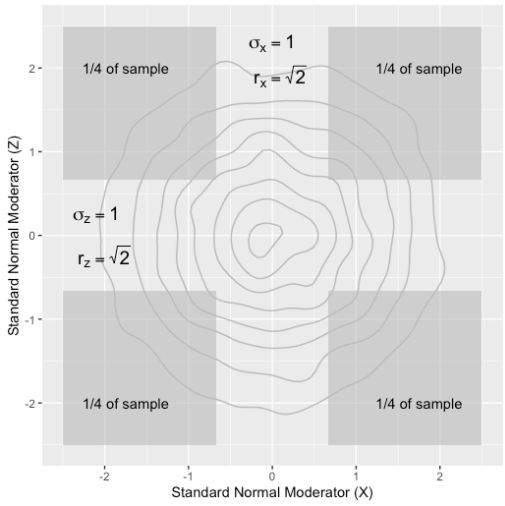
include an additional site-level moderator, Z, in which case the potential outcomes in the two groups can be written,

$$Y_{ij}(T_j = 0) = \gamma_{00} + \gamma_{10}X_j + \gamma_{20}Z_j \quad (10)$$

$$Y_{ij}(T_j = 1) = \gamma_{01} + \gamma_{11}X_j + \gamma_{21}Z_j$$

with interest now in the differences $\gamma_3 = \gamma_{11} - \gamma_{10}$ and $\gamma_5 = \gamma_{21} - \gamma_{20}$. Again, as a result of random assignment and the additive property of variances, the design that is optimal for estimation of the individual slopes is also optimal for estimation of these differences (i.e., interactions). McClelland and Judd (1993) explore this two-variable case and show that the optimal design is one in which (a) the two factors are orthogonal and (b) the variances of each covariate are maximized. This amounts to dividing the total sample evenly across these four corners. In Figure 3 we illustrate this when X and Z follow independent standard normal distributions. The result is $r_x = r_z = \sqrt{2}$.

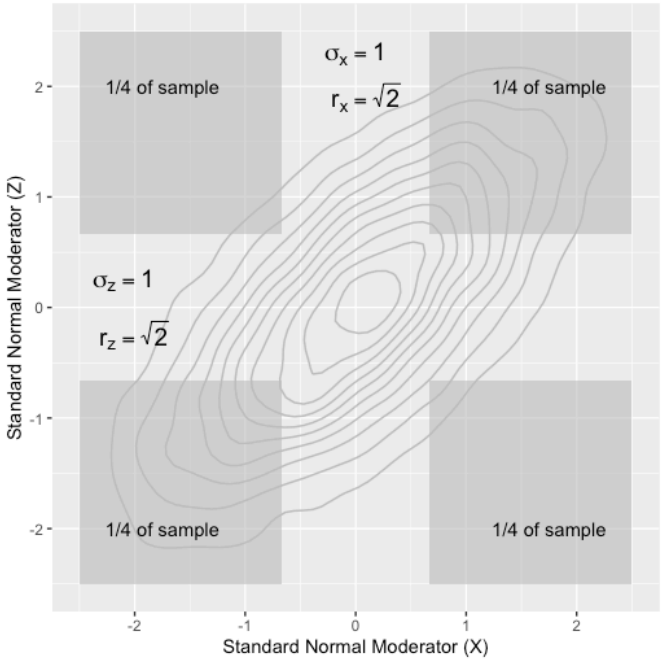
Figure 3. Contour plot of independent standard normal moderators with optimal sample allocation



McClelland and Judd – and much of the experimental design literature – has historically focused on cases in which the levels of the factors are completely under the control of the

researcher, as is the case in most laboratory experiments. In field experiments, however, researchers have considerably less control. For example, many site-level moderators are correlated with one another, making it more difficult to find sites in all four corners as shown above. We illustrate this in Figure 4, where now the covariates X and Z are both standard normal covariates that are highly correlated ($\rho = .7$). What becomes immediately clear is that there are many fewer sites in the upper-left and lower-right corner. In practice, this means that recruitment may be more difficult in these corners, since there are fewer sites to sample from. It also means that it may be impossible to select a sample in which X and Z are uncorrelated.

Figure 4. Contour plot of bivariate normal moderators with optimal sample allocation



It is straightforward to see that this problem will become more complex as the number of moderators increases, since for p covariates, the total number of corner points is 2^p, and since moderators are typically correlated in populations. While not common in the design of laboratory studies in psychology, this problem is one that has a long history in the design of

industrial experiments (e.g., see Smith, 1918). Using the more general *response surface* framework, the potential outcomes can be written more generally as

$$Y_{ij}(T_j = 0) = \gamma_{00} + \boldsymbol{X\gamma_0} \tag{11}$$

$$Y_{ij}(T_j = 1) = \gamma_{01} + \boldsymbol{X\gamma_1}$$

where now $\boldsymbol{X}$ is a matrix of moderators, $\boldsymbol{\gamma_0}$ and $\boldsymbol{\gamma_1}$ are the relationships between these moderators and outcomes under control and treatment respectively, and we are interested in estimation of $\boldsymbol{\gamma} = \boldsymbol{\gamma_1} - \boldsymbol{\gamma_0}$, a p dimensional vector of treatment by moderator interactions. Again, while this literature has not focused on the treatment effect moderation case, it is straightforward to see that a design that is optimal for estimation of $\boldsymbol{\gamma_0}$ and $\boldsymbol{\gamma_1}$ is also optimal for estimation of $\boldsymbol{\gamma}$.

These $\boldsymbol{\gamma}_i$ (i = 0,1) can be estimated using the least squares estimators $\boldsymbol{g}_i$, which are a function of the moderators $\boldsymbol{X}_j$,

$$V(\boldsymbol{g}_i) \propto (\boldsymbol{X'X})^{-\boldsymbol{1}}, \tag{12}$$

and the determinant $|(\boldsymbol{X'X})^{-\boldsymbol{1}}|$ is the generalized variance of these parameter estimates. Wald (1943) refers to the design that minimizes this generalized variance (or, conversely, maximizes $|(\boldsymbol{X'X})|$) as a D-optimal design. Kiefer (1961) shows that this design is both invariant to scale and minimizes the maximum variance of any value predicted outcome from the regression model. In the late 1950s to 1970s, iterative methods for D-optimal designs were introduced and theoretical properties were explored (e.g., Box & Draper, 1971; Fedorov, 2013; J. Kiefer, 1961, 1971; Jack Kiefer & Wolfowitz, 1959).

The one- and two-variable cases given previously can now be seen as special cases of this more general approach. For example, in the two-moderator case (X and Z), it can be shown that D = $|(\boldsymbol{X'X})| = S_x^2(S_z^2 - \rho S_x S_z)$, which is clearly maximized when both the variances of X and Z

are maximized and the covariance between them is zero. But the benefit of this D-optimal approach is that this extends to more complex cases easily, allowing for limitations in the set of possible designs (e.g., the case found in Figure 3). Unlike standard approaches in psychology and laboratory experiments, this approach does not require dividing covariates into 2^p corner points; instead, the design (sample) that achieves optimality can be determined using an iterative algorithm, of which the Federov-Wynn algorithm (or variants of it) is most common (see Fedorov, 2013; Wynn, 1972). These algorithms are implemented in software, including the **AlgDesign** (Wheeler, 2014) package in **R**, which will be illustrated later in our example analysis.

Compromise strategies for multiple goals

While the approach given previously is optimal for estimation of treatment by moderator interactions under specific functional forms, this optimality does not necessarily extend to estimation of other parameters or models. For example, recall that in Figure 2, the optimal design for a model with a single moderator is provided, under the assumption that the model is *linear*. This design puts half of the sample at either extreme of the distribution of X, making estimation of non-linear models impossible. Thus, the optimal design can depend heavily on the assumed functional form of the moderator relationship – which is difficult to know in advance.

Even more importantly, the design that is optimal for estimation of a moderator can be far from optimal for estimation of the ATE. To see why, imagine that sample selection was designed with the population ATE in mind (e.g., using stratified random sampling). The resulting sample would on average have similar moderator values as observed in the population (i.e., $E(X_j|S) = E\left(X_j|P\right)$), and the population standardized effect size (SES) could be estimated using the simple model,

$$Y_{ij} = \beta_0 + \beta_1 T_j + u_j + a_{ij} \quad (13)$$

where $a_{ij} \sim N(0, \sigma^2)$ and $u_j \sim N(0, \tau^2)$ and where $\delta_m = \beta_1 / \sqrt{\sigma^2 + \tau^2}$ is the standardized ATE. The variance of this standardized population ATE is (Raudenbush, 1997),

$$V(\widehat{\delta_m}) = \frac{n\rho + (1-\rho)}{(J-2)nP(1-P)} \quad (14)$$

where $\rho = \frac{\tau^2}{\tau^2 + \sigma^2}$ is the intra-class correlation (ICC). In comparison, if a different sample selection procedure is used – e.g., the one given in this paper – the resulting sample could differ on average from the population on these moderators (i.e., $E(X_j|S) \neq E(X_j|P)$), thus requiring adjustments. Tipton (2013) shows that if a propensity score reweighting estimator is used, this typically results in an inflated variance, with the inflation proportional to the degree of difference between the distributions of moderators (used in the adjustments) in the sample and population. Similarly, if a regression adjustment approach were used instead, statistical power would be impacted by the reduction of degrees of freedom from inclusion of additional moderators in the model. Even without adjustments, the approach developed in this paper *increases* the degree of residual between site heterogeneity (τ^2), which increases the ICC and variance, and decreasing statistical power.

In order to guard against both types of model-dependence – and to meet the goals found in most field experiments – we suggest *combining* the optimal sample selection methods for the ATE and moderators. To do so, the D-optimal criterion is used to *augment* a sample selection approach focused on estimation of the ATE. Zaslavsky, Zheng, and Adams (2008) explore this approach in a design-based framework in survey estimation in which instead of deterministic sampling, D-optimality is used to develop probabilities of selection. They show that designing a survey to estimate both the average *and* regression coefficients leads to a design that guards

against misspecification. This result does not require probability sampling to hold, however – since it is easy to see that a design that focuses on estimation of a population ATE will result in intermediate design-points (i.e., values of X and Z) that augment the extreme values found in the fully D-optimal design.

This approach is possible in existing software, including the **AlgDesign** package in **R**. In order to determine the optimal proportion p of a sample selected for estimating moderators, the relative efficiency of the augmented design can be compared to that of both the ATE-only design (before augmentation) and the only-D-optimal strategy (without ATE selection). In the next section, we explore this via an extended example.

Example: Success for All Evaluation

An evaluation of the *Success for All* (SFA) elementary school reading program was conducted in a cluster-randomized field experiment of $n = 41$ elementary schools between 2001 – 2003 (Borman et al., 2005). Unlike most studies, the list of schools that took part in the evaluation are listed in the published evaluation. For this paper, we define the population as Title I elementary schools in the U.S. using the *Common Core of Data* for 2002-3.[3] For this example, we focus on five covariates highlighted in the evaluation: total school enrollment, racial/ ethnic composition of students (% African American, % Hispanic), socio-economic status (% FRL), and urbanicity (urban, rural, other).

Method

The purpose of this example is to compare how different sample selection methods affect estimation of the population ATE and of moderator by treatment interactions in a real study. To

[3] 37 of the 41 schools were recruited into the study in 2002-3, while the remaining 4 were recruited in 2001-2.

do so, we compare three different approaches: (1) the actual sample selected in the SFA study; (2) a stratified random sample of the same size; and (3) a D-optimal sample. In addition, we explore compromise strategies, wherein $n - k$ sites in the sample are selected using (2) and the remaining k sites are selected to improve the D-optimality of the design, for $k \in (5, 10, 15, 20, 25, 30, 35)$. For the stratified random sampling, following Tipton (2014b) we created five strata based on population standardized versions of these covariates using k-means cluster analysis; these five strata explained 83% of the variation in these covariates across strata. All analyses are conducted in **R** using the **AlgDesign** package.

For each of these samples, we report three measures of optimality: D (the determinant), B (an index of similarity between the sample and population; see Tipton, 2014a), and the average r_x value across the five covariates. The B-index takes values between 0 and 1, with 1 indicating that the sample is exactly a miniature of the population on the five covariates. Tipton shows that B is inversely proportional to the increase in variance from reweighting when estimating the ATE, and that these adjustments are most effective when B > 0.80.

Results

In Figure 5, results are indicated for all sample selection procedures studied, with the stratified random sample size ($n - k$) indicated on the x-axis and values of B and relative values of D and the average r_x indicated on the y-axis. On the far left, the entire sample is selected for optimal estimation of moderators (i.e., D-optimality). On the near right (i.e., $n - k = 41$), the entire sample is selected using stratified random selection (optimal for estimation of the ATE). On the very far right, points indicate values in the actual SFA study (selected based on convenience). Note that in the figure, values of D and r_x are scaled relative to their largest values, with larger values indicating samples that are more efficient for estimation of moderators.

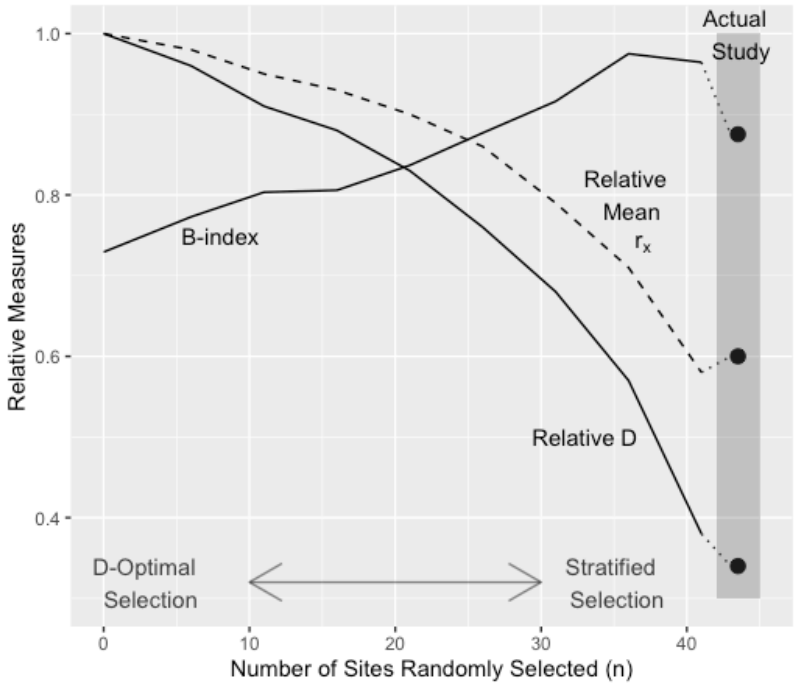
Clearly, B and D move in opposite directions – that is, the sample that is ideal for estimation of the population ATE (high B, $n - k = 41$) is far from optimal for estimation of moderators. In comparison, the sample optimal for moderators (high D, $n - k = 0$) is not far from optimal for the ATE; its B value is 0.72, indicating that while post-stratification reweighting would be necessary to estimate the ATE, this reweighting involve only a small efficiency loss. Similarly, Figure 5 shows that as D decreases, the average relative r_x value also decreases. While not indicated in the graph, the D-optimal sample has an average $r_x = S_x/\sigma_x = 1.7$, compared to a mean $r_x = 1.1$ for the stratified random sample (B-optimal).

In between these extremes, Figure 5 indicates that compromise strategies offer clear improvements. For example, examine the situation in which about half of the sample is selected to estimate the population ATE and the remainder is selected to improve moderator estimation ($n - k = 20$). The resulting sample is very similar to the population (B = 0.83) and estimation of moderators is about 80% as good as in the D-optimal design. Perhaps more importantly – the figure indicates that there are clear benefits to augmenting a sample designed to estimate the ATE with only a handful of sites selected to optimize detection of moderators. For example, by moving from $k = 0$ to $k = 5$ sites recruited for improved estimation of moderators, D is increased by 33% and the relative mean r_x by about 10%.

In order to understand how this works, in Figure 6 the distributions of these moderators are given in the population (top row), in the D-optimal sample (middle row) and in the actual SFA sample (bottom row); note that while not shown, results for the stratified random sample are similar to the top row. These covariates provide insight regarding three different variable types: a skewed variable (Total Students), a bounded variable (Proportions Black, Hispanic, FRL), and a categorical variable (Rural, Town/Suburb, Urban). Here a few trends are clear. For the

continuous variables, relative to the population, the D-optimal sample improves precision by selecting the sample from the most extreme sites. This means that the sample over-represents very large schools, schools that are more than 80% Hispanic and Black, and schools that are 0% FRL. For the categorical variable, the D-optimal sample is close to evenly divided across the three levels (thus over-representing rural schools). It is important to note, however, that this sample is not *exactly* at the extremes, as might occur in a factorial experiment, since such a sample is not possible given the relationships between covariates observed in the population.

Figure 5. Optimality indicators by sample procedure for SFA example

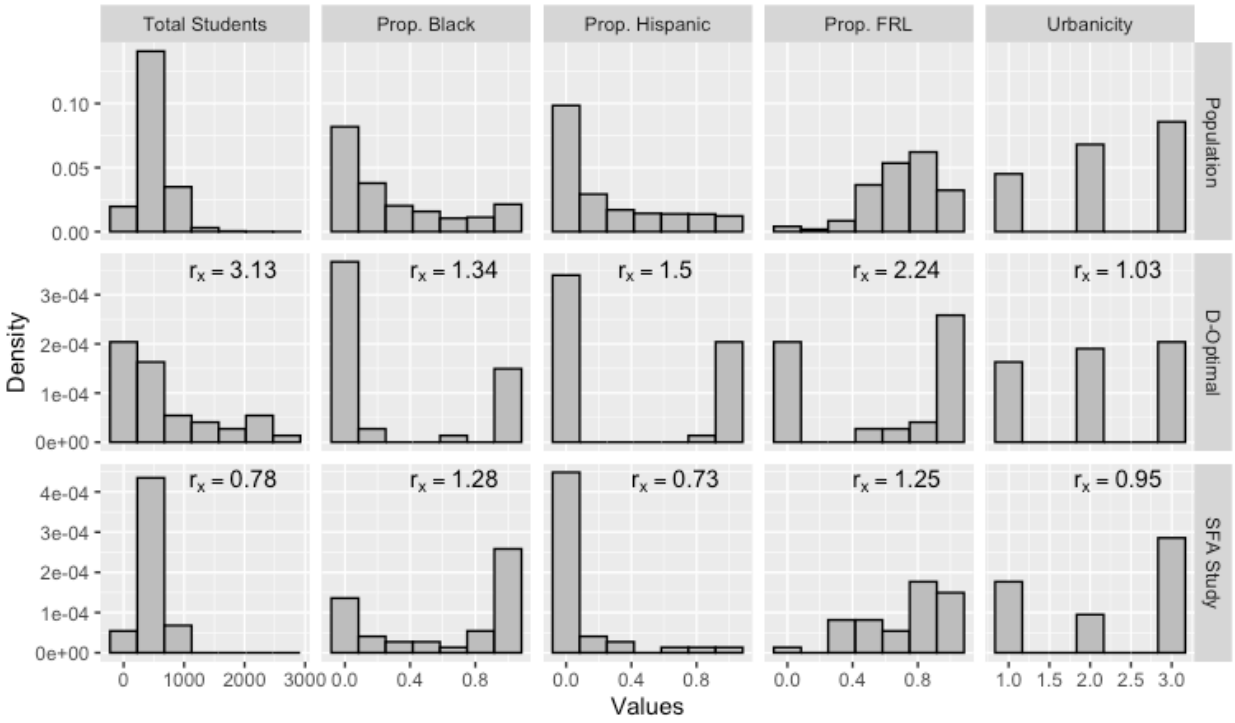


Note: The total sample is n = 41 schools. When the number of units selected using stratified random sampling is less than 41, the remaining sample is selected using D-optimality as the criterion. The far-right values indicated by "Actual study" are for the convenience sample included in the SFA study. The B-index values are naturally between 0 and 1, while D and r_x values are scaled based upon the largest value observed.

While not indicated in Figure 6, compromise samples offer a mixture of the first and second rows. Sites that are added to improve D-optimality not surprisingly occur at the extremes, for example adding in sites with high proportions of Black or Hispanic students and/or with very

large sample sizes. As these extremes are added, the standard deviations of the covariates increase relative to the population (i.e., r_x), thus reducing the MDES observable in such a study and increasing statistical power.

Figure 6. Covariate distributions by sampling procedure for SFA example



Note: The top row includes all schools in the Title I population, while the other two rows include samples of n = 41 schools. FRL = free- or reduced-price lunch (an indicator of low-socio-economic status), and for Urbanicity, 1 = rural, 2 = town or suburb, and 3 = urban.

Finally, note that in Figure 5 the actual SFA sample (far right) is not optimal for either estimation of the ATE (lower B-values) or moderators (lower relative D- and r_x-values). In particular, two variables are less heterogeneous than similarly sized random samples ($r_x = .78$ for "School Size" and $r_x = .73$ for "% Hispanic"). In comparison, the D-optimal samples for these two variables have much more extreme heterogeneity ($r_x = 3.13$ and 1.50, respectively). This means that shifting to a sample that was more representative of the population ($r_x \approx$

1) could have *decreased* the smallest effect sizes that could have been detected (MDESD) by about 30% for these moderators (and thus increased power), while shifting to a D-optimal strategy could have offered further (and larger) improvements. Finally, it is important to keep in mind that the SFA study was actually both more representative of its target population (B = .87) and more heterogeneous than most field experiments in education (recall Figure 1); in other studies, these improvements in MDESDs and power would likely be more dramatic.

Conclusion

In the first part of this paper, we showed that in order to move beyond the focus on *average* effects in randomized trials to understanding *moderators* of these effects, the samples that are included in randomized trials need to be more heterogeneous than is currently standard. By increasing this heterogeneity, we show that even when using the same sample sizes as are currently common in studies, standard errors and thus minimum detectable effect sizes for moderator effects can be greatly decreased. In the second part of this paper, we showed that it is possible to increase this heterogeneity using existing methods for selecting design runs in response surface models, but applied to sample selection. Furthermore, we showed that this heterogeneity can be increased even when this strategy is implemented for only a fraction of the sample. In this section we address some potential questions that researchers might have about implementing this strategy.

Population data limitations. The approach developed here requires population level data on site-level moderators. However, in many applications – particularly in education – these site level moderators are limited to *distal* measures, such as aggregates of student demographics and some measures of school context or resources. In contrast, the *proximal* measures researchers care about – e.g., baseline achievement, instruction, or school climate – may only be

available in smaller, non-representative datasets and not in any population frames. One strategy that might prove promising here would be to leverage findings from this auxiliary data regarding the relationship between the distal and proximal measures to project estimates of these proximal measures for all sites in the population. For example, it may be that school climate (unobserved in the population) is related to student-teacher ratios, school size, and school achievement levels (all of which are observed); thus, recruitment could be based upon a predicted version of school climate.

Causality and model specification. The approach developed in this paper assumes that the estimation model correctly specifies the true model. If this is not met, the moderator relationship estimated may be confounded with other moderators; for example, it may appear that the effect of an intervention is larger for urban schools than rural schools, when the urban schools included in the study also served more students and were in larger school districts. It is possible, then, that the urbanicity effect estimated is actually the effect of school and district size. For this reason, it is especially important the researchers think carefully about possible sources of moderation *in advance* and, even then, are careful in their interpretations, remembering that in general, moderator effects are *descriptive* not causal.

Feasibility of implementation. Throughout this paper we have assumed that it is possible for researchers to target and recruit more heterogeneous samples than they currently do. While it is true that most studies to date recruit based on convenience, given increased interest in the applicability of the results of randomized trials to policy, there has been strong movement towards targeted recruitment with generalization in mind (e.g., Tipton & Matlen, 2019). In practice, one problem that is likely to arise is that the most optimal sites (selected based on D-optimality) may refuse to participate once approached. This is where the ability to augment a

design becomes particularly useful: the refusals can be removed from the population data (as possible runs) and the augmentation procedure will produce a new set of optimal sites. Certainly, after several iterations this may mean that the resulting sample is not *as* optimal as the initially specified D-optimal sample; we doubt, however, that it will be worse than the sample that would have resulted without paying attention to heterogeneity at all.

Sequential studies. The methods developed here have focused on the role of sample heterogeneity in a single field experiment. However, questions of moderation are often addressed sequentially through planned conceptual replication studies. These can occur in a single lab – where the same intervention is studied in one context and then another – or across labs. While this paper does not speak directly to this type of study, the concepts and tools developed here could be easily extended to this case.

References

Borman, G. D., Slavin, R. E., Cheung, A., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2005). Success for All: First-year results from the national randomized field trial. *Educational Evaluation and Policy Analysis*, *27*(1), 1–22.

Box, M. J., & Draper, N. R. (1971). Factorial designs, the| X′ X| criterion, and some related matters. *Technometrics*, *13*(4), 731–742.

Chhin, C. S., Taylor, K. A., & Wei, W. S. (2018). Supporting a Culture of Replication: An Examination of Education and Special Education Research Grants Funded by the Institute of Education Sciences. *Educational Researcher*, *47*(9), 594–605.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*(3), 249–253.

Dong, N., Kelcey, B., Spybrook, J., & Maynard, R. A. (2007). *PowerUp!-Mediator: A tool for calculating statistical power for causally-defined mediation in cluster randomized trials.(Version 0.4)[Software]*.

Dong, Nianbo, Kelcey, B., & Spybrook, J. (2018). Power Analyses for Moderator Effects in Three-Level Cluster Randomized Trials. *The Journal of Experimental Education*, *86*(3), 489–514.

Fedorov, V. V. (2013). *Theory of optimal experiments*. Elsevier.

Fellers, L. (2017). *Developing an approach to determine generalizability: A review of efficacy and effectiveness trials funded by the Institute of Education Sciences*. Columbia University.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations—Gelman—2008—Statistics in Medicine—Wiley Online Library. Retrieved April 15, 2019, from https://onlinelibrary-wiley-com.turing.library.northwestern.edu/doi/abs/10.1002/sim.3107

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical Benchmarks for Interpreting Effect Sizes in Research. *Child Development Perspectives*, *2*(3), 172–177. https://doi.org/10.1111/j.1750-8606.2008.00061.x

Hong, G. (2015). *Causality in a social world: Moderation, mediation and spill-over*. John Wiley & Sons.

Institute of Education Sciences. (2018). *REQUEST FOR APPLICATIONS Education Research Grants CFDA Number: 84.305A*. Retrieved from https://ies.ed.gov/funding/pdf/2019_84305A.pdf

Kern, H. L., Stuart, E. A., Hill, J., & Green, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness*, *9*(1), 103–127.

Kiefer, J. (1961). Optimum designs in regression problems, II. *The Annals of Mathematical Statistics*, *32*(1), 298–325.

Kiefer, J. (1971). The role of symmetry and approximation in exact design optimality. In *Statistical decision theory and related topics* (pp. 109–118). Elsevier.

Kiefer, Jack, & Wolfowitz, J. (1959). Optimum designs in regression problems. *The Annals of Mathematical Statistics*, *30*(2), 271–294.

Olsen, R. B., Bell, S. H., & Nichols, A. (2018). Using Preferred Applicant Random Assignment (PARA) to Reduce Randomization Bias in Randomized Trials of Discretionary Programs. *Journal of Policy Analysis and Management*, *37*(1), 167–180.

Olsen, R. B., Orr, L. L., Bell, S. H., & Stuart, E. A. (2013). External Validity in Policy Evaluations That Choose Sites Purposively: External Validity in Policy Evaluations. *Journal of Policy Analysis and Management*, *32*(1), 107–121. https://doi.org/10.1002/pam.21660

Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: A critical reexamination and new recommendations. *Psychological Methods*, *10*(2), 178.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*(2), 173.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Sage.

Smith, K. (1918). On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika*, *12*(1/2), 1–85.

Spybrook, J., Kelcey, B., & Dong, N. (2016). Power for detecting treatment by moderator effects in two-and three-level cluster randomized trials. *Journal of Educational and Behavioral Statistics*, *41*(6), 605–627.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *174*(2), 369–386.

Stuart, E. A., Olsen, R. B., Bell, S. H., & Orr, L. L. (2012). Estimates of External Validity Bias When Impact Evaluations Select Sites Purposively. *Society for Research on Educational Effectiveness*.

Tipton, E., & Matlen, B. J. (2019). Improved generalizability through improved recruitment: Lessons learned from a large-scale randomized trial. *American Journal of Evaluation*. https://doi.org/10.1177/1098214018810519

Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, *38*(3), 239–266.

Tipton, E. (2014a). How generalizable is your experiment? An index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics*, *39*(6), 478–501.

Tipton, E. (2014b). Stratified sampling using cluster analysis: A sample selection strategy for improved generalizations from experiments. *Evaluation Review*, *37*(2), 109–139.

Tipton, E, Hedges, L., Vaden-Kiernan, M., Borman, G., Sullivan, K., & Caverly, S. (2014). Sample selection in randomized experiments: A new method using propensity score stratified sampling. *Journal of Research on Educational Effectiveness*, *7*(1), 114–135.

Tipton, E, & Peck, L. R. (2017). A design-based approach to improve external validity in welfare policy evaluations. *Evaluation Review*, *41*(4), 326–356.

VanderWeele, T. (2015). *Explanation in causal inference: Methods for mediation and interaction*. Oxford University Press.

Wheeler, B. (2014). AlgDesign: Algorithmic experimental design. R package version 1.1–7.3. *Retrieved October*, *2014*, 15.

White, M. C., Rowan, B., Hansen, B., & Lycurgus, T. (2019). Combining archival data and program-generated electronic records to improve the usefulness of efficacy trials in education: General considerations and an empirical example. *Journal of Research on Educational Effectiveness*, *12*(4), 659–684.

Wynn, H. P. (1972). Results in the theory and construction of D-optimum experimental designs. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 133–147.

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., … Dweck, C. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, *573*(7774), 364–369.

Zaslavsky, A. M., Zheng, H., & Adams, J. (2008). Optimal sample allocation for design-consistent regression in a cancer services survey when design variables are known for aggregates. *Survey Methodology*, *34*(1), 65.