Designing Probability Samples to Study Treatment Effect Heterogeneity

Elizabeth Tipton
Teachers College, Columbia University

David S. Yeager
University of Texas at Austin

Ronaldo Iachan
ICF International

Barbara Schneider
Michigan State University

DRAFT: November 30, 2016
(Comments welcome)

To appear in:
Lavrakas, P.J. (Ed.). *Experimental Methods in Survey Research: Techniques that Combine Random Sampling with Random Assignment.* New York: NY: Wiley.

Author Note

Writing of this chapter was supported by the Spencer Foundation and, for the second author, by the Raikes Foundation, the William T. Grant Foundation, the National Institute of Child Health and Human Development under award number R01HD084772, and a fellowship from the Center for Advanced Study in the Behavioral Sciences (CASBS). The National Study of Learning Mindsets (PI: Yeager; Co-Is: Dweck, Walton, Crosnoe, Schneider, & Muller) was made possible through methods and data systems created by the Project for Education Research That Scales (PERTS), data collection carried out by ICF International, and received support from the Raikes Foundation, the William T. Grant Foundation, the Spencer Foundation, the Bezos Family Foundation, the Character Lab, the Houston Endowment, Angela Duckworth (personal gift), and the President and Dean of Humanities and Social Sciences at Stanford University.

Please address correspondence to Elizabeth Tipton, 525 W 120th Street, Box 118, New York, NY 10027 (email: tipton@tc.columbia.edu) or to David S. Yeager, 108 E. Dean Keeton, Austin, TX, 78712 (email: dyeager@utexas.edu).

Designing Probability Samples to Study Treatment Effect Heterogeneity

Population-based survey experiments are increasingly common in the social and behavioral sciences, particularly in psychology, political science, economics and sociology. Much of this growth comes from the ability to randomize participants to different conditions via Internet-delivered surveys administered to panels of probability sampled adults (e.g., the Time Sharing Experiments for the Social Sciences; TESS).[1] Survey experiments are particularly amenable to experimental hypotheses regarding individuals' attitudes, beliefs, or intended behaviors because questions or vignettes can be randomly-assigned early in a survey and outcomes can be compared on later questions. Although less common, survey-administered manipulations have also been used to "nudge" individuals (Thaler & Sunstein, 2008) or change "mindsets" (Dweck, 2006) in ways that increase socially-beneficial behaviors such as voting or college graduation, assessed via official registries like the validated vote file (e.g., Bryan, Walton, Rogers, & Dweck, 2011, Studies 2 and 3), or the national student clearinghouse (e.g., Yeager, Walton, et al., 2016, Study 1).

Studies that nest a random-assignment experiment within a probability sample are well-positioned to estimate average treatment effects that generalize to a well-defined population.[2] This is in contrast to more typical experiments (e.g. psychology laboratory studies or field trials in education), which are often conducted with samples of convenience (see Olsen, Orr, Bell, & Stuart, 2013). Experiments conducted in convenience samples can provide initial evidence against the null hypothesis, but they are poorly-positioned for assessing generalizability if there is

[1] See www.tessexperiments.org

[2] Throughout the chapter, we assume that non-response in probability samples is ignorable or addressable through post-stratification. While our analysis does not depend on this assumption, there is some support for the idea that it is often defensible. Some find that non-response in surveys over the years has not led to lower accuracy in those surveys (e.g. Keeter, Kennedy, Dimock, Best, & Craighill, 2006; Yeager et al., 2011). However those studies were used marginal distributions on non-weighted variables as a measure of validity; to our knowledge there has been no evaluation of the effects of survey non-response on treatment effect sizes over time.

any variation in treatment effects across groups—that is, *treatment effect heterogeneity* (see Tipton, 2013). For example, if younger respondents with lower educational attainment are less responsive to a treatment, but a recruited sample over represents older, highly-educated individuals, then the average effect estimated in a sample will be larger than that in the population.

Unfortunately, too little is known about treatment effect heterogeneity, and what is known has typically come from the analysis phase of research, not the design phase. Most experiments are designed to have power to detect *average* effects, and are only rarely well powered for detecting *differences* in treatment effects across subgroups (i.e., moderators). This is because moderator effects are statistically less precise than the estimate of the average effect, and this is exacerbated when subgroups are rare. As we will discuss, under-powered moderation analyses can result in high Type 1 *and* Type 2 errors (also see Gelman, 2014).

We argue that probability sampling, as typically implemented, does not automatically produce samples that are suitable for studying treatment heterogeneity, even when samples are high-quality and have low non-response bias. We propose that it is better to *design* the probability sample from the outset with the goal of understanding sources of treatment effect heterogeneity. By doing so, researchers are thus able to not only estimate an average effect, but also identify for *whom, under what conditions, and why* a treatment may work best (see Bryk, 2009; Shadish, Cook, & Campbell, 2001). This type of knowledge is central to the development of a theory of causal mechanism – the highest goal of science (see Cook, 1993).

In the remainder of this chapter, we explain a new approach that survey samplers can use when designing probability samples for survey experiments where there is a possibility of treatment heterogeneity. We begin by explaining why probability samples are preferred to non-probability samples for estimating two quantities (or *estimands*): (1) population average treatment effects and (2) treatment effects within subgroups. The chapter furthermore explains why typical probability sampling methods that optimize statistical power for the *average* effect in a

population do not necessarily optimize statistical power for the *subgroup* effects of interest—especially when one's interest is in estimating effects within a rare subgroup.

Next, this chapter explains why even large, well-constructed, highly-representative probability samples with randomized treatments can produce *confounded* analyses of differences across subgroups. Specifically, when one interacts a manipulated treatment with a measured, non-manipulated variable (a moderator), then one faces the same set of confounds as any other observational study, despite randomization of the treatment (e.g., Morgan & Winship, 2014; Shadish et al., 2001). For instance, two competing moderators (e.g. race and education) might be highly correlated with each other and with a host of other unmeasured, third-variable confounds (like neighborhood segregation); a moderation analysis, even with both in a model, can have difficulty disambiguating which (if any) has causal power. Furthermore, the off-diagonal cases that could be useful for disambiguating the two (e.g., people who are both minority and highly educated) might be strongly under-represented even in a typical probability sample. This basic insight—that moderators are often observational and therefore confounded, threatening moderation analysis—has been almost completely ignored in the experimental literature on subgroup analyses and interaction effects (though see Tucker-Drob, 2011; Vanderweele, 2015).

We recommend a focus on the the development of sampling *designs* that are informed by a *theory of treatment effect heterogeneity*. Our recommended approach requires researchers to identify important subgroups and hypotheses regarding *to whom, where, and under what conditions* an intervention may work best *a priori*, rather than post hoc (see Gelman, 2014; Rothman & Greenland, 2005). In many cases, these potential moderators are not the usual demographics. Demographics are not always the most relevant to what enhances or precipitates a treatment effect. Instead, theoretically-relevant moderators are often rich contextual factors, or a person's behavioral history. Sometimes measures of these moderators may need to be developed using available population data, because they are not always available. These measures can then be used to stratify the population. Ultimately, the allocation of the sample to these strata can be

based on power analyses for not only the average treatment effect, but also subgroup impacts and comparisons between these subgroups.

We illustrate our proposed approach using an empirical case study of a survey-administered behavioral-science intervention: The U.S. National Study of Learning Mindsets. This experiment evaluates the effects of an Internet-based, self-administered intervention on achievement-oriented behavior over time in a national probability sample of 76 U.S. public high schools. They manipulation is delivered via survey, and the outcome data are collected from official school registrars at the end of the school year. By using a novel stratification method – which we show reduces confounding between two theoretically-developed contextual factors and improves subgroup statistical power— it will be possible to test not only if the intervention has an effect on high school achievement on *average*, but also if it has differential effects across school contexts. Answers to such questions directly inform causal, mechanistic theory. Although this case study has many unique features – like the use of high school students, the availability of rich auxiliary information for strata construction, and assessment of behavioral outcomes – the same basic logic could apply to more traditional survey experiments.

Probability Samples Facilitate Estimation of Average Treatment Effects

We start with a review of basic causal inference for the estimation of the average treatment impact in a population. We then expand to address treatment heterogeneity. As has been well documented (e.g. Morgan & Winship, 2014; Shadish et al., 2001), as a result of random assignment to treatment, the baseline characteristics of the experimental and control conditions are equivalent, on expectation. Thus, on expectation, any differences in the outcomes of the two groups can be attributed to the presence of the treatment. This difference in average group outcomes in the sample is called the *Sample Average Treatment Effect* (SATE), and can be formally represented for units $i = 1, \ldots, n$ in the sample as the expected outcome Y when a

treatment is present (1) minus the expected outcome when it is not (0), where the subscript S indicates this is the average over the sample,

$$SATE = E_S[Y_i(1)] - E_S[Y_i(0)]. \quad (1)$$

The estimate of the average treatment effect produced in an experiment is unbiased for the SATE (Holland, 1986). Put another way, as a result of randomization to treatment, in a well-designed experiment with no attrition, there is no *treatment selection bias*.

Importantly, if treatment effects vary across units, this SATE may differ across different samples and populations (Imai, King, & Stuart, 2008; Olsen et al., 2013; Weiss, Bloom, & Brock, 2014). For this reason, researchers have increasingly conducted these types of experiments in national probability samples, since doing so allows for the direct estimation of the average treatment effect in the population (i.e., PATE; see Mutz, 2011). The PATE has great value. For instance, policymakers may wish to implement a public health message for an entire nation, and they may want to know whether, on average, the message may improve people's health behaviors.

We can define the PATE formally based on units $i = 1 \ldots N$ in an inference population with the subscript P indicating average over the population, as

$$PATE = E_P[Y_i(1)] - E_P[Y_i(0)]. \quad (2)$$

Unlike the SATE, estimation of the PATE based on a randomized experiment is only unbiased when in addition to randomization of treatment, the sample is selected from the inference population randomly and when non-response is ignorable (i.e. uncorrelated with any source of treatment heterogeneity; Olsen et al., 2013; Tipton, 2013). When probability sampling is not used, the estimate of the PATE may be biased as a result of *sample selection bias* (Allcott, 2015; Tipton, 2013).

Recently, researchers in education, medicine and economics have become increasingly concerned about sample selection bias and its effect on ATEs (e.g., Allcott, 2015; Cole & Stuart, 2010; Tipton, 2013). This stems from the fact that it is rare in the social and behavioral sciences

for researchers to be certain that treatments have identical effects across all individuals in a population (e.g., Gelman et al., 2015; Green & Kern, 2012; K. Rothman & Greenland, 2005; Vanderweele, 2015). As Gelman et al. (2015) state, it is "better to start with the admission of variation in the [treatment] effect and go from there" (p. 637). Indeed, effects vary in relation to implementation quality, participant characteristics (e.g. gender or minority status in psychology experiments), and also contextual factors (Allcott, 2015; Hulleman & Cordray, 2009; Weiss et al., 2014).

One way to understand variation in treatment effects across subgroups in the estimation of PATEs is to decompose the overall effect into subgroup ATEs:

$$\text{PATE} = \text{CATE}_1\pi_1 + \text{CATE}_2\pi_2 + \ldots + \text{CATE}_J\pi_J = \Sigma\text{CATE}_h\pi_h \quad (2)$$

Here CATE_h is the conditional (i.e., subgroup) ATE for group $h = 1, \ldots, H$, and π_h is the proportion of the population in subgroup h (where $\Sigma\pi_h = 1$). The PATE is then the weighted sum of these subgroup ATEs.

Decomposing the PATE in this way (Equation 2) highlights two facts pertinent to survey sampling. First, when the CATE does not vary across subgroups—for instance, when investigating a basic cognitive or biological process that works in the same manner for all humans—then any sample's estimate of the ATE will match the PATE, within sampling error. In such cases, probability sampling is not needed. Second, if the CATE *does* vary across subgroups, but the achieved sample has the same proportions of individuals who are in each subgroup as in the population, then the ATE will match the population ATE, within sampling error (Olsen et al., 2013). A properly constructed and weighted probability sample with ignorable non-response ensures the latter, resulting in an unbiased estimate of the PATE, within sampling error.

If, however, a probability sample is not employed, improperly constructed, or has non-ignorable non-response, then sample proportions in each of the H subgroups may differ from the population proportions (the π_h). When treatment impacts vary, then the resulting estimate based on the sample (the SATE) may be biased for the ATE for the population (the PATE).

In the last five years, this problem has received increased attention in the statistical and methodological literature, resulting in a new set of tools for making post-hoc adjustments to the average treatment impact estimator (e.g., Stuart, Cole, Bradshaw, & Leaf, 2011; Tipton, 2013; Tipton, 2014). In general, however, probability samples are preferred over weighted non-probability samples for making generalizations about effect sizes when there is heterogeneity (as we expect there nearly always is in typical social and behavioral research).

Treatment Effect Heterogeneity in Experiments

We next briefly review current approaches to studying this heterogeneity in the social and behavioral sciences. Identifying subgroups has tremendous pratical utility. Policymakers who know for whom, and under what conditions, a treatment is effective can target a treatment to subgroups where it may be most useful, and avoid delivering it in settings where it may be harmful (Cook, 1993; Solon, Haider, & Wooldridge, 2015). In addition, moderator analyses – which make comparisons between treatment impacts in subgroups – are critical for theory development because they point to causal mechanism.

Consider several prominent experimental manipulations that showed different treatment effects across subgroups:

- a manipulation of questionnaire wording is weaker or stronger for individuals with high versus low levels of education (a means of testing satisficing theory; e.g., Krosnick, 1999; Narayan & Krosnick, 1996);
- A "nudge" to motivate people to decrease energy use is more effective in households with large square footage or swimming pools, and less effective in households where even motivated individuals cannot decrease energy use (Allcott, 2015);
- different policy framings are more or less compelling to members of different political parties or ideologies (Entman, 2010); and

- motivational interventions delivered via Internet surveys change achievement-oriented behavior more or less for low-achieving students and/or racial/ethnic minority students (Paunesku et al., 2015; Yeager, Romero, et al., 2016).

Testifying to the importance of these moderator analyses for basic theoretical advances, our re-analysis of a random sample of psychology experiments published in premier journals in 2008 showed that 20% of all studies interacted a measured individual difference moderator with an experimental treatment or task (Open Science Collaboration, 2015).

Despite this interest in moderators and treatment heterogeneity, social and behavioral scientists have lamented the state of subgroup analyses in experimental research for decades. For instance, D. P. Green and Kern (2012) state, "in practice, the investigation of treatment effect heterogeneity in survey experiments often seems ad hoc and unstructured."

Such unstructured analysis can cause problems for cumulative science. For example, authors have noted the large number of cases in which moderation of a treatment effect by a demographic characteristic did not replicate upon further examination (Gelman, 2014; Gelman et al., 2015; K. Rothman & Greenland, 2005; Vanderweele, 2015). In a famous clinical example, researchers discovered unexpected (and eventually unreplicated) moderation by gender for effects of aspirin in preventing stroke death, leading doctors to withhold aspirin from women for over a decade (see R. L. Rothman et al., 2005).

More broadly, Nosek and colleagues (Open Science Collaboration, 2015) attempted to replicate 100 randomly sampled psychology experiments. In this analysis, only 1 out of 20 studies that examined differences in responsiveness to an experimental treatment across measured individual characteristics (e.g., race, gender, IQ) replicated a statistically significant interaction. By contrast, 46% of all other kinds of studies replicated a significant difference ($X^2(1)=8.84$, $p=.003$).

In response to such pessimism, a growing amount of research has proposed novel analytic techniques for interrogating treatment effect heterogeneity, including regression trees (Imai &

Strauss, 2011), Bayesian Additive Regression Trees (BART; D. P. Green & Kern, 2012), Generalized Additive Models (GAM; Feller & Holmes, 2009), and semi-parametric differencing estimators (Horiuchi, Imai, & Taniguchi, 2007). These approaches are all carried out during the analysis phase, however, and do not require a strong theory when selecting potential moderators; instead, as data-driven approaches, they test nearly all possible moderators, with the focus on finding the best predictive model.

As statisticians consistently note, the key to strong inferences is not the model, but "design, design, design" (Bloom, 2010). Surprisingly, almost no research has explained the implications of concerns about treatment effect heterogeneity tests for the *design* of survey samples or experiments from the start (though see Tipton, 2015). We propose a design-oriented approach that is *theory*-driven, requiring researchers to design new studies based on hypotheses generated from prior work, thus resulting in a more cumulative model for science.

Estimation of Subgroup Treatment Effects

More formally, returning to equation (2), researchers may wish to estimate the conditional average treatment effect, or CATE, for subgroup 1 as well as the CATE for subgroup 2, and so on. They may then be interested in determining whether the differences between treatment effects in those subgroups is non-zero. Without planning carefully, estimates of subgroup impacts and comparisons between them can be both biased and imprecise.

Problems with sample selection bias. One might imagine that as long as a sample had a sufficient number of individuals from each subgroup, then it would be possible to obtain a generalizable CATE effect size *for that subgroup*. That is, if there were no heterogeneity in effect sizes within the subgroups—if men are men, women are women, whites are whites—then it should not make a difference whether the subgroup in a sample was recruited through probability methods or not. This position has long been taken in psychological research. That is, researchers have compared arbitrary samples of subgroups in their responsiveness to experimental

manipulations: western college students vs. east-Asian college students (Markus & Kitayama, 2010), white college students vs. African-American college students (Steele & Aronson, 1995), low-income people versus high income people (Mullainathan & Shafir, 2013), and so son. This approach has become standard in political science as well, particularly as this field has migrated toward the use of national non-probability samples for research on differential effects of messages within Republican and Democratic subgroups (Green & Kern, 2012).

According to this logic, as long as one could obtain enough representatives from a given subgroup, through whatever means, then it is possible to estimate that subgroup's treatment effect. Alas, there can be significant heterogeneity in effect sizes *within* demographic subgroups and therefore methods for acquiring a subgroup can produce bias.

This can be seen theoretically be extending Equation 2 to focus instead on a particular CATE for subgroup *h*,

$$CATE_h = [CCATE_{h1}\pi_{h1} + CCATE_{h2}\pi_{h2} + \ldots + CCATE_{hK}\pi_{hK} = \Sigma CCATE_{hk}\pi_{hk}]/\pi_h, \quad (3)$$

where here the subgroup CATE (e.g., for women) can itself be decomposed into $k = 1 \ldots K$ subgroup ATEs (what we call conditional-conditional ATEs, CCATEs), with the proportion of the population in these doubly-defined subgroups equal to π_{hk} (where now $\Sigma\pi_{hk} = \pi_h$). Equation 3 shows that just as when estimating the PATE (2), bias can result if the population and sample proportions (the π_{hk}) are not identical and treatment impacts are not constant within the defined subgroup. Thus, in general, simply having enough representatives *from* a given demographic subgroup does not ensure that a sample is representative *of* the demographic subgroup—if there is any heterogeneity in treatment impact within the group. This is one reason that probability sampling is important for experiments, since when implemented well (and with ignorable non-response), this bias is removed, at expectation.

Problems with small sample sizes. Putting aside this issue of sample selection bias, the second issue in estimating subgroup ATEs has to do with small sample sizes. This concern arises regardless of the method used to recruit the sample, probability and non-probability. Given recent

innovations in statistical methodology regarding the detection of treatment effect heterogeneity (e.g. regression trees, Feller & Holmes, 2009; D. P. Green & Kern, 2012; Horiuchi et al., 2007; Imai & Strauss, 2011) and the (often large) sample sizes often found in survey experiments, it is easy to overlook the importance of adequate subgroup sample sizes when understanding possible sources of treatment effect heterogeneity.

This problem becomes particularly likely as the number of subgroups studied (H) increases, thus resulting in smaller and smaller sample sizes in each group. For example, imagine if subgroup effects are desired not only in terms of racial or ethnic minority status but also by education level. This would mean estimating the subgroup CATE for African Americans with a four-year degree or higher; in the 2016 Current Population Survey this group accounted for only 2% of the U.S. population,[3] or just 20 respondents in a typical 1,000 person phone survey.

Small subgroup sample sizes can lead to two problems. First, small sample sizes typically lead to low statistical power, thus making it likely that one will erroneously conclude that a treatment was not effective for a subgroup (a Type II error). This is particularly likely given that most experiments are powered only for estimation of the SATE, not subgroup effects (K. Rothman & Greenland, 2005). For instance, in the example above, the minimal detectable effect for highly-educated African-Americans, with 80% power, would be d=1.16, a massive effect as far as survey manipulations go. In practice, this lack of a statistically significant effect within a subgroup can be mistaken for ineffectiveness.

Second, somewhat ironically, small sample sizes – and the resulting low power – might lead researchers to search for subgroup effects across a variety of potential groupings, thus producing statistically significant but non-replicable results (a Type I error) (Gelman, 2014). This is the multiple comparisons problem (Gelman et al., 2015; C. L. Green, 1999; K. Rothman & Greenland, 2005; Simmons, Nelson, & Simonsohn, 2011). A famous example of this is the ISIS-2

[3] Analyses conducted at http://www.census.gov/cps/data/cpstablecreator.html

trial (testing aspirin vs. placebo in acute myocardial infarction). Aspirin was effective for patients born under the signs of Libra and Gemini, but no others (see K. Rothman & Greenland, 2005). Results such as these could reasonably be obtained via post-hoc flexibility in data analysis, or *p-hacking* (Simmons et al., 2011). As a result, even a probability sample can produce non-replicable findings if the sample did not afford adequate power for a subgroup and therefore invited p-hacking (Gelman, 2014).

The design-oriented approach that we develop in this chapter reduces these problems by requiring researchers to articulate, from the very beginning, potential subgroups for whom estimating a CATE might be of interest. Information on the distribution of the related variables (e.g., education level, racial composition) is then gathered in the population, and power analyses are conducted based upon this information. For example, if the sample will include 1,000 individuals, but a CATE estimate is desired for a subgroup accounting for only 10% of this population, a power analysis would be conducted based on a sample of size 100 (=1000*.10). In many cases this will lead to a need for over-sampling (particularly with rare subgroups). Importantly, as we will show in the example, this means that when the PATE estimate is also desired from the same sample, a post-stratification adjustment will be necessary (otherwise, the simple, sample-based estimate will be biased).

Moderator Analyses for Understanding Causal Mechanism

As we have stated, the goal of mechanistic, theory-driven science is to not only understand *if* an intervention changes outcomes for particular subgroups, but also *why* or *how* (Cook, 1993). If we begin from this position, the detection of subgroup and moderator effects becomes not simply *descriptive* – as occurs when researchers report results by demographic subgroups without theory, in an actuarial sense – but explanatory, leading to a theory of the *causal mechanism*.

For example, one may conclude that an intervention has larger effects for women than for men – thus eliciting the question: Why? Is it something inherent about gender that leads to these differential effects? This is where the benefits of identifying possible moderators in advance are particularly large, since very often the focus is on contextual and social factors that may be more theoretically relevant to the mechanism of the intervention. In this section, we address two concerns that occur when testing these moderator effects – confounder bias and statistical power.

Problems of confounder bias in probability samples. In a probability sample, the estimate of a subgroup CATE is unbiased. However, just because the treatment impacts are found to vary in relation to a particular measured variable (e.g., education level) does not mean that that the moderator *causes* treatment effects to differ.

Consider that demographic characteristics are not, in and of themselves, causes of moderated treatment effects. Instead, they are proxies for unobserved material, social, or psychological realities that titrate a person's responsiveness to a given experimental manipulation (Vanderweele, 2015). Recall the example of moderation of questionnaire design effects by educational attainment (Narayan & Krosnick, 1996). Researchers did not have a theory that a *diploma* causes moderation of a questionnaire manipulation; instead, educational attainment covaries with qualities such as IQ, need for cognition, social class, openness to experience, and more, which can make people pay greater or less attention to novel information (Narayan & Krosnick, 1996). Likewise, when assessing moderation by race and ethnicity, researchers rarely start an analysis with a theory about inherent qualities of skin pigment or of self-labels; instead researchers may expect that race covaries with access to resources or experiences of discrimination or stereotyping over the lifespan (Spencer, 2006), for instance those that affect trust in institutions (Smith, 2010) or performance in school (Steele, 1997).

The issues of confounding we are raising here have long been discussed and acknowledged in observational or quasi-experimental studies of the "X causes Y" sort (Morgan & Winship, 2014; Shadish et al., 2001). Researchers would not say that education causes longer life

solely on the basis of a correlation between the two. Instead, researchers routinely attempt to control or adjust for baseline differences between the groups they wish to compare, so as to isolate the effect of a category or status.

We argue that researchers need to address these same problems of confounding in the "M causes the effect of X on Y to be larger (or smaller)" case. This means not only defining a handful of relevant subgroups, but also the development of a theory of treatment effect heterogeneity – that is, a mechanism that potentially *causes* impacts to vary – and the development of statistical and design-oriented approaches to directly address concerns with identifying that mechanism.

Additional confounder bias problems with non-probability samples. In non-probability samples, concerns with confounder bias in moderator analyses are even larger. Just as the estimates of subgroup effects can be biased in non-random samples, so too are *differences* in subgroup effects. This is because the correlation between a demographic characteristic and an unobserved moderator may vary across the population, and thus may differ in a non-random sample compared to in a larger population.

Recent audits have investigated this possibility, comparing the covariance between demographic variables and unobserved characteristics in seven non-probability samples as compared to two probability samples and government benchmarks (Yeager et al., 2011). Yeager et al. (2011) showed that the the probability samples were more demographically representative than the non-probability samples, and were more accurate when estimating non-demographic characteristics. Critically, when weights were then applied based on demographics, the accuracy on the variables not used in weighting was increased in the probability samples. By contrast, in the non-probability samples, accuracy on non-weighted characteristics did *not* improve with weighting; in some cases, accuracy was worse (Yeager et al., 2011). This directly illustrates that even well-constructed and properly weighted *non*-probability samples do not necessarily accurately represent the latent, unobserved characteristics that typically covary with measureable demographics in a population (for another perspective, see Chapter [Klar and Leeper]). Hence,

using data from large non-probability surveys does not, in our view, yet address the issues of confounding and generalizability that we raise here.

Problems with statistical power. Until now, we have focused on issues of bias when *attributing* tests of moderation in the development of a theory of the causal mechanism. As with subgroup impact estimation, however, with moderator analyses another issue arises: statistical power. In moderation analyses this concern with power is even larger than it is when estimating subgroup CATEs, since the resulting test involves comparisons between *estimated* quantities. Recall, for example, that in the simplest case (in which subgroups are defined at the same level as the PSU and two subgroups are compared)[4],

$$V(CATE_1 - CATE_2) = V(CATE_1) + V(CATE_2). \quad (4)$$

Equation 4 illustrates that the variance of a difference between subgroups is – in the best case – twice as large as the variance of a subgroup-specific CATE.

If the goal is to minimize the variance of this difference for a fixed total sample size (n) – thus increasing power – a well-known result (see Raudenbush & Liu, 2000) is that the best strategy is to divide the sample evenly across these subgroups (i.e., $n/2$ to estimate $CATE_1$, $n/2$ to estimate $CATE_2$). If multiple comparisons are desired – as when all pair-wise subgroup comparisons are of interest – the end result is a strategy in which not only are some subgroups over-sampled, but furthermore the sample size in each of the subgroups is the same. As noted previously, this uniform sampling across subgroups biases the estimate of the population ATE, which means that post-stratification weighting will be required.

To see why post-stratification weighting may be problematic, however, let us now examine the variance of such an estimator[5],

---

[4] Certainly, complex sampling will result in an additional covariance term here.

[5] This variance estimator assumes the subgroups are independent. In more complex samples they will not be and the variance formula will need to take these covariances into account. We focus on the simplest case to highlight the roles that p_h and π_h play.

$$V(\Sigma\pi_h CATE_h) = \Sigma\pi_h^2 V(CATE_h) \approx \Sigma(\pi_h/p_h)\pi_h * V(PATE),$$

where π_h is the proportion of the population in subgroup *h* and p_h is the proportion of the sample, V(PATE) is the variance of the PATE estimate if no post-stratification is required (i.e., $p_h = \pi_h$), and where the last approximation is due to Tipton (2013). In this estimator, the total sampling variance is thus a function of the ratio of population and sample proportions, leading to a larger variance whenever an allocation scheme other than proportional allocation is implemented. This particularly occurs if there are very small subgroups in the population (i.e., small π_h) but equal allocation is used across subgroups to ensure power for moderator tests (i.e., $p_h = 1/H$).

The end result is that power for detecting differences in subgroup treatment effects (i.e., having a significant interaction comparing two CATEs) is lower than for detecting a simple effect of the treatment within the subgroup itself (estimating a single CATE). Thus the optimal strategy for powering for the detection of moderation can be far from optimal for estimating the population ATE. We return to this tension in the next section and in an example from a national experimental study.

Stratification for Studying Heterogeneity

In this section, we outline a general approach to sample design for survey experiments interested in estimating three parameters: the average treatment impact in the population (PATE), subgroup treatment effects (CATEs), and pre-defined differences in treatment impacts across subgroups. Throughout we focus on designs in which observations are nested in clusters of some type, with a focus on treatment effect heterogeneity defined at the cluster level. In the experiment used as a motivating example in this chapter, these clusters are schools. In other survey experiments, clusters might be regions, villages, counties, demographic subgroups, or firms.

Overview. The approach we develop builds off one proposed by Tipton (2014) and Tipton et al. (2014) for estimation of the average treatment impact in field experiments in

education research. The approach is a version of stratified sampling, though the definition of the strata deviates from the traditional approach found in surveys.

To begin, an inference population must be clearly defined– for example, it might include all high schools in the United States – and information on these clusters in the population must be available. For example, in the education context, this information can come from the Common Core of Data and include information on school-level demographics. When participants are drawn from an existing probability sample panel, such as the GfK knowledge panel or the NORC Amerispeak panel, this may include information on person-level demographics, cognitive abilities, or socio-demographic questions answered on a profile survey, that can be aggregated to important subgroups (e.g., political party membership). When one's goal is to conduct experiments with adults who are not yet part of a panel, then in the future it may be possible to obtain richer micro-data on people's past behavior or characteristics (for instance, whether they voted in past elections, and what their party identification was, from the validated voter file) and then invite people to participate within strata.

Strata are commonly used in surveys to reduce residual variation when the measures of interest vary substantially across strata. In most surveys, these measures of interest are the *outcomes*. For example, in education experiments, strata are typically defined by the percent of students in free or reduced lunch, a measure of socio-economic status, which predicts schools' test scores (Battle & Lewis, 2002). In this usage, strata are employed only for reducing variation in the sample, and, while taken into account in the variance calculation, are not the focus of the analysis.

In contrast to the typical approach, we argue that the covariates that matter for stratification in survey experiments are instead those that explain *variation* in *treatment impacts* (Tipton, 2014). Since these cannot be known *a priori*, researchers can identify possible treatment effect moderators based upon both previous empirical research and a theoretical understanding of the mechanism of the intervention. In terms of practicality, however, not all potential moderators

can be included, since these must be available in existing population level data. For example, demographic variables are often available, whereas variables related to attitudes and beliefs are not. Once defined, these variables are then used to stratify the population.

When there are multiple variables, stratifying is more difficult, and is limited by the total number of experimental units; for example, a study including 100 sites could have at most 100 strata (and likely, many fewer). Tipton (2014) shows that one approach is to reduce the dimensionality to a single variable through the use of k-means cluster analysis[6]. In this approach, *k* is the total number of strata that study resources can accommodate, and the k-means procedure results in *k* clusters (strata) that are maximally heterogeneous. Tipton et al. (2014) show that another approach, propensity scoring, can also be used and is particularly effective in "synthetic" generalizations. These are generalizations in which the population to whom researchers want to generalize cannot take part in the study; for example, this arises when an effect is desired for the population of schools currently using a program.

In this previous research by Tipton (2014) and Tipton et al. (2014), the goal of stratification is to gather a sample that is compositionally similar to the inference population so that the SATE estimated in the study is an unbiased estimate of the PATE. In this chapter, however, we shift focus to an alternative approach to stratification that adds a concern with *treatment contrasts*.

Operationalizing moderators. In the survey experiment context, after researchers have identified a set of potential moderators, a next question is how to operationalize these. For example, socio-economic status may be of interest, in which case researchers must develop a measure that operationalizes this construct. This might involve the use of a *latent variable* approach, pooling data from several sources, for instance through structural equation modeling.

[6] Tipton (2014) shows that k-means can be used with both continuous and categorical variables through choice of the distance metric.

Latent variables can involve the inclusion of data from multiple – and sometimes novel – data sources; the NSLM, for example, included data on school quality collected on a website aimed at parents (GreatSchools.org). As noted above, this approach is most useful when the moderators of interest are contextual variables available in public data.

Orthogonalizing moderators. A second concern is the degree to which the hypothesized moderators covary, or, put another way, to what degree the effect of one moderator is *confounded* by another possible moderator. For example, in school data, the proportion of minority students in a school is highly correlated with the academic rigor and achievement level of a school. If the population is *not* stratified on these variables, then it is likely that in the resulting sample they will be so highly correlated as to make it difficult or impossible to test hypotheses regarding them separately. Being able to distinguish between these potential causes is important, since it allows for greater understanding of the causal mechanism. To orthogonalize moderators, it is helpful to have adequate sample not only of the marginal distribution cells but also the off-diagonals.

Stratum creation and allocation. Population units should then be divided into strata based upon this set of moderators. The next step is to allocate the population to these H strata, noting which proportion of the population is found in each stratum (i.e., π_h, $h = 1 \ldots H$). An important question in all stratified sample designs is how to allocate the n units in the sample to these H strata (i.e, the $p_h = n_h/n$).

If the goal is only to estimate the average treatment impact, the optimal strategy is to allocate the sample proportionally (i.e., $p_h = \pi_h$). This strategy, however, may not be optimal for testing contrasts between the treatment impacts in different strata. This approach is particularly problematic when the sample size n is small to moderate and some of the strata being tested contain only a very small proportion of the population. In fact, as shown above, if pair-wise contrasts are desired between all strata, the optimal strategy for moderator analyses is to instead divide the sample evenly across strata (i.e., $p_h = 1/H$).

In practice, the competing goals of estimating the population ATE and testing comparisons between strata must be balanced. If the allocation is not proportional (i.e., $p_h \neq \pi_h$), however, this means that in order to estimate the average treatment impact, a reweighting approach will be needed (see Equation 4). While the technical details are beyond the scope of this chapter, in order to determine the 'ideal' stratum allocation, these three separate parameters and their competing demands for power must be taken into account. In practice this often means prioritizing hypotheses into 'confirmatory' and 'exploratory', giving greater preference for allocation schemes for 'confirmatory' analyses than others (Open Science Framework, 2016).

Nesting a Randomized Treatment in a National Probability Sample: The NSLM

To illustrate the issues addressed in this chapter, we turn to an example based upon the National Study of Learning Mindsets (NSLM) study. So-called "mindset interventions" (e.g. Aronson, Fried, & Good, 2002; Paunesku et al., 2015) teach students new beliefs that can increase their motivation to learn—in particular, they teach that "smartness" is not a fixed quantity, but can be developed with effort over time (Dweck, 2006; Yeager & Dweck, 2012). Psychologists have advocated for their broader use in policy and practice (e.g., Rattan, Savani, Chugh, & Dweck, 2015)**.** The goal of the NSLM is to determine if brief mindset interventions can increase student motivation and improve learning outcomes for high school students throughout the U.S., especially during the difficult transition to high school (Yeager, Romero, et al., 2016). The procedures described here were validated in a successful pilot intervention conducted in a convenience sample, which raised low-achieving students' grades (Yeager, Romero, et al., 2016).

The NSLM provides a motivating example for this chapter for three reasons. First, questions regarding the generalizability of mindset interventions have received recent policy interest (e.g., see Executive Order No. 13707, 2015). Second, the NSLM is one of only a handful of educational and social welfare experiments to include both probability sampling and random

assignment (see Olsen et al, 2013). Third, and most critically, the NSLM illustrates well the theoretical points raised in this chapter so far: why a convenience sample is inadequate for estimating treatment effects, why typical probability samples are inadequate, why typical moderation analyses are confounded, and how a theory of heterogeneity can inform better design that addresses these issues.

Design of NSLM

Population frame. The inference population was based on a school sampling frame created by ICF International that included both information from the Common Core of Data (CCD), a file of public schools obtained from the National Center for Education Statistics (NCES), and data from Market Data Retrieval (MDR). The MDR data files augment and update CCD data for public schools. This inference population was defined to include only schools serving grades 9 – 12 (e.g., excluding K – 12 schools), since a focus was on the effect of a program for students transitioning to a new school environment. Only public 'regular' schools were included, thus excluding private, charter, Bureau of Indian Affairs, Department of Defense, and other exceptional schools. This resulted in a frame that included over 12,000 schools.

Two-stage design. The sample included two stages of selection. First, the >12,000 schools were divided into primary sampling units (PSUs). In some cases these aligned with school district boundaries, while in other cases they involved combining smaller districts. This created 4,693 (PSUs). Methods for probability proportional to size (PPS) sampling were then defined in order to first select PSUs and then schools. This was done to reduce cost of recruitment (which was done face-to-face). Within most schools, a census of students was recruited. In some schools, where computer lab availability was limited, a random sample of students was recruited.

Non-response. While the intervention in the NSLM is brief, it requires school resources, including working computers and the ability of a school to get all 9th grade students through the computer lab in adequate time, and the evaluation requires extraction of sensitive student records.

In anticipation of school-level non-response, ICF selected a stratified random sample of 140 high schools, with the expectation that more than 70 would agree to participate in the study; 76 did.

Experimental design. The experiment begins at the beginning of high school, in the fall semester, and is delivered through a computer program. Within each of the schools, 9th grade students were randomized to treatment condition. Randomization was conducted within the software itself. This person-level random-assignment also allowed for separate treatment impacts to be estimated in each school.

Treatment. Students randomized to the treatment received a brief mindset intervention (< 2 hours) (Yeager, Romero, et al., 2016); students randomized to the control condition received a brief, comparable non-mindset program focused on the transition to high schools (also < 2 hours). The study involves two doses of the program, with at least 2 weeks between doses. "Intent to treat" analyses are conducted with students regardless of whether they received the second dose.

Outcomes. Outcomes are survey responses regarding motivation (assessed during the second computerized session), as well as outcomes from administrative records gathered at the end of the academic year, such as average grade point average and proportion earning D or F averages in core classes.

Developing a Theory of Treatment Effect Heterogeneity

Goals. In addition to estimating a population ATE, the goal of the NSLM from the outset was to determine what kinds of schools show weaker or stronger treatment effects. This focus on treatment effect heterogeneity led to the considerations and design developed in this chapter, with a particular concern on issues of confounding in moderator analyses.

Potential moderators. In order to address concerns with confounder bias in the moderator analyses, a list of possible moderators was developed at the study outset and, based upon theory, two moderators were selected: school achievement and school minority

composition. More specifically, these were selected based upon the theory that a mindset message should boost motivation and achievement when (a) the mindset message counter-acts some other message that is suppressing motivation, and when (b) students attend schools where motivation matters for learning. Hence, the mindset intervention, which teaches that "smartness" is not fixed but can be developed, might be more effective (a) for students whose intelligence may be impugned by negative stereotypes (Steele, 2011), which in the U.S. is often students of color (specifically Black or Hispanic/Latino students), and (b) in schools that reward at least some level of motivation (i.e., not the worst schools) but where motivation is not already maximal (i.e. not the best schools).

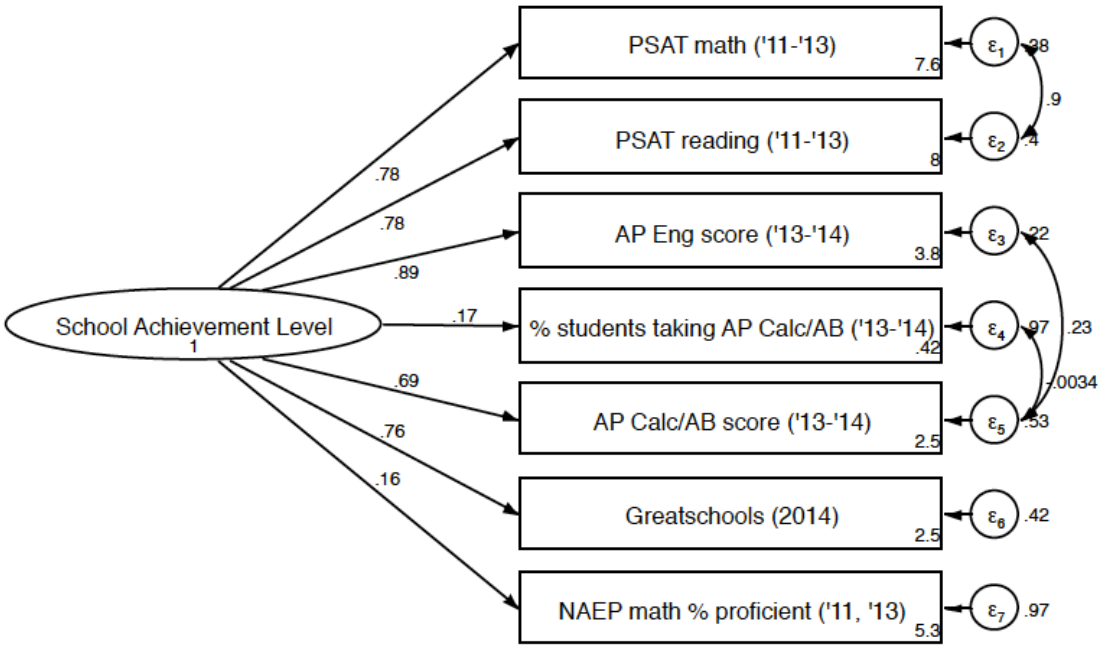
Operationalizing school achievement level. As noted, intervention effects might vary in relation to school achievement level, including features of the students (e.g., test scores, motivation), teachers, and school (e.g., offering AP classes, quality). Information of this type, however, was not readily available in the CCD, which ICF used to create the sample selection plan. This meant that the team needed to develop such a measure.

The creation of a school achievement variable involved pooling data from several outside sources. The first of these sources was GreatSchools.org, which provides within-state information on schools (and rankings from 0 - 10) based upon school test scores, as well as other features (e.g., parent and community ratings, absenteeism, programs offered). The second source was high school average PSAT scores and Calculus AB and English (Literature and Language) AP participation rates, which were provided from the College Board. The third source was state proficiency levels for math and reading for 8th grade (gathered from the National Assessment of Education Progress [NAEP]). Notably, while PSAT and NAEP scores all identify features of student achievement, the offering of AP courses speaks instead to decisions made at the institutional level in schools.

With these three data sources combined, a structural equation model was used to estimate a latent "school achievement" variable. The model and factor loadings are given in Figure 1

below. Based on this variable and the associated loadings, a "school achievement" value was estimated for each school; this variable was then standardized. Finally, the school achievement variable was divided into strata. Based upon a theory that the treatment impact would be largest in 'average' schools, three strata were created based on the 25th and 75th percentiles.

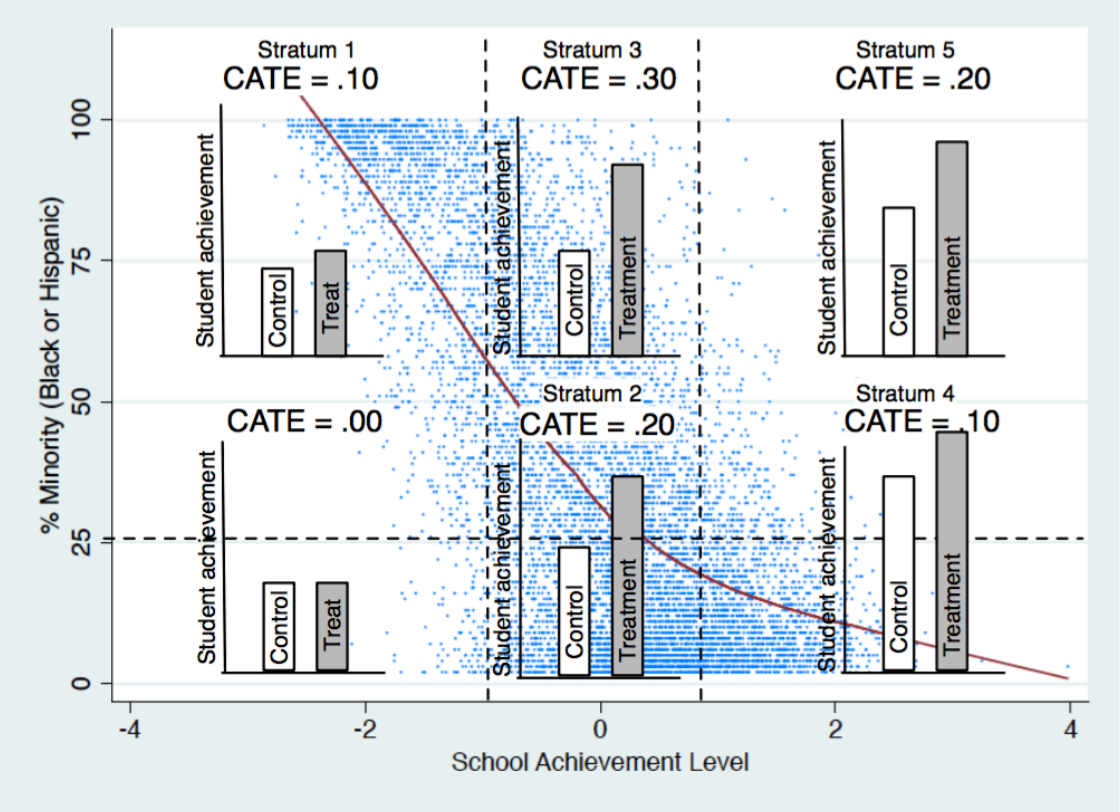
Figure 1. A latent variable model of school-level school achievement, estimated from various indicators of test scores, advanced achievement, and community ratings. Measures coming from the same data source were allowed to be correlated to adjust for shared method variance and to improve model fit; RMSEA<.10.



Orthogonalizing for minority composition. The obvious empirical challenge is that school minority composition is confounded with school average achievement. See Figure 2, which visually depicts this. The scatterplot shows the sampling frame of more than 12,000 regular U.S. public schools in 2014, and association between the % minority students (Black or Hispanic/Latino), on the y-axis, and, on the x-axis, the average school achievement level

(variable described below in Figure 2; r = -.67). A red line depicts the loess smoothing curve for the bivariate relation. Note the paucity of "off-diagonal" cases; very few schools have below-median minority composition and bottom 25% achievement level (the bottom left cell) and very few schools have above-median minority composition and top 25% achievement level (the top right cell).

Figure 2. CATE hypotheses by strata defined by school achievement by % minority, including correlation between moderators.



Given this concern, minority composition (% black or Hispanic/Latino) was divided into two groups based upon the median value in the population (26%). The strata for minority composition were then crossed with those for school achievement. This resulted in six possible strata.

Developing the Final Design

Stratum creation. The resulting design included 2 x 3 = 6 strata. In order to develop the final design, hypotheses regarding the CATEs in each stratum were specified. In Figure 2, overlain on top of this bivariate plot are theorized control and treatment group levels of student achievement—that is, hypothesized CATEs. Notice that, from left to right, the control group rises regardless of minority composition, because higher-achieving schools are expected to have higher achievement. And yet the size of the treatment *contrast* is thought to be largest in the middle range of school achievement, because of our hypothesis that in the lowest-achieving schools in the U.S. motivation is swamped by more basic concerns (like safety or curricular resources), while for the highest-achieving schools, all students may be nearly-maximally motivated.

In Figure 2, when comparing top to bottom, control condition levels are expected to be weaker in high-minority schools due to stereotype threat, which is the worry about contending with negative stereotypes about one's group (Steele, 2011). However, because the mindset intervention presumably alleviates a portion of the consequences of stereotype threat (Aronson et al., 2002), then the treatment *contrast* is hypothesized to be larger in high-minority schools versus low-minority schools. Adding these two moderation effects together, the largest treatment contrast was expected for high-minority, medium-quality schools (Stratum 3 in Figure 2).

Based on this theory, it was decided that two of the strata could be combined, the 'high' and 'low' minority by 'low' context strata, since it was hypothesized that effects would be similar in both and since testing these differences were not a high priority. This resulted in five strata, given in Table 1.

Stratum allocation. The first row of Table 1 provides the proportion of the population of schools in each of the five strata. Importantly, while these are mostly evenly distributed, the fifth stratum – containing 'high' minority 'high' context schools – included only a very small proportion of the population (approximately 5%; also recall Figure 2). A typical probability sample could easily have included far too few of these schools to reliably test hypotheses.

In the second row of the table, the number of schools (out of 140) is given that would have been included using proportional allocation. Given the goals of pair-wise comparisons between strata, the third row indicates the allocation that would have been best for stratum contrasts (equal proportions across strata). Here the differences are largest with respect to the fifth stratum again, which would have moved from 5 to 15 schools, a five-fold increase. The fourth row indicates the final allocation used for sample selection. This allocation offers a compromise between the previous two, taking into account expected CATEs in each stratum (which were higher, based on theory, in some strata than others).

Implementing sample selection. Implementing the stratified selection design was not straightforward because the strata were defined based upon *school* characteristics, but the sample selection process involved first the selection of PSUs. The final design included the selection of 70 sample PSUs with probabilities proportional to size (PPS) at the first stage; at the second stage, 140 sample schools were selected with equal probabilities within strata (i.e., two schools from each sample PSU). The measure of size assigned to each school was eligible 9th grade enrollment.

Final sample. Sample selection for the study took place over the Summer and Fall of 2016. Overall, 76 schools agreed to take part. The final numbers of schools per strata can be located in the bottom row of Table 1. Even with non-response, this increased the proportion of the sample in the High-High stratum (the most rare subgroup) threefold. The analysis of the study results are currently in process, with preliminary results expected in Summer 2017.

Table 1. Strata definitions and allocations in NSLM

School achievement level	**Low**	**Med**	**Med**	**High**	**High**
Minority %	**Low/High**	**Low**	**High**	**Low**	**High**
Population of schools (100%)	25%	27%	23%	20%	5%
Proportional Allocation (n=140)	35 (25%)	39 (27%)	32 (23%)	28 (20%)	7 (5%)
Optimal for Strata Contrasts (n=140)	28 (20%)	28 (20%)	28 (20%)	28 (20%)	28 (20%)
Final Plan (n = 140)	28 (20%)	34 (24%)	32 (23%)	19 (14%)	26 (19%)
Final Sample (n = 76)	13 (17%)	25 (33%)	18 (24%)	9 (12%)	11 (15%)

Discussion

The NSLM is unique in several ways compared to other survey experiments. However, we argue that the focus on treatment effect heterogeneity and the design-based perspective are important in *all* survey experiments. This is particularly important as behavioral experiments are rapidly increasing in their prevalence in public policy, economics, and political science (Wilson & Juarez, 2015). As we conclude this chapter, we therefore focus on five discussion points for researchers designing survey experiments.

Estimands. Survey experiments typically focus on estimation of the ATE in a population. However, as we have argued throughout, if treatment impacts do in fact vary, then the average effect is not adequate for answering many questions regarding an intervention. Instead, testing hypotheses regarding this variation offers greater insight into *how* and *why* an intervention changes outcomes.

Contextual effects matter. In much of the survey experiment literature, the focus is on *individual* behaviors and treatment heterogeneity related to these features. However, we argue that in these situations – just as in education – context matters and should be studied. Individual behaviors are often affected by social conditions, including local culture, region, and organizational affiliations. Additionally, very often the individual is not the decision maker when implementing an intervention outside a study. For example, in political science studies focused on Get-Out-the-Vote campaigns, resources may be allocated at the local, village, regional, or state level, and as such moderators associated with these contexts matter.

Plan for moderators. Much of recent innovations regarding treatment effect heterogeneity and generalizability in both experiments conducted in probability samples and convenience samples have focused on post-hoc analysis methods. In contrast, the approach developed here – in following recent work on generalization more broadly (e.g., Tipton, 2014; Tipton et al., 2014; Tipton & Miller, 2015) – asks researchers to begin their study with a discussion of these moderators. There are three benefits to this. First, it is possible that even in large studies, there are small subgroups that while *representing* a very small fraction of the population are important for *testing* hypotheses regarding moderator effects. Second, by identifying these at the beginning of the study, better pre-treatment data can be gathered both from other sources, as well as *from the study sample* itself. Third, by shifting to 'confirmatory' and 'exploratory' hypotheses, problems from multiple testing and data fishing are circumvented. This follows best practice in the design of experiments more generally, particularly in clinical trials, and speaks, too, to reproducibility issues in social science studies.

Designing strata. While standard in the design of surveys, strata are typically included with the goal of reducing variance of the estimators; these are typically related to measures of geography or basic demographics. In this chapter, we argued that in survey experiments strata should be designed instead with respect to variables that potentially explain variation in treatment

impacts. Creating these strata can take work – and may require the collection or combination of new sources of data.

Importantly, in some cases, scientific hypotheses are highly correlated, and in these cases, stratifying on multiple covariates can greatly increase the ability to unconfound the effects. In the NSLM example, unplanned moderation analysis of either of racial composition or school achievement would have been confounded in critical ways. Indeed, although racial-composition and school achievement level are positively correlated, the direction of the moderation, at least in the top 75% of schools, is opposite (see Figure 2). Without understanding one or the other, then the two relations might have cancelled out.

Power concerns. In this chapter, we argue that when designing a survey experiment to study treatment effect heterogeneity, compromises will be required in order to accommodate the multiple estimands of focus. This is because the design that is optimal for one estimand (e.g., the PATE) is not always optimal for another estimand (e.g., moderators). As we showed in the NSLM, with only 76 clusters, moderator analyses would have been greatly under-powered had prior planning not taken these into account in the study design.

Conclusion

In this chapter, we have argued for a new design-based approach to studying treatment effect heterogeneity and causal mechanism in survey experiments. The focus of this approach is on the development of a sample selection procedure that enables estimation of three parameters – the population average treatment effect, subgroup impacts, and contrasts between subgroups. This approach was developed in relation to a real example – that of the NSLM– that included76 public high schools throughout the United States. Our hope is that future survey experiments will learn from the NSLM design, leading researchers to develop not only generalizable estimates of average treatment effects, but also better theory regarding the mechanism behind these effects.

References

Allcott, H. (2015). Site selection bias in program evaluation. *The Quarterly Journal of Economics*, *130*(3), 1117–1165. https://doi.org/10.1093/qje/qjv015

Aronson, J. M., Fried, C. B., & Good, C. (2002). Reducing the effects of stereotype threat on African American college students by shaping theories of intelligence. *Journal of Experimental Social Psychology*, *38*(2), 113–125. https://doi.org/10.1006/jesp.2001.1491

Battle, J., & Lewis, M. (2002). The increasing significance of class: The relative effects of race and socioeconomic status on academic achievement. *Journal of Poverty*, *6*(2), 21–35.

Bloom, H. S. (2010). *Nine lessons about doing evaluation research: Remarks on accepting the Peter H. Rossi Award*.

Bryan, C. J., Walton, G. M., Rogers, T., & Dweck, C. S. (2011). Motivating voter turnout by invoking the self. *Proceedings of the National Academy of Sciences*, *108*(31), 12653–12656. https://doi.org/10.1073/pnas.1103343108

Bryk, A. S. (2009). Support a science of performance improvement. *Phi Delta Kappan*, *90*, 597–600.

Cole, S. R., & Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *American Journal of Epidemiology*, *172*(1), 107–115. https://doi.org/10.1093/aje/kwq084

Cook, T. D. (1993). A quasi-sampling theory of the generalization of causal relationships. *New Directions for Program Evaluation*, *1993*(57), 39–82. https://doi.org/10.1002/ev.1638

Dweck, C. S. (2006). *Mindset: The new psychology of success*. New York, NY: Random House.

Entman, R. M. (2010). Media framing biases and political power: Explaining slant in news of Campaign 2008. *Journalism*, *11*(4), 389–408. https://doi.org/10.1177/1464884910367587

Feller, A., & Holmes, C. C. (2009). Beyond toplines: Heterogeneous treatment effects in randomized experiments. *Unpublished Manuscript, Oxford University*.

Gelman, A. (2014). The connection between varying treatment effects and the crisis of unreplicable research a Bayesian perspective. *Journal of Management*, 149206314525208.

Gelman, A., Su, Y.-S., Yajima, M., Hill, J., Pittau, M. G., Kerman, J., … Su, M. Y.-S. (2015). Package "arm." *Data Analysis Using Regression and Multilevel/Hierarchical Models*.

Green, C. L. (1999). Ethnic evaluations of advertising: Interaction effects of strength of ethnic identification, media placement, and degree of racial composition. *Journal of Advertising*, 49–64.

Green, D. P., & Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opinion Quarterly*, *76*(3), 491–511. https://doi.org/10.1093/poq/nfs036

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960. https://doi.org/10.2307/2289064

Horiuchi, Y., Imai, K., & Taniguchi, N. (2007). Designing and analyzing randomized experiments: Application to a Japanese election survey experiment. *American Journal of Political Science*, *51*(3), 669–687.

Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness*, *2*(1), 88–110. https://doi.org/10.1080/19345740802539325

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *171*(2), 481–502.

Imai, K., & Strauss, A. (2011). Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political Analysis*, *19*(1), 1–19. https://doi.org/10.1093/pan/mpq035

Keeter, S., Kennedy, C., Dimock, M., Best, J., & Craighill, P. (2006). Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey. *Public Opinion Quarterly*, *70*(5), 759–779. https://doi.org/10.1093/poq/nfl035

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*(1), 537–567. https://doi.org/10.1146/annurev.psych.50.1.537

Markus, H. R., & Kitayama, S. (2010). Cultures and selves: A cycle of mutual constitution. *Perspectives on Psychological Science*, *5*(4), 420–430. https://doi.org/10.1177/1745691610375557

Morgan, S. L., & Winship, C. (2014). *Counterfactuals and causal inference*. Cambridge University Press.

Mullainathan, S., & Shafir, E. (2013). *Scarcity: Why having too little means so much*. Macmillan.

Mutz, D. C. (2011). *Population-based survey experiments*. Princeton, NJ: Princeton University Press.

Narayan, S., & Krosnick, J. A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly*, *60*(1), 58–88. https://doi.org/10.1086/297739

Olsen, R. B., Orr, L. L., Bell, S. H., & Stuart, E. A. (2013). External validity in policy evaluations that choose sites purposively. *Journal of Policy Analysis and Management*, *32*(1), 107–121. https://doi.org/10.1002/pam.21660

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*(6), 657–660. https://doi.org/10.1177/1745691612462588

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716-1-aac4716-8. https://doi.org/10.1126/science.aac4716

Open Science Framework. (2016). The $1 million preregistration challenge by the center for open science. Retrieved from https://osf.io/peut2/

Paunesku, D., Walton, G. M., Romero, C., Smith, E. N., Yeager, D. S., & Dweck, C. S. (2015). Mindset interventions are a scalable treatment for academic underachievement. *Psychological Science*, *26*(6), 284–293. https://doi.org/10.1177/0956797615571017

Rattan, A., Savani, K., Chugh, D., & Dweck, C. S. (2015). Leveraging mindsets to promote academic achievement policy recommendations. *Perspectives on Psychological Science*, *10*(6), 721–726.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*(2), 199.

Rothman, K., & Greenland, S. (2005). Causation and causal inference in epidemiology. *American Journal of Public Health*, *95*(S1), S144–S150.

Rothman, R. L., Malone, R., Bryant, B., Shintani, A. K., Crigler, B., Dewalt, D. A., … Pignone, M. P. (2005). A randomized trial of a primary care-based disease management program to improve cardiovascular risk factors and glycated hemoglobin levels in patients with diabetes. *The American Journal of Medicine*, *118*(3), 276–284. https://doi.org/10.1016/j.amjmed.2004.09.017

Shadish, W. R., Cook, T. D., & Campbell, T. D. (2001). *Experimental and quasi-experimental designs for generalized causal inference* (2nd ed.). Boston, MA: Wadsworth Publishing.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Smith, S. S. (2010). Race and trust. *Annual Review of Sociology*, *36*(1), 453–475. https://doi.org/10.1146/annurev.soc.012809.102526

Solon, G., Haider, S. J., & Wooldridge, J. M. (2015). What are we weighting for? *Journal of Human Resources*, *50*(2), 301–316.

Spencer, M. B. (2006). Phenomenology and ecological systems theory: Development of diverse groups. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology, Vol. 1: Theoretical models of human development* (6th ed.). New York: Wiley & Sons.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*(6), 613–629. https://doi.org/10.1037/0003-066X.52.6.613

Steele, C. M. (2011). *Whistling Vivaldi: How stereotypes affect us and what we can do*. New York, NY: W. W. Norton & Company.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*(5), 797–811. https://doi.org/10.1037/0022-3514.69.5.797

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. https://doi.org/10.1111/j.1467-985X.2010.00673.x

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

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. https://doi.org/10.3102/1076998612441947

Tipton, E. (2014). How generalizable is your experiment? An index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics*, *39*(6), 478–501. https://doi.org/10.3102/1076998614558486

Tipton, E. (2015). *Planning for generalization with stratified selection: Design parameters and sample size requirements for use in power analysis*. Presented at the Annual Meeting of the Association for Public Policy Analysis and Management, Miami, FL.

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. https://doi.org/10.1080/19345747.2013.831154

Tipton, E., & Miller, K. (2015). Generalizer [Web-tool]. Retrieved from http://www.generalizer.org

Tucker-Drob, E. M. (2011). Individual differences methods for randomized experiments. *Psychological Methods*, *16*(3), 298–318. https://doi.org/10.1037/a0023349

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

Weiss, M. J., Bloom, H. S., & Brock, T. (2014). A conceptual framework for studying the sources of variation in program effects. *Journal of Policy Analysis and Management*, *33*(3), 778–808. https://doi.org/10.1002/pam.21760

Wilson, T. D., & Juarez, L. P. (2015). Intuition is not evidence: Prescriptions for behavioral interventions from social psychology. *Behavioral Science & Policy*, *1*(1), 13–20.

Yeager, D. S., & Dweck, C. S. (2012). Mindsets that promote resilience: When students believe that personal characteristics can be developed. *Educational Psychologist*, *47*(4), 302–314. https://doi.org/10.1080/00461520.2012.722805

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., & Wang, R. (2011). Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, 1–39. https://doi.org/10.1093/poq/nfr020

Yeager, D. S., Romero, C., Paunesku, D., Hulleman, C. S., Schneider, B., Hinojosa, C., … Dweck, C. S. (2016). Using design thinking to improve psychological interventions: The case of the growth mindset during the transition to high school. *Journal of Educational Psychology*, *108*(3), 374–391. https://doi.org/10.1037/edu0000098

Yeager, D. S., Walton, G. M., Brady, S. T., Akcinar, E. N., Paunesku, D., Keane, L., … Dweck, C. S. (2016). Teaching a lay theory before college narrows achievement gaps at scale. *Proceedings of the National Academy of Sciences*, *113*(24), E3341–E3348. https://doi.org/10.1073/pnas.1524360113