Dry Runs and "PWRD" Aggregation: Two New Methods for Extracting Power from Careful Observation of a Randomized Controlled Trial's Context

by

Timothy P. Lycurgus

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in the University of Michigan
2022

Doctoral Committee:

Associate Professor Ben B. Hansen, Chair
Assistant Professor Johann Gagnon-Bartsch
Professor Edward Ionides
Research Associate Professor Brady West

Timothy P. Lycurgus
tlycurgu@umich.edu
ORCID iD: 0000-0002-7600-4011

DEDICATION

To my father, Peter Lycurgus, for sparking my love of statistics through our shared passion for baseball.

ACKNOWLEDGMENTS

I would like to express my immense gratitude to my advisor Ben Hansen for his mentorship and support during my time in graduate school. His knowledge, patience, and time were instrumental during this process and I am forever grateful.

I would like to thank my committee members, Professors Johann Gagnon-Bartsch, Edward Ionides, and Brady West, for their time and feedback on my dissertation. I would like to thank the entire Statistics Department, especially the support staff, who made this process immeasurably easier.

I would like to thank Mark Fredrickson and Josh Errickson for the many discussions on these projects along with all of the advice they provided. I would like to thank my friends Roger Fan, Jack Goetz, and Byoung Jang for their friendship and support throughout my time at Michigan.

I would like to thank Daniella Raz for her support and humor, making these last few years fun and enjoyable despite a global pandemic.

Lastly, I would like to thank my family for their years of love and support. They encouraged my initial love for statistics through days studying baseball boxscores and the back of baseball cards.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

ABSTRACT

Well-conducted field experiments, broadly construed to contain both randomized controlled trials and quasi-experiments, involve extensive planning with substantive deliberation. Such deliberation has the potential to fuel and strengthen the analysis stage of the study. Each field experiment is unique in its own manner, from the subgroups on which effects are expected to concentrate to the design of the study itself. Reliance on off-the-shelf methods to analyze field experiments may exclude this potentially valuable information that, if handled properly, would provide a greater opportunity to detect an effect. In this dissertation, we propose two novel methods that look to extract information unique to a specific study and translate it into additional power. We demonstrate these methods on a large-scale education intervention aimed at correcting the stalled reading trajectories of early elementary students.

The first method, Power-maximizing Weighting for Repeated-measurements and Delayed-effects (i.e. PWRD) aggregation, converts the theory of change behind a class of education interventions into a test statistic that maximizes the asymptotic relative efficiency (Pitman, 1948) over standard methods, thus providing greater power. The scheme emphasizes cohorts and years-of-follow-up on which effects are expected to accrue with appropriate attention paid to the relative precision of estimates within cohorts. We find through a simulation study mirroring the design of the reading intervention that this method provides stark gains in power over methods common to education research. While PWRD aggregation increases power, confidence interval estimation is more difficult. To alleviate this problem, we partition our parameter space into three regions: equivalence, superiority, and inferiority. In the first, we employ PWRD aggregation to provide the greatest opportunity to detect an effect. In the latter two regions, we employ a method common to the domain in which the research occurs such that when we are able to detect an effect, interpretation of the point estimate and confidence interval proceeds in a standard fashion.

The second method we propose is a dry run simulation scheme that creates a pseudo-experiment replicating the initial randomized trial in a manner that preserves blinding to impact estimates. This procedure, which uses real rather than synthetic data, provides a sandbox in which various models may be tested to discover the model specification that most precisely estimates an artificially imposed treatment effect. The dry run method, similar in motivation to cross-validation and uniformity trials (Rosenbaum, 2018), allows the statistician advising field experiments to estimate

expected losses for each of a variety of methods, enabling them to elect a novel or unfamiliar method if it demonstrably outperforms methods more familiar to the broader team. When applied to the reading intervention that motivated dry runs, results from this method challenged preconceived notions about covariate choice, suggesting we control for covariates beyond pre-test scores.

CHAPTER 1

Introduction

Though causal inference has a long history, the central question of the discipline under the modern framework —proposed by Donald Rubin in 1974 —is whether and how an outcome differs under the presence of a treatment as opposed to in its absence (Rubin, 1974). To answer this question, researchers in the social sciences often rely on randomized controlled trials (RCTs) or quasi-experiments. Randomized trials are generally viewed as the "gold standard" in establishing causal relationships. Due to random treatment assignment, they are typically free from bias from confounding variables and thus, the observed gap between the treatment and the control is often a reasonable estimate of the treatment effect.

However, these "field experiments" may be prohibitively expensive and thus involve detailed planning and substantive deliberation, which has the ability to strengthen the analysis of the study. Each of these field experiments is unique, from the study design to the mechanism by which the treatment is hypothesized to provide a benefit or detriment. While standard, off-the-shelf methods offer good analysis strategies in general, such techniques may exclude potentially valuable information about the study and therefore may not be the optimal method when studying a specific randomized trial or quasi-experiment. In this dissertation, we propose two novel methods —*p*ower-maximizing *w*eighting for *r*epeated measurements with *d*elayed effects (PWRD) aggregation and dry-run simulations—that look to extract information unique to a randomized trial or quasi-experiment and convert it into statistical power. We demonstrate the benefits of these methods both through simulation studies and through their application on problems of substantive importance in health and education.

These methods emerge from the analysis of a large-scale cluster-randomized trial assessing an education intervention, BURST[R]: Reading (BURST), that aimed to correct the stalled learning trajectories of early elementary students (Rowan et al., 2019). The intervention, equipped with a well-developed theory of change, provided supplemental learning to a subset of students once they required intervention; thus, the effect was delayed (students must have tested into the intervention before receiving a potential benefit) and scattered across the study population (not every student required supplemental instruction). As a consequence, the effect was not uniform across those students assigned to the treatment, but concentrated to a greater or lesser extent on certain cohorts and in certain years-of-follow-up. While we found this intervention to be unsuccessful at assisting students in their ability to read, BURST served as the motivating example for many of the methods presented in this dissertation which seek to leverage such information about the study to conduct more precise and powerful outcome analysis.

1.1 PWRD Aggregation

In Chapter 2, we introduce a method of effect aggregation that converts the theory of change behind this intervention into statistical power. Our method, Power-maximizing Weighting for Repeated-measurements and Delayed-effects (i.e. PWRD) aggregation, provides greater weight to those cohorts and occasions of follow-up that are best-positioned to demonstrate an effect of the intervention, while accounting for the relative precision of estimates within these subgroups. This scheme maximizes the asymptotic relative efficiency of our test statistic (Pitman, 1948) when the theory of change holds, which in turn maximizes our ability to detect an effect. A simulation study mirroring the design of BURST lends evidence to this claim. Simulations show that PWRD aggregation provides substantial increases in one's ability to detect an effect when that theory of change holds over methods common in education such as mixed models. We additionally find that when the theory of change does not hold, the detriment in terms of power to using PWRD aggregation is minimal.

PWRD aggregation is related in concept, although not in its aims, to instrumental variables estimation (Bloom, 1984; Angrist et al., 1996) and principal stratification (Frangakis and Rubin, 2002). The connection to the principal stratification method of Sales and Pane (2021) is perhaps most clear. Sales and Pane (2021) apply principal stratification with the goal of estimating separate effects for latent subgroups determined through their dosage. We too estimate separate effects for different subgroups, yet those subgroups are determined based on projected differing dosages and our method's ultimate goal is to aggregate those separate effects into a single test statistic. As with instrumental variables estimation in experimental settings without full compliance, PWRD aggregation also typically assumes that the effect is proportional to the dosage received. In contrast to instrumental variables, however, our method is fully compatible with intention-to-treat (ITT) estimation. In fact, PWRD aggregation may be viewed as ITT analysis with an "as-treated" flavor: every treatment observation is incorporated into our outcome analysis regardless of whether they received the treatment itself, yet we implicitly attach greater importance to those observations most likely to have received stronger dosages of the treatment. While this may lead to fears of exploitation rather than exploration on our part, note that this method is centered around the theory of change of an intervention which must be determined prior to outcome analysis. Furthermore, our method attaches greater importance to those observations that we *expect* to demonstrate a greater treatment effect; in practice, this may be a faulty assumption.

1.2 Confidence Intervals for PWRD Aggregation

A standard method for constructing confidence intervals with effects proportional to the dosage is to estimate a confidence interval for that proportionality constant (see (Rosenbaum et al., 2010, p.135-6)). While PWRD aggregation implicitly assumes that the effect will be proportional to the dosage, our aggregation scheme is not compatible with the method outlined in Rosenbaum et al. (2010). Furthermore, PWRD aggregation was constructed with hypothesis testing, rather than confidence intervals, in mind and as a consequence, the standard point estimate for PWRD aggregation may not be easily interpretable to less technical audiences. In Chapter 3, we propose a novel method of constructing confidence intervals in tandem with PWRD aggregation through adaptation of the three-sided hypothesis testing scheme of Goeman et al. (2010). We partition the parameter space into regions of equivalence (i.e. no effect), inferiority (a negative effect), and superiority (a positive effect). Within equivalence, we apply PWRD aggregation, giving us the greatest opportunity to yield a non-zero effect. Outside the equivalence region, we apply a standard method to ensure our confidence interval and point estimates are easily interpretable. A key step behind this method requires partitioning the parameter space into the three regions described previously, so Chapter 3 additionally provides guidance for setting the thresholds of the equivalence region. This includes proposing a threshold that, like PWRD aggregation, leverages the benefits of PWRD aggregation in terms of asymptotic relative efficiency versus commonly implemented methods. This sets the bounds of the equivalence region such that the power advantage of PWRD aggregation will be maximized. Chapter 3 demonstrates this method through the same simulation study described in Chapter 2 and on Medicaid expansion through the Affordable Care Act.

1.3 The Dry Run Simulation Scheme

Research has shown that incorporating covariates in outcome analysis may improve precision, both in education settings and elsewhere in the social sciences. Yet which covariates should be incorporated remains an open debate. Some argue that merely controlling for pre-test scores will fully account for baseline differences between the treatment and the control and thus is sufficient covariate adjustment (Bloom et al., 2007). Others suggest that incorporating additional covariates may further improve one's precision (Raudenbush, 1997). Ultimately, there is no one-size-fits-all approach and the set of covariates selected a priori may not provide researchers with the most powerful test in a given experiment.

To address this issue, we present an adaptation of the dry-run simulation scheme of Wyss et al.

(2017) in Chapter 4. Wyss et al. (2017) initially proposed their dry run method as an evaluation strategy for a prognostic score model's ability to control for confounding in observational studies. We extend this method to randomized trials, not with the aim of assessing prognostic score models, but with the aim of estimating the future performance of models in outcome analysis. In addition, we enhance the method through an additional subsampling step that lessens the risk of overfitting. Briefly, this method replicates the full randomized trial solely using observations assigned to the control which creates a pseudo-experiment that preserves blinding to impact estimates. Repeating this process then provides many different simulated "realizations" of the study on which various models may be tested to determine which specification most precisely estimates some artificially imposed treatment effect.

Dry runs in their simplest form may be viewed as a variant on cross-validation: rather than dividing our sample into training and test sets to estimate a model's predictive performance, we divide our control data into pseudo-treatment and pseudo-control groups to estimate the precision of different model specifications. Cross validation typically iterates through different folds to examine alternative training and testing sets. Similarly, dry runs iterate through different sets of pseudo-treatment and pseudo-control groups to examine alternative pseudo-experiments. Dry runs are also akin in motivation to uniformity trials (Rosenbaum, 2018, p.33), a technique popular in the 1920s and 1930s in which plots of land were randomly divided into treatment and control groups yet all received the control condition. This allowed researchers to empirically test how much the treatment and control could differ through random chance. Dry runs may be viewed as a uniformity trial embedded within a randomized trial: the control is divided into pseudo-treatment and pseudo-control groups, yet no observation received the treatment in actuality. This then allows the researcher to examine model performance without any imposed treatment. Where our method deviates from uniformity trials, however, is that we additionally allow for artificially imposed treatment effects on the pseudo-groups originating from control data in order to compare model performance in the presence of a treatment.

This procedure, compatible with both standard randomized trials incorporating simple random assignment as well as cluster randomized trials or randomized trials embedded within survey designs, facilitates selection of covariate adjusted models in a manner that uses real rather than synthetic data. In addition to covariate selection, it also allows researchers to estimate expected losses for a variety of methods, from the methods standard to education research (e.g. hierarchical linear models (Raudenbush and Bryk, 2002; Bryk and Raudenbush, 1987) or ordinary least squares) to novel methods like a Peters-Belson approach that models covariate adjustment strictly on control observations (Peters, 1941; Belson, 1956). The statistician advising field experiments may use dry-runs to select one of those more novel methods so long as it demonstrably outperforms the standard methods familiar to the broader team. In our application on BURST, this technique revealed a distinct advantage to ignoring certain preconceived articles of education methodology, suggesting we control for covariates beyond pre-tests.

CHAPTER 2

PWRD Aggregation

2.1 Introduction

Many large-scale randomized controlled trials (RCTs) and high-quality quasi-experiments are conducted only after careful vetting in national funding competitions. In the United States, a leading competition for education efficacy studies is the Institute of Education Sciences's (IES) Education Research Grants program, which aims to contribute to education theory by informing stakeholders of learning interventions' costs and benefits. "Strong applications" to the program are expected to detail and justify an intervention's "theory of change" (NCER, 2020, p.48): How and why does a desired improvement in outcomes occur as a consequence of the intervention?

This paper introduces a novel scheme, PWRD aggregation of effects, converting theories of change into statistical power. Given an efficacious program, a correct theory of change, and measurements indicating which students stand to benefit, this *p*ower-maximizing *w*eighting for *r*epeated measurements with *d*elayed effects method can increase the probability of detecting program benefits, in some cases dramatically. It is compatible with the range of clustering accommodations and covariate adjustment techniques that are commonly used for analysis of education RCTs. It maintains the canonical intention-to-treat (ITT) perspective on program benefits. While applicable to studies with or without measures of implementation, with single as well as multiple occasions of follow-up, it maximizes its advantage when there are baseline or post-treatment measures of intervention delivery or availability, in combination with primary outcomes measured on varying numbers of occasions.

We illustrate our scheme on an IES Education Research Grant-funded efficacy trial of an intervention for early elementary students at risk of falling behind in learning to read. This intervention, BURST[R]: Reading (BURST), aims to detect and correct deflections from what would otherwise be students' upward trajectory in reading ability. The theory of change for BURST posits this "trajectory correction" arises by providing targeted instruction to students whose progress has deviated from the expected course (e.g. tested below a certain benchmark). Thus, effects are delayed—students do not immediately obtain an effect but must first receive targeted remediation—and non-uniform in that only students with stalled reading abilities are affected. As a consequence of this theory of change, the treatment effect will be anything but constant; if the intervention works in the hypothesized manner, its effects will be greatest at those ages and for those subgroups for which student learning has already begun to stall. Accordingly, beginning from estimates of the average treatment effect (ATE) calculated separately for different subgroups and occasions of follow-up, as well as information about the extent of stalled progress at each occasion, PWRD aggregation combines effect estimates not only with attention to their mutual correlations, but also with attention to their expected sizes relative to one another. These expectations are determined by a carefully structured set of alternative hypotheses, which PWRD aggregation in turn adduces from the environing theory of change.

In underlying concept if not in its goals, the method relates to instrumental variables estimation (Bloom, 1984; Angrist et al., 1996; Baiocchi et al., 2014) and principal stratification (Frangakis and Rubin, 2002; Page, 2012; Sales and Pane, 2019). But whereas Sales and Pane (2021), for example, use principal stratification to estimate separate effects for latent subgroups distinguished in terms of dosage level, we marshal related considerations to inform aggregation of effects across manifest subgroups receiving or likely to receive differing doses. For recent evaluation methodology using dosage information in other manners (e.g. to determine fidelity of implementation or to define the causal parameter of interest) see Schochet (2013) and White et al. (2019).

2.1.1 Literature Review

Extensive literature addresses the issue of causal effect estimation for time-varying exposures but largely through methods other than effect aggregation. Instrumental variables (IV) (Bloom, 1984) are one prominent example. This technique, utilized broadly in economics literature, is gaining influence in other fields as well. Briefly, researchers use regression-based instrumental variables in scenarios where the explanatory variable of interest is correlated with the error of a regression, potentially through measurement error, omitted variable bias, or confounding. By applying a valid instrument, i.e. a variable that itself is not predictive of the outcome but is conditionally correlated with predictors, researchers can consistently estimate the causal effect of that predictor despite its correlation with the error.

While not their primary aim, IV approaches are not entirely incompatible with intention-to-treat (ITT) analysis or randomized trials. Sussman and Hayward (2010) actually refer to instrumental variable analysis in randomized trials as a contamination-adjusted intention-to-treat (CAITT) analysis where treatment assignment serves as the instrument. Under this framework, the ITT estimator is then adjusted by the proportion of participants who receive the treatment. IV analysis in this setting can be referred to as as a contamination adjusted intention-to-treat analysis because the two-stage least squares estimator is equivalent to the ITT estimate prior to its scaling by the proportion of compliers (Baiocchi et al., 2014). Nonetheless, this scaling marks a departure from standard intention-to-treat analysis. Under certain conditions, this IV estimand will be identical to the complier-average causal effect (CACE) (Angrist et al., 1996; Baiocchi et al., 2014), which measures the average effect of treatment in the subgroup of compliant individuals, i.e. individuals who adhered to their treatment assignment. To illustrate, among those assigned to the treatment, the CACE only examines the subgroup who actually received the treatment.

IV approaches are applicable in scenarios with partial compliance as well; here, the researcher tests the hypothesis that the effect is proportional to the dose of treatment received (Rosenbaum et al., 2010). Under this framework, ITT analysis rejects a hypothesis of no effect if and only if the IV method also rejects the hypothesis of no effect. IV analysis and the CACE have both been extended to the longitudinal setting, allowing for repeated observations and subjects with incomplete observations over time.

Another area of research addressing the issue of causal effect estimation for time-varying exposures revolves around three related but distinct methods: the g-computation algorithm formula (i.e. the "g-formula"), inverse probability of treatment weighting (IPTW) of marginal structural models, and g-estimation of structural nested models (Robins, 1986; Robins et al., 1992) of which instrumental variables is a form (Hernán and Hernández-Díaz, 2012). These three methods fall under the general umbrella of "g-methods" and will provide identical estimates of the treatment effect under certain conditions.

Much of this literature is constructed with sequentially randomized experiments with differing treatment regimes across time in mind, similar to how students in BURST who test into the intervention receive a different treatment regimen than those who do not. For example, let us allow Z_{it} to denote the treatment received by individual i in time t with $Z_{it} = 1$ signifying receiving the treatment. Then $\bar{Z}_i$ is the treatment regime throughout the length of the experiment; we could observe $\bar{Z}_i = (1, 1, \ldots, 1)$ for continuous exposure, $\bar{Z}_i = (1, 0, \ldots, 0)$ if they are only exposed to the treatment in the first time period, or some more complicated regime.

One formulation for time-varying exposures allows us to apply marginal structural models to test a null hypothesis of no effect versus an alternative hypothesis that the outcome Y increases linearly as a function of the individual's cumulative exposure to the treatment, $\sum_t Z_{it}$. For example, we could test this hypothesis by using ordinary least squares with IPTW (Robins et al., 2000). Note that with respect to BURST, we do not work under an assumption of increasing effect as a function of exposure (e.g. an effect of the form βD where D represents dosage), but under the assumption of increasing probability of exposure (e.g. $\beta\mathbb{P}(D = 1)$). Under the first, individuals in the treatment group could receive any effect of size βd for $d \in \{0, D\}$ whereas under the second, the effect is binary in the sense that those who are exposed receive an effect of size β and those who are not receive an effect of size 0. Nonetheless, this distinction is small and the parallel to g-methods is readily apparent. For a more in depth review of g-methods, see Fitzmaurice et al.

(2008).

Both instrumental variables and the broader class of g-methods generally formulate outcome analysis in a manner such that the mode of estimation (e.g. a difference in means or a regression coefficient) is consistent for a specific target parameter. In contrast, our goal is to conduct outcome analysis with the intention of providing a foundation for hypothesis tests about the value of a specific target parameter; as a consequence, this formulation possesses power against alternative methods. Additionally, IPTW and other flavors of g-method techniques do not fully adhere to principles of intention-to-treat analysis; PWRD aggregation, however, fully respects this form of analysis.

2.1.2 Roadmap

In this chapter, we first discuss the connection of longitudinal data in education settings to interventions with supplemental instruction to correct stalled learning trajectories. After, we use the theory of change behind this class of interventions to define assumptions under which PWRD aggregation will be power-maximizing. We then explicitly present the formulation for PWRD aggregation weights. In Section 2.3, we present a simulation study mirroring BURST design to show PWRD aggregation performance in comparison with commonly used methods under various assumptions. In Section 2.4, we then illustrate how PWRD aggregation compares with those same methods for BURST itself. Finally, in Section 2.5, we conclude by summarizing how PWRD aggregation provides researchers with a tool that will best help them detect an effect for interventions with supplemental instruction.

2.2 Method

2.2.1 Review: Comparative studies with repeated measurements of the outcome

In educational settings assessing the efficacy of interventions, students frequently enter and exit studies at different points. For example in BURST, we examined a reading intervention on early elementary students across four years. Depending on their grade at the study's outset, the number of observations on each student varied from one to four. Table 2.1 illustrates this phenomenon for BURST's first of four total cohorts.

	Grade at Entry	**Year 1**	**Year 2**	**Year 3**	**Year 4**
Cohort 1	**3**	3	-	-	-
	2	2	3	-	-
	1	1	2	3	-
	0	K	1	2	3

Table 2.1: Progression of Cohort 1 through the four years of the BURST study.

Data sources for similarly structured efficacy trials will incorporate an analogous design, with varying numbers of observations on any given participant. Thus, the method chosen to handle multiple observations is of great importance not only in BURST but in other longitudinal settings as well. The simplest outcome analysis might sidestep this debate entirely by solely examining outcomes when students exit the study (e.g. 3rd grade observations in BURST). For Cohort 1 in Table 2.1, this entails using data from the diagonal and discarding the remaining data. This method, herein termed "exit observation" analysis, treats the student rather than the student-year as the unit of analysis. Exit observation analysis is appropriate to such models as

$$Y_{ij3} = \beta_0 + \tau Z_{ij3} + \beta X_{ij3} + \epsilon_{ij3} \quad \big(\mathbb{E}(\epsilon_{ij3}) = 0;\ \mathrm{Var}(\epsilon_{ij3}) = \sigma^2\big), \tag{2.1}$$

where Y_{ij3} denotes the outcome of student i in school j in the third grade, X represents a set of demographic covariates, and Z denotes the treatment status. An example of this method may be found in Simmons et al. (2008). In addition to its simplicity, exit observation analysis provides one notable benefit: an easily defined and identified overall average treatment effect, i.e. $\mathbb{E}[Y_{ij3}^{(Z=1)} - Y_{ij3}^{(Z=0)}]$.

However, complications emerge. According to BURST's theory of change, students are more likely to benefit when they participate in the intervention for a longer period. Therefore, we are less likely to observe an effect in Cohort 1.3 than in Cohort 1.0, and treating these two groups equally may hinder a researcher's ability to detect an effect. Table 2.2 demonstrates this occurrence in BURST where we observe larger differences between unadjusted treatment and control means as students participate in the study for a greater length of time.

	Entry Grade	**Entry Year**	**Exit Year**	δ
Cohort 1	**3**	5.2	5.2	-
	2	-1.3	-0.4	0.9
	1	0.3	3.2	2.9
	0	-7.0	2.1	9.1

Table 2.2: Differences in mean reading scores between treatment and control groups for the first of four cohorts of students. δ refers to the difference in the differences during the first year of participation and the final year of participation.

In addition to the aforementioned drawback, researchers often simply prefer to use all of their available data. Perhaps the easiest way to handle repeated measurements is to fit a linear model predicting student-year observations from independent variables identifying the time of follow-up before estimating standard errors of these coefficients with appropriate attention to "clustering" by student or by school; in mixed modeling and general estimating equations literature, this is known as the linear model with "working independence structure" (Fox, 2015; Laird, 2004). These analyses effectively attach equal weight to each student-year observation and thus we refer to them as "flat" weights. In combination with least squares, flat weighting delivers minimum-variance unbiased coefficient estimates under the model that

$$Y_{ijk} = \beta_0 + \tau Z_{ijk} + \beta X_{ijk} + \epsilon_{ijk} \quad \big(\mathbb{E}(\epsilon_{ijk}) = 0;\ \mathrm{Var}(\epsilon_{ijk}) = \sigma^2\big), \tag{2.2}$$

where the disturbances $\{\epsilon_{ijk} : i, j, k\}$ *are all independent of one another*. The model is said only to have "working" independence structure because even if in actuality the disturbances are not mutually independent, its least squares estimates remain unbiased under (2.2), while clustering ensures consistency of standard errors by taking into account heterogeneity across groups. Model (2.2) differs from the exit-observations-only model (2.1) in allowing multiple values of k for each student i; in BURST, k ranges from one to four under flat weighting. An example of flat weighting may be found in Meece and Miller (1999).

With multiple observations per student, (2.2) may be realistic but independence of its disturbances is not; as a result, flat weighting is inefficient. Instead of adopting this scheme, many researchers apply mixed effects models like hierarchical linear models (Bryk and Raudenbush, 1987; Raudenbush and Bryk, 2002) when conducting outcome analysis. This third option implicitly chooses a middle ground between flat weighting and exit observation analysis. Mixed effects models allow for some correlation between observations but not complete correlation. In parallel with (2.1) and (2.2), we may represent the two-level mixed effects model appropriate to analysis of BURST within the single regression equation

$$Y_{ijk} = \beta_0 + \tau Z_{ijk} + \beta X_{ijk} + \mu_j + \epsilon_{ijk} \quad \big(\mathbb{E}(\epsilon_{ijk}) = 0;\ \mathrm{Var}(\epsilon_{ijk}) = \sigma^2\big),$$

where we adopt the same structure as with flat weighting, including independence of $\{\epsilon_{ijk} : (i, j, k)\}$, but now incorporate random effects $\{\mu_j : j\}$ at the school level where $\mu_j \sim N(0, \nu)$. This allows researchers to account for unobserved heterogeneity by school. Other formulations might incorporate an additional random effect at the student-level. For examples of studies that apply mixed effects models, see Ethington (1997), Guo (2005), and Lee (2000).

One notable drawback arises when applying the two methods incorporating complete, longitudinal data. Exit observation analysis allowed us to articulate a well-defined overall average treatment effect: the expected difference in outcomes among third grade students. Making use of the complete data removes that possibility. The overall average treatment effect still represents an expected difference in outcomes between treatment and control students, but students contribute to that ATE in varying quantities depending on the length of time they participated in the study and perhaps the intraclass correlation (ICC).

The presence of clustered observations, either within schools or within students, has implications beyond regression-based modeling decisions. Within-group dependence, perhaps arising due to the presence of panel data or random assignment of blocks of units, complicates standard error estimation as well. BURST data exhibit within-group dependence as a consequence of both these phenomena: treatment assignment occurred by school and we have repeated observations on multiple students. Thus, both classical and heteroskedasticity-robust standard error calculations (Huber, 1967; White, 1980) are inappropriate. Nonetheless, dependent observations within BURST are grouped into mutually exclusive and non-overlapping clusters where every observation within the cluster received the same treatment assignment, allowing us to calculate standard errors that are robust to heterogeneity by group. For this purpose, we employ the "cluster robust" standard errors outlined in Pustejovsky and Tipton (2016), who in turn extended the work of Bell and McCaffrey (2002).

2.2.2 PWRD Aggregation

The three estimation methods presented in Section 2.2.1 all possess certain benefits. For example, exit observation analysis allows for a well-articulated overall average treatment effect and flat weighting allows researchers to use all of their data. Mixed effects models are particularly applicable in education settings with treatment assigned to clusters of units. Nonetheless, all three methods fail to take into account which observations will best allow researchers to detect a treatment effect according to the intervention's theory of change. In this section, we introduce an aggregation method that, similar to mixed effects models, is intermediate to flat weighting and exit observation analysis yet in contrast to each of those methods, leverages the theory of change to determine which observations are most likely to demonstrate a treatment effect.

To simplify the presentation of PWRD aggregation, we first illustrate our method on students who were in kindergaten during the first year of the study (i.e. Cohort 1.0 in Table 2.1) for a collection of schools that implemented the intervention with some fidelity. These students participated in BURST for the entire study and thus, had the greatest opportunity to benefit from the intervention. Implementation is a post-treatment variable so we do not recommend results from this subset to serve as an estimate of the effectiveness of BURST (presented in Section 2.4), but rather we use this subset as an example that best-serves to illustrate the intuition and process behind PWRD aggregation.

To implement PWRD aggregation, we need to estimate a separate treatment effect for each subgroup of interest. In the case of BURST, subgroup refers to the cohort year-of-follow-up because the theory of change suggests that students were more likely to have received targeted remediation when they had participated in the intervention for longer. We treat years-of-follow-up differently for the various cohorts because schools may implement the intervention differently over time. These treatment effect estimates for Cohort 1.0 during each year-of-follow-up are presented in Table 2.3. PWRD aggregation then serves as the tool by which we aggregate the four estimated effects into a single estimate for hypothesis testing. Note that while this aggregated treatment effect need not correspond to a simple average of individual treatment effects, it serves to address one basic research question of interest in all experiments and quasi-experiments: was there a treatment effect? This formulation simultaneously allows us to sidestep the debate as to what the best parameterization of the treatment effect should be, while making use of the full, longitudinal data in a fashion best suited to detect that effect.

PWRD aggregation is particularly beneficial in terms of power versus extant alternatives in trajectory correction interventions. In these interventions, students only receive the treatment once their performance stalls, resulting in effects that are scattered and delayed rather than concentrated and instantaneous. Prior to this occurrence, students receive the same instruction they otherwise

Cohort 1	**Coef.**	**S.E.**
Year 1	2.3	19.6
Year 2	-9.7	22.6
Year 3	8.7	8.5
Year 4	12.8	10.9

Table 2.3: Estimated change in outcome in each year-of-follow-up for a subset of Cohort 1.0

would have received if no intervention took place. As a consequence, the theory of change maintains that students only obtain an effect once they have received the supplemental instruction. It follows that the longer an individual participates in an intervention of this nature, the greater the likelihood their reading performance will require trajectory correction. We observe this phenomenon in Cohort 1.0 in BURST.

Years in BURST	**Tested In**
1	66.8%
2	75.4%
3	76.7%
4	79.3%

Table 2.4: The proportion of students in Cohort 1.0 who have "tested in" to BURST to receive supplemental instruction by how long they have participated in the study.

Table 2.4 illustrates how students enrolled in the study for a greater length of time are more likely to have required trajectory correction, and thus, are more likely to have benefitted from the intervention. Consequently, the theory of change posits that the expected size of the effect in cohort year g will be proportional to the percentage of students in a given cohort-year g who were eligible for supplemental instruction by that point in time, i.e. proportional to $\mathbf{p}_0 := \big(p_{0g} : g\big)$, where $p_{0g} := \mathbb{P}(\text{An individual in cohort-year } g \text{ is eligible to receive the supplemental instruction})$. PWRD aggregation takes advantage of this structure by attaching greater importance to observations from students in their fourth year in BURST than to observations from earlier periods of the study. Nonetheless, the expected size of the effect as estimated through $\hat{\mathbf{p}}_0$ is not the only consideration of PWRD aggregation. Rather, the estimated relative covariances (i.e. $\hat{\Sigma}$) between the treatment effect estimates also factor into our method. Thus in our example on Cohort 1.0, PWRD aggregation does not merely utilize $\hat{\mathbf{p}}_0$, but also:

$$\hat{\Sigma} = \begin{pmatrix} 81.2 & 59.2 & 1.6 & -8.8 \\ 59.2 & 106.8 & 13.7 & 11.6 \\ 1.6 & 13.7 & 21.4 & 13.8 \\ -8.8 & 11.6 & 13.8 & 25.8 \end{pmatrix}. \tag{2.3}$$

PWRD aggregation is constructed under the potential outcomes framework of Rubin (1974), Holland (1986), and Splawa-Neyman et al. (1990) and for the class of intention-to-treat estimators (Gupta, 2011; Montori and Guyatt, 2001). We let $Z = 1$ denote those who were assigned to the treatment and $Z = 0$ denote those assigned to the control. Y is our outcome of interest.

We can then define Δ_g as the parameter representing the treatment effect during cohort-year g, i.e.,

$$\Delta_g = \mathbb{E}(Y_g^{(Z=1)} - Y_g^{(Z=0)} | G = g).$$

Note that our unit of observation is at the student-year level rather than at the student-level. We then define our overall average treatment effect as a linear combination of parameters for the treatment effect in each cohort-year Δ_g, i.e. $\sum_g \omega_g \Delta_g$, where ω_g denotes the relative weight attached to each treatment effect Δ_g.

In other words, we calculate separate ITT estimates for each cohort during each year-of-follow-up. PWRD aggregation then looks to uncover the specific linear combination, i.e. the specific ω, that maximizes the power of tests based on the following aggregated statistic:

$$\hat{\Delta}_{agg} := \sum_g \omega_g \hat{\Delta}_g. \tag{2.4}$$

We know that ω will account for the expected size of the effect ($\mathbf{p}_0$) and the relative precision of each Δ_g (Σ), but to find the specific linear combination that maximizes our power to detect an effect, we first make multiple assumptions about the nature of the treatment, given the theory of change behind the trajectory correction intervention with targeted remediation holds:

Condition 2.2.1 *Individuals who receive supplemental instruction as a result of the intervention at time j receive an effect $\tau \geq 0$ between j and t_i, where t_i denotes the time at which individual i exits the study. Individuals who do not receive supplemental instruction are unaffected.*

Condition 2.2.2 *Effect τ received by individual i at time j is retained by individual i in full throughout the duration of the study, i.e. from $[j, t_i]$.*

Condition 2.2.1 is an extension of the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980). Briefly, SUTVA states that the treatment received by one individual will not affect the potential outcomes of other individuals. SUTVA generally refers to the treatment not affecting the potential outcomes of individuals in the control group. With respect to BURST, we argue this additionally implies individuals testing into the intervention to receive targeted remediation will not affect the potential outcomes of individuals in the treatment who do not test in but instead remain in the classroom. This corresponds to a situation where there is no interference across individuals (Sobel, 2006). We illustrate this distinction in Figure 2.1. With no interference, those students in treatment schools who receive the supplemental instruction do not affect the potential outcomes of those who remain in the classroom to receive standard instruction.

A technical condition simplifies the development by excluding pathological cases.

Condition 2.2.3 $\mathrm{Cov}(\hat{\Delta}) = n^{-1}\Sigma$*, with* Σ *a positive-definite symmetric matrix.*

From these conditions, we now construct PWRD aggregation.

Proposition 2.2.4 *Consider test statistics of the form $\sum_g \omega_g \hat{\Delta}_g$, $\sum_g \omega_g \hat{\Delta}_g - \sum_g \omega_g \delta_{0g}$, or $\widehat{V}^{-1/2}(\sum_g \omega_g \hat{\Delta}_g - \sum_g \omega_g \delta_{0g})$ where δ_0 is a vector of hypothesized values of Δ, $\Delta := (\Delta_g : g)$ and $n\widehat{V} \to_p c$, $c > 0$. Additionally, take the family of hypotheses $K_\eta : \Delta = \eta \mathbf{p}_0$. Under Conditions 2.2.1, 2.2.2, and 2.2.3, and for tests of $H_0 = K_0$ against any alternative K_η, $\eta \geq 0$, aggregation weights ω that satisfy the following formula will maximize the asymptotic relative efficiency of the*

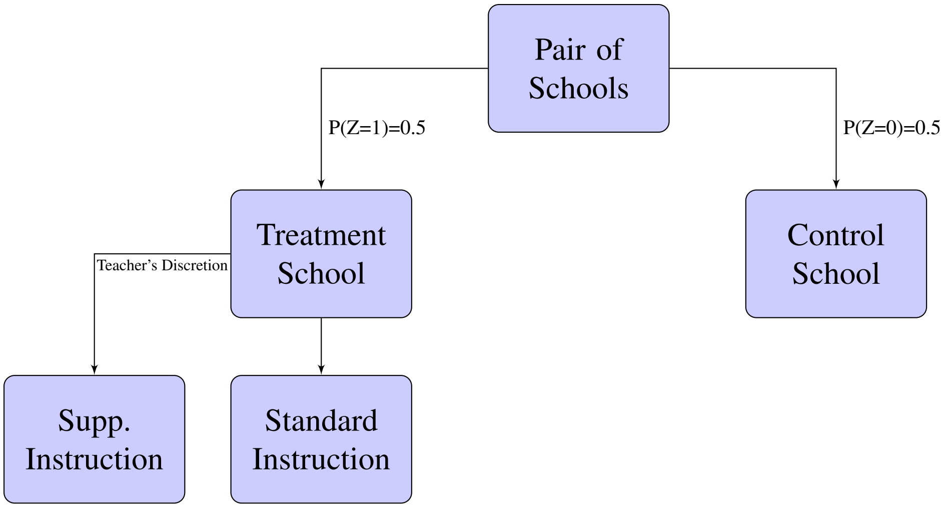


Figure 2.1: BURST design for a pair of schools.

above test statistics:

$$\omega = (\Sigma^{-1}\mathbf{p}_0)_+ \Big/ \sum_j (\Sigma^{-1}\mathbf{p}_0)_{+_j}, \tag{2.5}$$

where $(\Sigma^{-1}\mathbf{p}_0)_+$ *denotes the element-wise maximum of* $(\Sigma^{-1}\mathbf{p}_0)$ *and* $\mathbf{0}$*, and* $(\cdot)_{+_j}$ *denotes the* j*th element of* $(\cdot)_+$ *such that* $\omega'\mathbf{1} = 1$.

In other words, the slope (Pitman, 1948) of test statistics described in Proposition 2.2.4 may be maximized by weights proportional to both the expected size of the effect for each cohort-year $\mathbf{p}_0$ and also the relative precisions between each cohort-year effect estimate Σ: $\omega \propto \Sigma^{-1}\mathbf{p}_0$. Note that any test statistic of the form:

$$\frac{\sum_g \omega_g \hat{\Delta}_g - \sum_g \omega_g \delta_{0g}}{\widehat{V}^{1/2}}, \tag{2.6}$$

such as the t-statistic combining estimates $\hat{\Delta}_g$ with fixed weights ω_g, will be covered by Proposition 2.2.4.

By maximizing the test slope, PWRD aggregation provides test statistics with greater asymptotic relative efficiency, represented by the square of the quotient of two test slopes, than alternative test statistics. Improving relative efficiency by 20% corresponds with a 20% reduction in the sample size required to achieve the same level of power (Van der Vaart, 2000). Consequently, by maximizing the slope of the test statistic, PWRD aggregation maximizes the power of tests $H_0 = K_0$ against any K_η, $\eta \geq 0$, i.e. for effects that are proportional to the dosage. In other words, we may view PWRD aggregation as maximizing the signal-to-noise ratio for test statistics of the form presented in Equation 2.6, which includes t-statistics. Given the theory of change holds, test statistics incorporating PWRD aggregation weights will provide researchers with a greater opportunity to detect an effect of the intervention.

Very generally, we derive these weights by taking the gradient of the test slope of (2.4) (the argument for the other forms of test statistic being similar) with respect to ω. After setting this term equal to zero and simplifying through a grouping of scalar quantities, we obtain PWRD aggregation weights. We additionally add a constraint to ensure that our aggregation weights are non-negative. For a proof, see Appendix B.1.

Neither $\mathbf{p}_0$ nor Σ are directly observed, but both can be estimated easily. We estimate $\mathbf{p}_0$ through the proportion $\hat{\mathbf{p}}_0$ observed among students assigned to the control. Note that this is the probability of *ever* having tested in to receive supplemental instruction rather than the probability of having tested into the treatment during that year. This alleviates fears of selection bias due to post-treatment conditioning because once a student tests in once, each subsequent observation for that student is designated as having tested in as well. Thus, the treatment received by a student in year t, does not affect their weight in year $t + 1$. Estimating Σ requires a slightly more elaborate calculation centered around control-group residuals. Briefly, we fit a model onto control observations predicting the outcome and controlling for potential confounders. We then fit a second model estimating the residuals generated by the previous model, solely controlling for each cohort-year. The subsequent cluster-robust covariance matrix serves as $\hat{\Sigma}$.

PWRD aggregation combines with standard techniques to address complexities of study design such as block randomization and assignment to treatment conditions by cluster, such as the school or the classroom, rather than by the individual student. We scale the "bread" component of Huber-White sandwich estimators of the variance using a similar method as that presented by Pustejovsky and Tipton (2016). With these cluster-robust standard errors, we are then able to conduct Wald tests to reject or accept the null hypotheses previously presented.

Covariate adjustment may be incorporated while estimating each individual Δ_g either through design-based approaches outlined in Lin et al. (2013), Hansen and Bowers (2009), or Middleton and Aronow (2015), or through more conventional model-based formulations. While not constructed around attributable effects (Rosenbaum, 2001), we can extend PWRD aggregation into that setting with minor adjustments.

2.2.3 Considerations when the Theory of Change Fails

When the theory of change holds, PWRD aggregation maximizes the test slope and thus, the corresponding power for the family of hypotheses $K_\eta : \Delta = \eta \mathbf{p}_0$. That is, when the treatment effect is proportional to the dosage received, PWRD aggregation maximizes power. Nonetheless, there may be fears that when the theory of change does not hold and the effect is not proportional to the dosage received, using PWRD aggregation will have adverse effects on outcome analysis. For example, there may be worries that PWRD aggregation will lead to seriously biased impact estimates. However, PWRD aggregation is used for testing rather than for estimation so this fear is unfounded. There may instead be the following issues:

- When there is no effect of the intervention, PWRD aggregation leads to incorrect Type I errors.
- When the effect accrues in a different fashion than the theory of change hypothesizes, PWRD aggregation loses power versus alternative methods.

To alleviate the first issue, we prove that when a few technical conditions hold, PWRD aggregation will maintain proper Type I error rates rather than over or under-rejecting a null hypothesis of no effect. For a greater examination of Type I errors with PWRD aggregation, see Appendix A.

To address the second fear, we show through simulations in Section 2.3 that when the theory of change fails to hold, PWRD aggregation either leads to a trivial amount of power lost or, at times, provides substantially greater power. For an example of the latter, we present the following scenario that often arises in education research.

2.2.3.1 Addressing within-cluster interference

We have interpreted the BURST theory of change to hold that a student's outcomes may depend on her own treatment assignment but not that of any other student — that is, that the experiment was free of *interference* (Cox, 1958; Sobel, 2006). As applied to students within a school, this may be simplistic. A school possesses finite resources, so its adopting a supplemental instruction regime may transfer resources away from students not receiving the supplement. In this scenario, Condition 2.2.1 no longer holds: students not targeted for a BURST supplement may suffer an instructional detriment, with adverse effects on their learning.

Addressing such *spillover effects* within a classroom or school is an area of active methodological research (Fletcher, 2010; Vanderweele et al., 2013; Gottfried, 2013), often calling for specialized methods or other accommodations (Sobel, 2006; Rosenbaum, 2007; Vanderweele et al., 2013; Bowers et al., 2018). To address the common scenario of spillover within but not across clusters, where clusters denote experimental units as assigned to treatment conditions, the PWRD aggregation method applies without change. Specifically, we may relax Condition 2.2.1 in favor of the following:

Condition 2.2.5 *Individual i receiving supplemental instruction due to the intervention at time j gains non-negative effect τ between j and t_i. Individuals who do not receive the supplemental instruction may experience an effect, positive or negative, so long as the overall effect of all students is positive in aggregate.*

From Condition 2.2.5, we now present Proposition 2.2.6, a corollary to Proposition 2.2.4:

Proposition 2.2.6 *Under Conditions 2.2.2, 2.2.3, and 2.2.5, the following aggregation weights ω will maximize the slope of test statistics of the form $\sum_g \omega_g \hat{\Delta}_g$, $\sum_g \omega_g \hat{\Delta}_g - \sum_g \omega_g \delta_{0g}$, or*

$\widehat{V}^{-1/2}(\sum_g \omega_g \hat{\Delta}_g - \sum_g \omega_g \delta_{0g})$ *for the family of hypothesis tests and alternative hypotheses elaborated in Proposition 2.2.4:*

$$\omega = (\Sigma^{-1}\mathbf{p}_0)_+ \Big/ \sum_j (\Sigma^{-1}\mathbf{p}_0)_{+_j}.$$

According to Proposition 2.2.6, PWRD aggregation maintains its advantage in the presence of spillover within clusters, so long as the interference is compatible with a suitable adjustment of the theory of the intervention. This is the situation arising in BURST: a greater proportion of a school's students directly receiving the intervention corresponds with a lower proportion of those students being at risk of corresponding adverse spillover; its theory of change must hold that benefits accruing to the first group exceed any detriment toward the latter in aggregate.

The derivation of Proposition 2.2.6 follows the same structure as the derivation of Proposition 2.2.4 found in Appendix B.1.

2.3 Simulations

In order to demonstrate how PWRD aggregation performs in comparison to flat weighting and mixed effects models when the theory of change works as intended, we construct a simulation study mirroring the design of BURST. We generate student outcomes to compare statistical power across different scenarios using the following two-level model:

$$\begin{aligned} Y_{ijk} &= \beta_0 + \beta_1 \text{Grade}_{ijk} + \mu_k + \epsilon_{ijk} \\ \mu_k &= \gamma_0 + \nu_k \end{aligned}, \tag{2.7}$$

with $\nu_k \sim N(0, \xi)$. The outcome of student i in year-of-follow-up j at school k is a function of the grade of the student and the random intercept of the school at which the student is enrolled, μ_k. Note that fixed effects like race, gender, socio-economic status, and others could be added to this process, but were excluded as we have presented PWRD aggregation without covariate adjustment. Once we generate these outcomes, we perform the following two steps. First, we flag outcomes that fall below a given threshold as having tested into the intervention. Once a student tests in, all of their subsequent observations are flagged as well. The threshold changes by grade to adjust for natural improvement with age. Second, we impose artificial treatment effects on students within treatment-schools and find the corresponding power across iterations of this data generation.

We compare three variations of treatment effects in this simulation study. Under the first, all treatment observations flagged as having tested into the intervention receive some constant, positive effect τ. Under the second, flagged treatment observations receive a constant, positive effect τ and unflagged treatment observations, i.e. individuals in the treatment who do not test into the intervention, receive a constant negative effect $-p\tau$ where $p \in (0, 1]$. The third version of treatment effect imposes $\tau \sim N(l, 2.5 * l)$ for some l to all treatment observations.

To mirror BURST, we generate 32,000 student-year observations across 26 pairs of schools with students divided roughly evenly across kindergarten through third grade. We assess the power provided by each of the models across 1,000 iterations of this simulation study for each artificially imposed effect size. Power for a given effect size is determined by calculating how often a model rejects a null hypothesis of no effect at the 5% level out of the 1,000 iterations. We use cluster-robust standard errors with clusters at the school level from the `clubSandwich` package in R (Pustejovsky, 2017; Pustejovsky and Tipton, 2016).

2.3.1 Simulation Results

We now present results from these simulations across the three variations of imposed treatment effect described previously. For reference, the standard deviation of the outcome variable is 23.5. Following guidance from Kraft (2020), we will denote effect sizes smaller than 0.05σ (1.2 points in our simulation study) as *small*, those between 0.05σ and 0.2σ (4.7 points) as moderate, and those greater than 0.2σ as large. Across 1,260 effect sizes on reading outcomes from 495 randomized controlled trials, the mean effect size is 0.17σ (4 points) and the 90th percentile is 0.5σ (11.8 points) (Kraft, 2020). Thus, our simulation study examines these three methods on effect sizes that frequently appear in reading interventions.

2.3.1.1 Effect 1

Figure 2.2 shows the power for 1000 replications of the synthetic experiment across three analytical schemes: PWRD aggregation, flat weighting, and a mixed effects model specified according to (2.7), but with an independent variable representing the treatment. It is immediately apparent that PWRD aggregation outperforms the other two methods, especially for medium effect sizes under which we observe a 35-50% increase in power. This is unsurprising as PWRD aggregation attaches greater importance to student-year observations most likely to have received an effect from the intervention and down-weights the remaining observations. Power as observed when the effect is 0 is simply the empirical size of the test; thus the left side of the plot indicates that use of the PWRD method did not negatively affect Type I error rates.

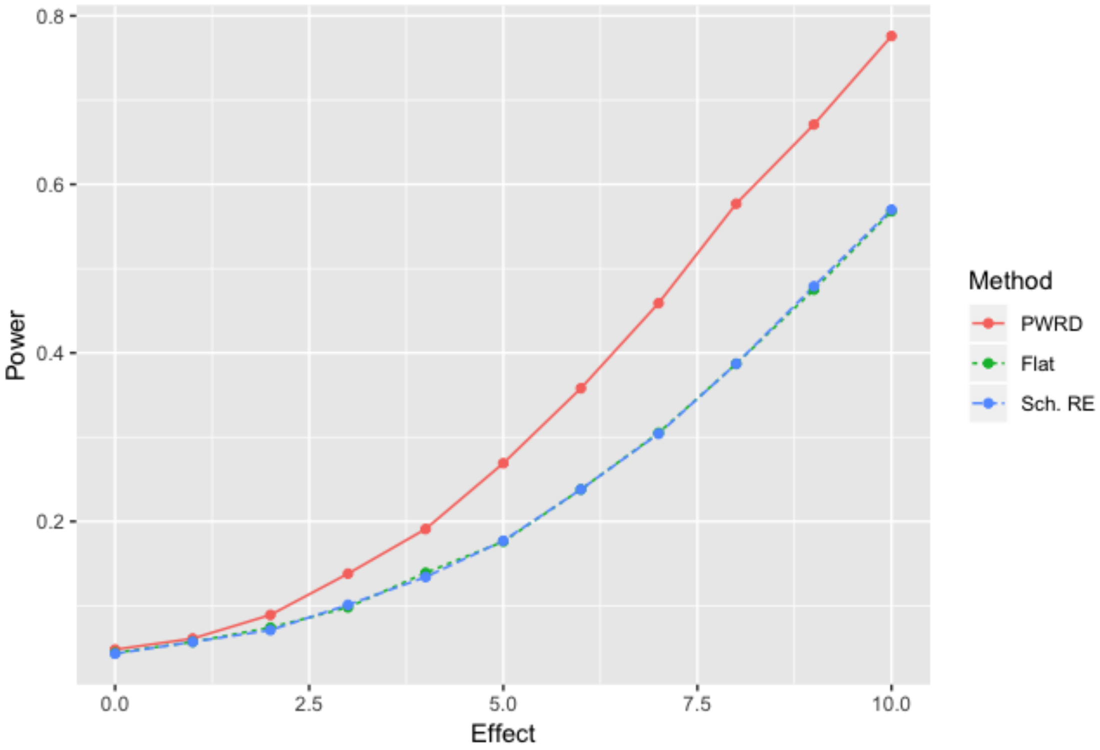


Figure 2.2: Power for the three methods under Effect 1, i.e. across increasing effect sizes.

It is natural to ask whether the gains in power present in Table 2.2 hold across different levels of correlation of observations within a school. To examine this we conducted additional simulations holding the imposed effect constant, but varying the intraclass correlation (ICC). We present these results in Figure 2.3.

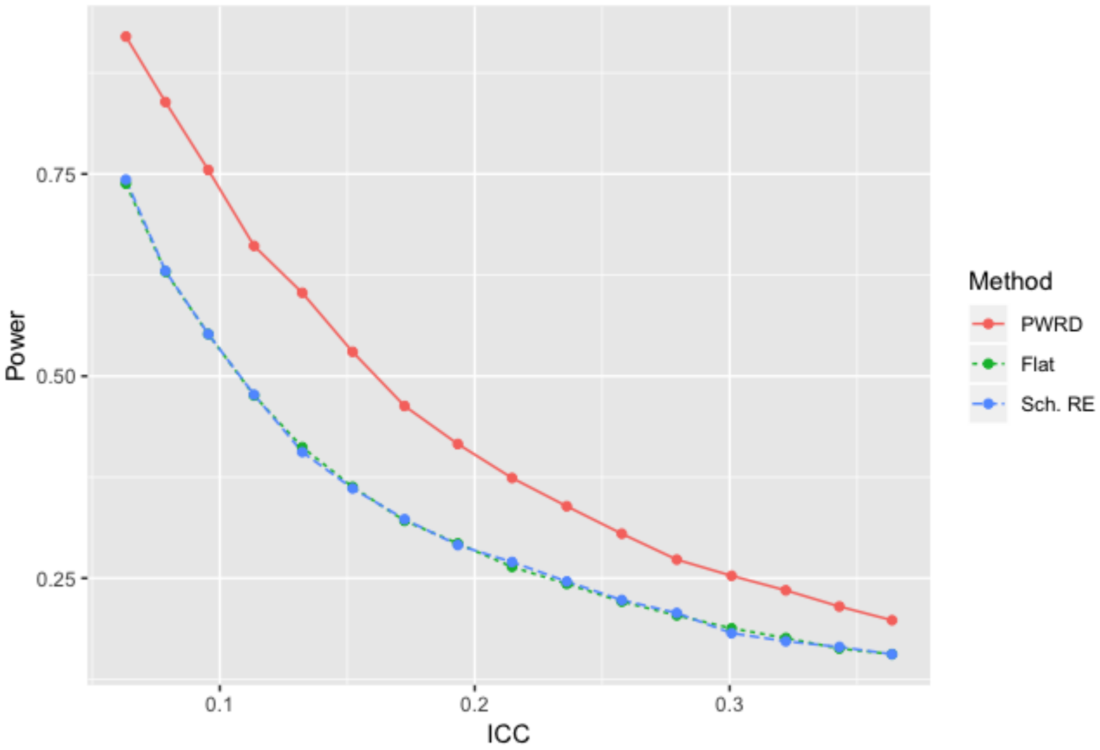


Figure 2.3: Power for the three methods under Effect 1 with increasing intraclass correlations.

In Figure 2.3, PWRD aggregation consistently outperforms flat weighting and mixed effects models across ICCs that typically arise in educational settings (Hedges et al., 2007). For intraclass correlations between 0.1 and 0.2, PWRD aggregation provides 35-45% more power than the competitors. That gap decreases for larger ICCs, although this is at the upper range of reasonable ICC values. Furthermore, we still obtain a 25% improvement in power.

2.3.1.2 Effect 2

We now relax the assumption that students who do not receive targeted remediation through the intervention are unaffected. Instead, we impose a negative effect that is in magnitude 40% of the positive effect imposed on students who receive the supplemental instruction. This is a scenario where there is interference within a school, corresponding to replacing Condition 2.2.1 with Condition 2.2.5 and thus Proposition 2.2.4 with Proposition 2.2.6. We chose 40% to ensure the overall effect is positive in aggregate.

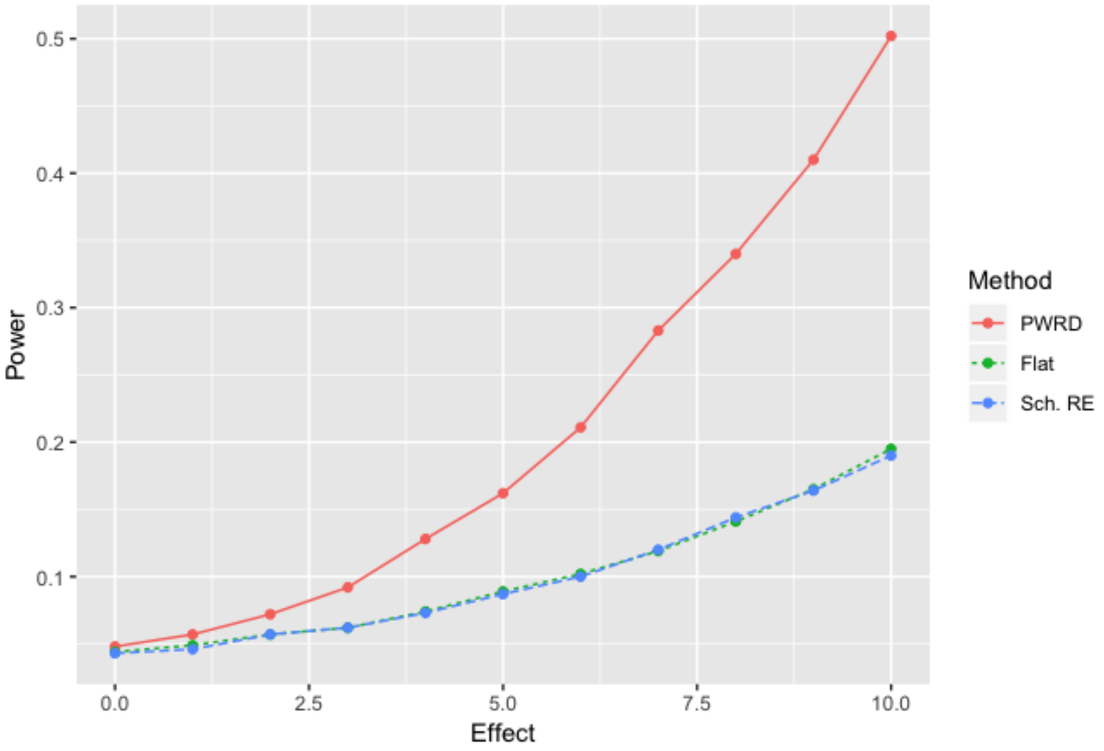


Figure 2.4: Power for the three methods under Effect 2, i.e. across increasing effect sizes when Condition 2.1 does not hold.

In Figure 2.4, we observe that under the relaxed assumption, PWRD aggregation performs even better in comparison to the other two methods than it did under the standard assumptions. This relative gain in power is expected. We weight down effect estimates that are more likely to incorporate students with *negative* effects, attaching greater importance to those more likely to have received a *positive* effect. Neither flat weighting nor our school random effects model perform a similar function and their power to detect an effect is substantially reduced as a consequence. For small effect sizes, PWRD aggregation increases power by nearly 20% and this gap only widens as the effect size increases. For example, our method more than doubles the power of mixed effects models and flat weighting for large effect sizes.

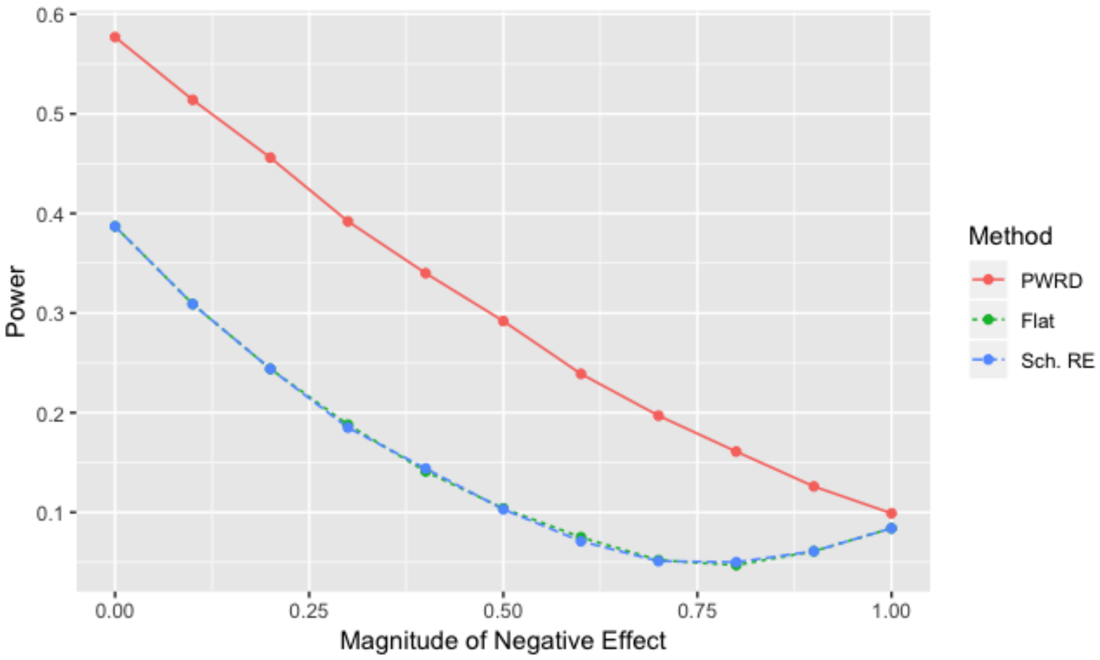


Figure 2.5: Power for the three methods under Effect 2 with increasing negative effects. Here we add a positive effect of size 8 to students in the intervention and a negative effect that increases from 0% to 100% of the positive effect.

The phenomenon present in Figure 2.4 holds when the magnitude of the negative effect varies as well. We observe this in Figure 2.5. Under this scenario, the size of the benefit remains constant. Instead, the adverse effect for those treatment students who do not test into the intervention varies from 0% of the benefit to 100% of the benefit. PWRD aggregation provides a persistent 15-20 percentage point advantage in power for negative effects up to 60% of the positive effect before narrowing out. This corresponds to at least a 40% improvement in power for all magnitudes of the negative effect; under certain circumstances, our method provides double the power. When the negative effect is equal in magnitude to the positive effect, PWRD aggregation no longer provides gains in power.

2.3.1.3 Effect 3

We now examine what occurs in cases where the theory behind interventions of this sort entirely fails. This does not necessarily mean the intervention does not provide a benefit, just that it does not work as hypothesized by the theory of change. Here, we impose an artificial treatment effect on all treatment observations such that $\tau_{ijk} \sim N(l, 2.5 * l)$ for $l = 1, \ldots, 10$. Note that while the aggregate effect is still positive, any given student may be negatively affected. Furthermore, effects are neither stacked nor persistent across time. We present these results in Figure 2.6.

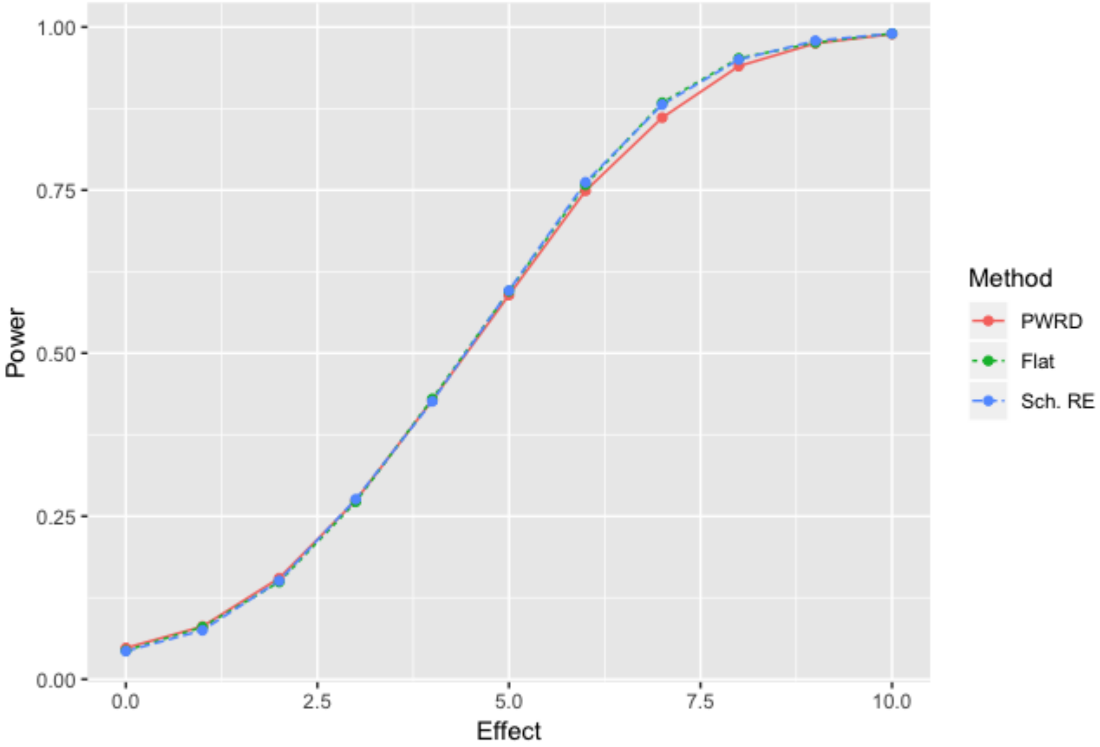


Figure 2.6: Power for the three methods under Effect 3, i.e. across increasing effect sizes when none of the conditions hold.

We immediately observe that while the school random effects model and flat weighting slightly outperform PWRD aggregation, this improvement is minimal and never exceeds 3%. For effect sizes greater than 6 (roughly 0.25σ), we are able to reject frequently under any of the three schemes. From these simulations, it is clear that our method provides substantial gains in power in situations where the theory behind the intervention holds. When the theory does not hold, we see marginal decreases in our ability to detect an effect. These results hold in smaller samples as well, providing gains to power when the theory of change holds and a minimal loss in power when the theory of change is incorrect. For a deeper examination, see Appendix C.

2.4 PWRD analysis findings

This section presents results for BURST, both on Cohort 1.0 and on the overall randomized trial using PWRD aggregation and commonly applied alternative methods. The theory behind BURST was presented in Section 2.2. Nonetheless, its data structure merits additional discussion to clarify analysis in this section. We utilized a large-scale cluster randomized trial to test the efficacy of BURST, a reading intervention designed to assist early-elementary students at risk of falling below grade-level proficiency. The experiment was block-randomized at the school level with 26 total blocks, 24 of which were pairs of schools. The remaining two blocks were a triplet of schools, in which two schools were assigned to treatment, and a singleton. The singleton originally belonged to a pair until the school assigned to the control attrited. Nearly every school was matched within its school district. Across these 52 schools, we observed 27,000 unique students on 1–4 occasions each, for a total of 52,000 student-year observations. As discussed in Section 2.2.1, the length of time for which each student participated in the study depended on the grade and year during which they entered the study. While we encountered some missing data, we had demographic information (race, gender, age, free lunch status, etc.) for the vast majority of students. In addition, we had *Dynamic Indicators of Basic Early Literacy Skills* (DIBELS) scores and end-of-year assessment scores for each student. DIBELS, a widely used reading assessment, served as the diagnostic by which students were designated to receive targeted instruction and additionally functioned as a pre-test. The end-of-year assessments were our primary outcome of interest.

2.4.1 BURST Cohort 1.0

In this section, we begin by showing how the aggregation weights $\hat{\omega}$ were generated before presenting the results themselves. In order to calculate $\hat{\omega}$, we need to estimate $\mathbf{p}_0$ and Σ. We know from Section 2.2.2 that we estimate $\mathbf{p}_0$ using the proportion of control students who tested in to receive supplemental instruction for each year-of-follow-up. These values are presented in Table 2.4. We then calculate Σ with the grouping of control-group residuals described in greater detail in Section 2.2.2. For this example, $\hat{\Sigma}$ may be found in (2.3). We then formulate:

$$\hat{\omega} = (\Sigma^{-1}\mathbf{p}_0)_+ \Big/ \sum_j (\Sigma^{-1}\mathbf{p}_0)_{+_j} = (0.25, 0, 0.32, 0.43).$$

Note that while more students were eligible for supplemental instruction by the second year than in the first, the relative precision of the estimate in the second year-of-follow-up and its mutual correlations with the other estimates were prohibitively large. Thus, PWRD aggregation determined that outcome analysis would be best served by attaching no weight to those observations.

We then employ a Peters-Belson (Peters, 1941; Belson, 1956) approach to estimating the average treatment effect both under standard analyses like flat weighting and mixed effects models with a random effect at the school level, and also under PWRD aggregation incorporating $\hat{\omega}$ described above. Briefly, Peters-Belson methods apply covariate adjustment to the control group rather than to the treatment and control simultaneously. That control-adjusted model is then used to predict treatment outcomes. The differences between the fitted and observed values serve to estimate the average treatment effect. Results are presented in Table 2.5.

Method	**Est.**	**S.E.**	**t value**	**Sig.**	**Test Slope**
Exit	9.88	9.72	1.02	-	-
Flat	2.50	10.61	0.24	-	0.070
Sch. RE	-1.10	10.47	-0.11	-	0.071
PWRD	8.87	6.89	1.28	-	0.109

Table 2.5: BURST results on a subset of Cohort 1.0 for various methods, including PWRD aggregation. Note that we do not present the test slope for exit observation analysis, as this method does not make full use of the data.

As we can see, none of the methods are able to detect an effect of the intervention, although PWRD aggregation provides the greatest test statistic. In this scenario, exit observation analysis also performs relatively well, perhaps because students in their fourth year-of-follow-up, i.e. in third grade, were best situated to benefit from BURST.

2.4.2 BURST[R]: Reading

We now conduct the same analysis as previously, yet using the complete data from BURST. For PWRD aggregation, we now calculate separate effect estimates and aggregation weights for each cohort-year. As with analysis on Cohort 1.0, we employ a Peters-Belson approach to outcome analysis. Results are presented in Table 2.6.

Method	Est.	S.E.	t value	Sig.	Test Slope
Exit	0.40	2.76	0.15	-	-
Flat	-0.09	4.17	-0.02	-	0.152
Sch. RE	-3.70	3.91	-0.95	-	0.162
PWRD	-0.34	3.03	-0.11	-	0.216

Table 2.6: BURST for various methods, including PWRD aggregation. Note that we do not present the test slope for exit observation analysis, as this method does not make full use of the data.

None of these methods detect an effect of BURST on student achievement: unfortunately, this program appears not to have provided a benefit. A possible explanation for the lack of an effect is that schools possess limited resources; more students required supplemental instruction than schools had the ability to serve at levels recommended by the theory of change (Rowan et al., 2019). Thus, schools had to ration resources and make choices about depth of implementation versus breadth of implementation. These factors, along with many others, may have contributed to BURST not providing a reading benefit. Despite the theory of change not holding, PWRD aggregation still provides valid standard errors and a valid hypothesis test. This additionally remains the case when the intervention provides detrimental effects to students.

Nonetheless, if the theory of change held true, the relative Pitman efficiency of PWRD aggregation versus flat weighting and mixed effects modeling was 2.02 and 1.78 respectively. This suggests that we would have required over 52 and 40 additional schools in BURST in order to achieve the same power we possessed under PWRD aggregation. Note that while in this case, exit observation analysis came the closest to detecting a positive effect, it too did not come close to rejecting the null hypothesis.

For an example where PWRD aggregation detects an effect where other methods fail to reject a null hypothesis of no effect, see the Massachusetts healthcare reform example in Appendix D.

2.5 Discussion

In this chapter, we have presented a novel method of aggregation that converts an intervention's theory of change into statistical power for a broad class of interventions. This method is compatible both on its own and in tandem with commonly used methods of analysis including regression based techniques.

The strategy of using a regression coefficient to conduct a hypothesis test is standard in settings across the social sciences. Nonetheless, these conventional regressions may not prove optimal in any given scenario because they fail to account for which observations are most likely to benefit from the treatment. PWRD aggregation, constructed to maximize the test slope for a family of hypotheses, offers a solution by providing greater efficiency and thus, greater power than extant methods when the theory of change holds.

A suitable theory of change will generally be available in educational settings because they play featured roles in competitive funding proposals. We demonstrated extraction of PWRD aggregation from a theory of change likely to be typical of interventions providing supplemental instruction. In it and similar circumstances, our method gives researchers the best possible opportunity to detect an effect.

While PWRD aggregation is optimal when its supporting theory of change holds, no benefit is gained when that theory is incorrect. Nonetheless, this scheme does not greatly hamper one's ability to detect an effect in this situation. We believe PWRD aggregation can be extended to many other scenarios, both experimental and quasi-experimental, with longitudinal data and a treatment that accrues heterogeneously across observations. In each of these scenarios, similar aggregation weights can be formulated around the theory of change that will maximize power.

CHAPTER 3

Confidence Intervals for Effects Proportional to Dosage

3.1 Introduction

Randomized trials and observational studies in the social sciences often result in effects that accrue heterogeneously across subgroups of the study population. For example, one flavor of education intervention provides supplemental instruction to a subset of students who require corrections to their otherwise stalled learning trajectories. In studies of this type, according to the theory of change, the groups which are most likely to require supplemental instruction are most likely to demonstrate a benefit from the intervention as well. Lycurgus and Hansen (2021) introduce a mode of hypothesis testing structured around these subgroups to increase power for such interventions. Briefly, the method emphasizes cohorts and years-of-follow-up on which effects are expected to concentrate, with appropriate attention to the relative precision of these estimated effects. Simulations found that when the theory of change behind the intervention held, this method provided much greater power to detect an effect (at times twice the power) than commonly used techniques. For a more detailed presentation of this method, Power-maximizing Weighting for Repeated-measurements and Delayed-effects (i.e. PWRD) aggregation, see Chapter 2.

While designed with supplemental instruction interventions in mind, PWRD aggregation may be extended to settings outside of education so long as there is a) some measure of the dosage received by different subgroups and b) the effect accrues in a manner that is proportional to that dosage. To illustrate, Massachusetts extended Medicaid eligibility to adults with incomes below 138% of the federal poverty limit in 2006. PWRD aggregation is a natural fit when conducting county-level analysis of this reform's effect on mortality (see Appendix D). Health insurance is the vehicle by which mortality would be reduced, so counties with greater numbers of low-income residents newly eligible for health insurance stood to reap a larger mortality benefit from this expansion than wealthier counties with fewer residents gaining health insurance.

Despite the broad applicability of PWRD aggregation, this method was designed to maximize power for hypothesis testing, rather than to be applied when estimating an overall average treatment effect across subgroups or constructing a confidence interval around that estimate. As a consequence, confidence interval construction is not as simple as inverting hypothesis tests that employ PWRD aggregation and retaining values that are accepted by the test.

In this chapter, we propose multiple methods that may be used in tandem with PWRD aggregation to construct confidence intervals. These allow researchers to benefit from PWRD aggregation when conducting a hypothesis test of no effect while still attaching interpretable confidence intervals to the estimated effect. The first method is motivated through the three-sided hypothesis testing strategy proposed in Goeman et al. (2010), where the parameter space is partitioned into three regions. Different hypothesis tests are simultaneously applied depending on the partition. Those hypothesis tests may then be inverted to construct a confidence interval. The next method takes one of the underpinnings of PWRD aggregation—that the effect is proportional to the level of dosage received—and attaches a confidence interval on that proportionality constant in a manner similar to Rosenbaum et al. (2010, p.135-6).

In this chapter, we use Medicaid expansion through the Affordable Care Act (ACA) as a test case to illustrate how the combination of PWRD aggregation and these confidence interval techniques provides researchers with an invaluable tool when conducting outcome analysis. Similar to Massachusetts healthcare reform, counties with greater numbers of residents newly eligible for Medicaid, i.e. those counties with greater dosage levels, should reap a larger benefit from Medicaid expansion than those counties with fewer newly eligible individuals given mortality is, in fact, reduced. We demonstrate how it is possible to realize gains in power through PWRD aggregation while simultaneously attaching easily interpretable confidence intervals bounded away from the null hypothesis.

3.1.1 Roadmap

We begin with a review of PWRD aggregation, first introduced in Chapter 2. Then in Section 3.2.2, we adapt the three-sided hypothesis testing method into a confidence interval scheme that may be used in tandem with PWRD aggregation. This method requires partitioning the parameter space, so we provide guidance as to how that division should occur. After, we present a confidence interval method for the proportionality constant as both an alternative to the three-sided confidence interval scheme and also as a check on the validity of PWRD aggregation. In Section 3.3, we adapt the simulation scheme presented in Lycurgus and Hansen (2021) (and in Section 2.3) to the three-sided confidence interval method, showing power and confidence intervals under different partitions of the parameter space. We then demonstrate these methods on Medicaid expansion through the ACA in Section 3.4. We conclude with a discussion summarizing the benefits and drawbacks of using PWRD aggregation in tandem with these confidence interval techniques.

3.2 Method

3.2.1 Review of PWRD Aggregation

PWRD aggregation was first introduced in Chapter 2 and additionally in Lycurgus and Hansen (2021) with the intention of increasing power in education interventions where students only receive the treatment if they fall sufficiently behind their peers. The motivation behind PWRD aggregation is that some cohorts or groups of students—e.g. those cohorts with a greater proportion of students who require supplemental instruction—are more likely to manifest a treatment effect than others. Consequently, those cohorts receive greater emphasis in hypothesis tests for primary outcome analysis attempting to detect a benefit of the intervention.

Briefly, PWRD aggregation is constructed around the potential outcomes framework of Rubin (1974), Holland (1986), and Splawa-Neyman et al. (1990) and the class of intention-to-treat estimators. We define Δ_g as the treatment effect for group g, i.e. $\Delta_g = \mathbb{E}(Y_g^{(Z=1)} - Y_g^{(Z=0)}|G = g)$. The overall average treatment effect is formulated by aggregating each Δ_g into a single estimate, i.e. $\sum_g \omega_g \Delta_g$, where ω_g denotes the relative weight of each Δ_g. PWRD aggregation looks to uncover which ω_g, i.e. which specific linear combination, will provide the greatest power.

Take the family of hypotheses $K_\eta : \Delta_g = \eta \mathbf{p}_0$, where $\mathbf{p}_0 \coloneqq \big(p_{0g} : g\big)$, $p_{0g} \coloneqq \mathbb{P}($An individual in group g is eligible to benefit from the treatment), and η represents some proportionality constant. Under appropriate conditions (see Conditions 2.2.1, 2.2.2 & 2.2.3), the asymptotic relative efficiency (first introduced in Section 2.2), and consequently the power, of test statistics with the form $\widehat{V}^{-1/2}(\sum_g \omega_g \hat{\Delta}_g - \sum_g \omega_g \delta_{0g})$ where δ_0 is a vector of hypothesized values of Δ, $\Delta \coloneqq (\Delta_g : g)$ and $n\widehat{V} \rightarrow_p c$, $c > 0$, will be maximized by the following aggregation weights ω:

$$\omega = (\Sigma^{-1}\mathbf{p}_0)_+ \Big/ \sum_j (\Sigma^{-1}\mathbf{p}_0)_{+_j}. \tag{3.1}$$

where $\Sigma \coloneqq \mathrm{Cov}\{(\hat{\Delta}_g : g)\}$ and $(\cdot)_+ \coloneqq \max(\cdot, 0)$. Here, $(\cdot)_+$ denotes the element-wise maximum of $(\Sigma^{-1}\mathbf{p}_0)$ and $\vec{0}$ and $(\cdot)_{+_j}$ denotes the jth element of $(\cdot)_+$. This normalizes ω such that $\omega'\mathbf{1} = 1$. In summary, the test slope of test statistics described previously, such as the t-statistic, will be maximized by the aggregation weights defined in Equation 3.1 (Equation 2.4 in Chapter 2).

While originally constructed with education interventions in mind, PWRD aggregation may be extended to other scenarios where the treatment accrues heterogeneously across groups of observations, so long as a measure of the extent to which each group stands to benefit is available. For example, we intend to illustrate the performance of PWRD aggregation on restricted-use mortality data to estimate the effect of Medicaid expansion on mortality. Here, the measure of the extent to which each subgroup stands to benefit is the proportion of newly eligible residents in a county.

3.2.2 Confidence Interval Construction

PWRD aggregation is shown to provide substantial gains in power when applied in appropriate situations (see Section 2.3). Nonetheless, the method is best suited for hypothesis testing rather than for confidence interval estimation because $\hat{\Delta}_{PWRD} := \sum_g \hat{\omega}_g \hat{\Delta}_g$ does not represent an easily interpretable treatment effect estimate. Thus, inverting hypothesis tests using PWRD aggregation will not provide an interpretable confidence interval of the overall average treatment effect.

To combat this, we adapt the approach to constructing confidence intervals introduced in Goeman et al. (2010). There, the authors present two methods. The first coarsens the parameter space into three regions, which they define as regions of inferiority, superiority, and equivalence. The regions not rejected by the three-sided hypothesis test serve as the confidence region. When testing $H_0 : \Delta = 0$, this is equivalent to a confidence interval on the sign of Δ. For instance, rejecting the superiority and equivalence regions points towards a negative effect and thus, a negative sign on Δ.

The second method allows for construction of confidence intervals with specific values of Δ in mind. Briefly, the method tests each of $H_{0,m}$ where the appropriate test for m is determined through its membership in a given partition. The confidence interval may be formed by inverting each of these hypothesis tests and retaining values of m that are not rejected. Note that the definition of a confidence interval solely necessitates a valid level α test for each value m, not that the same test is performed across the entire domain. Goeman et al. (2010) use this to justify different sided tests in different partitions of the parameter space. For example, when $m < -\phi$, where ϕ represents some previously chosen threshold demarcating the equivalence region, they apply a left-sided test. When $m > \phi$, they apply a right-sided test and when testing values of m within their equivalence region, they employ a two-sided test.

We expand upon this idea by additionally employing different estimates of Δ in different partitions. When testing equivalence, we employ PWRD aggregation through $\hat{\Delta}_{PWRD}$ to give us the greatest opportunity to detect a treatment effect. Outside of that partition, we use a method from a class of methods $f_{std}(X, Y, Z)$ that are standard to the field in which the research is conducted.

We denote the treatment effect estimate provided by a method from $f_{std}(X, Y, Z,)$ as $\hat{\Delta}_{std}$. When working with education data, for example, $f_{std}(X, Y, Z)$ may denote the class of hierarchical linear models that are common in that field. Employing these standard methods outside of the equivalence region ensures our point estimate and confidence interval are easily interpretable. This provides us with the following confidence interval bounds when conducting a t-test where s denotes the standard error and t_α denotes the critical value for significance level α:

$$l = \begin{cases} \hat{\Delta}_{std} + st_\alpha & \text{if} \quad \hat{\Delta}_{PWRD} < -\phi - st_\alpha \\ -\phi \text{ (inclusive)} & \text{if} \quad -\phi - st_\alpha \leq \hat{\Delta}_{PWRD} \leq -\phi - st_{\alpha/2} \\ \hat{\Delta}_{PWRD} + st_{\alpha/2} & \text{if} \quad -\phi - st_{\alpha/2} < \hat{\Delta}_{PWRD} < \phi - st_{\alpha/2} \\ \phi & \text{if} \quad \phi - st_{\alpha/2} \leq \hat{\Delta}_{PWRD} \end{cases},$$

$$u = \begin{cases} -\phi & \text{if} \quad \hat{\Delta}_{PWRD} \leq -\phi - st_{1-\alpha/2} \\ \hat{\Delta}_{PWRD} + st_{1-\alpha/2} & \text{if} \quad -\phi - st_{1-\alpha/2} < \hat{\Delta}_{PWRD} < \phi - st_{1-\alpha/2} \\ \phi \text{ (inclusive)} & \text{if} \quad \phi - st_{1-\alpha/2} \leq \hat{\Delta}_{PWRD} \leq \phi - st_{1-\alpha} \\ \hat{\Delta}_{std} + st_{1-\alpha} & \text{if} \quad \phi - st_{1-\alpha} < \hat{\Delta}_{PWRD} \end{cases}.$$

We use PWRD aggregation when testing within the equivalence region and a standard analysis when testing within the inferiority and superiority regions. Confidence intervals are constructed by employing the appropriate estimate of Δ and the appropriate one or two-sided test within each region. Despite applying one-sided tests in the inferiority and superiority partitions, our confidence intervals will be bounded on both sides. One of the bounds will be the standard bound from the one-sided test. The second bound is set at the edge of the rejected equivalence region rather than at positive or negative infinity.

We motivate this method in a manner similar to Caughey et al. (2021), who extend randomization tests to test bounded null hypotheses rather than solely sharp null hypotheses. They argue that one-sided rejection of the sharp null that the treatment effect τ equals some constant δ also implies rejection of any null hypothesis under which τ is bounded on one side by its corresponding δ. To simplify, they introduce bounded null hypotheses where $\tau \leq \delta$ and argue that one-sided rejection of a null of $\tau = \delta$ implies one-sided rejection of $\tau \leq \delta$. Despite using classical methods rather than randomization inference, the connection to our scenario is readily apparent: the hypotheses we test remain interpretable as a bounded null hypothesis, $H_0 : \Delta < \phi$, where our bounds are denoted through ϕ.

In addition to the earlier conditions for PWRD aggregation, this method requires one more condition adapted from Caughey et al. (2021): the test statistic must be effect increasing (Rosenbaum, 2002).

Condition 3.2.1 *The test statistic, $t(\cdot, \cdot)$ must be effect increasing, i.e. $t(z, y + z \cdot \tau + (1 - z)\psi) \geq t(z, y)$ for any $\tau \geq 0 \geq \psi$.*

In other words, take the vector of outcomes Y and the vector of treatment assignments Z. An effect increasing statistic will be increasing in Y_i when $Z_i = 1$ and decreasing in Y_i when $Z_i = 0$. This holds trivially for the class of statistics compatible with PWRD aggregation centered around $\sum_g \omega_g \hat{\Delta}_g$, where $\hat{\Delta}_g$ may be estimated through $\frac{\sum_i Z_i Y_i}{\sum_i Z_i} - \frac{\sum_i (1-Z_i) Y_i}{\sum_i (1-Z_i)}$. Clearly, $\sum_i Z_i(Y_i + \tau) - \sum_i (1 - Z_i)(Y_i - \psi) \geq \sum_i Z_i Y_i - \sum_i (1 - Z_i) Y_i$ whenever $\tau \geq 0 \geq \psi$. Thus, these statistics are effect increasing.

Caughey et al. (2021) show that when the test statistic is effect increasing for any constant ϕ, the corresponding randomization p-value $p_{Z,\phi}$ for the sharp null H_ϕ will remain valid for testing the bounded null $H_{\leq \phi}$. We argue this will similarly hold when the test statistic is effect increasing for classical inference and formalize this in Proposition 3.2.2.

Proposition 3.2.2 *If the test statistic $t(\cdot, \cdot)$ is effect increasing, then for any constant ϕ, the corresponding classical p-value p_ϕ for the sharp null H_ϕ is also valid for testing the bounded null $H_{\leq \phi}$.*

In other words, when testing the bounded null $H_{\leq \phi}$ (i.e. equivalence), $\mathbb{P}(p_\phi \leq \alpha) \leq \alpha$ for any $\alpha \in [0, 1]$ under $H_{\leq \phi}$. For a short proof, see Appendix E.

3.2.2.1 Selection of partition thresholds

When testing for equivalence, a natural threshold of interest examines the null hypothesis $H_0 : \Delta = 0$. Yet when implementing the method of Goeman et al. (2010), confidence intervals will stretch from some upper or lower bound to 0 (the edge of the equivalence region), even when equivalence is rejected. Researchers may instead prefer a confidence interval bounded away from zero. For this to occur, we need to select some threshold ϕ, $\phi > 0$ when partitioning the parameter space into three. While this ensures the confidence interval will be bounded away from zero, this also lessens the power to reject the equivalence region, i.e. to reject $H_{\leq\phi} : \Delta \leq \phi$. PWRD aggregation offers a solution. Under PWRD aggregation, it is possible to set $\phi > 0$ while simultaneously providing power to detect non-equivalence, i.e. to reject $H_{\leq\phi} : \Delta \leq \phi$, that is at least as large as the power to reject $H_0 : \Delta = 0$ using a method from $f_{std}(X, Y, Z)$.

Given we are able to reject equivalence, the following method allows researchers to select ϕ such that the confidence interval will be bounded as far from zero as possible while providing comparable power to reject $H_{\leq\phi} : \Delta \leq \phi$ as a standard method from the class of models $f_{std}(X, Y, Z)$ yields when testing $H_0 : \Delta = 0$.

1. Calculate the asymptotic relative efficiency (PE) (Pitman, 1948) of PWRD aggregation versus a standard t-test.

2. Estimate the minimum detectable effect size (MDES) for a desired level of power $(1 - \beta)$ under a standard t-test, i.e. $MDES_{1-\beta}$.

3. $\phi_{PWRD} := (\sqrt{PE} - 1)MDES_{1-\beta}$.

We observe how this works in Figure 3.1. As the asymptotic relative efficiency grows, the minimum detectable effect size for PWRD aggregation as a proportion of the minimum detectable effect size from a standard mode of analysis from $f_{std}(X, Y, Z)$ decreases. For example, when the asymptotic relative efficiency is 1, the two methods have an identical MDES. Alternatively, an asymptotic relative efficiency of 4 suggests that the MDES using PWRD aggregation is half the magnitude of the MDES under a standard method from $f_{std}(X, Y, Z)$. This allows us to shift the threshold of the equivalence region away from zero while still providing comparable (or greater) power to reject $H_{\leq\phi} : \Delta \leq \phi$ than we would possess when testing $H_0 : \Delta = 0$ and applying a method from $f_{std}(X, Y, Z)$. Note that the MDES is typically determined prior to the analysis stage. Thus, the threshold ϕ_{PWRD} is dependent on decisions made before outcome analysis. For a complete derivation of ϕ_{PWRD}, see Appendix F.

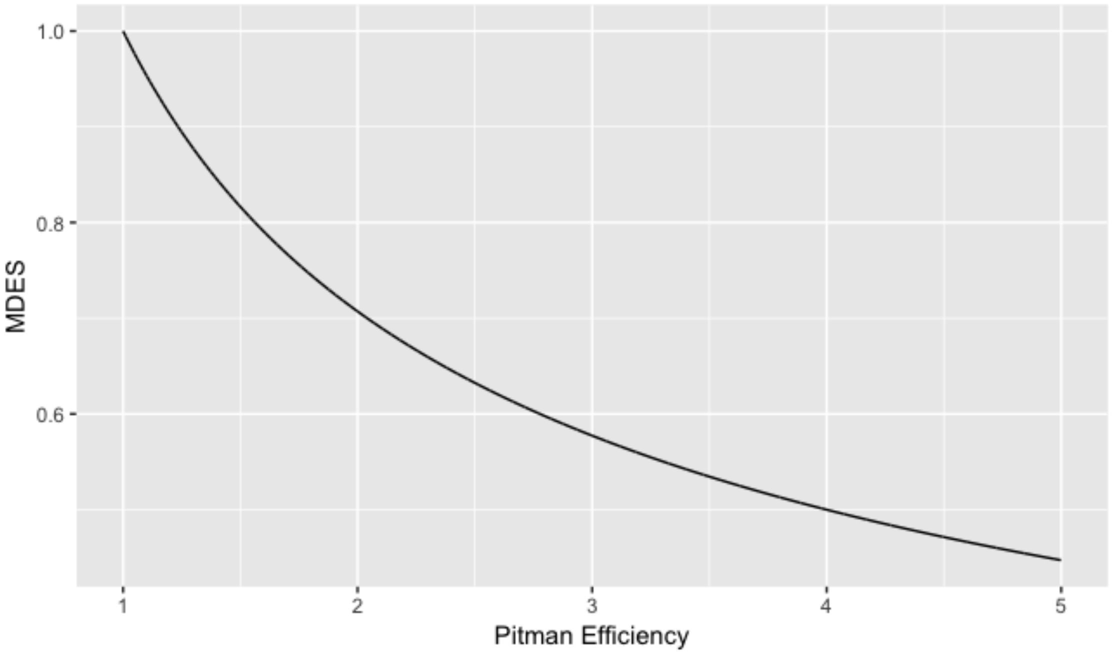


Figure 3.1: This figure shows relative improvements in MDES for PWRD aggregation over a standard analysis for asymptotic relative efficiencies (i.e. relative Pitman efficiencies). When the asymptotic relative efficiency is 4, the MDES for PWRD aggregation is 0.5 that of the MDES for a standard method.

This threshold ϕ_{PWRD} then allows researchers to weigh providing meaningful confidence intervals bounded away from zero with power considerations. For example, selecting $\phi = 0$ (and thus testing $H_0 : \Delta = 0$) realizes the complete gain in power from PWRD aggregation but the interval will stretch from zero to some upper or lower bound. Conversely, selecting $\phi = \phi_{PWRD}$ will provide a confidence interval bounded as far from zero as possible while providing identical power to reject $H_{\leq\phi} : \Delta \leq \phi$ as a standard mode of analysis from $f_{std}(X, Y, Z)$ yields to reject $H_0 : \Delta = 0$

at the level of power $(1-\beta)$ used when calculating the MDES. Under this threshold, PWRD aggregation will provide less power to reject $H_{\leq\phi} : \Delta \leq \phi$ than $f_{std}(X,Y,Z)$ yields when testing $H_0 : \Delta = 0$ for effect sizes smaller than the MDES yet greater power to reject $H_{\leq\phi} : \Delta \leq \phi$ when the effect size is larger. Thus, selecting a MDES for a lower level of power (leading to a bound closer to zero, albeit greater power) may best serve a researcher's interests. Alternatively, selecting $\phi \in (0, \phi_{PWRD})$ leads to an intermediate scenario: greater power to reject $H_{\leq\phi} : \Delta \leq \phi$ than $f_{std}(X,Y,Z)$ yields when testing $H_0 : \Delta = 0$ and intervals bounded away from zero.

Threshold	**Bound**	**Power at MDES**
$\phi = 0$	0	$(1-\beta)_{PWRD}$
$0 < \phi < \phi_{PWRD}$	$(0, \phi_{PWRD})$	$\big((1-\beta)_{std}, (1-\beta)_{PWRD}\big)$
$\phi = \phi_{PWRD}$	ϕ_{PWRD}	$(1-\beta)_{std}$

Table 3.1: Power and confidence interval bounds for various choices of ϕ

It is also possible to use the effect size to select a threshold ϕ that bounds the equivalence region. The standard approach to interpreting effect sizes uses the following guidelines: (1) effect sizes of 0.2-0.5 standard deviations of the outcome are small effect sizes; (2) those from 0.5 to 0.8 are medium effect sizes; and (3) any effect of 0.8 standard deviations or greater is a large effect size (Cohen, 2013). Researchers in education initially adopted these interpretations for effect sizes; at one point, the What Works Clearinghouse (WWC) determined that "substantively important" effect sizes arise at 0.25 standard deviations or larger. Nonetheless, Hedges et al. (2007) propose that effect sizes smaller than 0.20 may still be of substantive importance. WWC guidelines have adjusted accordingly. This reversal by WWC may have arisen because education research typically results in modest effects. To illustrate, Cheung and Slavin (2016) and Fryer Jr (2017) analyze results from randomized trials in education and find average effect sizes substantially beneath Cohen's cutoff.

In response, Kraft (2020) proposes new thresholds for effect size categorization. Effect sizes of less than 0.05 standard deviations are small, effects from 0.05 to 0.20 standard deviations are moderate, and effects greater than 0.20 standard deviations are large. From these thresholds, the researcher may select ϕ at either $\phi = 0.05\sigma$ or $\phi = 0.2\sigma$ depending on the magnitude of the effect they would like their confidence interval to cover, where σ denotes the standard deviation of the outcome.

For example, selecting $\phi = 0.05\sigma$ guarantees that, given we are able to detect non-equivalence, the confidence interval will strictly cover effect sizes of at least moderate magnitude. Similarly, the confidence interval will strictly cover large effect sizes with a threshold of $\phi = 0.2\sigma$ when we are able to reject the equivalence region.

3.2.2.2 Alternative Confidence Intervals

We previously presented a method of confidence interval construction for PWRD aggregation that adapts the three-sided confidence intervals of Goeman et al. (2010). That scheme provides a confidence interval for the overall average treatment effect Δ with power at least as large as under a standard analysis. Nonetheless, this method fails to account for different effect sizes on different groups of observations which was the initial motivation behind PWRD aggregation.

Instead, researchers may be interested in a confidence interval on η, the proportionality constant present in our family of hypotheses: $K_\eta : \Delta_g = \eta \mathbf{p}_0$. Through the duality between hypothesis tests and confidence intervals, this is comparable to testing whether the effect is proportional to dosage, an underlying tenet of PWRD aggregation. Thus, this confidence interval implicitly tests the assumptions underlying PWRD aggregation. Implementing this method and finding a significant, non-zero proportionality constant prior to employing PWRD aggregation ensures the validity behind the hypothesis testing scheme. PWRD aggregation may still be valid in certain situations when the confidence interval around the proportionality constant contains zero, but the researcher should proceed with caution. An additional benefit of this method is that the confidence interval on η implicitly allows for separate confidence intervals for the effect size on separate groups or cohorts, i.e. confidence intervals for each Δ_g.

To calculate a confidence interval around η, perform the following steps (Rosenbaum et al., 2010, p.135-6):

1. Calculate the adjusted response: $a_{ij} = Y_{ijk} - \eta p_{0g} \mathbb{1}_{(Z=1)}$.

2. Test a null hypothesis of no effect on the adjusted responses.

3. Record whether the null hypothesis is accepted or rejected.

By iterating through these three steps across different hypothesized values of η, it is possible to obtain a confidence interval for η where the confidence interval is the set of hypothesized values of η that are accepted. With this confidence interval, it is also possible to construct separate confidence intervals for each Δ_g through $[\eta_{LB}, \eta_{UB}]p_{0g}$, where η_{LB} and η_{UB} represent the lower and upper bounds for the confidence interval on η and p_{0g} represents the proportion of individuals in group g who stood to benefit from the intervention.

3.3 Simulations

To demonstrate how varying thresholds for ϕ provide different levels of power and varying confidence bounds, we extend the simulation study present in Lycurgus and Hansen (2021) (additionally found in Section 2.3). This study, mirroring the design of a reading intervention for early-elementary students (BURST in Chapters 2 and 4, generated student outcomes with the following two-level model:

$$\begin{aligned} Y_{ijk} &= \beta_0 + \beta_1 \text{Grade}_{ijk} + \mu_k + \epsilon_{ijk} \\ \mu_k &= \gamma_0 + \nu_k \end{aligned}, \tag{3.2}$$

where $\nu_k \sim N(0, \xi)$. The outcome Y of student i in school j in time-period k is a function of their school and grade. Those with sufficiently low scores are flagged as eligible to receive supplemental instruction from the intervention.

We then impose three variations of treatment effects on students in treatment schools. Under the first, each treatment observation flagged as eligible receives some constant, positive effect τ. Under the second, flagged treatment observations receive a constant, positive effect τ and unflagged treatment observations receive a constant negative effect $-p\tau$ where $p \in (0, 1]$. The third version of treatment effect imposes an effect $\tau \sim N(l, 2.5 * l)$ for some l to each treatment observation.

The first two variations of imposed treatment effect may be viewed as scenarios where the theory of change holds —the effect will be proportional to the number of individuals who stood to benefit from the intervention. Under these scenarios, we have maximized the asymptotic relative efficiency of PWRD aggregation over the standard method and thus, should be able to set $\phi > 0$ without losing power to reject $H_{\leq\phi} : \Delta \leq \phi$ (i.e. equivalence) in comparison to the power to reject $H_0 : \Delta = 0$ provided by a method drawn from $f_{std}(X, Y, Z)$. Scenario 3 demonstrates a violation in the theory of change; effects are no longer related to the proportion who stood to benefit. We should lose power to reject $H_{\leq\phi} : \Delta \leq \phi$ by adopting PWRD aggregation and expanding the size of the equivalence region. But how much power is lost?

3.3.1 Effect 1

Figure 3.2 illustrates the power for various thresholds across increasing effect sizes when the theory of change holds. As expected, PWRD aggregation with a threshold of $\phi = 0$ (thus testing $H_0 : \Delta = 0$) provides the most power. This is the base scenario under which PWRD aggregation is asymptotically efficient. To examine the performance of $\phi = \phi_{PWRD}$ and $\phi = \phi_{PWRD/2}$, we created our threshold using an MDES for $\beta = 0.50$, i.e. we select the minimum detectable effect size that will provide 50% power to reject $H_0 : \Delta = 0$. If our method works as intended, PWRD aggregation with $\phi = \phi_{PWRD}$ will provide 50% power to reject $H_{\leq\phi} : \Delta \leq \phi$ and a method from $f_{std}(X, Y, Z)$ will provide 50% power to reject $H_0 : \Delta = 0$ at the same effect size. This is, in fact, what happens. The standard method provides greater power to reject $H_0 : \Delta = 0$ than PWRD aggregation yields when testing $H_{\leq\phi} : \Delta \leq \phi$ until an imposed effect of roughly nine, at which point both methods attain 50% power. For effect sizes greater than nine, PWRD aggregation with a threshold at $\phi = \phi_{PWRD}$ provides more power. A threshold at $\phi = \phi_{PWRD/2}$ overtakes the power provided by the standard method at a much smaller level. Interestingly, in this case $\phi = \phi_{PWRD/2}$

tracks closely with a threshold set at $\phi = 0.05\sigma$, i.e. the threshold that demarcates the boundary between small and moderate effect sizes

Selecting a threshold $\phi = 0.2\sigma$ decreases power to reject $H_{\leq\phi} : \Delta \leq \phi$ substantially and may be unreasonable as a consequence. For these simulations, we let σ denote the standard deviation of the gain score rather than the standard deviation of covariate adjusted student scores. While the WWC recommends reporting effect sizes using the second, these standard deviations will be larger (Clearinghouse, 2020, p.58), and thus, rejecting equivalence using a threshold calculated from the student-level standard deviation would be prohibitively difficult. For simulations comparing power for thresholds calculated using gain score standard deviations versus student-level standard deviations, see Appendix G. In addition, note that each threshold greater than zero provides slightly deflated Type I errors. Unsurprisingly, larger thresholds provide smaller Type I errors. This is consistent with what we expect to see from Caughey et al. (2021) and from Proposition 3.2.2. The p-value for the sharp null will be conservative when testing the bounded null rather than rejecting too frequently.

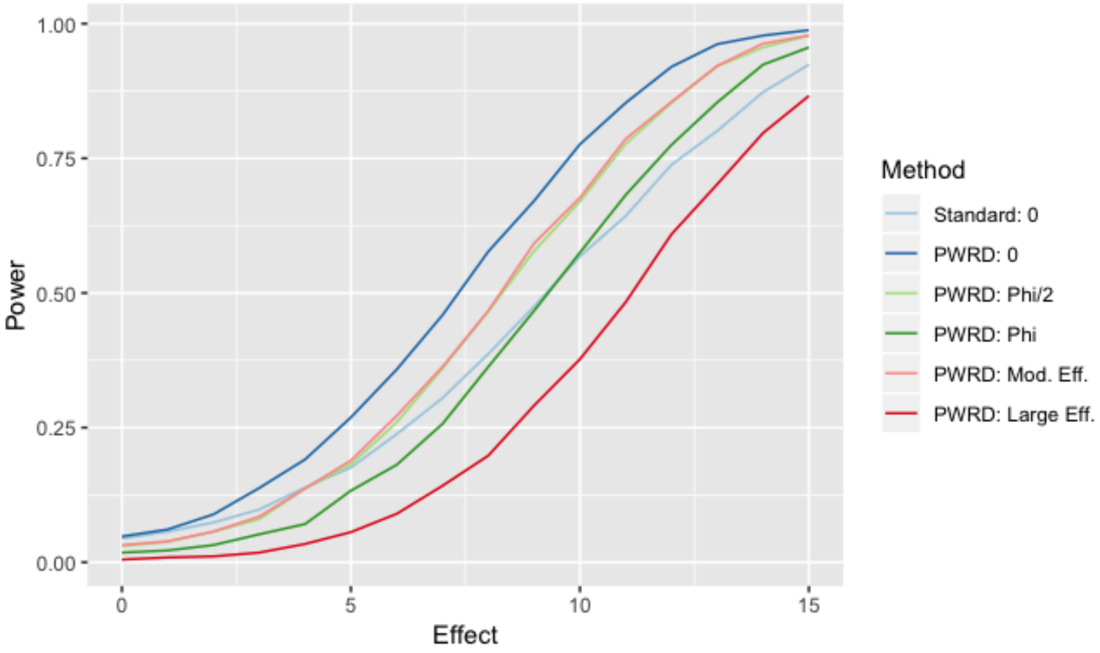


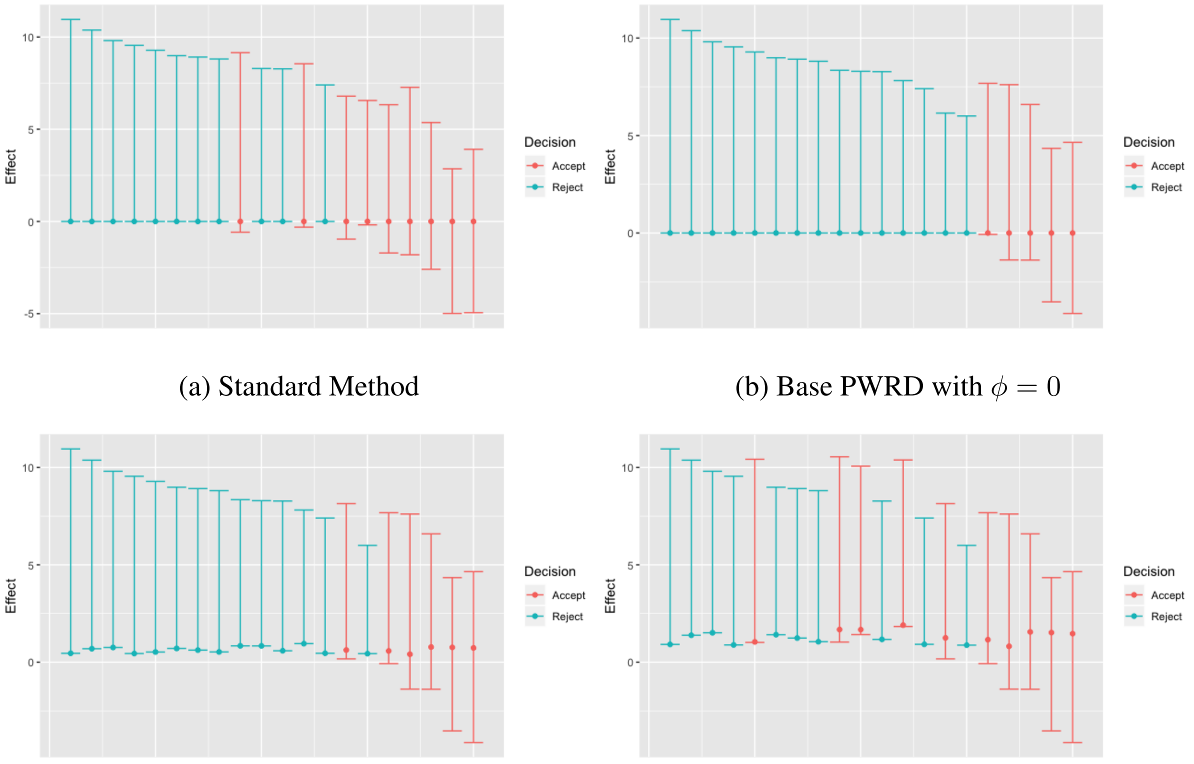
Figure 3.2: Power under various thresholds for Effect 1, i.e. across increasing effect sizes.

Figure 3.3 presents 20 confidence intervals from these simulations provided by the standard method and by PWRD aggregation under varying thresholds. Each subfigure presents the intervals in the same order—the third interval from the left in Figure 3.3d corresponds to the same simulation iteration as the third interval from the left in each of the other subfigures. Note that, under the base scenario and with $\phi_{PWRD/2}$, PWRD aggregation rejects equivalence more frequently than the standard method. The intervals constructed using ϕ_{PWRD} reject equivalence less frequently, but this is merely a function of the randomly selected intervals.

In Figures 3.3a and 3.3b, the lower bounds of intervals that reject equivalence are set at zero as both are testing $H_0 : \Delta = 0$. Lower limits are bounded away from zero in Figures 3.3c and 3.3d because the equivalence regions are set at $\phi_{PWRD/2}$ and ϕ_{PWRD}, respectively. Note that the rejection thresholds in Figure 3.3c, denoted by the points, are exactly half the magnitude of the rejection thresholds in Figure 3.3d. Thus, intervals 5, 9, 10, and 12 are rejected under $\phi_{PWRD/2}$ but not under ϕ_{PWRD}. These intervals (and some others) are entirely greater than zero yet we are unable to reject the equivalence region. In other words, we would have been able to detect non-equivalence when testing $H_0 : \Delta = 0$, yet we were not able to detect non-equivalence for $H_{\leq\phi} : \Delta \leq \phi$. This arises because the threshold ϕ lies within our interval, so while the 95% confidence interval is entirely non-zero, the entire equivalence region cannot be rejected. Instead, we have evidence that $\Delta > 0$ but not that $\Delta > \phi$. Despite failing to reject equivalence, the method still provides an interval comparable to those intervals constructed with alternative thresholds.

3.3.2 Effect 2

Figure 3.4 illustrates the power for various thresholds across increasing effect sizes when those who receive supplemental instruction benefit and those who do not are adversely affected. This, too, is consistent with a related theory of change and thus, we should see benefits to PWRD aggregation. Similar to Figure 3.2, PWRD aggregation does provide a benefit and that benefit is relatively larger than under Effect 1. Thresholds set at $\phi = \phi_{PWRD}$ and $\phi = \phi_{PWRD/2}$ both provide large gains in power to detect non equivalence when compared to the power when testing $H_0 : \Delta = 0$ yielded



(a) Standard Method

(b) Base PWRD with $\phi = 0$

(c) PWRD with $\phi = \phi_{PWRD}/2$

(d) PWRD with $\phi = \phi_{PWRD}$

Figure 3.3: Confidence intervals with varying thresholds and methods for Effect 1. The point represents the threshold that demarcates the equivalence region.

by a standard method from $f_{std}(X, Y, Z)$. Furthermore, the power to reject $H_{\leq\phi} : \Delta \leq \phi$ provided by PWRD aggregation surpasses the standard method's ability to reject $H_0 : \Delta = 0$ at a smaller imposed effect than under Effect 1. The difference is particularly stark when those individuals who receive the supplemental instruction see a large benefit and individuals who do not receive the supplemental instruction are adversely affected.

Note that the power to reject $H_{\leq\phi} : \Delta \leq \phi$ provided by PWRD aggregation using ϕ_{PWRD} no longer intersects the power to reject $H_0 : \Delta = 0$ from $f_{std}(X, Y, Z)$ at the same effect size, but well before that point. This occurs because the threshold was determined using assumptions from the original theory of change; in particular, the threshold assumed that those who do not receive supplemental instruction are unaffected (i.e. no interference is present (Sobel, 2006)). Nonetheless, PWRD aggregation still provides large gains in power over a standard method from $f_{std}(X, Y, Z)$

because a related theory of change holds. If the threshold were calculated using assumptions from the revised theory of change, we expect the two methods would intersect at the same effect size as they did under Effect 1.

Setting the threshold at $\phi = 0.05\sigma$ also provides a boost to power that once again tracks closely to the power provided by $\phi = \phi_{PWRD/2}$. This suggests that it is reasonable to select a threshold using σ when σ is calculated using the standard deviation of the gain score.

As with Effect 1, setting a threshold at $\phi = 0.2\sigma$ greatly reduces power to reject $H_{\leq\phi} : \Delta \leq \phi$. Rejecting at this threshold requires a very large effect size to overtake the power to reject $H_0 : \Delta = 0$ provided by a standard method. While this occurs when both methods attain 25% power, this is more so attributable to the standard method struggling to detect this form of effect than because $\phi = 0.2\sigma$ provides a reasonable threshold.

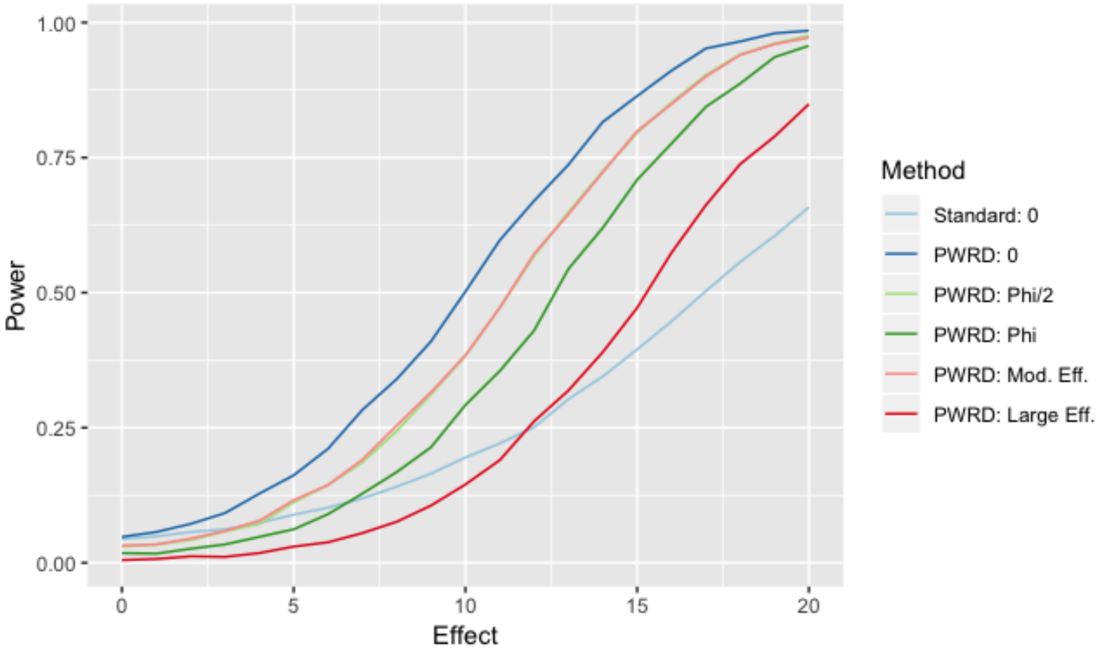
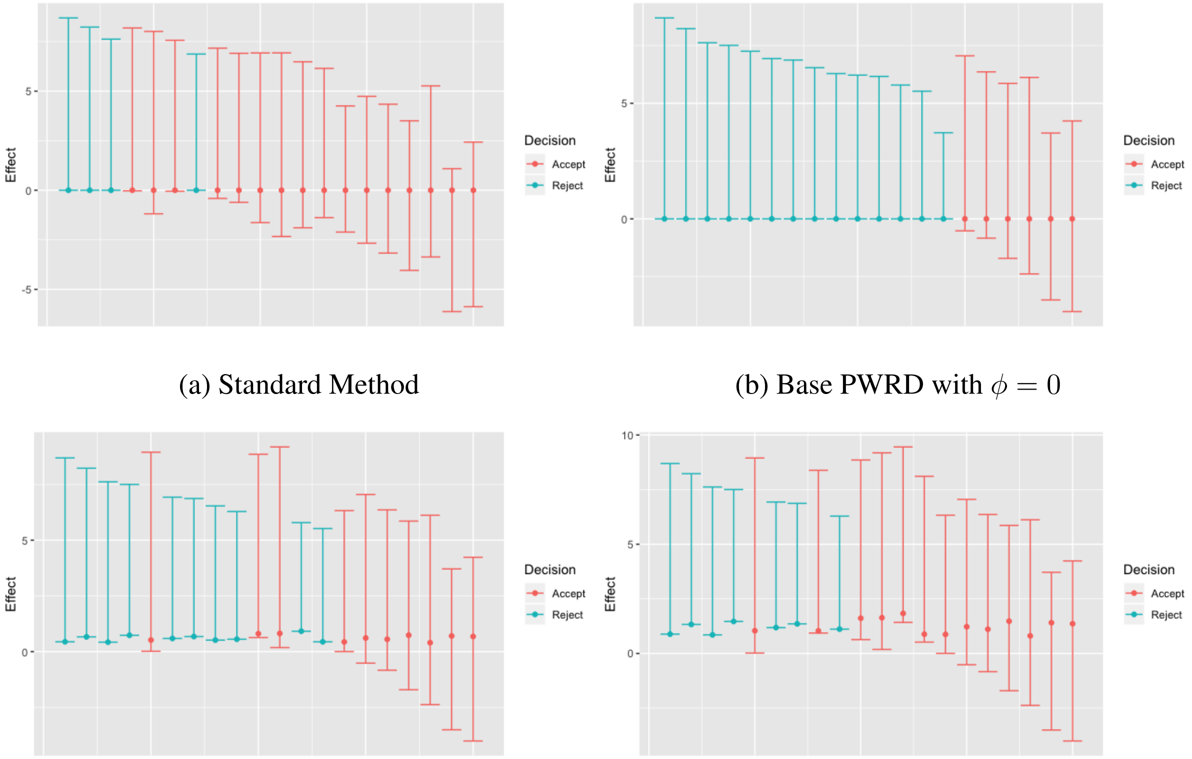


Figure 3.4: Power under various thresholds for Effect 2.

Figure 3.5 presents confidence intervals obtained through the standard method, along with through PWRD aggregation with varying thresholds for Effect 2. As with Figure 3.3, the lower bounds for rejected confidence intervals in Figures 3.5a and 3.5b are zero. Setting $\phi > 0$ provides confidence intervals bounded away from zero at the expense of some power to detect non-equivalence compared to PWRD aggregation with $\phi = 0$, yet still yields more power to reject $H_{\leq\phi} : \Delta \leq \phi$ than the power to reject $H_0 : \Delta = 0$ yielded by a standard method. The relative improvement in power for moderate to large effect sizes is greater than under Effect 1. As with Effect 1, using the most aggressive threshold for ϕ leads to accepting equivalence for intervals that are entirely non-zero. Similarly, doubling the magnitude of the equivalence threshold from Figure 3.5c to that present in Figure 3.3c costs the researcher the ability to reject equivalence in intervals 8, 12, and 13 despite these intervals strictly covering non-zero values.



(a) Standard Method

(b) Base PWRD with $\phi = 0$

(c) PWRD with $\phi = \phi_{PWRD}/2$

(d) PWRD with $\phi = \phi_{PWRD}$

Figure 3.5: Confidence intervals with varying thresholds and methods for Effect 2. The point represents the threshold that demarcates the equivalence region.

3.3.3 Effect 3

Figure 3.6 illustrates the power for various thresholds across increasing effect sizes when the theory of change fails. Under this scenario, PWRD aggregation should be detrimental to one's ability to reject $H_0 : \Delta = 0$ and that detriment should grow when setting thresholds further from zero, thus testing $H_{\leq\phi} : \Delta \leq \phi$. We do in fact observe this. PWRD aggregation with a threshold at $\phi = 0$ yields less power to reject $H_0 : \Delta = 0$ than the standard method, although this gap is never greater than 3%.

A threshold at $\phi = \phi_{PWRD/2}$ decreases power to reject the entire equivalence region by roughly 25% for moderate effect sizes but by under 15% for large effect sizes. While harmful in relative terms, this gap is smaller in absolute terms, as it corresponds to at most a 12 percentage point difference in power provided. The gulf widens when selecting a threshold at $\phi = \phi_{PWRD}$. Under small effect sizes, a standard method from $f_{std}(X, Y, Z)$ yields over double the power to reject $H_0 : \Delta = 0$ as PWRD aggregation yields when testing $H_{\leq\phi} : \Delta \leq \phi$ using $\phi = \phi_{PWRD}$. For moderate effect sizes, this decreases to 50% greater power and narrows to under a 20% gap for large effect sizes. While a substantial difference, this is never larger than a 20 percentage point deficit in power.

Unsurprisingly, partitioning based on the effect size, i.e. $\phi = 0.05\sigma$ or $\phi = 0.2\sigma$, greatly harms power to detect non-equivalence when the theory of change fails. The thresholds $\phi = 0.05\sigma$ and $\phi = \phi_{PWRD/2}$ are generally comparable in this simulation study. Thus, we observe a similar loss in power to reject $H_{\leq\phi} : \Delta \leq \phi$ using $\phi = 0.05\sigma$ as we did when using $\phi = \phi_{PWRD/2}$ when in comparison to the power yielded by a standard method drawn from $f_{std}(X, Y, Z)$ to reject $H_0 : \Delta = 0$. The gap is much larger with $\phi = 0.2\sigma$. For small effect sizes, the standard method yields over three times the power of PWRD aggregation with $\phi = 0.2\sigma$. Even for large effect sizes, the standard method offers 35% more power. In terms of absolute power lost, selecting $\phi = 0.2\sigma$ costs researchers 36 percentage points of power to detect non-equivalence for some effect sizes.

Figure 3.7 shows confidence intervals provided under the standard method, along with under PWRD aggregation for varying thresholds. As with Figures 3.3 and 3.5, the lower bounds for

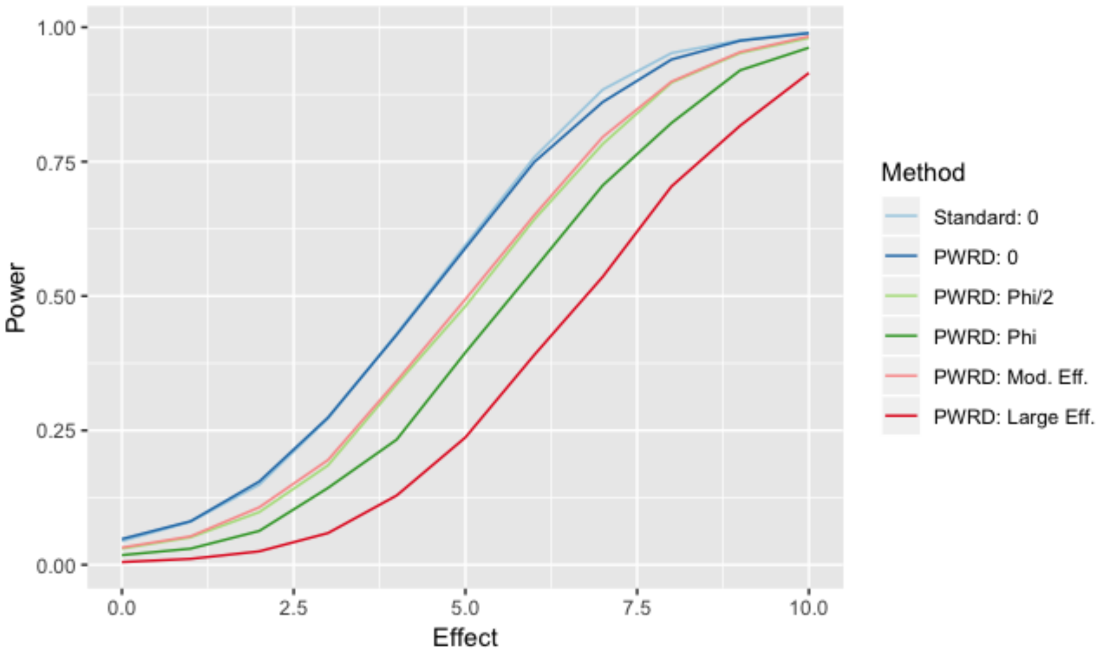
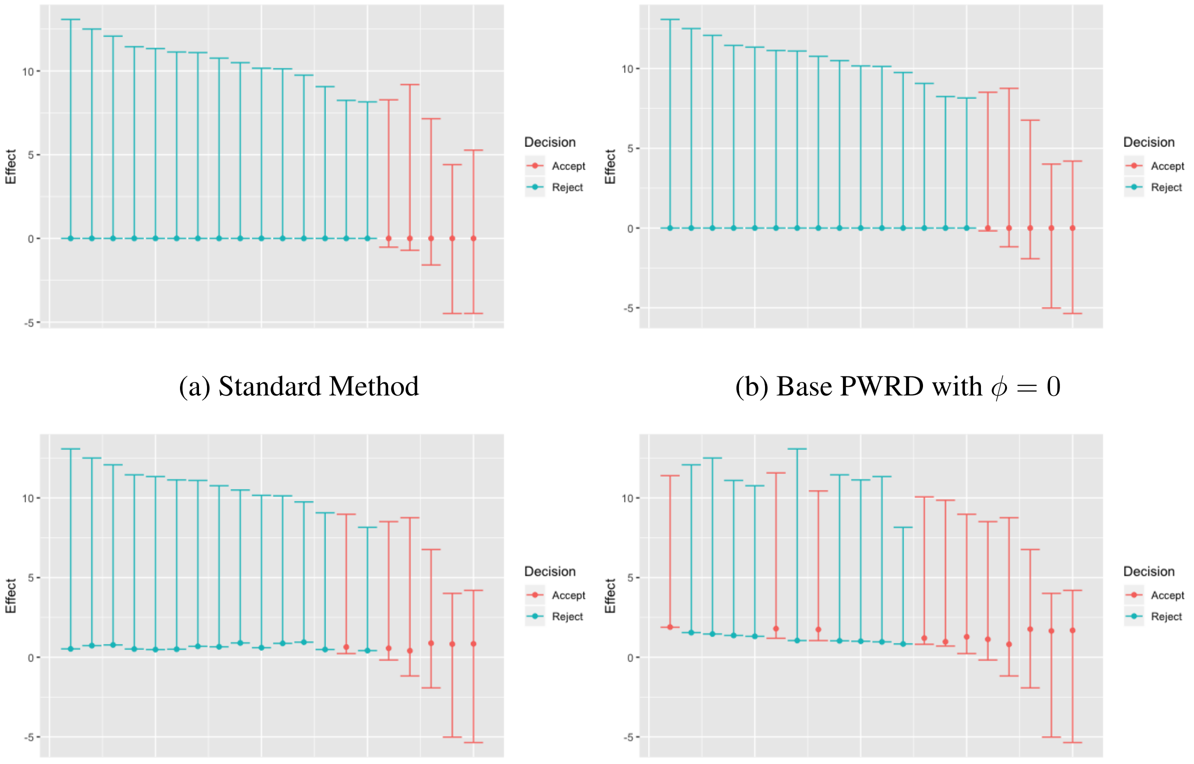


Figure 3.6: Power under various thresholds for Effect 3.

rejected confidence intervals in Figures 3.7a and 3.7b remain at zero. Setting $\phi > 0$ generates confidence intervals bounded away from zero at the expense of some power to detect non-equivalence. Note that under Effect 3, both $\phi = \phi_{PWRD}$ and $\phi = \phi_{PWRD/2}$ lead to intervals greater than 0 where equivalence is accepted. The gap in power to detect non-equivalence provided by $\phi = \phi_{PWRD}$ in Figure 3.7d compared to the other three figures is stark.

From these results, it is clear that PWRD aggregation harms a researcher's ability to detect non-equivalence when the theory of change fails. This detriment is very small under the base version of PWRD aggregation with $\phi = 0$ (and testing $H_0 : \Delta = 0$), but steadily widens as the bounds of the equivalence region widen. Nonetheless, setting ϕ in relation to ϕ_{PWRD} may still represent a fair tradeoff. A moderate amount of power to detect non-equivalence is lost using this threshold when compared to PWRD aggregation with a threshold at $\phi = 0$. However, confidence intervals for the effect estimate will include non-trivial bounds when the equivalence partition is rejected. Setting ϕ in relation to the effect size is riskier, particularly when the standard deviation of the outcome is large.



(a) Standard Method

(b) Base PWRD with $\phi = 0$

(c) PWRD with $\phi = \phi_{PWRD}/2$

(d) PWRD with $\phi = \phi_{PWRD}$

Figure 3.7: Confidence intervals with varying thresholds and methods for Effect 3. The point represents the threshold that demarcates the equivalence region.

To ensure PWRD aggregation does not fail in this manner, researchers may benefit from first constructing a confidence interval on the proportionality constant, η. This provides an initial check of whether the theory of change holds, and thus whether PWRD aggregation may safely be employed. When the interval for the proportionality constant is centered close to zero, the researcher has evidence to avoid implementing PWRD aggregation. This is the case with Effect 3: each student benefits on average regardless of their level of dosage so the effect is not proportional to the dosage received. For example, let us look at Effect 3 with an imposed effect of size 7. Among all simulations where PWRD aggregation failed to detect non-equivalence, 74% of the intervals on the proportionality constant included zero, suggesting a standard method drawn from $f_{std}(X, Y, Z)$ would have provided a greater opportunity to reject equivalence. Furthermore, when solely examining simulations where the standard method rejected equivalence and PWRD aggregation did not, i.e. those situations when PWRD aggregation directly had adverse consequences, the interval on the proportionality constant included zero one third of the time. Thus, only implementing PWRD aggregation when the interval on the proportionality constant does not include zero would cut the gap in power yielded by PWRD aggregation and the standard method by one third.

3.4 Medicaid Expansion through the Affordable Care Act

We now demonstrate three-sided confidence interval construction using our adapted scheme on the Affordable Care Act's Medicaid expansion. Beginning in 2014, states were allowed to expand Medicaid through the Affordable Care Act, vastly increasing access to health insurance in states that implemented this policy. Under the expansion, all adults with incomes below 138% of the federal poverty limit were newly eligible for health insurance through Medicaid whereas before, eligibility depended on multiple factors like age and assets. Nonetheless, many states chose not to implement the policy.

Numerous studies have examined the effect of Medicaid expansion. However, the results are mixed with some finding it reduced mortality (Borgschulte and Vogler, 2020), others finding no effect (Black et al., 2019), and still some finding mortality reductions in particular subgroups (Miller et al., 2019; Swaminathan et al., 2018). We look to add to this literature by leveraging an understudied feature: the newly eligible. Medicaid directly targets low-income individuals who would otherwise be unable to afford health insurance. Thus, we believe that given health insurance serves as the vehicle by which mortality would decrease, the benefits to expanding Medicaid will disproportionately accrue in counties that contain larger numbers of newly eligible individuals.

To assess the effect Medicaid expansion had on mortality, we use county-level full matching (Hansen, 2004; Rosenbaum, 1991) using the `optmatch` package in R (Hansen and Klopfer, 2006). This matching uses propensity scores and propensity score calipers where counties in states that expanded Medicaid are matched with counties in states that opted not to expand. We additionally attach penalties on pairings that are either not physically adjacent or are poor matches;

poor matches are determined through the match's Mahalanobis distance in comparison to the Mahalanobis distances of adjacent counties. The propensity score accounts for important covariates that are likely associated with a state expanding Medicaid (e.g. income, political leaning, education levels, etc.) along with general pre-ACA mortality. This analysis solely examines working age adults (ages 20-64), removing both children (who have different eligibility requirements) and older adults who qualify for Medicare. For a more in depth explanation of the matching strategy, see Mann et al. (2021).

To analyze Medicaid expansion and the resulting effect on healthcare amenable mortality, we apply a variation of PWRD aggregation that divides counties into six distinct brackets based on the proportion of adults that would be newly eligible for Medicaid under expansion. We then incorporate interactions between treatment status and the newly eligible bracket along with a host of demographic covariates into a generalized linear mixed effects model for the negative binomial family with random effects for the matched sets. From this model, we obtain six effect estimates (i.e. one for each bracket), $\hat{\Delta}_b$, where $b \in (1, \ldots, 6)$. We aggregate those estimates into a single test statistic, $\hat{\Delta}_{PWRD} = \sum_b \hat{\omega}_b \hat{\Delta}_b$, where $\hat{\omega}$ depends on the proportion of adults who would be newly eligible for Medicaid, calibrated by the relative precision of estimates $\hat{\Delta}_b$. For a more thorough description of our intended analysis, see Lycurgus et al. (2021).

3.4.1 Selecting ϕ

To begin, we partition the parameter space into three distinct regions by selecting ϕ. The natural threshold in this situation is $\phi = 0$ because we would like to determine whether there is a mortality benefit to Medicaid expansion. Nonetheless, we would like to construct confidence intervals bounded away from zero. Thus when selecting ϕ, we first determine the maximum value of ϕ that will provide comparable power when testing $H_{\leq\phi} : \Delta \leq \phi$ as a standard analysis yields testing $H_0 : \Delta = 0$. That is, we would like to select ϕ_{PWRD}.

To calculate ϕ_{PWRD}, we calculate the asymptotic relative efficiency between PWRD aggregation and a standard method drawn from $f_{std}(X, Y, Z)$. In this case, the standard method implements the model described previously, yet with a single indicator denoting treatment status rather than treatment status interacted with the bracket of newly eligible individuals. We estimate test slopes for PWRD aggregation and the standard method to be 12.5 and 9.61, respectively. The corresponding asymptotic relative efficiency is 1.69. We then employ the minimum detectable effect size determined in advance of estimation. The minimum detectable effect for 80% power to reject $H_0 : \Delta = 0$ when applying a method from $f_{std}(X, Y, Z)$ and testing at the 95% level with 25 degrees of freedom is 0.04. This translates to requiring at least a 4% reduction in healthcare amenable mortality. We then apply our asymptotic relative efficiency with the MDES to set $\phi = (\sqrt{1.69} - 1) \cdot 0.04 = 0.012$.

This divides our parameter space for Δ into three distinct partitions: $[0, 1 - \phi)$, $[1 - \phi, 1 + \phi]$, and $(1 + \phi, \infty)$. These regions correspond with a 1.2% or larger reduction in mortality, a 1.2% reduction to a 1.2% increase in mortality, and at least a 1.2% increase in mortality. We may then conduct three simultaneous hypothesis tests for Δ while retaining conservative type I error control.

3.4.2 Confidence Interval Construction

We then construct confidence intervals as described in Section 3.4.2. We begin by inverting two-sided hypothesis tests using PWRD aggregation within the region $[1 - \phi, 1 + \phi]$. We obtain the results presented in Figure 3.8.

It is clear that our interval lies entirely outside of this partition. We invert hypothesis tests using a standard mode of analysis in the remaining regions, $[0, 1 - \phi)$ and $(1 + \phi, \infty)$, using right-sided and left-sided tests respectively within those regions. We obtain the following confidence interval: $(0.911, 0.988)$, with a point estimate of 0.942. In other words, we estimate expanding Medicaid through the ACA reduced healthcare amenable mortality by 5.8%, with a 95% confidence interval on that estimate of $(1.2\%, 8.9\%)$. As mentioned in Section 3.2.2, the lower bound corresponds with the edge of our equivalence region thus providing an interval finitely bounded on both sides.

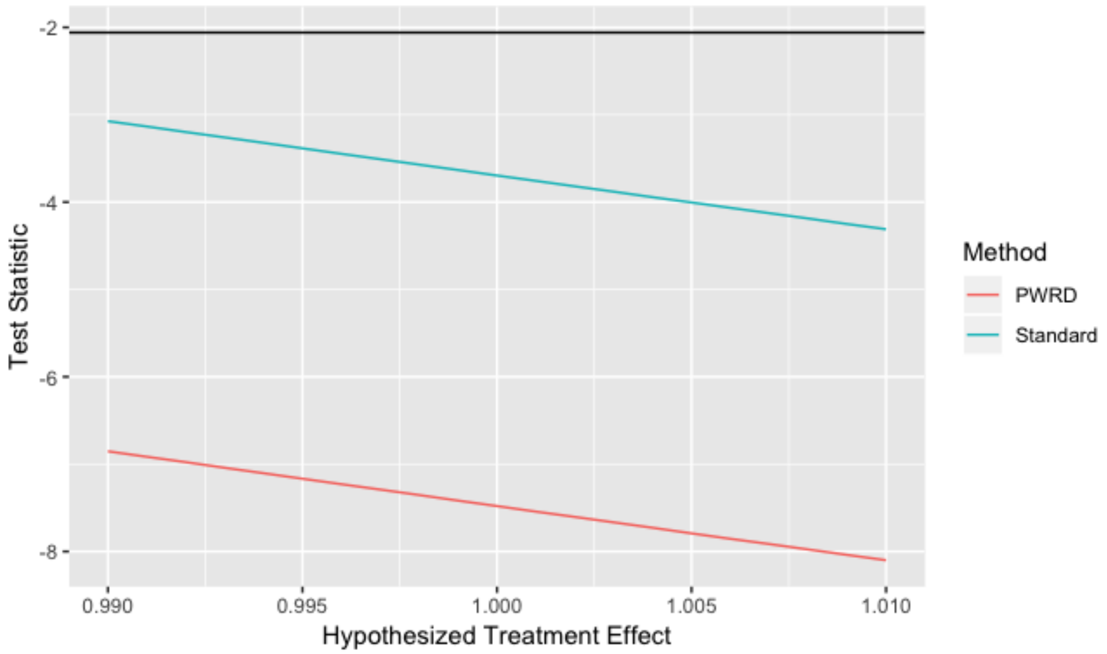


Figure 3.8: Test statistics under the standard method and PWRD analysis within the equivalence region. The horizontal line denotes the value at which we reject equivalence. This occurs using both methods.

3.4.3 Confidence Intervals for Different Values of ϕ

To illustrate how changing the thresholds affects confidence intervals, Figure 3.9 presents confidence intervals that we would have obtained under alternative thresholds for ϕ. Note that for $\phi \in [0, 0.083]$, the upper bound of the confidence interval remains fixed at 0.089, an 8.9% decrease in mortality, as does the point estimate of 0.058, a 5.8% reduction in mortality. The lower bound changes and is equal to the value of ϕ. For any value of $\phi > 0.083$, both the lower and upper bounds remain fixed at $[0.083, 0.142]$ as we fail to reject equivalence in this scenario. This interval is computed using PWRD aggregation since we are technically within our equivalence region and corresponds to the intervals in our simulation study that fail to reject $H_\phi : \Delta \leq \phi$ despite solely containing non-zero values. Note that this does not mean that 0, i.e. no effect, falls within our interval. It merely suggests that we were unable to reject the entire equivalence region.

This is a rather extreme example. PWRD aggregation provided substantially greater power than

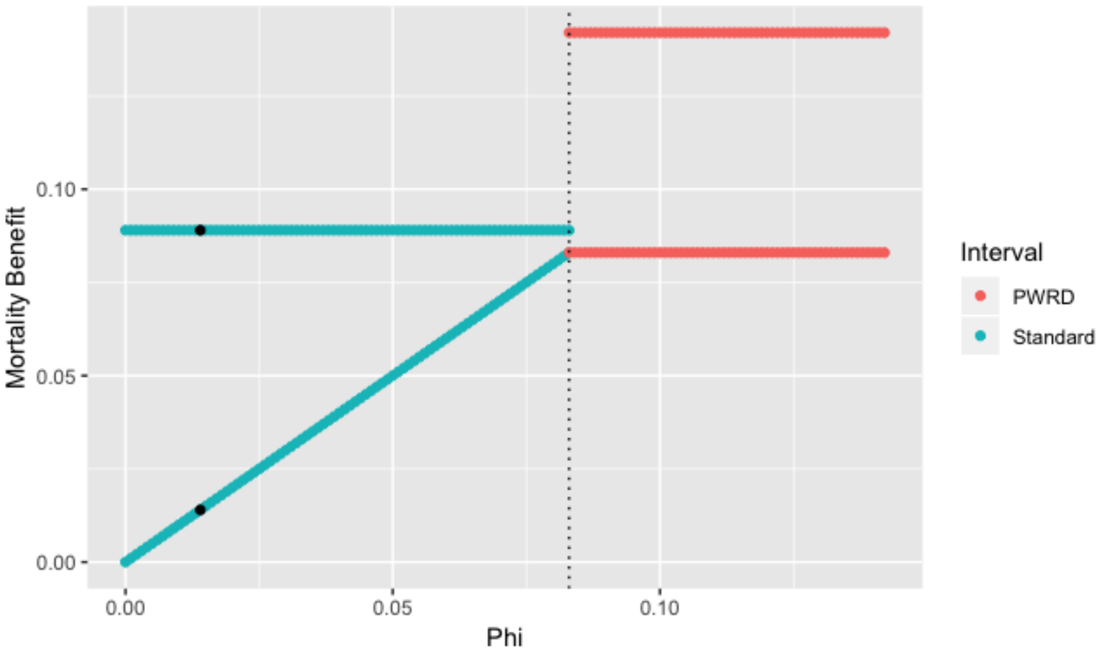


Figure 3.9: Confidence intervals for various thresholds of ϕ. The black points represent the interval bounds for the ϕ presented in Section 3.4.2. We reject equivalence and use a standard mode of analysis to the left of the vertical line, i.e. for any $\phi < 0.083$. For $\phi > 0.083$, we accept equivalence and use the PWRD aggregate to estimate our interval.

would have been available otherwise and we are able to reject equivalence at thresholds far beyond what reasonably may have been selected a priori. Thus, we only fail to reject equivalence when ϕ is roughly five standard deviations from zero. For any level of ϕ beneath that value, we reject equivalence and use a standard analysis in the inferiority and superiority regions.

3.4.4 Supermajority White Analysis

We previously demonstrated PWRD aggregation and the resulting confidence intervals when examining the effect Medicaid expansion had on healthcare amenable mortality. In this scenario, PWRD aggregation provided a substantial increase in power over the standard method. This proved to be unnecessary as a standard analysis would have rejected the null hypothesis of no effect.

To demonstrate a less extreme scenario, we conduct the same analysis, but only on counties that are deemed "supermajority white", which we define as the set of counties above the 97.5th weighted quantile for the proportion of the county's population that was white (Mann et al., 2021). For this analysis, our model is fit on every county in a matched set with a supermajority white county. Nonetheless, through an additional interaction term differentiating supermajority white counties from the others, only supermajority white counties themselves are incorporated into our test statistic. We use a three-sided approach to construct a confidence interval estimating the effect Medicaid expansion had on healthcare amenable mortality in supermajority white counties.

We find the test slopes to be 7.05 and 5.64 for PWRD aggregation and the standard method, respectively. This leads to an asymptotic relative efficiency of 1.56. With a minimum detectable effect size of 0.07 (with 80$\%$ power), we select a threshold of $\phi = (\sqrt{1.56} - 1) \cdot 0.07 = 0.018$. PWRD aggregation (but not a standard method) allows us to reject the equivalence region of $[1 - \phi, 1 + \phi]$. By inverting hypothesis tests within the inferiority and superiority regions, we obtain a confidence interval of $(0.897, 0.982)$ with a point estimate of 0.938. This suggests that supermajority white counties that expanded Medicaid experienced a 6.2% reduction in mortality from 2015-2018 compared to supermajority white counties that did not expand Medicaid. The 95$\%$ confidence interval on this estimate is $(1.8\%, 10.3\%)$.

Note that in this scenario, we would not have been able to reject equivalence at $\phi = 0.017$ despite rejecting a null hypothesis of $\Delta = 0$. Under a standard analysis with neither PWRD aggregation nor three-sided confidence intervals, the 95$\%$ confidence interval would have stretched from a 1.2% reduction in mortality to a 11.2% reduction in mortality. This estimate entails both wider bounds and a smaller minimum effect.

3.4.5 Confidence Intervals for Effects Proportional to Dosage

We now construct confidence intervals for the proportionality constant, working under the assumption that the effect is proportional to dosage. First, we calculate adjusted responses for each county. Let M_{ij} represent the mortality in county i in year j and p_b denote the proportion of adults in a county belonging to bracket b who would be newly eligible for Medicaid under expansion through the ACA. The adjusted mortality may then be written as:

$$a_{ij} = M_{ij} - \eta p_b \mathbb{1}_{(Z=1)}, \tag{3.3}$$

where η denotes the proportionality constant. For a given value of η, $\eta \in (-1, 1)$, we calculate the adjusted mortality responses through Equation 3.3. We then test a null hypothesis of no effect using a two-sided t-test at the 5% level and record whether we reject or accept the null hypothesis for that proportionality constant η. We iterate through this process for each possible η from -1 to 1 and keep the values for η where the hypothesis test fails to reject the null. We obtain a point estimate of $\eta = 0.55$ (i.e. the η that provided a test statistic of zero), with a 95% confidence interval of $\eta = [0.27, 0.83]$. This suggests that if 10% of adults in a county stood to benefit from Medicaid expansion, that county would see a $0.55 * 10 = 5.5\%$ reduction in healthcare amenable mortality. Table 3.2 provides the proportion of newly eligible residents by bracket, along with bracket specific confidence intervals on the reduction in healthcare amenable mortality under the assumption that the effect accrues proportionally to dosage. These results provide evidence that the proportionality assumption behind PWRD aggregation does, in fact, hold and it was proper to use that method when conducting outcome analysis. These results are consistent with the idea that counties with lower proportions of residents who were newly eligible for Medicaid benefitted less from the expansion. For example, Bracket 1, with roughly 1% of residents newly eligible,

Bracket	**Newly Eligible**	**LB:** $\eta = 0.27$	**Est:** $\eta = 0.55$	**UB:** $\eta = 0.83$
1	0.01	0.00	0.01	0.01
2	0.07	0.02	0.04	0.06
3	0.10	0.03	0.05	0.08
4	0.12	0.03	0.07	0.10
5	0.15	0.04	0.08	0.12
6	0.20	0.06	0.11	0.17
Agg.	0.11	0.03	0.06	0.09

Table 3.2: The estimated effects on healthcare amenable mortality for each of the six brackets under different values of η, the proportionality constant.

experienced a reduction in mortality of under 1%. Bracket 6, on the other hand, likely saw more than a 10% reduction in mortality due to the expansion.

In addition, note that the confidence interval provided on the aggregated estimate is very similar to the confidence interval from the three-sided method. That method provided a point estimate of a 5.8% reduction in mortality with a 95% confidence interval stretching from 1.2% to 8.9%. Here, we obtained a point estimate of 6% with an interval stretching from 3% to 9%. The smaller lower bound with the three-sided method results from the threshold ϕ selected in Section 3.4.1.

We now calculate an interval on the proportionality constant for the supermajority white analysis. Following the steps outlined above, we estimate the proportionality constant to be $\eta = 0.56$ with a 95% confidence interval of $\eta = [0.11, 0.99]$. This interval is substantially wider than the interval yielded by the overall analysis of healthcare amenable mortality. Nonetheless, this is understandable. The sample size for supermajority white counties is 496 compared to 2996 counties overall and the issue is more pronounced when looking at population size since supermajority white counties tend to be smaller (as a result of being more rural). The full results are presented in Table 3.3. Interpretation follows as before.

Bracket	**Newly Eligible**	**LB:** $\eta = 0.11$	**Est:** $\eta = 0.56$	**UB:** $\eta = 0.99$
1	0.03	0.00	0.02	0.03
2	0.07	0.01	0.04	0.07
3	0.10	0.01	0.06	0.10
4	0.12	0.01	0.07	0.12
5	0.15	0.02	0.09	0.15
6	0.20	0.02	0.11	0.20
Agg.	0.11	0.01	0.06	0.11

Table 3.3: The estimated effects on healthcare amenable mortality in supermajority white counties for each of the six brackets under different values of η, the proportionality constant.

Note that the confidence interval on the overall effect stretches from a 1.2% reduction in mortality to an 11% reduction in mortality with a point estimate at 6.3%. The three-sided method resulted in an interval of $(1.8\%, 10.3\%)$ and an estimate of a 6.2% reduction in mortality. Once again, these two methods provide similar confidence intervals on the overall reduction in mortality.

In this case, the three-sided method through PWRD aggregation provided a larger lower bound of the effect than did the dosage interval.

3.5 Discussion

In this chapter, we present two methods for confidence interval construction that are compatible with PWRD aggregation. The first is a novel adaptation of the three-sided testing scheme of Goeman et al. (2010) that allows for implementation of PWRD aggregation for increased power to detect non-equivalence while attaching easily interpretable and valid confidence intervals on the overall effect estimate. The second, working under the assumption of an effect proportional to the dosage, calculates a confidence interval on the proportionality constant. This additionally serves as a check on PWRD aggregation. When the proportionality constant is non-zero, the assumptions behind PWRD aggregation likely hold and the method may be safely implemented. When the interval on the proportionality constant contains zero, the assumptions may hold (in which case PWRD aggregation still provides a benefit), but the researcher should proceed with caution.

Simulations demonstrate that PWRD aggregation in tandem with the three-sided confidence interval presented in Section 3.2 will provide additional power over a standard flavor of analysis, given assumptions behind PWRD aggregation hold. Nonetheless, there may be a detriment, albeit a small one, in situations where the standard analysis would have detected an effect as well. In this scenario, the lower bound of the confidence interval may be closer to zero at times than it otherwise would have been. Thus, unsurprisingly, PWRD aggregation provides the greatest benefit in settings where there is an effect, but the signal is drowned out by the noise. In this scenario, PWRD aggregation allows the researcher to detect the effect and additionally bound the confidence interval away from zero while other methods may not detect any effect. From our simulations, the benefit to calculating the interval on the proportionality constant first becomes apparent as well. When the simulated effect was not proportional to dosage, i.e. when assumptions failed, calculating the proportionality interval first and only proceeding with PWRD aggregation if the interval did not contain zero reduced the difference in power between PWRD aggregation and the standard method by one third.

We demonstrated these methods on a case study examining whether expanding Medicaid through the Affordable Care Act led to a reduction in mortality. This scenario was appropriate for PWRD aggregation since counties with larger numbers of newly eligible individuals likely stood to benefit more from the policy. Yet the scenario also necessitated a method of estimating the overall potential reduction in mortality. PWRD aggregation allowed us to detect a highly significant reduction in mortality both overall and on the subset of supermajority white counties. Using our three-sided confidence interval method, we were able to estimate the magnitude of that reduction to be roughly 6% both overall and solely examining supermajority white counties. The confidence intervals on the proportionality constant led to similar estimates of the reduction in mortality as a result of Medicaid expansion through the ACA, demonstrating the cohesion between the two methods when assumptions behind PWRD aggregation hold.

CHAPTER 4

Power Enhancement for Cluster Randomized Trials via Dry Runs

4.1 Introduction

Randomized controlled trials (RCTs) are considered the gold-standard among applied research organizations because of their strong internal validity and general lack of bias in large samples. Consequently, RCTs have become increasingly popular over the last few decades across diverse disciplines like medicine and criminology. They have especially flourished in the field of education. From 1980 to 2016, researchers conducted over 1000 randomized trials to answer education-related questions; three quarters of these trials were conducted in the final ten years of that period (Connolly et al., 2018).

Despite their widespread use, much debate remains as to the best method for outcome analysis in education settings. For example, some propose that because of the lack of bias in well-balanced randomized trials, covariate adjustment is entirely unnecessary and a simple difference-in-means suffices for estimating the treatment effect. Others suggest that while researchers can substantially improve precision by controlling for pre-tests, further covariate adjustment remains unnecessary (Bloom et al., 2007). A third line of thought argues that while pre-tests often improve precision, other covariates may also help in different scenarios (Raudenbush, 1997). Nonetheless, that merely leads to questions of which covariates should be incorporated.

Even among those who favor covariate adjustment, questions about model type abound. The prevailing technique is to fit a hierarchical linear model to account for the clustered nature of many designs in education, incorporating random effects at the school, classroom, or student-level, or potentially some combination of the three (Raudenbush and Bryk, 2002; Bryk and Raudenbush, 1987). Others, however, opt for standard covariate adjusted linear regression (see Meece and Miller (1999) and Simmons et al. (2008)). Ultimately, as argued in Bloom et al. (2007), there is no one-size-fits-all approach; instead different studies require different covariate adjustment and perhaps, different modeling entirely. The question is then how to best select covariates and model type.

One common strategy to assist in variable selection is the LASSO, a penalized regression technique first proposed in Tibshirani (1996) that sets certain coefficients with less predictive power to zero. While it has been adapted for use in randomized trials (e.g. in Bloniarz et al. (2016)), it is applied less frequently in education settings as it may select covariates without theoretical backing and remove others that are grounded in theory. On the other hand, lists of mandatory covariates specified in analysis plans may exclude important predictors while including less relevant variables. This is because researchers frequently have inadequate prior knowledge as to which covariates will provide the most precise estimate of the effect (Pocock et al., 2002). Instead, some researchers merely choose covariates that are imbalanced between treatment and control groups, and others still adopt a stepwise variable selection procedure (Pocock et al., 2002). Stepwise variable selection, however, tends to overestimate how well the model fits the data, among other issues, and should be performed with caution (Hurvich and Tsai, 1990).

As an alternative to one of these methods, we introduce a "dry run" simulation method that allows for model and covariate selection by sampling strictly from the control data. This method may be viewed as a form of cross-validation adapted to an experimental context. In cross-validation, models are trained on a randomly divided portion of the data before being tested on the remaining data. By iterating through this process, it is possible to estimate the predictive performance of the tested models. In our dry run simulation scheme, certain observations are sampled from the control data and assigned to the "pseudo-treatment" while the remaining observations remain part of the "pseudo-control". Notably, each observation originates from the control data and therefore, has not been exposed to any treatment effect. Thus, iterating through this procedure of generating a "pseudo-experiment" allows for comparisons of different model specifications in terms of mean-squared-error, power and bias among others.

This may also be viewed as a form of *uniformity trial* (Rosenbaum, 2018, p.33), a process popular in the 1920s and 1930s where plots of land were divided into treatment and control, yet all received the control. After a uniformity trial, researchers possess each y_c, i.e. every potential outcome under the control. This allows for researchers to learn empirically how much the treatment and control could differ under the presence of no treatment effect. A full uniformity trial is impossible for a randomized trial: a subset of observations receive the treatment and thus, we do not observe their potential outcomes under the control. Nonetheless, the dry run process serves as a uniformity trial *within* the randomized trial. Inferences drawn from the dry run uniformity trial from a subset of the y_c's should be informative about the full, theoretical uniformity trial.

This method allows us to avoid incorporating any assumption of the form $Y = X\beta + \epsilon$ into our model selector. This provides us with an alternative method to covariate selection strategies common among field trialists like stepwise variable selection using R^2. To illustrate the advantage of avoiding model selectors reliant on the above form, we present a trivial example. We generate outcomes Y using the following model:

$$Y = 0.9Z + \beta X^2 + \epsilon, \tag{4.1}$$

where Z is a binary variable indicating treatment status and $\epsilon \sim N(0, 1)$. Figure 4.1 presents these observations and the regression $Y \sim Z + X\beta$ fit to this data when $\beta = 1$. The above model provides an R^2 of 0.92 suggesting a very good fit. Nonetheless, potential outcomes are clearly not linear in X. In addition, the estimates of β, the treatment effect, and even the intercept are imprecise. While data were generated without an intercept, the model in Figure 4.1 estimates an intercept of -4. Worse, the treatment effect is estimated to be 0.47, far undershooting the true value of 0.90 and limiting our ability to detect a treatment effect. As the coefficient β increases in

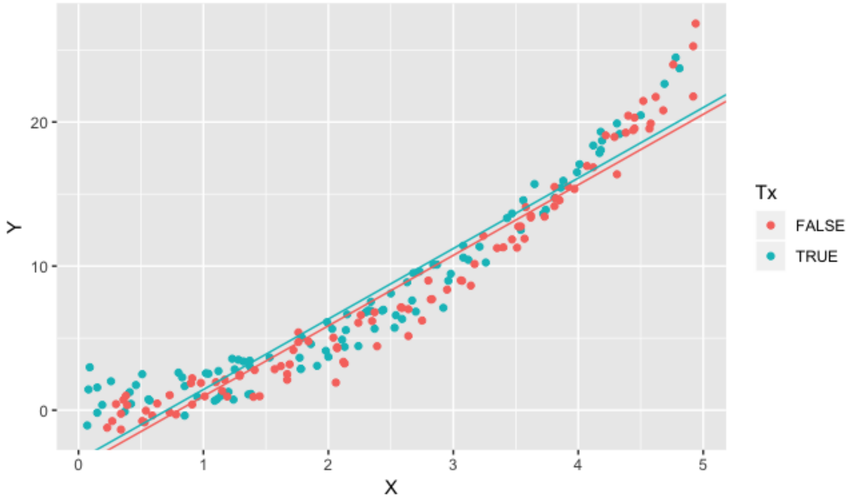


Figure 4.1: Linear regressions of the form $Y \sim Z + X\beta$ for data generated from the non-linear model specified in Equation 4.1, where $\beta = 1$.

magnitude, the model's R^2 increases yet the estimated coefficients grow increasingly imprecise.

Our alternative method is motivated through the dry run simulations proposed in Wyss et al. (2017) for nonrandomized studies. We construct pseudo-treatment and pseudo-control clusters strictly using observations assigned to the control. By pseudo-treatment and pseudo-control, we mean separate clusters from the control data that we randomly "assign" to the treatment and the control. If performed properly, the pseudo-experiment should resemble the true experiment providing a sandbox on which we can test various models and sets of covariates on real rather than contrived data to determine how best to maximize power and precision.

This simulation scheme is similar to standard power analyses often conducted in the design stage of randomized trials, which are typically performed to ensure that the randomized trial has a sample size large enough to detect an effect, conditional on the presence of a treatment effect. Generally, these are closed-form power analyses, although power analyses with simulated data appear in the literature as well (Black et al., 2019; Croke et al., 2016; Hannon et al., 1993). Adopting a simulation-based power analysis rather than a closed-form approach removes the need to fully model the data generating process, a particularly difficult task with clustered standard errors often present in education research.

Simulated power analyses usually take two forms: those with synthetic data and those with real data. The first is an entirely artificial process and possesses many of the same drawbacks as a closed-form analysis. The researcher must make assumptions about both the data generation process and the covariance matrix. On the other hand, synthetic data provide one notable advantage: a known treatment effect. Since the data are entirely artificial, any treatment effect present must have been artificially incorporated by the researcher. Thus, the magnitude of that effect is known.

However, the dry run simulation scheme presented in this chapter best corresponds to the power analyses conducted in Black et al. (2019) and Hannon et al. (1993), which employ real data. Black et al. (2019) achieve this by using and modifying pre-treatment data (which may be unavailable in many cases) whereas Hannon et al. (1993) bootstrap their data. Neither of these methods offers precise control over the simulated treatment effect. Our simulation scheme can make this promise. Every simulated data point is constructed from observations drawn from the control group and thus, has not been exposed to the treatment. As a result, our method possesses the same benefit of a power analysis with synthetic data without needing to make any assumptions about the data generation process. Furthermore, dry runs create treatment and control groups that are realistic in terms of covariate balance.

4.1.1 Roadmap

In Section 4.2 of this chapter, we outline the general process for creating our sandbox simulation method. First, we present the base scenario where the randomized control trial uses simple random assignment to designate treatment versus control groups. We then extend the method to the cluster randomized trials common in education research. We demonstrate this process on both a small, contrived data set, and on a large-scale, IES-funded cluster-randomized trial. Section 4.4 illustrates benefits and possible applications of this simulation method with reference to the aforementioned IES study. In particular, we demonstrate how our scheme assists with model selection by providing precision estimates for different model specifications and effect sizes. We then examine whether inferences drawn from this method are valid using re-randomization inference on the complete data. Finally in Section 4.5, we finish with a discussion summarizing how this dry run sandbox simulation scheme helps researchers increase precision and power through model and covariate selection.

4.2 The Dry Run Simulation Method

In this section, we discuss how pseudo-clusters are generated and assigned a treatment status. We provide steps outlining the process on a randomized trial with simple random assignment and a cluster randomized trial. We then discuss how this method fits into randomized trials embedded within surveys.

4.2.1 Dry Runs for Randomized Trials with Simple Random Assignment

We first illustrate the dry run simulation scheme on the most straightforward scenario: a randomized trial with simple random assignment. Under simple random assignment, each observation or participant has the same probability of assignment to the treatment. Thus, the steps necessary to generate a pseudo-experiment solely using control data are simple. Let n_c and n_t denote the number of observations in the randomized trial assigned to the control and treatment respectively. Researchers then have the following two options for sampling students into pseudo-groups:

- **Sample without replacement:** Form the pseudo-treatment group by sampling $n_c\left(\frac{n_t}{n_t+n_c}\right)$ observations without replacement from the n_c control observations. In other words, if 40% of observations were randomly assigned to the control in the randomized trial, 40% of the control observations will be assigned to the pseudo-treatment within the pseudo-experiment. The remaining observations are grouped into the pseudo-control.

- **Sample with replacement:** Form the pseudo-treatment group by sampling n_t observations with replacement from the n_c control observations. Then, form the pseudo-control group by sampling n_c observations with replacement from the n_c control observations.

This setting, particularly when sampling without replacement, most closely adheres to cross validation since each observation has the same likelihood of appearing in the pseudo-treatment versus pseudo-control groups. As with cross-validation, the dry run method iterates through different realizations (although unlike cross-validation, we do not *require* each observation to belong to the pseudo-treatment during a single iteration) and estimates the squared error on each iteration. Those errors are then averaged to estimate the mean-squared-error for the randomized trial.

We briefly illustrate this process on the 2006 Massachusetts Healthcare Reform. This reform, on which the Affordable Care Act was modeled, employed a three-pronged approach to providing universal health care coverage to Massachusetts residents: expansion of Medicaid, subsidized private health insurance, and an individual mandate. To analyze the benefits, if any, this legislation had on mortality, Sommers et al. (2014) employs propensity score methods to construct a control group of counties that closely resemble counties in Massachusetts, i.e. those counties that received the treatment. For more background, see Appendix D.

While this is a quasi-experimental setting rather than a randomized trial with simple random assignment, they share many similar characteristics. Furthermore, once the control group is formulated, analysis proceeds as if the treatment were assigned randomly. Thus, we implement the dry run method outlined above on this example after constructing the control group by applying a revised propensity score that accounts for county population.

There are 14 counties in Massachusetts and 512 counties in our control group, so we sample 526 counties with replacement from the 512 control counties and randomly assign 14 of them to the pseudo-treatment. We then fit two models to this pseudo-experiment: a negative binomial model controlling for demographic covariates and fixed effects at the state level and the same model but with random effects at the state level. Since each county belongs to the true control group, the treatment effect is necessarily zero. Thus, we can estimate the bias and error of each model. We iterate through this process 500 times, assigning 14 counties randomly to the treatment at each iteration. Across these 500 iterations, we estimate the bias and root mean-squared-error of the two models (RMSE). Results are presented in Table 4.1. We see that both models are unbiased

	Bias	**RMSE**
Fixed Eff.	0.000	0.0376
Random Eff.	0.000	0.0375

Table 4.1: Dry run results across 500 iterations.

and while the random effects model is slightly more efficient, its performance with the model incorporating fixed effects is comparable.

4.2.2 Dry Runs for Cluster Randomized Trials

We previously discussed how to implement dry runs in the base scenario: randomized trials with simple random assignment. Nonetheless many interventions, particularly those in the social sciences, involve cluster random assignment. For example, an entire school or classroom is assigned to the treatment or control rather than assigning students individually. Thus, while each *school* may have the sample probability of assignment to the treatment, those schools themselves often possess different demographic profiles. The dry run method must be adapted to account for these potential discrepancies and to ensure that the pseudo-treatment cluster resembles the true treatment cluster and the pseudo-control cluster resembles the true control cluster.

To divide control observations into two pseudo-clusters, we perform the following steps:

1. **Fit a propensity score model within each block:** Use a parametric binary regression model, such as a propensity score model (Rosenbaum and Rubin, 1983), to summarize baseline differences among observations belonging to the treatment or control cluster. This estimates the probability of belonging to the treatment within a block of observations. We fit a new propensity model within each block rather than employing one model across all blocks.

2. **Form the first cluster using propensity scores:** Create pseudo-cluster A by sampling from the control observations in a manner proportional to their propensity scores. This may be performed in one of two ways:

- Sample with replacement from the control observations, where the probability of placing an observation into pseudo-cluster A is equivalent to the propensity score, PS_{ij}, for that observation calculated in the previous step.

- Sample without replacement from the control observations. The sampling probability is performed in a manner such that the odds of assignment into pseudo-cluster A are proportional, although not equivalent, to the propensity scores from the preceding step. To elaborate, calculate p_{ij} (the probability that observation i block j is placed in pseudo-treatment cluster A) as follows:

$$p_{ij} = \frac{\exp(d + \theta_{ij})}{1 + \exp(d + \theta_{ij})}.$$

In this case, θ_{ij} represents the log-odds of propensity score PS_{ij} of observation i belonging to the treatment cluster within match j, i.e. $\theta_{ij} = \log(\frac{PS_{ij}}{1-PS_{ij}})$. In addition, d is a constant chosen such that:

$$\sum_{i=1}^{n_{c_j}} p_{ij} = \frac{n_{t_j}}{n_{c_j} + n_{t_j}} n_{c_j},$$

where n_{c_j} and n_{t_j} denote the number of observations in the control and treatment clusters within block j respectively. Assigning observations to pseudo-cluster A using this scaled quantity p_{ij} (with the correct constant d) rather than the propensity score ensures that the proportion of observations within A is equivalent in expectation to the proportion of observations within the actual treatment cluster for each block. That is, the expected number of observations in pseudo-cluster A will equal $\frac{n_{t_j}}{n_{t_j}+n_{c_j}} n_{c_j}$.

To illustrate, take a block of classrooms with 50 total students. The treatment classroom of 20 students is matched with a control classroom of 30 students. 40% of students in this block belong to the treatment classroom so we would like 40% of students in the pseudo-block to belong to the pseudo-treatment classroom. This process ensures that, in expectation, the pseudo-treatment classroom will consist of 12 of the 30 students from the control classroom, or 40% of the control students.

3. **Form the second cluster:** If sampling with replacement was chosen in the previous step, form pseudo-cluster B by sampling with replacement from the same set of observations, but with a probability equal to $(1 - PS_{ij})$ rather than PS_{ij}, i.e. the propensity of an observation belonging to the control cluster.

   If sampling without replacement was chosen in the previous step, form pseudo-cluster B by grouping together all the remaining observations that were not selected into pseudo-cluster A.

The above steps will divide the control observations in a block into two pseudo-clusters that largely mimic the true clusters in the block. In an education context, a “block” refers to clusters (e.g. classrooms, schools, or school districts) that are matched together based on common characteristics and randomly assigned to the treatment or control. Note that we use “block” and “matched set” interchangeably throughout this chapter.

While the pseudo-clusters within a pseudo-block will resemble the true clusters within the actual block, small discrepancies in covariate balance between the actual treatment and control clusters will widen in the dry run analysis. For example, if the true treatment and control clusters are 55% and 50% white respectively, the pseudo-clusters may be 57% and 47% white respectively. To combat this, researchers can randomly apply “treatment” and “control” labels after dividing observations into two groups, which is similar to how treatment is reassigned at each permutation during re-randomization inference. Random assignment within each pair yields perfectly balanced blocks across many iterations of the simulation whereas trial and error of different treatment assignment probabilities can recreate the relative covariate imbalance found in the actual experiment.

For the purpose of this chapter, we choose the first option: within each pair, a pseudo-cluster has a 50% chance of receiving the treatment. While this will leave perfectly balanced groups across all iterations of the simulation scheme, we will still observe minor covariate imbalances within any given iteration of our simulations. This corresponds neatly with the theory behind randomized trials. Although any specific experiment will likely possess minor discrepancies in covariate balance between treatment and control clusters, those discrepancies disappear over an infinite number of realizations because each cluster is equally likely to receive the treatment.

Naturally, researchers may worry that this simulation method will lead to invalid inferences after model selection, either as a consequence of overfitting or inflated Type I errors. The first issue may arise because we directly use control data to select our model, which may over-adapt idiosyncrasies of the control data onto the treatment data as well. Many different non-parametric resampling tools are available to protect against this worry, from the jackknife (Quenouille et al., 1949; Quenouille, 1956) to the bootstrap (Efron et al., 1979). In this chapter, we choose to apply the subsampling method outlined in Politis et al. (1999) where only a subset of blocks are selected at each iteration of the dry run process. We additionally address fears of over-rejection by leveraging a unique aspect of the cluster randomized trial serving as the motivating example. For a more complete examination of complications following model selection, see Section 4.4.4.

4.2.3 Dry Runs for Experiments Embedded within Surveys

The dry run method for cluster randomized trials is also compatible for cluster randomized trials embedded within complex survey designs involving a single stage of cluster sampling where the same clusters serve as both sampling units and as units of assignment in the experiment. In this scenario, it suffices to take precisely the sampling weights that govern the original survey and use them in tandem with the inverse probability of assignment weights during estimation of model performance. Under this simplest formulation, each observation within a given cluster may possess the same weight although adjustments for factors like nonresponse will typically lead to different sampling weights even within the same cluster.

Nonetheless, dry runs do not require that the same clusters serve as both the sampling units and the units of assignment in the experiment. For example, schools could be sampled as clusters but have intact classrooms within the schools randomly assigned to the treatment or control. In this scenario, the inverse probability of assignment weights could differ for classrooms within a school despite the entire school potentially possessing the same sample inclusion weights. However, the sample inclusion weights and the inverse probability of assignment weights can still be used in tandem during the estimation of model performance.

More complex designs with multiple stages of cluster sampling should be compatible with dry runs as well. To illustrate, let us consider mode tests within the National Assessment of Educational Progress (NAEP). Often referred to as “the nation’s report card”, NAEP has provided insight into how American students are performing in mathematics and reading since 1969. These mode tests examine aspects of switching the delivery of the test from paper exams to computer exams. For example, a mode test might estimate how consistent the scores are from year to year when transitioning from paper to computer exams. Sampling for NAEP involves three stages of selection (Rust and Johnson, 1992). 94 primary sampling units (PSUs) are sampled out of roughly 1,000 total PSUs. Within each PSU, metropolitan statistical areas (MSAs), non-MSA counties, or contiguous non-MSA counties are sampled. Finally, schools are sampled within each MSA. This three-stage design leads to varying sample inclusion weights so analysis of the embedded randomized trial must account for those unequal weights in order to attain unbiased effect estimates.

As with one stage of cluster sampling, the same clusters need not serve as both the sampling units and the units of assignment. For a mode test within NAEP, different treatments (e.g. paper tests versus computer tests) may be applied to different classrooms within the same school. This may in turn lead to different probabilities of assignment for observations within the same school. That being said, it should once again suffice to take the sampling weights from the original survey that already account for different probabilities of selection at different stages of cluster sampling and use them in tandem with different inverse probability of assignment weights during the estimation stage of dry runs.

Note that dry runs are fully compatible with complex survey designs when estimating bias and mean-squared-error but may require adjustments when estimating power. Prior to estimating power, standard variance estimation approaches would need to be updated to account for complex sampling features. This will ensure standard errors are valid during hypothesis testing, allowing for proper power estimates.

4.3 Illustration of Pseudo-Cluster Construction

In this section, we begin by illustrating how pseudo-clusters are generated for cluster randomized trials on a small, contrived dataset. This should detail the process and intuition behind the method in an easy to understand manner. We then illustrate pseudo-cluster generation on one matched set from a large scale, cluster randomized trial assessing a reading intervention designed to assist early elementary students.

4.3.1 Example of construction on a small, contrived dataset

We first demonstrate the steps outlined in Section 4.2.2 on a small, contrived dataset to illustrate how to divide blocks, i.e. matched sets, into pseudo-clusters. Let Block A consist of two classrooms matched together, one of which received the treatment whereas the other received the control. Table 4.2 illustrates how the pseudo-cluster generation process occurs when students are divided based on race and sex.

We fit a logistic regression to these 22 students to find the propensity of a student belonging to the treatment classroom. In Table 4.2, this value is denoted $\mathbb{P}(A1)$. We then use those propensity scores to assign the 10 control students to pseudo-classrooms $A1$ and $A2$ by sampling without replacement, weighted by that propensity score. This assignment is denoted by "Classroom" in Table 4.2. We also sample with replacement, weighted by the propensity score. These classroom demographics are found in Table 4.3.

It is unlikely that this method will perfectly recreate the demographic breakdowns from the toy example in any given iteration due to variability in the data. For example, white students were oversampled in classroom $A1$ relative to the initial treatment classroom in the above example. Nonetheless, we attain rates of white and female students comparable to the true rates across multiple realizations of this process.

Classroom	**# Students**	**% Female**	**% White**	
Treatment	12	50%	67%	
Control	10	50%	60%	

Student	**Classroom**	**Sex**	**Race**	$\mathbb{P}(A1)$
1	A2	Female	White	0.57
2	A1	Male	White	0.57
3	A1	Female	White	0.57
4	A1	Male	White	0.57
5	A1	Female	White	0.57
6	A2	Male	White	0.57
7	A2	Female	Non-White	0.50
8	A1	Male	Non-White	0.50
9	A1	Female	Non-White	0.50
10	A2	Male	Non-White	0.50

Table 4.2: Breakdown of demographic characteristics of a pair of classrooms along with how individual control students may be divided into pseudo-classrooms A1 & A2

Pseudo-Classroom	**# Students**	**% Female**	**% White**
A1	12	50%	75%
A2	10	50%	60%

Table 4.3: One realization of the pseudo-cluster generation process.

Under this initial formulation when sampling with replacement, classroom sizes in the pseudo-school remain the same as in the actual school. As a result, our classrooms in the toy example are always of size 12 and 10. However, the researcher has the option to incorporate variation in cluster sizes. For example, allowing 20% variation from the true classroom size allows different realizations of classroom sizes; classroom $A1$ could contain anywhere from 8 to 12 students and classroom $A2$ could contain 9 to 14 students. In expectation, we would still observe classroom sizes of 10 and 12 students, respectively.

4.3.2 Example of construction on BURST data

We now extend the previous contrived example to real data. In particular, we allow propensity scores calculated on more covariates, incorporate more than one matched set, and allow matched sets of triplets rather than solely pairs of schools.

4.3.2.1 Motivating Example

The data used in the subsequent sections come from a research project conducted jointly with Amplify, Inc., a digital education company that provides assessments and analytics for data-driven instruction along with digital curriculum. In particular, the aim of the research project was to evaluate the efficacy of BURST[R]: Reading. BURST is a personalized reading program designed to improve literacy among early elementary students who are at risk of falling behind their peers, henceforth referred to as Tier 2 students. BURST uses a proprietary algorithm to assign these Tier 2 students to small groups where they receive additional instruction based on their needs as identified through *Dynamic Indicators of Basic Early Literacy Skills* (DIBELS) test scores—a widely used early-literacy assessment provided by Amplify, Inc.

4.3.2.2 Description of BURST Data Set

We now briefly describe the data used in the following sections. We have observations on nearly 27,000 unique students over 52 schools and 4 years for a total of over 52,000 observations (henceforth called records). Each student participated in the study for somewhere between one and four years depending on their grade and the year they entered the study. While we encounter some missing data, we have demographic information (Race/Gender/Date of Birth/Free Lunch Status/etc.) on the majority of students. Furthermore, our dataset contains both DIBELS scores and end-of-year assessment scores for the vast majority of participants. Along with those scores, we have the date each student took their end-of-year assessment and their date of birth so we can calculate their age on the test date. These end-of-year assessments are our primary outcome of interest.

The 52 schools were divided into 24 pairs of schools, 1 triplet of schools, and one singleton based on common, school-wide characteristics such as race, enrollment size, socio-economic status, and pre-intervention reading and math proficiencies. When available, these rates were averaged over the three pre-intervention years. Matching was performed using the R package `nbpMatching` for optimal nonbipartite matching to group schools prior to random assignment. Nearly every match occurred within a school district. One school in each pair was randomly assigned to the treatment. Within the triplet, two schools randomly received the treatment. The singleton originally belonged to a pair of schools before the school assigned to the control attrited.

4.3.2.3 Burst Pseudo-School Creation

Within each matched set of schools (either pairs or triplets), we fit a propensity score model to determine the probability that a given student within that block belonged to the treatment school. We possess multiple observations on many students, but only allow each student to contribute once to our propensity score model. The propensity score model incorporates baseline characteristics such as gender, race, and socioeconomic status, as well as their DIBELS pre-test score. For more guidance on determining which baseline covariates should be incorporated into the propensity score model, see Section 4.4.1.1. Once each student receives a propensity score, all students in treatment schools are removed and the control students are sampled with replacement into pseudo-schools $A1$ and $A2$. Note that the student's complete set of records will be placed into the same pseudo-school. For example, if we observe a control student in BURST across three years, those three records are sampled as a group into either pseudo-school $A1$ or pseudo-school $A2$.

This process is repeated for each matched set, with appropriate attention paid to the triplet such that we have two pseudo-treatment schools rather than just one. Note that within each matched set, our dry run simulation scheme strictly samples students from the control. As a consequence, the singleton is entirely discarded because it lacks a corresponding control school. Table 4.4 provides a demographic comparison for the treatment and control school in Block A across the covariates incorporated into our propensity score model.

Table 4.5 presents the output of one realization of pseudo-school division in Block A. As

School	# Students	Female	White	Pre-Test	Free Lunch	ELL
Treatment	1420	46%	40%	85	68%	6%
Control	1094	48%	33%	106	69%	6%

Table 4.4: Demographics of schools in Block A.

School	# Students	Female	White	Pre-Test	Free Lunch	ELL
A1	1141	46%	38%	93	69%	5%
A2	1094	48%	31%	119	69%	6%

Table 4.5: One realization of the pseudo-school generation process for Block A.

expected, the contrived schools and the observed schools largely possess similar demographic profiles. Nonetheless, contrived school means shift slightly towards the observed control school means. This is a consequence of solely sampling from the control school. To illustrate with a trivial example, if the control school were 100% white and the treatment school were 0% white, both pseudo-schools would be 100% white. Nonetheless, we do not expect large gaps in practice, as high-quality randomized trials are well-balanced in terms of covariates. The gap between the pseudo-schools also mirrors the gap between the true treatment and control schools for the majority of covariates, with pre-test score the sole exception.

4.4 Applications through Simulations and Permutations

In this section, we apply the framework outlined in Section 4.2 on the BURST reading intervention in order to demonstrate how dry run sandbox simulations can assist with model and covariate selection. We illustrate how to select a model that provides both high power and precision determined through root mean-squared-error by utilizing this dry run simulation scheme. In turn, this allows us to answer two primary questions of interest:

- Do meaningful benefits to precision accrue when estimating student achievement by adjusting for student demographics over and above pre-tests in cluster randomized trials of lower-grade education interventions?

- In the context of BURST, can Peters-Belson type adjustment strategies (Peters, 1941; Belson, 1956; Cochran, 1969) —i.e. techniques that apply covariate adjustment to the control group rather than to the treatment and control simultaneously—offer precision comparable to ordinary least squares with covariate adjustment and hierarchical linear models with covariate adjustment?

We begin by demonstrating the particulars of the pseudo-school division process. We then proceed to use the dry run simulation method to answer the aforementioned questions.

4.4.1 Pseudo-School Division

The first step when conducting a dry run analysis is to divide control observations into pseudo-schools within each block. We provided a detailed walk-through of the process in Section 4.2.2. Researchers have flexibility at three different points:

1. What covariates should be included in the propensity score model?

2. Should we sample students with or without replacement?

3. Should we apply subsampling and if so, what size should we select for our subsampling blocks?

4.4.1.1 Considerations for the Propensity Score Model

We begin with the first question: which covariates should we incorporate into our propensity score model? What Works Clearinghouse (WWC) (Clearinghouse, 2020) serves as a natural resource for research design questions when analyzing education experiments and quasi-experiments. However, they provide little guidance for covariate adjustment besides stressing the importance of pretest scores. Generally, variables that may explain the outcome or selection into the treatment or control groups are of interest (Fan and Nowell, 2011). In education research, these typically include student demographics, school characteristics, income, and parental education level (Barth et al., 2008; Nguyen et al., 2006), although the particular covariates available will generally be situation-specific (Domingue and Briggs, 2009). One school of thought suggests that all of these available covariates should be included (Powell et al., 2020). An alternative method of selecting covariates to incorporate into each propensity score model is to solely include covariates that are not well-balanced between the treatment and control. These are the covariates that best differentiate the two schools in the matched set and thus, are important additions to a propensity score model.

For BURST, we choose to select all demographic covariates in our possession that are not time-variant. This includes pre-test score (as recommended from WWC), gender, race, socio-economic status, English Language Learner status, and an indicator denoting a learning disability. Some of these covariates, e.g. pre-test scores or English Language Learning status, likely help to explain the outcome whereas others like race may differentiate treatment and control schools. We do not incorporate school-specific covariates, although we implicitly account for school size through our sampling methods.

We additionally omit time-dependent variables for two reasons. First, each school should possess roughly similar distributions of students by grade. Thus, controlling for grade in our propensity score model is unnecessary. Second, we fit our propensity score model on students rather than student-years, which ensures each student will have one propensity of belonging to the treatment school rather than multiple depending on their grade. This facilitates sampling the complete set of student records into the same pseudo-school rather than having the same student appear in different schools depending on their grade. Student age is not incorporated into our model as a consequence. Note that in the rare case where schools in the same matched set come from distinct states, age may need to be incorporated. Different states use different age cutoffs for grade eligibility so age would serve as an important confounder.

We then fit a propensity score model by applying a generalized logistic regression with a

Student-t prior distribution on each coefficient using the R package `arm`. This Bayesian logistic regression model largely prevents extreme propensity scores, ensuring that every student will be assigned to each pseudo-school with some frequency.

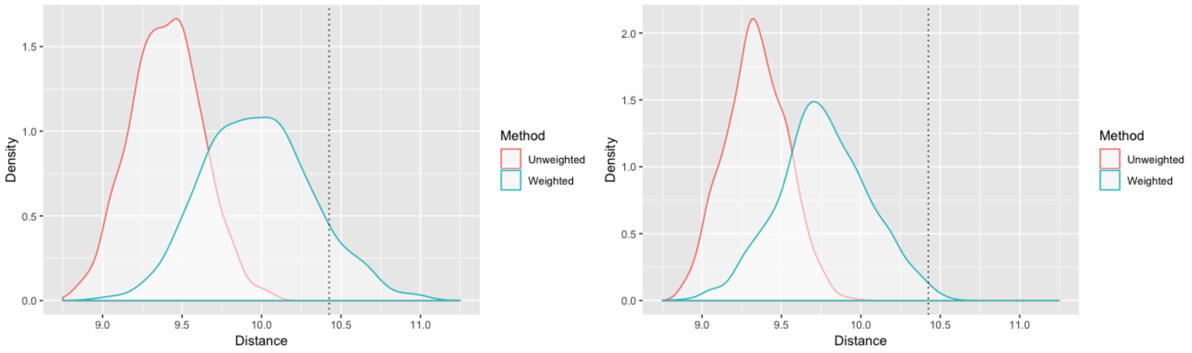
4.4.1.2 Pseudo-School Sampling Method

Once we have constructed propensity scores for each control observation and discarded the treatment observations, we need to determine how to sample students into pseudo-schools. From Section 4.2.2, we know we can perform this step either with or without replacement. When selecting between the options, there are two primary considerations:

- The size of the pseudo-schools, where closer to the true size is preferable.
- The discrepancy in covariate balance of the pseudo-schools, where we would like to recreate the gap in covariate balance found in the actual randomized trial.

The importance of the first consideration is readily apparent. We would like to accurately estimate the power and precision each method provides the researcher. Sample size affects those issues. As a consequence, we want to mimic school size as closely as possible. The relevance of the second issue may not be as salient. Mirroring randomized trials, we know that each block of schools will be perfectly balanced in expectation because we permute the treatment assignment within that block. Nonetheless, any given realization of the randomized trial will likely possess some covariate imbalance, a feature we would like to replicate in our dry run analysis.

Sampling without replacement will roughly halve our sample size; control observations are divided into pseudo-treatment and pseudo-control schools. Sampling with replacement, on the other hand, does not have this drawback. However, sampling with replacement may yield pseudo-schools that are too similar to one another with respect to covariate balance. To study this possibility, we examine the average of the within-block Mahalanobis distance between pseudo-schools—we calculate this within each matched set before taking the mean of each of these distances.



(a) Sampling without replacement.

(b) Sampling with replacement.

Figure 4.2: Average Mahalanobis distances across 1000 iterations of pseudo-school division. For comparison, we include distances for an unweighted version where each student has a 50% probability of assignment to a given school rather than using the probabilities from propensity score models. The true average Mahalanobis distance between Treatment and Control schools is 10.42.

Figure 4.2 presents Mahalanobis distances across 1000 iterations for with and without replacement sampling schemes. We also provide these distances for the unweighted scenario where we discard propensity scores and sample students into schools with 50% probability. From these calculations, we see an immediate benefit to dividing students based on propensity scores rather than through random division. Using propensity scores provides both greater separation on average and a larger range of Mahalanobis distance separation. This holds both with and without replacement.

Figure 4.3 presents Mahalanobis distances from propensity score sampling with and without replacement, and we observe a slightly greater separation in the latter. In certain iterations of pseudo-school division, both methods attain a separation greater than the separation of 10.42 observed in the actual experiment, but this occurs more frequently when sampling without replacement. Nonetheless, the two distributions are largely overlapping. This overlap, in tandem with substantially larger sample sizes, leads us to select sampling with replacement for pseudo-school division in our dry run analysis.

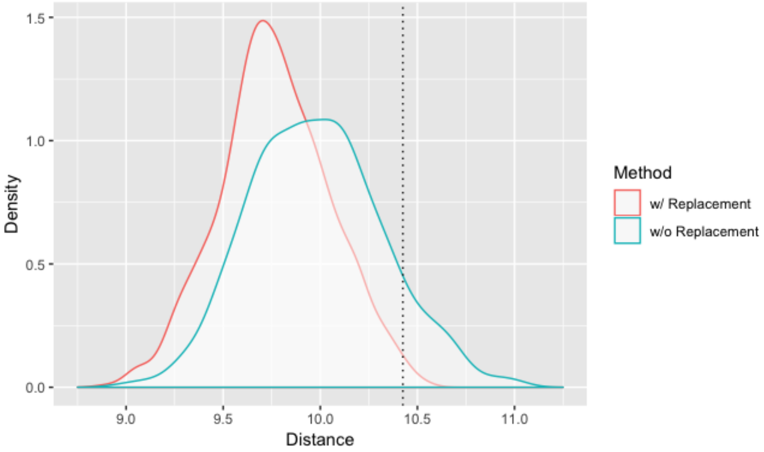


Figure 4.3: Comparison of average Mahalanobis distances across 1000 iterations of pseudo-school division when sampling with and without replacement. The true average Mahalanobis distance between Treatment and Control schools is 10.42.

4.4.1.3 Subsampling Block Size Determination

In this analysis, we choose to apply subsampling as outlined in Section 4.2.2, of which a brief overview may be necessary. Generally, subsampling takes b observations without replacement from some sample n to form a new sample. A statistic, such as an estimate of an average treatment effect, is then calculated on that subsample and the process is repeated for a new subsample. Both subsampling and bootstrapping generate new samples by drawing observations from the initial data in order to estimate the sampling distribution of that statistic. Nonetheless, subsampling is less assumption laden as it is valid under weak assumptions on b whenever the statistic has a limiting distribution. Bootstrapping requires the distribution of the statistic to be locally smooth as a function of the unknown model.

The intuition behind incorporating subsampling into this simulation scheme is rather simple. Subsampling occurs without replacement. Thus, each subsample of size b from a sample of size n is merely a sample of size b from the larger population. The parallel to randomized trials is readily apparent. Each subsample may be viewed as a unique sample from the population; consequently, different subsamples correspond to different realizations of the experiment itself. This serves to address fears as to whether one model's superiority in this simulation scheme occurs solely due to randomness unique to a particular sample.

To implement subsampling, we must first determine the proper block size b. Asymptotic conditions typically require $b \to \infty$ and $b/n \to 0$ as $n \to \infty$ but that leaves a wide range of potential block sizes (Politis et al., 1999). One such possible block size is $n/\log(n)$ but even this formulation leads to a broad set of possible sizes. Applying a base-ten logarithm more than doubles the block size attained from applying a natural logarithm. Rather than selecting arbitrarily within this range, we implement the minimum volatility method outlined in Politis et al. (1999):

1. For $b = b_{small}$ to $b = b_{big}$, compute a subsampling interval at the desired confidence level, resulting in endpoints $I_{b,low}$ and $I_{b,high}$.

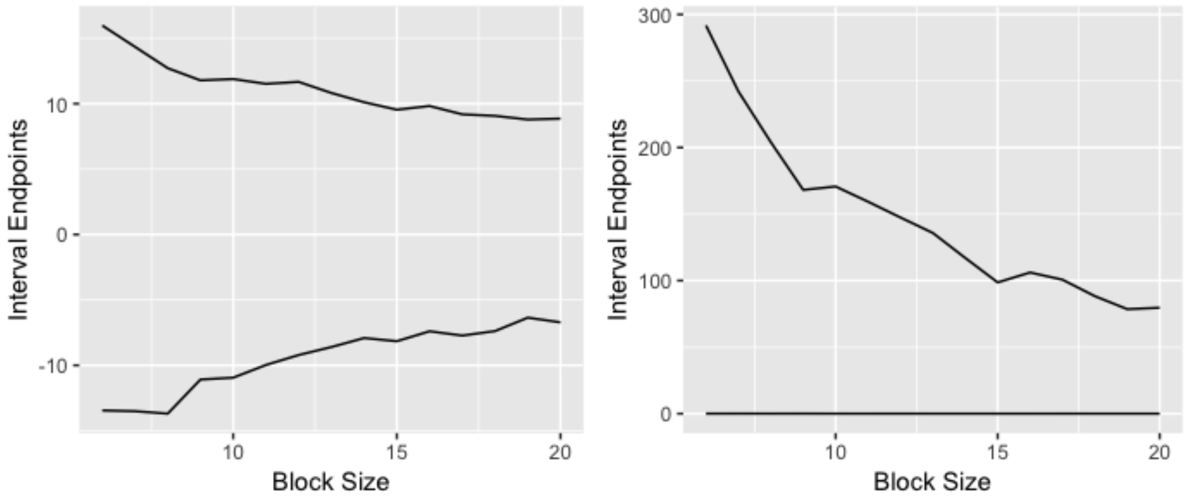
2. For each b, compute volatility index VI_b, i.e. the standard deviation of the endpoints within neighborhood k of b. VI_b denotes the standard deviation of lower bounds $\{I_{b-k,low}, \ldots, I_{b+k,low}\}$ plus the standard deviation of upper bounds $\{I_{b-k,high}, \ldots, I_{b+k,high}\}$. Setting $k = 2$ or $k = 3$ is standard. For our analysis, we use the first.

3. Select block size b with the smallest volatility index.

When implementing this algorithm, we chose to conduct analysis on our simplest model—the difference in Hajek estimators discussed in greater depth in Section 4.4.2. The difference in Hajek estimators will serve as our "null model" in future sections. We examined block sizes from $n/\ln(n)$ to $n/\log_{10}(n)$, i.e. from 8 to 18.

In Figure 4.4, we present interval width across different subsample sizes for our point estimate and for the error of our point estimate. Note that we are not trying to minimize the width of the interval, but rather minimize the variability of the interval from block size to block size. Thus, Figure 4.4 suggests the least variability occurs for block sizes between 15 and 18. Calculating the Volatility Index allows us to select a specific value.

Table 4.6 presents the Volatility Index for all potential subsample sizes. For intervals of both the point estimate and the mean-squared-error, the volatility is smallest for $b = 16$. We intend to



(a) Point Estimates.

(b) Squared Errors.

Figure 4.4: Confidence Interval volatility for different block sizes. We would like to select the block size that provides the least volatility in interval length. In this figure, that will be the block size with the flattest slope.

now perform our simulation method using subsampling with blocks of size 16 in each iteration.

4.4.2 Models for Comparison

With our pseudo-school sampling method fully formulated, we now use dry runs to compare models for analysis of BURST. Ideally, we would like to select a model that provides an accurate estimate of the treatment effect while also possessing maximum power. However, as outlined in Section 4.1, there are many competing theories as to which strategy best achieves this. Some suggest looking at the difference-in-means between treatment and control groups. Others recommend controlling for pre-test results, working under the assumption that this will adequately account for baseline differences between students in the treatment and control groups (Bloom et al., 2007). A different school of thought suggests including additional covariates may further improve precision. We aim to assess that claim in this section.

However, adding covariates further complicates the picture. Which method of covariate adjustment provides the most precision? Standard methods point towards ordinary least squares or

Block Size	Point Est. VI	MSE VI
8	3.19	52.2
9	2.81	34.3
10	2.16	21.2
11	1.50	14.6
12	1.91	20.8
13	1.74	24.2
14	1.55	20.4
15	1.07	15.3
16	0.77	10.5
17	1.07	11.0
18	0.97	12.4

Table 4.6: Subsampling Volatility Indices of different block sizes for point estimates and squared errors.

hierarchical linear models. Yet, recent literature proposes group-specific covariate adjustment instead, i.e. covariate adjustment based on the control group (Peters, 1941; Belson, 1956), or separate covariate adjustment for the control and treatment groups (Lin et al., 2013).

To examine these questions, we compare the performance of the following 8 models:

- Model 0: Difference in treatment and control outcomes as estimated through Horvitz-Thompson estimators (Horvitz and Thompson, 1952).
- Model 1: Difference in treatment and control outcomes as estimated through Hajek estimators (Hajek, 1971).
- Model 2: Linear model examining the outcome on the treatment, controlling for pre-test scores.
- Model 3: Linear model examining the outcome on the treatment, controlling for pre-test scores along with other demographic covariates (age, race, gender, socioeconomic status, etc.).
- Model 4: A mixed effects linear model with random effects at the school level and fixed

effects for the covariates included in Model 3 (Raudenbush and Bryk, 2002).

- Model 5: A Peters-Belson technique incorporating the covariates from Model 3. Briefly, this technique models outcomes for the control and uses that model to predict treatment outcomes. The difference between the fitted and observed values is used to estimate the treatment effect.
- Model 6: A linear model with full treatment by covariate interactions (Lin et al., 2013).
- Model 7: A covariate-adjusted linear model with fixed effects at the block level (Schochet, 2008), a method proportional to weighting clusters by the harmonic mean of the number of treatment students and number of control students in a given matched set (Kalton, 1968).

The first two models are not models per se, but rather difference-in-means estimators and as a result, do not incorporate covariate adjustment. The difference in treatment and control outcomes as estimated through Horvitz-Thompson estimators, i.e. Model 0, should provide an unbiased estimate and thus, serves as a check that dry runs provide valid results. If Model 0 yields a biased point estimate, dry runs may be invalid. We do not expect Model 0 to serve as a true competitor to the other seven models. Model 2 adds to Model 1 by examining the difference in Hajek estimators after controlling for pre-test scores. Model 3 through Model 7 all incorporate substantial covariate adjustment. Models 3 and 4 are standard methods of performing this adjustment whereas Model 5 and Model 6 use Peters-Belson type adjustment strategies. Model 7 adds to Model 3 by including fixed effects at the block level to weight by the size of a given matched set.

We compare these models by generating different realizations of the experiment through pseudo-school creation and treatment assignment permutation. For each of those realizations, we fit the above models and examine their attributes. In the base scenario under which no students receive the treatment, a precise model should not detect a treatment effect. We then artificially impose a treatment effect on students belonging to pseudo-treatment schools; ideally each model will detect the magnitude of that imposed effect. To compare these different models, we begin by examining the following two scenarios:

- No treatment effect added on average, i.e. each treatment student receives an effect generated from a $N(0, 80)$ distribution. This differs from a sharp null of no effect where $Y_{T_i} = Y_{C_i} \; \forall \, i$.
- All students who tested into the intervention to receive supplemental instruction receive a treatment effect generated randomly from a $N(32, 80)$ distribution.

In Section 4.4.2.2 we will compare model performance under two additional scenarios: one where the treatment effect is correlated with student-level characteristics and the other where the treatment effect is correlated with the school itself.

We present the results for the first two scenarios in Table 4.7. Note that bias is generally viewed as $\mathbb{E}(\hat{\mu}) - \hat{\mu}$, i.e. the difference between the parameter itself and an estimate of that parameter. Here, however, we refer to bias as the mean of $\hat{\mu} - \mu$ where $\mu = \sum_{i \in T} y_{T_i} - y_{C_i}$. In other words, we are trying to estimate the attributable effect (Rosenbaum, 2001; Hansen and Bowers, 2009). Bias and RMSE results shift slightly under the standard form for bias but generally results still hold. For a vignette that illustrates how to implement this process, see **here**.

		No Effect		Imposed Effect	
		Bias	**RMSE**	**Bias**	**RMSE**
Unadjusted	**Model 0**	-0.05	33.89	-0.10	34.20
	Model 1	1.10	5.67	0.98	5.63
Pre-Test	**Model 2**	1.49	4.61	1.38	4.56
Adjusted	**Model 3**	1.07	3.49	0.94	3.46
	Model 4	-1.39	5.56	-1.73	5.69
	Model 5	0.83	3.48	0.83	3.47
	Model 6	0.60	3.69	0.50	3.68
	Model 7	0.41	3.20	0.27	3.19

Table 4.7: Model Performance with 1000 simulations under no effect ($\tau \sim N(0, 80)$ added to each Tier 2 treatment student) and under an imposed effect ($\tau \sim N(32, 80)$ added to each Tier 2 treatment student).

A few points of interest are apparent from Table 4.7. First, the Horvitz-Thompson estimator in Model 0 suggests that this simulation scheme provides unbiased estimates. In addition, we observe that our simplest models, those that are fully unadjusted or merely control for a student's pre-test, are generally more biased and less precise than models with covariate adjustment. Among those covariate adjusted models, ordinary least squares performs similarly to the group specific covariate adjustment strategies (i.e. Models 5 and 6). Furthermore, inclusion of fixed effects at the block level provides an additional improvement to both bias and RMSE. The hierarchical linear model is an outlier among these covariate adjusted models with performance similar to that of the entirely unadjusted difference in Hajek estimators. This phenomenon merits closer consideration in the following section.

To assist with model comparison, Table 4.8 provides standard errors for the differences in RMSE between each pair of models. Generally, the standard errors for models in comparison with Model 1 (the unadjusted, difference in Hajek estimators) and Model 4 (the random effects model) are larger than the standard errors for the difference in RMSEs between the covariate adjusted models. This is unsurprising, as Models 1, 2, and 4 have particularly large variation in their performance from iteration to iteration.

	M1	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**
M1	-						
M2	0.14	-					
M3	0.13	0.11	-				
M4	0.14	0.13	0.12	-			
M5	0.13	0.11	0.10	0.12	-		
M6	0.13	0.11	0.10	0.12	0.10	-	
M7	0.12	0.11	0.09	0.12	0.09	0.09	-

Table 4.8: Standard errors for the difference in RMSE between each of the seven models examined through dry runs.

4.4.2.1 Random Effects and the Hierarchical Linear Model

The base hierarchical linear model incorporates complete demographic covariate adjustment and attaches a random effect at the school level. We re-fit this model twice more, once with random effects at the block level, i.e. a random effect for each matched set, and once with random effects at both the school and block level. Table 4.9 presents these results.

	Bias	**RMSE**
School Cluster REs	-1.39	5.56
Block Clusters REs	0.19	3.04
Block/School REs	0.00	5.22

Table 4.9: Model performance with 1000 simulations for a mixed model with random effects at the school level, the block level, or block level random effects nested within school level random effects. We impose an effect of $N(0, 80)$ on students in treatment schools.

Including random effects at the block level shows a marked improvement over random effects at the school level. RMSE drops by roughly 40% and bias improves by an even greater amount. The model with school random effects nested within block random effects performs similarly to the initial model and well-illustrates the "bias-variance tradeoff". This model is entirely free of bias but yields a substantially larger RMSE than the model with block random effects. This demonstrates how introducing some bias can greatly reduce variance.

Why do these results occur? One explanation may simply be that this is the nature of the dry run simulation method. Students in both pseudo-schools in a given pseudo-block were initially drawn from the same actual school; as a result, the hierarchical linear model may detect that the natural grouping occurs at the pseudo-block level. On the other hand, some literature recommends applying random effects for each matched set in observational studies settings. This suggests these results may not be a consequence of the dry run design but rather an inference to be drawn from the method (Smith, 1997). The majority of schools within each matched set in BURST were drawn from the same school district and students frequently transferred schools; thus, some students appear in both treatment and control schools depending on the year. As a result, these dry run simulation results may accurately recommend inclusion of random effects at the block rather than school level.

4.4.2.2 School and Student Specific Effects

We now compare model performance across two different realizations of imposed treatment effects, both of which may better represent the heterogeneous effects that often arise in education settings:

- The effect is school specific, i.e. every student who tests into the intervention in a subset of pseudo-treatment schools receives the effect whereas students in the remaining pseudo-treatment schools receive no effect.

- The effect is student specific, i.e. every pseudo-treatment student who tests into the intervention and receives free or reduced price lunch receives an effect and the rest are unaffected.

Imposition of an intervention rarely occurs exactly as intended. Furthermore, complicated interventions are more likely to be implemented to a varying extent across schools. By only imposing treatment effects on certain schools, we mimic the scenario in which certain schools faithfully implement the intervention and other schools fail to implement it properly or, potentially, at all. This replicates the implementation observed in BURST; many schools assigned to the treatment solely implemented the intervention in one of the four study years and some failed to implement it whatsoever (Rowan et al., 2019). It is also possible that interventions are more or less effective on certain subgroups of students. To examine this scenario, we impose the treatment effect on individuals of a certain subgroup and then determine which method is best at detecting the overall effect of the intervention.

In Table 4.10, the trends observed under the standard imposed effect in Section 4.4.2 largely persist. In comparison with no covariate adjustment, for example, strictly controlling for the pre-test improves precision at the expense of some additional bias under both school and student specific effects. Furthermore, we still realize substantial gains in RMSE and a smaller reduction in bias when applying additional covariate adjustment. Model 4, the hierarchical linear model, is the exception. Finally, the full covariate model with fixed effects at the block level remains the most precise and least biased model. It should be noted that all seven models perform slightly worse when compared to their performance under the standard imposed effect. This is unsurprising as the treatment effect is accruing in a more heterogeneous fashion than before; students must now fulfill an additional criterion on top of testing in to the intervention (i.e. free or reduced price lunch

		School Effect		Student Effect	
		Bias	**RMSE**	**Bias**	**RMSE**
Unadjusted	**Model 1**	0.88	5.66	1.17	5.73
Pre-Test	**Model 2**	1.22	4.61	1.37	4.69
Adjusted	**Model 3**	0.90	3.58	1.12	3.67
	Model 4	-1.93	6.13	-1.51	5.67
	Model 5	0.65	3.57	0.78	3.69
	Model 6	0.64	3.59	0.76	3.54
	Model 7	0.11	3.33	0.48	3.33

Table 4.10: Model Performance with 1000 simulations. Under the school-specific effect, eligible students in 50% of pseudo-treatment schools receive an imposed effect of $N(\mu, \sigma)$. Under the student-specific effect, eligible students in pseudo-treatment schools who receive free or reduced price lunch receive an imposed effect of $N(\mu, \sigma)$.

or belonging to a specific treatment school) in order to receive a benefit.

4.4.3 Covariate Selection using Dry Runs

For ease of presentation, the covariate adjusted models in Section 4.4.2 used the full set of covariates available to us: race, sex, pre-test scores, socio-economic status, normalized grade (a student's expected grade given their age), and indicators for students who are English language learners (ELL) or differently-abled. Nonetheless, this method also allows for selection of specific sets of covariates. We use a forward variable selection procedure to perform this process. We fit Model 3 (covariate-adjusted OLS) on 1000 iterations of the dry run process, controlling for each of the seven covariates one at a time. The covariate from the model with the lowest RMSE is then permanently added to our model and we refit these models with each of the six remaining covariates added one by one. The results are presented in Table 4.11.

Unsurprisingly, pre-test scores provide the largest gains to RMSE. Race and a student's normalized grade are also important covariates that yield substantial gains in precision. Socioeconomic status and the indicator for differently-abled students form the third tier of covariates: they improve precision, but by relatively small quantities. Interestingly, incorporating sex or an indicator for En-

	Number of Covariates						
	1	2	3	4	5	6	7
Pre-Test	4.61						
Race	4.95	3.87					
Grade	4.82	4.47	3.40				
Diff. Abled	5.45	4.60	3.86	3.33			
Low SES	5.47	4.60	3.83	3.34	**3.28**		
Sex	5.59	4.70	3.93	3.45	3.37	3.32	
Eng. Learner	5.11	4.63	3.94	3.59	3.53	3.45	3.50

Table 4.11: RMSEs for different covariate adjustment strategies. For a given number of covariates i and for covariate j, the covariates included are j and the blank entries for the given column. Thus for Sex with 3 covariates, the model incorporates Sex, Pre-Test, and Race.

glish language learners hurts our RMSE. The English language learner indicator harming RMSE may seem surprising at first. English language learners likely perform worse on reading exams than native speakers. Nonetheless, much of that information is soaked up by pre-test scores. Note that when controlling for a single variable, the ELL indicator provided a substantially lower RMSE than socioeconomic status, sex, and an indicator for differently-abled students. After controlling for pre-test, however, its RMSE falls in line with the other less informative covariates. These results suggest that the covariate adjusted models could be further improved by removing two of our covariates: ELL and sex.

4.4.4 Complications Following Selection

Naturally, concerns may arise that inferences drawn from dry run simulations could prove invalid across different iterations of the randomized trial. For example, it is easy to argue that while this particular realization of the experiment found covariate adjustment improved precision, an unadjusted model would perform better under other realizations. Here, “other realizations” speaks to different permutations of the treatment assignment, holding fixed the set of covariates X and the potential outcomes y_c and y_t.

This argument may be extended to suggest that the most precise method from dry runs might not even be the most precise method for the true realization when working with the complete data rather than merely control observations. For example, dry runs may select a method that over-adapts idiosyncracies of the control data onto the full data, leading to overfitting. This in turn has the potential to increase the Type II error rate. We addressed these concerns by subsampling matched sets of schools, allowing each iteration of the dry run analysis to represent a different pseudo-randomized trial. In the case of BURST, each iteration served as a different experiment of 16, rather than 26, matched sets. To assess whether dry runs remain at risk of overfitting despite the subsampling step, we leverage an aspect of BURST data: an estimated treatment effect of zero.

Analysis of the BURST reading intervention failed to uncover a treatment effect of any magnitude (Rowan et al., 2019). We make the assumption that our treatment effect estimate of zero is, in fact, correct. As a consequence, we can permute treatment assignments (while holding fixed X, y_c, and y_t) within each matched set of schools which yields different realizations of the randomized trial. This allows us to conduct the same dry run analysis on each permutation of the complete randomized trial. Table 4.12 presents the average RMSE across 1000 permutations of the treatment assignment along with the proportion of permutations where each model provided the lowest RMSE.

		Avg. RMSE	**Var(RMSE)**	**Best Mod.**
Unadjusted	**Model 1**	5.22	0.10	0%
Pre-Test	**Model 2**	3.69	0.54	0%
Adjusted	**Model 3**	2.61	0.23	11%
	Model 4	4.04	1.74	5%
	Model 5	2.55	0.20	6%
	Model 6	2.59	0.22	0%
	Model 7	2.36	0.17	77%

Table 4.12: Average RMSE across dry runs for 1000 permutations of the original treatment assignment. "Best Mod." refers to the proportion of permutations where the given model possessed the lowest RMSE.

These results are largely consistent with what we observed in the dry run analysis in Section 4.4.2. The entirely unadjusted model performs poorly, as does the model simply controlling for the pre-test. Among the covariate adjusted models, the standard linear model and its group-specific analogs perform comparably, whereas the hierarchical linear model generally provides a much larger RMSE and the model including fixed effects at the block level performs better. In fact, Model 7 possesses the lowest RMSE in over three quarters of permutations suggesting that its superiority in Section 4.4.2 was not dependent on idiosyncrasies of that particular realization of the intervention.

One other observation merits noting. Despite an average RMSE roughly 50% higher than Models 3 and 5, the hierarchical linear model has the lowest RMSE in nearly as many permutations as do those models. We believe this is due to the variance in the performance of Model 4. While the RMSEs of most of the other models are rather stable across permutations, the RMSEs of Model 4 (and to a lesser extent, Model 2) are much more volatile. This substantially greater variance for Model 4 allows for permutations when its performance is superior to the other six models. Conversely, it also provides the largest RMSE in 14% of permutations, a phenomenon that never occurs with the other covariate adjusted models.

Concerns about the validity of inference following model selection may persist despite the consistency of results across permutations of the randomized trial. While we solely use control data to select a method through dry run simulations, we still leverage data from our experiment to select the best method to analyze that experiment. As a consequence, we may have inflated or deflated Type I errors when testing a null hypothesis of no effect, leading to over or under-rejection of the null. To assess this possibility, we use the same permutation strategy elaborated previously (including the assumption that our estimated average treatment effect of 0 is correct), but now check the Type I error rate by determining how frequently we reject a null hypothesis of no effect on the overall permuted data when implementing the best model as selected through dry runs.

In other words, we permute the treatment assignment within each matched set on the true data. On that permuted data, we use dry runs to select a best model as determined through RMSE. We then fit that model on the permuted, true data and determine whether that model rejects or accepts a null hypothesis of no effect.

We use the same 1000 permutations described previously and calculate an overall Type I error rate of 0.049, just beneath our desired threshold of 0.05. With only 1000 permutations, 0.05 lies well within the margin of error so it is clear that we have not substantially increased or decreased our Type I error rate. As a consequence, we believe we are protected against fears of inflated or deflated Type I error rates due to dry run model selection.

4.5 Discussion

In this chapter, we have presented a novel method to assist with model and covariate selection in randomized trials. Ideally, researchers would like to select an analysis strategy that provides a precise estimate of the treatment effect while avoiding post-hoc justification for their choices. We believe the method presented in this chapter provides the tool by which researchers can achieve this goal. Our dry run simulation scheme, similar in motivation to uniformity trials, reconstructs the design of the study by sampling observations solely from the control group. Each reconstruction presents an opportunity to test the power and precision of various models after imposing assorted treatments (including no treatment effect) onto the generated data. This allows researchers to make informed decisions about which model will provide the best opportunity to detect a treatment effect.

Our method provides an alternative to other model and covariate selection strategies typical in the analysis stage of field experiments and can assist at both the larger model level (e.g. OLS versus Peters-Belson versus mixed models) and at the covariate level (e.g. different levels of covariate adjustment). This method is compatible with standard randomized trials as well as cluster randomized trials and trials embedded within complex survey designs. We demonstrated this process on a large, IES-funded cluster randomized trial for an early elementary reading intervention. Due to the nature of that randomized trial—we found the intervention provided no treatment effect, neither positive nor negative—we were able to test the validity of this method by comparing performance under different permutations of the treatment assignment. While there is some minor deviation, results largely hold and retain proper Type I error rates, protecting against claims of post-selection inference.

Thus, the dry run method serves as a simple, easy to implement tool assisting with model and covariate selection for field trialists. Researchers can construct many different pseudo-randomized trials, forming a "sandbox" on which a variety of model specifications may be compared in terms of their precision and power to detect an effect.

CHAPTER 5

Conclusion

In this dissertation, we have proposed two novel methods which aim to extract information unique to a randomized controlled trial or quasi-experiment and translate it into statistical power and precision. The first, PWRD aggregation, converts the theory of change behind an intervention into a test statistic that maximizes its relative efficiency over standard methods which in turn, provides greater power. We showed the advantages of PWRD aggregation when the theory of change is valid both through a simulation study and through a study examining the effects of a reading intervention on early elementary students. We then provide a method of constructing confidence intervals that leverages the power advantages behind PWRD aggregation while providing intervals and point estimates that adhere to standard, widely used methods. The second method we propose in this dissertation is the dry run simulation scheme. This procedure, using real rather than synthetic data, creates a pseudo-experiment mimicking the initial randomized trial that preserves blinding to impact estimates. These pseudo-experiments then form a sandbox on which various models may be compared to discover the model specification that best estimates the treatment effect.

5.1 Future Work

Both methods may be implemented in their current state. Nonetheless, there are promising avenues of research that extend both of these methods in different fashions. PWRD aggregation, for example, may be extended safely to experiments embedded within surveys so long as the sample design is simple enough (e.g. one stage of cluster sampling) to be summarized by sample inclusion weights. The extension would simply incorporate the inclusion probabilities into the model that estimates the effects for each of the various subgroups. Nonetheless, how to modify PWRD aggregation to admit expansion to experiments embedded in complex surveys with multiple stages of cluster sampling remains an open question. This is particularly the case with variance estimation, as standard errors need to additionally account for complex sampling features. In addition, this dissertation shows that PWRD aggregation is fully compatible with classical inference and we believe it is compatible with randomization inference as well. However, we have yet to formally extend PWRD aggregation to randomization inference. Another extension for PWRD aggregation is to develop software allowing applied researchers to implement the method in analysis of their randomized trials.

Dry runs have some similar and some alternative lines of future research. Like PWRD aggregation, this method too requires development of software for easy implementation of the dry run scheme. Another different line of future research would involve strengthening the connection of dry runs serving as a uniformity trial within a randomized trial. Unlike standard uniformity trials, within dry runs we only possess some of the potential outcomes under the control rather than all of them. However, the dry runs embedded uniformity trial should remain informative about the true uniformity trial, as the experimental control is simply a subsample of the study population. Yet, dry runs exist in a finite population setting whereas subsampling literature assumes each observation is independent and identically distributed. Thus, the underlying assumption fails. Future work may generalize subsampling to finite populations, including generalizing results from Bardenet and Maillard (2015), to better fit the theoretical justification for dry runs.

APPENDIX A

PWRD Aggregation and Type I Errors

In Section 2.2.2, we demonstrated how PWRD aggregation maximizes the test slope and thus, the corresponding power for the family of hypotheses $K_\eta : \Delta = \eta \mathbf{p}_0$. That is, when the treatment effect is proportional to the dosage received, PWRD aggregation maximizes power. Here, we remove that assumption and all assumptions about the form of the treatment effect. We do require joint limiting Normality of $\hat{\Delta}$ and a consistent estimator of its covariance.

Condition A.0.1 *The estimator* $\widehat{\mathrm{Cov}}(\hat{\Delta})$ *is consistent for* $\mathrm{Cov}(\hat{\Delta})$*, in the sense that* $\|n\widehat{\mathrm{Cov}}(\hat{\Delta}) - \Sigma\|_2 \to_P 0$*, where* Σ *is as in Condition 2.2.3.*

Condition A.0.2 $\sqrt{n}(\hat{\Delta} - \Delta) \to_d N\big(\mathbf{0}, \mathrm{Cov}(\Delta)\big)$.

With Conditions 2.2.3, A.0.1 and A.0.2, we formulate a simple proposition about the distribution of the test statistic specified in Equation 2.6.

Proposition A.0.3 *Take fixed aggregation weights* w*. Under the null hypothesis* H_0 *and when Conditions 2.2.3, A.0.1, and A.0.2 hold,*

$$\frac{\sum_g w_g \hat{\Delta}_g - \sum_g w_g \delta_{0g}}{(w' \mathrm{Cov}(\hat{\Delta}) w)^{1/2}} \to_d N(0, 1).$$

Proposition A.0.3 states that with a consistent estimator of the covariance and an estimator that is asymptotically multivariate normal, the test statistic specified in Equation 2.6 with fixed aggregation weights w will converge to a standard multivariate normal distribution. For finite sample sizes n, this test statistic should approximately follow a t-distribution with $n - k$ degrees of freedom, where k represents the number of estimated parameters. Note that the denominator, $\widehat{V}^{1/2}$, present in Equation 2.6 and Section 2.2.2 at large denotes the quadratic form of estimated covariances of $\hat{\Delta}$. PWRD aggregation requires statisticians provide a covariance estimator with consistency guarantees, i.e. Condition A.0.1.

While Proposition A.0.3 allows us to determine the asymptotic distribution of test statistics with the form in Equation 2.6 for fixed aggregation weights w, PWRD aggregation does not incorporate fixed weights. Rather, two components of PWRD aggregation, $\hat{\mathbf{p}}_0$ and $\hat{\Sigma}$, are random variables. Consequently, the aggregated statistic $\sum_g \hat{\omega}_g \hat{\Delta}_g$ includes an auxiliary statistic: $\hat{\omega}_g$. Addressing additional variation of this type generally requires analysis through stacked estimating equations, a technique not readily compatible with the best-in-class clustered standard error estimation of Pustejovsky and Tipton (2016). Thus, our standard error scales the covariance between each $\hat{\Delta}_g$ by aggregation weights $\hat{\omega}$, yet does not incorporate the covariance between each $\hat{\omega}_g$. To address this issue, we first present a mild condition on $\hat{\mathbf{p}}_0$.

Condition A.0.4 $\hat{\mathbf{p}}_0$ *is root-n consistent, i.e.* $\|\hat{\mathbf{p}}_0 - \mathbf{p}_0\|_2 = O_P(n^{-1/2})$.

As applied to the BURST study, Condition A.0.4 is immediate from the Weak Law of Large Numbers. Conditions 2.2.3, A.0.1, and A.0.4 allow us to circumvent our standard error not incorporating additional variation from $\hat{\omega}$ through Proposition A.0.5.

Proposition A.0.5 *Consider t-statistics of the form*

$$\frac{(\sum_g \hat{\omega}_g \hat{\Delta}_g - \sum_g \hat{\omega}_g \delta_{0g})}{(\hat{\omega}' \widehat{\mathrm{Cov}}(\hat{\Delta}) \hat{\omega})^{1/2}}, \tag{A.1}$$

where $\hat{\omega} = (\widehat{\mathrm{Cov}}[\hat{\Delta}]^{-1}\hat{\mathbf{p}}_0)_+ \Big/ \sum_j (\widehat{\mathrm{Cov}}[\hat{\Delta}]^{-1}\hat{\mathbf{p}}_0)_{+j} \in [0,1]$ *represents weights for PWRD aggregation. Under Conditions 2.2.3, A.0.1, and A.0.4, the difference between* (A.1) *and*

$$\frac{(\sum_g \omega_g \hat{\Delta}_g - \sum_g \omega_g \delta_{0g})}{(\omega' \mathrm{Cov}(\hat{\Delta}) \omega)^{1/2}},$$

where $\omega = (\Sigma^{-1}\mathbf{p}_0)_+ \Big/ \sum_j (\Sigma^{-1}\mathbf{p}_0)_{+_j}$, *is asymptotically negligible:*

$$\left[\frac{\sum_g \hat{\omega}_g \hat{\Delta}_g - \sum_g \hat{\omega}_g \delta_{0g}}{(\hat{\omega}' \widehat{\text{Cov}}(\hat{\Delta})\hat{\omega})^{1/2}} - \frac{\sum_g \omega_g \hat{\Delta}_g - \sum_g \omega_g \delta_{0g}}{(\omega' \text{Cov}(\hat{\Delta})\omega)^{1/2}}\right] \to_P 0. \tag{A.2}$$

Simply, Proposition A.0.5 states that the t-statistic centered around $\sum_g \hat{\omega}_g \delta_{0g}$, where $\hat{\omega} = (\hat{\Sigma}^{-1}\hat{\mathbf{p}}_0)_+ \Big/ \sum_j (\hat{\Sigma}^{-1}\hat{\mathbf{p}}_0)_{+_j}$, and scaled by a consistently estimated standard error will converge in probability to the "proto" t-statistic appearing in Prop. A.0.3 and covered by Prop. 2.2.4, which is centered around the parameter $\sum_g \omega_g \delta_{0g}$ and scaled by the sampling s.d. of $\sum_g \omega_g \hat{\Delta}_g$. As a consequence, hypothesis tests incorporating PWRD aggregation will maintain proper Type I error rates. Therefore, PWRD aggregation provides valid hypothesis tests even when the theory of change does not hold. The proof of Proposition A.0.5 can be found in Appendix B.2.

APPENDIX B

PWRD Aggregation Proofs

B.1 Proof of Proposition 2.2.4

Notation	**Description**
$\hat{\Delta}_g$	Estimated effect in cohort year-of-follow-up g
$\hat{\Delta}$	Vector of estimated effects of dimension $(1 \times G)$
Δ_g	$\mathbb{E}\hat{\Delta}_g$, i.e. true effect in cohort year-of-follow-up g
ω_g	Weight attached to cohort year-of-follow-up g
ω	Vector of weights of dimension $(1 \times G)$
$\sum_g \omega_g \hat{\Delta}_g$	Pooled estimated effect
Σ_Δ	Covariance of effects across cohort years g
p_{0g}	Probability control student in cohort year g received supp. instruction
$\mathbf{p}_0$	Vector of dimension $(1 \times G)$ of p_{0g}

Table B.1: Notation for Appendix B.1.

Consider the parameter $\Delta_{agg} = \mathbb{E}(\sum_g \omega_g \hat{\Delta}_g) = \omega' \Delta$ where Δ_g, and thus Δ_{agg}, follow a proportionality assumption, i.e. $\Delta_g \propto \eta p_{0g}$. The variance of $\omega' \Delta$ satisfies $\mathrm{Var}(\sum_g \omega_g \hat{\Delta}_g) = \omega' \Sigma_\Delta \omega$, where Σ_Δ denotes the covariance of effects across cohort-years g, and is assumed fixed at a common value across hypotheses K_η, $-\infty < \eta < \infty$.

Now examine the test statistic $\sum_g \omega_g \hat{\Delta}_g$, the argument for the other forms being similar. Our problem is to select $\omega = (\omega_1, \ldots, \omega_G) \geq 0$ that maximizes the test slope of $\sum_g \omega_g \hat{\Delta}_g$ which in turn will maximize the relative Pitman efficiency for PWRD aggregation versus alternative methods of aggregation given the theory of change is true. Following the definition of test slope provided in

(Van der Vaart, 2000, p.201):

$$h(\omega) = \frac{\Delta'_{agg}(0)}{\mathrm{Cov}_0^{1/2}(\omega'\hat{\Delta})} = \frac{\Delta'_{agg}(0)}{\left[\omega'\Sigma_\Delta\omega\right]^{1/2}}, \tag{B.1}$$

where $\Delta'_{agg}(0)$ denotes the derivative at zero of a function of the form $d \mapsto \Delta(d)$. The corresponding relative Pitman efficiency for different ω may be represented by $\big(h(\omega_1)/h(\omega_2)\big)^2$. The form of the two test statistics is identical; they merely incorporate different aggregation weights ω. Thus, it follows that finding ω_{opt}, where ω_{opt} maximizes the test slope, will also maximize the relative Pitman efficiency $\big(h(\omega_{opt})/h(\omega_{alt})\big)^2$. Under flat weighting, $\omega_{alt_g} := n_g/N$, where n_g denotes the number of observations in cohort-year g and N denotes the total number of observations.

B.1.1 Determining the Optimum ω_{opt}

We would like to determine which ω maximizes the test slope in (B.1). Under the assumption that $\Delta_g \propto \eta p_{0g}$, then $\Delta'_g(0) \propto p_{0g}$ as well. Thus, to determine which ω maximizes the test slope in (B.1), we maximize the following:

$$\max_\omega \frac{\omega'\mathbf{p}_0}{\mathrm{Var}^{1/2}(\omega'\hat{\Delta})}. \tag{B.2}$$

We first transform B.2 logarithmically which is equivalent to maximizing the following:

$$f(\omega) = \log(\omega'\mathbf{p}_0) - \frac{1}{2}\log(\mathrm{Var}(\omega'\hat{\Delta})). \tag{B.3}$$

To maximize, we take the gradient of $f(\omega)$ and set the gradient equal to the zero-vector, $\mathbf{0}$:

$$\nabla f(\omega) : \frac{\mathbf{p}_0'}{\omega'\mathbf{p}_0} - \frac{\omega'\Sigma_\Delta}{\omega'\Sigma_\Delta\omega} = \mathbf{0}.$$

Note that both $\omega'\mathbf{p}_0$ and $\omega'\Sigma_\Delta\omega$ are scalars, so we can rewrite this as follows:

$$(\omega'\mathbf{p}_0)^{-1}\mathbf{p}_0' - (\omega'\Sigma_\Delta\omega)^{-1}\omega'\Sigma_\Delta = \mathbf{0}.$$

We now rearrange the terms to solve for ω_{opt}:

$$\omega_{opt} = \left(\frac{\omega'\Sigma_\Delta\omega}{\omega'\mathbf{p}_0}\right)\mathbf{p}_0'\Sigma_\Delta^{-1}.$$

B.1.2 Estimation of ω_{opt}

From Slutsky's Theorem, we can then estimate ω_{opt} as follows:

$$\hat{\omega}_{opt} = \left(\frac{\omega'\Sigma_\Delta\omega}{\omega'\hat{\mathbf{p}}_0}\right)\hat{\mathbf{p}}_0'\Sigma_\Delta^{-1}. \tag{B.4}$$

If we allow $\alpha = \left(\frac{\omega'\Sigma_\Delta\omega}{\omega'\hat{\mathbf{p}}_0}\right)$, we can then rewrite this as $\hat{\omega}_{opt} = \alpha \cdot \hat{\mathbf{p}}_0'\Sigma_\Delta^{-1}$. To check this simplifies, plug $\alpha \cdot \hat{\mathbf{p}}_0'\Sigma_\Delta^{-1}$ back into ω in B.4. We have thus uniquely specified $\hat{\omega}_{opt}$. Furthermore, in principle we can define $\hat{\omega}_{opt}$ only up to a constant of proportionality such that $\hat{\omega}_{opt} = \hat{\mathbf{p}}_0'\Sigma_\Delta^{-1}$. Since Σ_Δ^{-1} is symmetric, we can rewrite this as $\hat{\omega}_{opt} = \Sigma_\Delta^{-1}\hat{\mathbf{p}}_0$.

B.1.3 ω_{opt} with a Non-Negativity Constraint

In Equation B.3, we wished to maximize $f(\omega) = \log(\omega'\mathbf{p}_0) - \frac{1}{2}\log(\mathrm{Var}(\omega'\hat{\Delta}))$. We now add in two constraints to prevent $\omega_g < 0$. In particular, we would now like to find $\max_\omega \log(\omega'\mathbf{p}_0) - \frac{1}{2}\log(\mathrm{Var}(\omega'\hat{\Delta}))$ such that $\omega_g \geq 0 \; \forall \; g$ and $\mathbf{1}'\omega = 1$. In other words, we would like to maximize ω such that each ω_g is non-negative and $\sum_{g=1}^{G}\omega_g = 1$.

This is equivalent to:

$$\max_\omega \log(\omega'\mathbf{p}_0) - \frac{1}{2}\log(\mathrm{Var}(\omega'\hat{\Delta})) - u'\omega + v'\omega.$$

We begin by looking at the KKT conditions (Karush, 1939; Kuhn and Tucker, 2014):

- **Stationarity**

$$(\omega'\mathbf{p}_0)^{-1}\mathbf{p}_0' - (\omega'\Sigma_\Delta\omega)^{-1}\omega'\Sigma_\Delta - u' + v' = \mathbf{0}.$$

Note: Both $(\omega'\mathbf{p}_0)^{-1}$ and $(\omega'\Sigma_\Delta\omega)^{-1}$ are scalar random variables, so for ease we redefine them as c_1 and c_2 respectively, i.e. $c_1\mathbf{p}_0' - c_2\omega'\Sigma_\Delta - u' + v' = \mathbf{0}$.

- **Complementary Slackness**

$$u'\omega = 0.$$

- **Primal Feasibility**

$$\omega \geq 0, \mathbf{1}'\omega = 1$$

- **Dual Feasibility**

$$u \geq 0$$

To solve this, we begin by eliminating u, giving us

$$v' - u' = c_2\omega'\Sigma_\Delta - c_1\mathbf{p}_0' \Rightarrow v' \geq c_2\omega'\Sigma_\Delta - c_1\mathbf{p}_0',$$

from stationarity, and

$$(c_1\mathbf{p}_0' - c_2\omega'\Sigma_\Delta + v')\omega = 0,$$

from complementary slackness. After rearranging, we see that

$$\mathbf{0} \leq \omega' \leq \frac{v' + c_1\mathbf{p}_0'}{c_2}\Sigma_\Delta^{-1}.$$

From this, we then argue that ω_{opt} is maximized by the following:

$$\omega_g = \begin{cases} \left(\frac{v'+c_1\mathbf{p}_0'}{c_2}\Sigma_\Delta^{-1}\right)_g & \text{if } v_g \geq -c_1 p_{0g} \\ 0 & \text{if } v_g < -c_1 p_{0g} \end{cases}.$$

In other words, $\omega_{opt}' = (\frac{v'+c_1\mathbf{p}_0'}{c_2}\Sigma_\Delta^{-1})_+$ where $\mathbf{1}'\omega = 1$. We can then estimate ω_{opt} following the same argument as in Appendix B.1.2.

B.2 Proof of Proposition A.0.5

To show Proposition A.0.5, we begin by showing that $\|\hat{\omega} - \omega\|_2 \to_P 0$. Writing $\hat{\Sigma} := n\widehat{\text{Cov}}(\hat{\Delta})$, Condition A.0.1 says $\|\hat{\Sigma} - \Sigma\|_2 = o_P(1)$. Because Σ is positive-definite (Condition 2.2.3), it is invertible and $\|\hat{\Sigma}^{-1}\| \to_P \|\Sigma^{-1}\|$. Applying sub-multiplicativity of the spectral norm to the algebraic identity $\hat{\Sigma}^{-1} - \Sigma^{-1} = \hat{\Sigma}^{-1}(\hat{\Sigma} - \Sigma)\Sigma^{-1}$,

$$\begin{aligned}\|\hat{\Sigma}^{-1} - \Sigma^{-1}\|_2 \leq & \|\hat{\Sigma}^{-1}\|_2 \|\hat{\Sigma} - \Sigma\|_2 \|\Sigma^{-1}\|_2 \\ & = O_P(1) o_P(1) O(1) = o_P(1).\end{aligned}$$

Combining this with $\|\hat{\mathbf{p}}_0 - \mathbf{p}_0\|_2 = O_P(n^{-1/2})$ by Condition A.0.4, $\|\hat{\Sigma}^{-1}\hat{\mathbf{p}}_0 - \Sigma^{-1}\hat{\mathbf{p}}_0\|_2 = o_P(1)O_P(1) = o_P(1)$. Separately $\|\Sigma^{-1}\hat{\mathbf{p}}_0 - \Sigma^{-1}\mathbf{p}_0\|_2 = O_P(1)O_P(n^{-1/2}) = O_P(n^{-1/2})$. Thus, $\|\hat{\Sigma}^{-1}\hat{\mathbf{p}}_0 - \Sigma^{-1}\mathbf{p}_0\|_2 = o_P(1)$. Now $\hat{\omega} = [\sum_j (\hat{\Sigma}^{-1}\hat{\mathbf{p}}_0)_{+j}]^{-1}(\hat{\Sigma}^{-1}\hat{\mathbf{p}}_0)_+$, and similarly $\omega = [\sum_j (\Sigma^{-1}\mathbf{p}_0)_{+j}]^{-1}\Sigma^{-1}\mathbf{p}_0$; through an application of the Continuous Mapping Theorem, $\|\hat{\Sigma}^{-1}\hat{\mathbf{p}}_0 - \Sigma^{-1}\mathbf{p}_0\|_2 = o_P(1)$ entails that the normalizing constant in the definition of $\hat{\omega}$ converges to the one in that of ω. As a result, $\|\hat{\omega} - \omega\|_2 \to_P 0$.

We adopt a similar argument for the denominator. $\|\hat{\omega}'\hat{\Sigma}\hat{\omega} - \hat{\omega}'\Sigma\hat{\omega}\|_2 = O_P(1)o_P(1)O_P(1) = o_P(1)$ and $\|\hat{\omega}'\Sigma\hat{\omega} - \omega'\Sigma\omega\|_2 = o_P(1)O_P(1)o_P(1) = o_P(1)$. Thus, $|\hat{\omega}'\hat{\Sigma}\hat{\omega} - \omega'\Sigma\omega| \to_P 0$, i.e. in (B.5) below the left denominator converges to the denominator at the right, a positive constant:

$$\frac{\sqrt{n}(\sum_g \hat{\omega}_g \hat{\Delta}_g - \sum_g \hat{\omega}_g \delta_{0g})}{(\hat{\omega}' n\widehat{\text{Cov}}(\hat{\Delta})\hat{\omega})^{1/2}} - \frac{\sqrt{n}(\sum_g \omega_g \hat{\Delta}_g - \sum_g \omega_g \delta_{0g})}{(\omega' n\text{Cov}(\hat{\Delta})\omega)^{1/2}}. \tag{B.5}$$

Noting that (B.5) is equivalent to the left-hand side of (A.2) in the statement of the Proposition, we just need to show that $\sqrt{n}\big[\sum_g \hat{\omega}_g \hat{\Delta}_g - \sum_g \hat{\omega}_g \delta_{0g} - \sum_g \omega_g \hat{\Delta}_g + \sum_g \omega_g \delta_{0g}\big] \to_P 0$, which is equivalent to showing $\sqrt{n}\big[\sum_g (\hat{\omega}_g - \omega_g)(\hat{\Delta}_g - \delta_{0g})\big] \to_P 0$. We have already demonstrated $\|\hat{\omega} - \omega\|_2 = o_P(1)$ and under the null distribution, $\|\hat{\Delta} - \delta_0\|_2 = O_P(n^{-1/2})$ through another application of the Weak Law of Large Numbers. Thus, $n^{1/2}[(\hat{\omega} - \omega)(\hat{\Delta} - \delta_0)] = o_P(1)$.

APPENDIX C

PWRD Aggregation in Smaller Samples

In this appendix, we demonstrate PWRD aggregation versus alternative methods using the same simulation study present in Section 2.3, yet on three pairs of schools rather than 26 pairs of schools. This illustrates how PWRD aggregation performs in much smaller sample sizes.

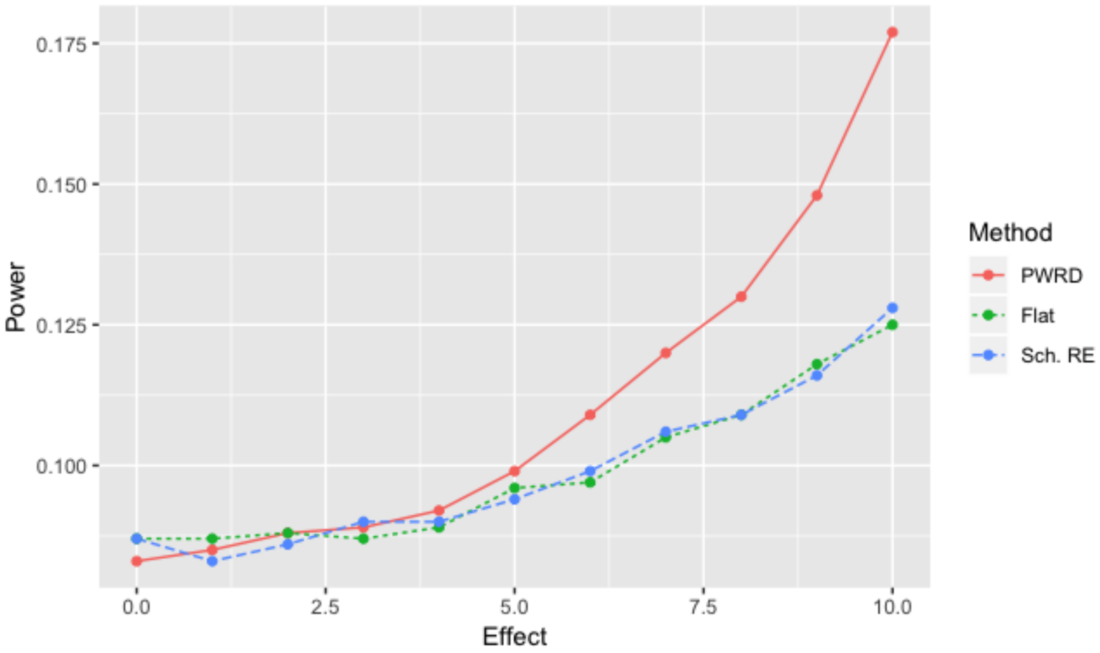


Figure C.1: Power for the three methods under Effect 1, i.e. across increasing effect sizes.

For Effect 1 in Figure C.1, PWRD aggregation provides a clear benefit over the other two methods common to education outcome analysis, although the results for all three methods are much noisier than with the full 26 pairs of schools. PWRD aggregation provides an even larger benefit in terms of power compared to flat weighting and the school random effects model under Effect 2 (presented in Figure C.2). Finally, under Effect 3 in Figure C.3, PWRD aggregation performs comparably to the other two methods, showing that even when the theory of change does not hold, implementing PWRD aggregation will not have large, adverse consequences in terms of power.

A good portion of the noise in these power curves is likely due to the form of the imposed effect. Under Effect 1, only students who test into the intervention receive a benefit. With only three matched sets, different iterations of this simulation study likely have greatly varying numbers of students who stood to benefit. This issue is exacerbated under Effect 2 due to the presence of a negative effect on those who do not receive the intervention but not an issue under Effect 3 where all students in the treatment receive a benefit on average.

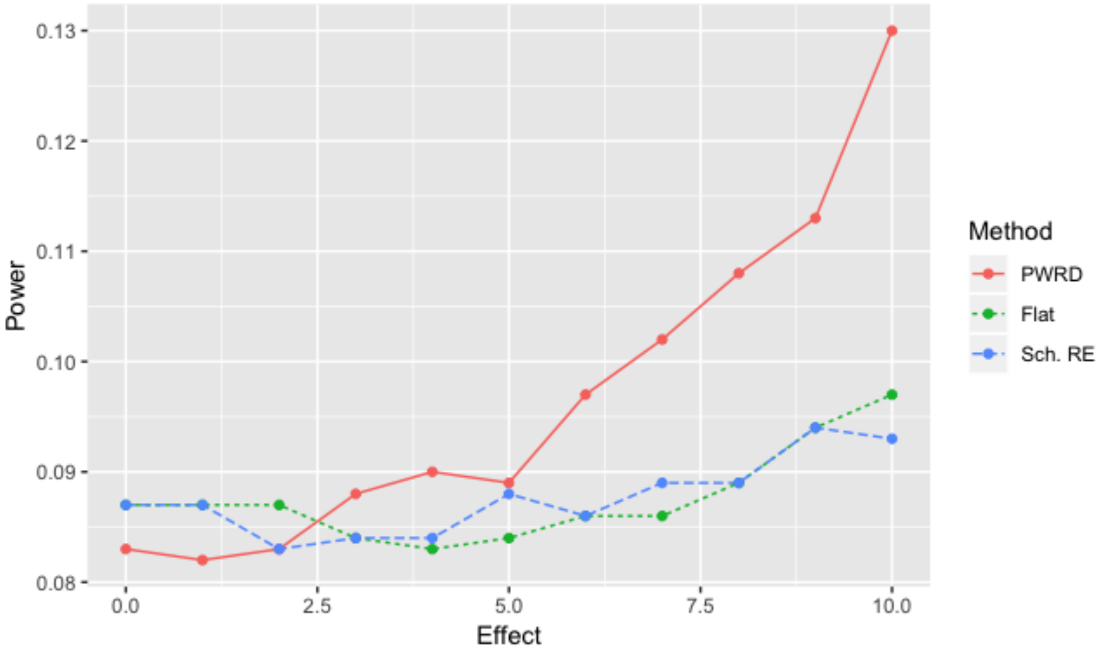


Figure C.2: Power for the three methods under Effect 2, i.e. across increasing effect sizes when Condition 2.1 does not hold.

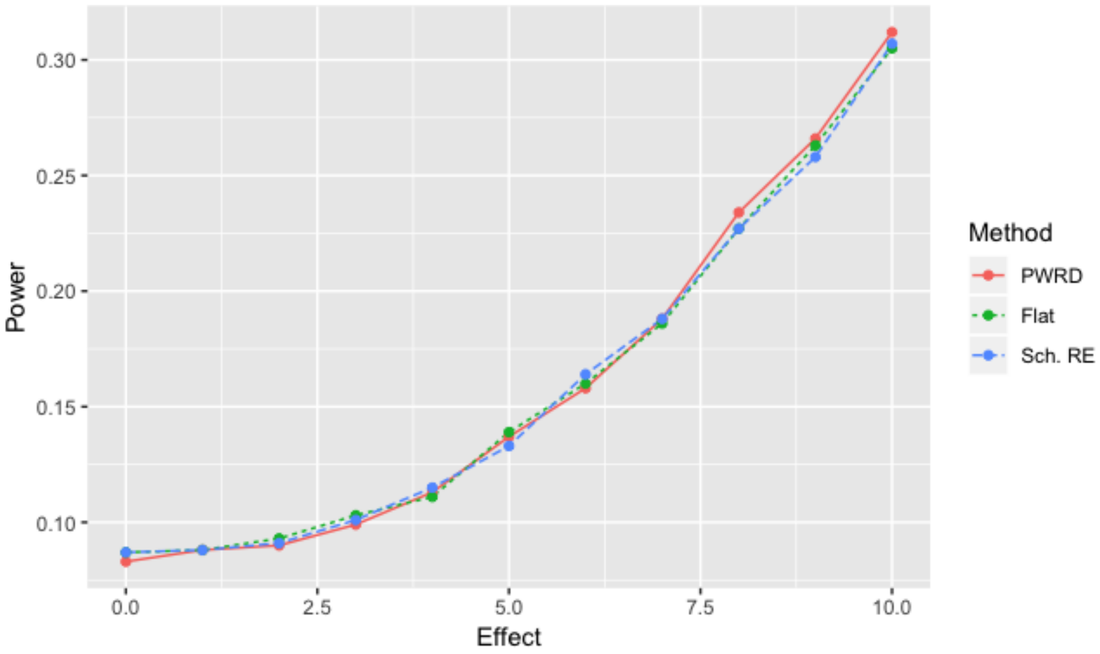


Figure C.3: Power for the three methods under Effect 3, i.e. across increasing effect sizes when none of the conditions hold.

APPENDIX D

PWRD Aggregation and Massachusetts Healthcare Reform

We draw from a 2006 health care initiative in Massachusetts to motivate the aggregation method presented in this paper. This reform, on which the Affordable Care Act was modeled, employed a three-pronged approach to providing universal health care coverage to Massachusetts residents: expansion of Medicaid, subsidized private health insurance, and an individual mandate. Though there is little doubt that access to healthcare increased due to this reform, it is less clear what benefits, if any, the legislation had on mortality (Sommers et al., 2014; Kaestner, 2016; Sommers et al., 2017).

Counties stood to gain from this coverage expansion to varying degrees, just as people did. To illustrate, we examine two Massachusetts counties, whose demographic information is presented in Table D.1.

County	**Poverty**	**Uninsured**	**Median Income**
Middlesex County, MA	7.2%	14.5%	$75494
Suffolk County, MA	17.1%	18.1%	$48683

Table D.1: A comparison of average rates of poverty and uninsurance, and average median income for 2001-2006.

Medicaid expansion and subsidized private health insurance primarily target low-income individuals. Middlesex County, home to Harvard University and MIT, is one of the wealthiest counties in the United States. Prior to health care reform, merely 7.2% of its residents were living in poverty and fewer than 15% were uninsured. In contrast, in Suffolk County, which includes most of Boston proper, uninsurance and poverty rates were much higher (17.1% and 18.1%, respectively). Consequently, if this Massachusetts healthcare reform decreases mortality, we would expect such a benefit to accrue more in Suffolk than Middlesex.

Based on this supposition—that counties with greater proportions of low-income residents will realize larger mortality benefits—we group each of Massachusetts' 14 counties into one of four brackets delineated by their 2006 poverty rates. We calculate separate effect estimates for each of these brackets. All else equal, we expect the effect to increase across brackets, with counties in the high-poverty bracket experiencing the largest mortality benefit from Massachusetts healthcare reform.

D.1 Analysis of Mortality

We follow an analytic procedure adapted from Sommers et al. (2014), who used negative binomial regression to model healthcare-amenable mortality as a linear function of the treatment and prior demographics. Our modified version is similar, but we incorporate four treatment variables rather than one, corresponding to the interaction of the Sommers et al. treatment variable with 2001-2006 poverty categories as shown in Figure D.1. This provides the effect estimates presented in Table D.2.

Poverty	**Coef.**	**S.E.**
Low	-0.001	0.023
Low/Moderate	-0.030	0.019
Moderate/High	-0.035	0.021
High	-0.035	0.019

Table D.2: Estimated change in health care-amenable mortality due to Massachusetts Health Care Reform, by 2001-2006 county level poverty rate.

The magnitude of the mortality benefit increases across the four brackets, while the standard

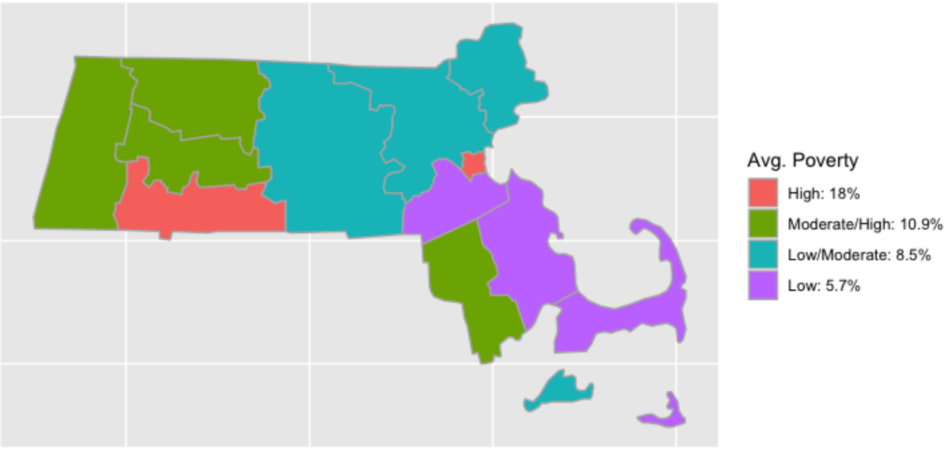


Figure D.1: Massachusetts counties by poverty bracket.

errors generally decrease. The following question then arises: what single aggregate of these coefficients is best suited to demonstrate a program benefit, if the reform in fact reduced mortality?

D.2 Massachusetts Healthcare Results

We previously presented four treatment effect estimates, one for each of the brackets based on pre-intervention poverty levels. To aggregate the four estimates into an overall estimate of the effect of the healthcare reform on mortality, we apply a variation of PWRD aggregation described in Section 2.2.2 and find a benefit to employing this scheme. Results are presented in Table D.3. All methods incorporate small sample corrections for the p-values (Pustejovsky and Tipton, 2016).

While Sommers et al. (2014) find significance at the 5% level, Kaestner (2016) failed to discover a benefit when they replicated the analysis but calculated p-values through a permutation inference technique. We found the significant result in Sommers et al. (2014) to be sensitive to analysis decisions on their part. After replicating their analysis and findings, we made the minor change of incorporating clustered standard errors with clusters at the county rather than the state level.

Method	**Test Slope**	**t value**	**Sig.**
Replication	1.67	-1.88	-
Mixed Effects	1.79	-2.05	-
PWRD	2.44	-2.45	*

Table D.3: Massachusetts Health Care Reform Results: the test slope and t-statistic for our replicated adjustment of Sommers et al. (2014), a mixed-effects adjusted version of Sommers et al. (2014), and our PWRD aggregation method.

Under this specification, significance disappears even without applying permutation techniques. The same holds for a mixed-models variant of Sommers et al. (2014), using state random effects rather than state fixed effects; this is also reported in Table D.3. Although not reported in the table, adapting this mixed effects specification to fit four poverty-bracket fixed effects that are then combined using PWRD aggregation also provides significance at the 5% level.

APPENDIX E

Proof of Proposition 3.2.2

Suppose the bounded null holds, i.e. $\Delta \leq \phi$ where Δ denotes the true average treatment effect and ϕ denotes the threshold of our equivalence region. Additionally, let us take some test statistic $t(Z, Y)$, a function of outcome vector Y and the random assignment Z. Let us additionally assume that $t(\cdot, \cdot)$ is effect increasing (see Condition 3.2.1).

We then need to show that the p-value under the bounded null, $p_{\leq\phi}$, is at least as large as the p-value under the sharp null, p_{ϕ}. This is equivalent to showing the test-statistic under the sharp null, $t_{\phi}(Z, Y)$, is at least as large as the test statistic under the bounded null, $t_{\leq\phi}(Z, Y)$. We begin by decomposing our test statistic under the bounded null.

$$\begin{aligned}\mathbb{P}(t_{\leq\phi}(Z,Y) \geq c|\Delta \leq \phi) = \mathbb{P}(t_{\leq\phi}(Z,Y) \geq c|\Delta = \phi) \cdot \mathbb{P}(\Delta = \phi) +& \\ \mathbb{P}(t_{\leq\phi}(Z,Y) \geq c|\Delta < \phi) \cdot \mathbb{P}(\Delta < \phi).&\end{aligned} \tag{E.1}$$

When $\Delta = \phi$, then $\mathbb{P}(t_{\leq\phi}(Z,Y) \geq c|\Delta \leq \phi) = \mathbb{P}(t_{\leq\phi}(Z,Y) \geq c|\Delta = \phi) = \mathbb{P}(t_{\phi}(Z,Y) \geq c)$. Thus, equality holds between the bounded null and the sharp null.

Now let us look at scenarios when $\Delta < \phi$. Now, we define $\tau = \phi - \Delta$; τ represents the difference between the true average treatment effect Δ and the hypothesized average treatment effect under the sharp null, ϕ. Let us additionally define the imputed treatment and control potential outcomes under the sharp null H_{ϕ} as follows, where $Y(1)$ and $Y(0)$ denote the observed outcomes for the treatment and control respectively:

$$\begin{aligned} Y_\phi(1) &= Z \cdot Y(1) + (1 - Z) \cdot (Y(0) + \phi) \\ Y_\phi(0) &= Z \cdot (Y(1) - \phi) + (1 - Z) \cdot Y(0). \end{aligned} \tag{E.2}$$

When $\Delta < \phi$, we can write the difference between the realized and potential outcomes as follows:

$$\begin{aligned} Y_\phi(1) - Y(1) &= Z \cdot Y(1) + (1 - Z) \cdot (Y(0) + \phi) - Y(1) \\ &= (Z - 1) \cdot Y(1) + (1 - Z) \cdot (Y(0) + \phi) \\ &= (1 - Z) \cdot (\phi - \Delta) > 0, \\ Y_\phi(0) - Y(0) &= Z \cdot (Y(1) - \phi) + (1 - Z) \cdot Y(0) - Y(0) \\ &= Z \cdot (Y(1) - Y(0) - \phi) \\ &= Z \cdot (\Delta - \phi) < 0. \end{aligned} \tag{E.3}$$

In other words, the true treatment observations are smaller than the imputed treatment observations and the true control observations are greater than the imputed control observations. Yet, if we add τ to each outcome, equality would hold. Thus, $t_{\leq\phi}(Z, Y + Z \cdot \tau) = t_\phi(Z, Y)$. Furthering this argument, $t_{\leq\phi}(Z, Y) < t_\phi(Z, Y)$ when $\Delta < \phi$ because the test statistic $t(Z, Y)$ is effect increasing. Therefore, $\mathbb{P}(t_{\leq\phi}(Z, Y) \geq c | \Delta < \phi) < \mathbb{P}(t_{\leq\phi}(Z, Y) \geq c | \Delta = \phi)$.

Thus, $\mathbb{P}(t(Z, Y) \geq c | \Delta \leq \phi) \leq \mathbb{P}(t(Z, Y) \geq c | \Delta = \phi)$ or in other words, the probability that our test statistic will be large enough to reject the bounded null, given $\Delta \leq \phi$, is no greater than the probability that our test statistic will be large enough to reject the sharp null. It trivially follows that using the p-value for the sharp null will provide valid tests when testing the bounded null.

APPENDIX F

Derivation of Optimal Threshold

We would like to determine the maximum bounds for the equivalence region at which we can reject a null hypothesis of no effect using PWRD aggregation with comparable power as we would possess under a standard mode of analysis. For the purposes of this derivation, we will be using t-statistics and their corresponding power.

Take the family of hypotheses described in Section 3.2 where $K_\eta : \Delta = \eta \mathbf{p}_0$. A standard mode of analysis would provide the following t-statistic:

$$t_{std} = \frac{\omega'_{std}(\hat{\eta}\mathbf{p}_0)}{\widehat{V}_{std}^{1/2}}, \tag{F.1}$$

with $\omega_{std_g} \coloneqq n_g/N$, $\omega_{std} \coloneqq (\omega_{std_g} : g)$, where n_g denotes the number of observations in cohort-year g and N denotes the total number of observations. We let $\widehat{V}_{std}^{1/2}$ denote the standard error using aggregation weights ω_{std}.

We would then like to calculate the minimum detectable effect size (MDES) for $1-\beta$ power with $\alpha = 0.05$ when applying the standard method:

$$Power \approx P\left(t > t^*_{(df,1-\alpha/2)} - \frac{\hat{\eta}(\omega'_{std}\mathbf{p}_0)}{\widehat{V}_{std}^{1/2}}\right),$$

and thus,

$$MDES_{std,1-\beta} \approx (t^*_{(df,1-\alpha/2)} + t^*_{(df,1-\beta)})\frac{\widehat{V}_{std}^{1/2}}{\omega'_{std}\mathbf{p}_0}.$$

Under PWRD aggregation, we write these quantities as follows:

$$t_{PWRD} = \frac{\omega'_{PWRD}(\hat{\eta}\mathbf{p}_0)}{\widehat{V}^{1/2}_{PWRD}}, \tag{F.2}$$

and

$$MDES_{PWRD,1-\beta} \approx (t^*_{(df,1-\alpha/2)} + t^*_{(df,1-\beta)})\frac{\widehat{V}^{1/2}_{PWRD}}{\omega'_{PWRD}\mathbf{p}_0}.$$

If the assumptions behind PWRD aggregation hold, then PWRD aggregation should be Pitman efficient and thus provide greater power for the same sample size. As a consequence, $MDES_{P,1-\beta}$ should be smaller than $MDES_{std,1-\beta}$ and we can shift the threshold for equivalence away from zero while, through PWRD aggregation, still obtaining comparable power to the standard analysis.

Note that under the assumption that $\Delta_g \propto \eta p_{0g}$, then $\Delta'_g(0) \propto p_{0g}$ as well. We can then estimate the test slope of PWRD aggregation and the standard analysis as $(\omega'_{PWRD}\mathbf{p}_0)/\widehat{V}^{1/2}_{PWRD}$ and $(\omega'_{std}\mathbf{p}_0)/\widehat{V}^{1/2}_{std}$ respectively. Thus, we can write the difference in minimum detectable effect sizes, or the maximum threshold for PWRD aggregation ϕ_{PWRD}, as

$$\phi_{PWRD} = \left(\frac{\widehat{V}^{1/2}_{std}(\omega'_{PWRD}\mathbf{p}_0)}{\widehat{V}^{1/2}_{PWRD}(\omega'_{std}\mathbf{p}_0)} - 1\right)MDES_{std,1-\beta}.$$

This simplifies to:

$$\phi_{PWRD} = (\sqrt{PE} - 1)MDES_{std,1-\beta}.$$

APPENDIX G

Considerations for Effect Size Based Thresholds

In this appendix, we provide power comparisons for thresholds selected using the standard deviation of both gain scores student performance. WWC recommends using student standard deviations rather than gain score standard deviations and notes that "the standard deviation of gain scores is typically smaller than the standard deviation of unadjusted posttest scores" (Clearinghouse, 2020, p.58). Nonetheless, the larger standard deviations resulting from using student-level standard deviations may result in thresholds that are too large for our purposes.

Figure G.1 presents power under Effect 1 from the simulation study presented in Section 2.3 for the standard method, along with PWRD aggregation using thresholds determined through the effect size. If we let σ_g and σ_s denote the standard deviations of the gain scores and student-level test scores respectively, then the vertical blue lines demarcate thresholds $\phi = 0.05\sigma_g$ and $\phi = 0.2\sigma_g$ and the vertical red lines demarcate thresholds $\phi = 0.05\sigma_s$ and $\phi = 0.2\sigma_s$.

Unsurprisingly, thresholds using gain scores rather than student-level scores provide greater power, as do thresholds using $\phi = 0.05\sigma$ compared to those using $\phi = 0.2\sigma$. Using $\phi = 0.05\sigma_s$, however, does seem to be a reasonable proposition. This threshold actually provides greater power than does $\phi = 0.2\sigma_g$. The real outlier among the four potential thresholds is $\phi = 0.2\sigma_s$. This threshold provides prohibitively less power than the other three and fails to provide comparable power to the standard method for any plausible effect size.

The trends present in Figure G.1 persist in Figure G.2 with Effect 2. Under this effect, three of the thresholds provide reasonable power with $\phi = 0.2\sigma_s$ the outlier once again. This final threshold

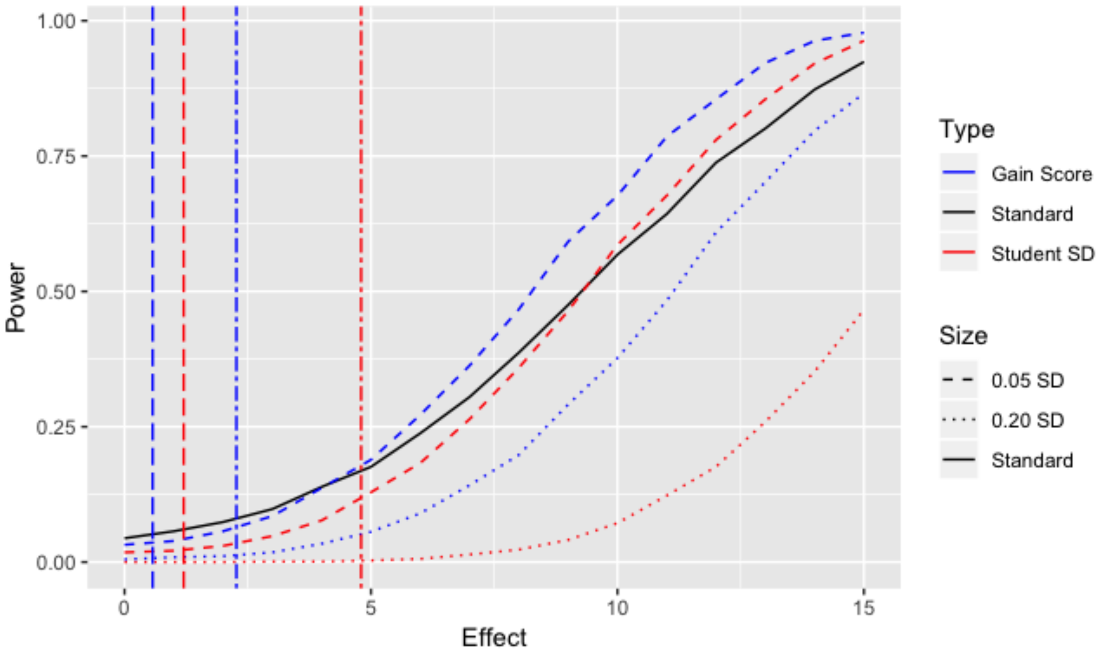


Figure G.1: Power under various thresholds for Effect 1. The vertical blue lines denote the threshold boundaries for $\phi = 0.05\sigma$ and $\phi = 0.2\sigma$ and the vertical red lines denote the boundaries for $\phi = 0.05\sigma$ and $\phi = 0.2\sigma$ determined through gain score standard deviations and student-level standard deviations respectively.

fails to provide any power to detect an effect until an effect size of over 10. While this counts as a "large" effect size, note that the effect size presented in these figures solely speaks to the size of the effect for those who receive the supplemental instruction. The average treatment effect across all students is much lower as many do not receive supplemental instruction or are adversely affected by the intervention.

Additionally note that while the power curve for $\phi = 0.2\sigma_g$ seems comparable to $\phi = 0.05\sigma_g$ and $\phi = 0.05\sigma_g$, it takes a far larger effect to overtake the power provided by the standard method. These two surpass the standard method with an imposed effect of size 7, whereas $\phi = 0.2\sigma_g$ does not surpass the standard method until an imposed effect of size 12.

Under Effect 3, each of these thresholds provides substantially less power than the standard method. The deficit is particularly pronounced for the two intervals requiring large effect sizes, i.e. $\phi = 0.2\sigma$.

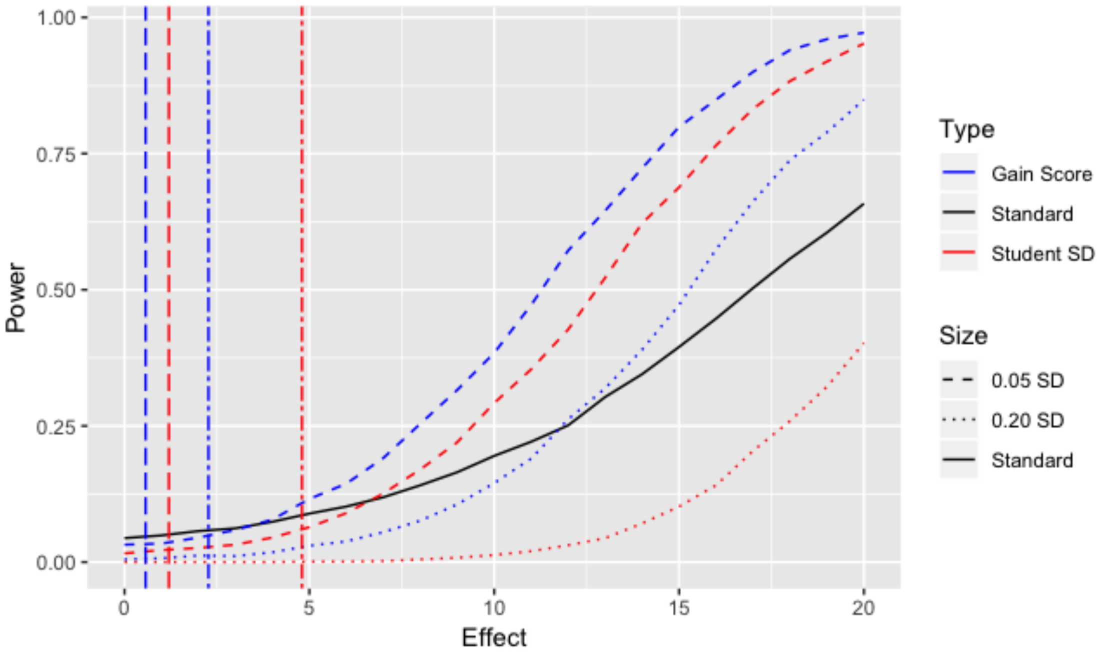


Figure G.2: Power under various thresholds for Effect 2. The vertical blue lines denote the threshold boundaries for $\phi = 0.05\sigma$ and $\phi = 0.2\sigma$ and the vertical red lines denote the boundaries for $\phi = 0.05\sigma$ and $\phi = 0.2\sigma$ determined through gain score standard deviations and student-level standard deviations respectively.

From these simulations it is evident why we chose to select thresholds using the standard deviation of the gain score rather than the standard deviation of student-level scores. While selecting a threshold of $\phi = 0.05\sigma_s$ is reasonable in situations where the theory of change holds, $\phi = 0.2\sigma_s$ is thoroughly unreasonable. While the power provided by that threshold will eventually surpass that of the standard method, this only occurs for implausibly large effect sizes.

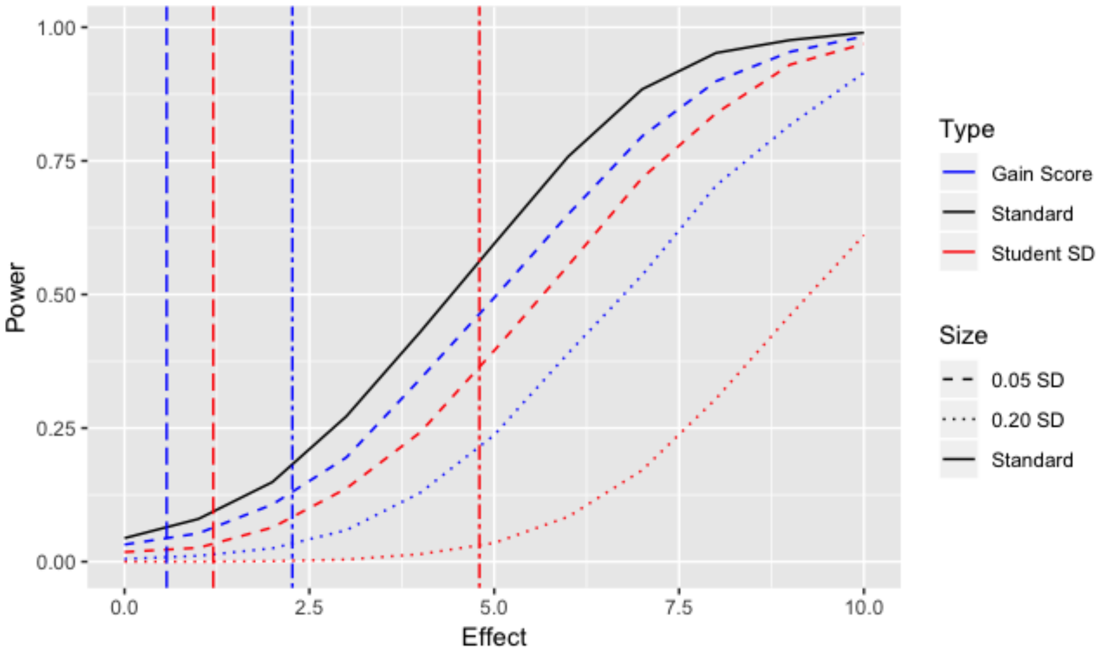


Figure G.3: Power under various thresholds for Effect 3. The vertical blue lines denote the threshold boundaries for $\phi = 0.05\sigma$ and $\phi = 0.2\sigma$ and the vertical red lines denote the boundaries for $\phi = 0.05\sigma$ and $\phi = 0.2\sigma$ determined through gain score standard deviations and student-level standard deviations respectively.

BIBLIOGRAPHY

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.

Baiocchi, M., Cheng, J., and Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in Medicine*, 33(13):2297–2340.

Bardenet, R. and Maillard, O.-A. (2015). Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3):1361–1385.

Barth, R. P., Guo, S., and McCrae, J. S. (2008). Propensity Score Matching Strategies for Evaluating the Success of Child and Family Service Programs. *Research on Social Work Practice*, 18(3):212–222.

Bell, R. M. and McCaffrey, D. F. (2002). Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples. *Survey Methodology*, 28(2):169–182.

Belson, W. A. (1956). A Technique for Studying the Effects of a Television Broadcast. *Applied Statistics*, pages 195–202.

Black, B., Hollingsworth, A., Nunes, L., and Simon, K. (2019). The Effect of Health Insurance on Mortality: Power Analysis and What We Can Learn from the Affordable Care Act Coverage Expansions. Technical report, National Bureau of Economic Research.

Bloniarz, A., Liu, H., Zhang, C.-H., Sekhon, J. S., and Yu, B. (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(27):7383–7390.

Bloom, H. S. (1984). Accounting for No-Shows in Experimental Evaluation Designs. *Evaluation Review*, 8(2):225–246.

Bloom, H. S., Richburg-Hayes, L., and Black, A. R. (2007). Using Covariates to Improve Precision for Studies That Randomize Schools to Evaluate Educational Interventions. *Educational Evaluation and Policy Analysis*, 29(1):30–59.

Borgschulte, M. and Vogler, J. (2020). Did the ACA Medicaid expansion save lives? *Journal of Health Economics*, 72:102333.

Bowers, J., Desmarais, B. A., Frederickson, M., Ichino, N., Lee, H.-W., and Wang, S. (2018). Models, methods and network topology: Experimental design for the study of interference. *Social Networks*, 54:196–208.

Bryk, A. S. and Raudenbush, S. W. (1987). Application of Hierarchical Linear Models to Assessing Change. *Psychological Bulletin*, 101(1):147.

Caughey, D., Dafoe, A., Li, X., and Miratrix, L. (2021). Randomization Inference beyond the Sharp Null: Bounded Null Hypotheses and Quantiles of Individual Treatment Effects. *arXiv preprint arXiv:2101.09195*.

Cheung, A. C. and Slavin, R. E. (2016). How Methodological Features Affect Effect Sizes in Education. *Educational Researcher*, 45(5):283–292.

Clearinghouse, W. W. (2020). What Works Clearinghouse Procedures Handbook Version 4.0. *US Department of Education, Institute of Education Science*.

Cochran, W. G. (1969). The Use of Covariance in Observational Studies. *Applied Statistics*, pages 270–275.

Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences*. Academic Press.

Connolly, P., Keenan, C., and Urbanska, K. (2018). The trials of evidence-based practice in education: a systematic review of randomised controlled trials in education research 1980–2016. *Educational Research*, 60(3):276–291.

Cox, D. (1958). *The Planning of Experiments*. John Wiley.

Croke, K., Hicks, J. H., Hsu, E., Kremer, M., and Miguel, E. (2016). *Does Mass Deworming Affect Child Nutrition? Meta-analysis, Cost-Effectiveness, and Statistical Power*. The World Bank.

Domingue, B. and Briggs, D. C. (2009). Using Linear Regression and Propensity Score Matching to Estimate the Effect of Coaching on the SAT. *Multiple Linear Regression Viewpoints*, 35(1):12–29.

Efron, B. et al. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26.

Ethington, C. A. (1997). A Hierarchical Linear Modeling Approach to Studying College Effects. *Higher Education-New York-Agathon Press Incorporated*, 12:165–194.

Fan, X. and Nowell, D. L. (2011). Using Propensity Score Matching in Educational Research. *Gifted Child Quarterly*, 55(1):74–79.

Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (2008). Generalized estimating equations for longitudinal data analysis. In *Longitudinal Data Analysis*, pages 57–92. Chapman and Hall/CRC.

Fletcher, J. (2010). Spillover effects of inclusion of classmates with emotional problems on test scores in early elementary school. *Journal of Policy Analysis and Management*, 29(1):69–83.

Fox, J. (2015). *Applied Regression Analysis and Generalized Linear Models*. Sage Publications.

Frangakis, C. E. and Rubin, D. B. (2002). Principal Stratification in Causal Inference. *Biometrics*, 58:21–29.

Fryer Jr, R. G. (2017). The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments. In *Handbook of Economic Field Experiments*, volume 2, pages 95–322. Elsevier.

Goeman, J. J., Solari, A., and Stijnen, T. (2010). Three-sided hypothesis testing: Simultaneous testing of superiority, equivalence and inferiority. *Statistics in Medicine*, 29(20):2117–2125.

Gottfried, M. A. (2013). The Spillover Effects of Grade-Retained Classmates: Evidence from Urban Elementary Schools. *American Journal of Education*, 119(3):405–444.

Guo, S. (2005). Analyzing grouped data with hierarchical linear modeling. *Children and Youth Services Review*, 27(6):637–652.

Gupta, S. K. (2011). Intention-to-treat concept: A review. *Perspectives in Clinical Research*, 2(3):109.

Hajek, J. (1971). Contribution to discussion of paper by D. Basu. *Foundations of Statistical Inference. Eds. VP Godambe and DA Sprott*, page 236.

Hannon, S. J., Martin, K., Thomas, L., and Schieck, J. (1993). Investigator Disturbance and Clutch Predation in Willow Ptarmigan: Methods for Evaluating Impact (Métodos para evaluar el impacto del disturbio causado por el investigador en la depredación de camadas de individuos de Lagopus lagopus). *Journal of Field Ornithology*, pages 575–586.

Hansen, B. B. (2004). Full Matching in an Observational Study of Coaching for the SAT. *Journal of the American Statistical Association*, 99(467):609–618.

Hansen, B. B. and Bowers, J. (2009). Attributing Effects to a Cluster-Randomized Get-Out-the-Vote campaign. *Journal of the American Statistical Association*, 104(487):873–885.

Hansen, B. B. and Klopfer, S. O. (2006). Optimal Full Matching and Related Designs via Network Flows. *Journal of Computational and Graphical Statistics*, 15(3):609–627.

Hedges, L. V., Hedberg, E., et al. (2007). Intraclass correlations for planning group randomized experiments in rural education. *Journal of Research in Rural Education*, 22(10):1–15.

Hernán, M. A. and Hernández-Díaz, S. (2012). Beyond the intention-to-treat in comparative effectiveness research. *Clinical Trials*, 9(1):48–55.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945–960.

Horvitz, D. G. and Thompson, D. J. (1952). A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, 47(260):663–685.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 221–233. Berkeley, CA.

Hurvich, C. M. and Tsai, C. (1990). The Impact of Model Selection on Inference in Linear Regression. *The American Statistician*, 44(3):214–217.

Kaestner, R. (2016). Did Massachusetts Health Care Reform Lower Mortality? No According to Randomization Inference. *Statistics and Public Policy*, 3(1):1–6.

Kalton, G. (1968). Standardization: A Technique to Control for Extraneous Variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 17(2):118–136.

Karush, W. (1939). Minima of functions of several variables with inequalities as side constraints. *M. Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago*.

Kraft, M. A. (2020). Interpreting Effect Sizes of Education Interventions. *Educational Researcher*, 49(4):241–253.

Kuhn, H. W. and Tucker, A. W. (2014). Nonlinear Programming. In *Traces and Emergence of Nonlinear Programming*, pages 247–258. Springer.

Laird, N. (2004). Analysis of Longitudinal and Cluster-Correlated Data. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, pages i–155. JSTOR.

Lee, V. E. (2000). Using Hierarchical Linear Modeling to Study Social Contexts: The Case of School Effects. *Educational Psychologist*, 35(2):125–141.

Lin, W. et al. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318.

Lycurgus, T. and Hansen, B. B. (2021). An Aggregation Scheme for Increased Power in Primary Outcome Analysis. *arXiv preprint arXiv:2107.13070*.

Lycurgus, T., Mann, C. Z., and Hansen, B. B. (2021). Protocol: Evaluating the Effect of ACA Medicaid Expansion on 2015-2018 Mortality Through Matching and Weighting. *Observational Studies*, 7.

Mann, C. Z., Hansen, B. B., Gaydosh, L., and Lycurgus, T. (2021). Protocol-Evaluating the Effect of ACA Medicaid Expansion on Mortality During the COVID-19 Pandemic Using County-level Matching. *Observational Studies*, 7(2):S1–S31.

Meece, J. L. and Miller, S. D. (1999). Changes in Elementary School Children's Achievement Goals for Reading and Writing: Results of a Longitudinal and an Intervention Study. *Scientific Studies of Reading*, 3(3):207–229.

Middleton, J. A. and Aronow, P. M. (2015). Unbiased Estimation of the Average Treatment Effect in Cluster-Randomized Experiments. *Statistics, Politics and Policy*, 6(1-2):39–75.

Miller, S., Johnson, N., and Wherry, L. R. (2019). Medicaid and Mortality: New Evidence from Linked Survey and Administrative Data. Technical report, National Bureau of Economic Research.

Montori, V. M. and Guyatt, G. H. (2001). Intention-to-treat principle. *Canadian Medical Association Journal*, 165(10):1339–1341.

NCER (2020). *Education Research Grant Program*. Washington, DC.

Nguyen, A. N., Taylor, J., and Bradley, S. (2006). The Estimated Effect of Catholic Schooling on Educational Outcomes Using Propensity Score Matching. *Bulletin of Economic Research*, 58(4):285–307.

Page, L. C. (2012). Principal Stratification as a Framework for Investigating Mediational Processes in Experimental Settings. *Journal of Research on Educational Effectiveness*, 5(3):215–244.

Peters, C. C. (1941). A Method of Matching Groups for Experiment with No Loss of Population. *The Journal of Educational Research*, 34(8):606–612.

Pitman, E. J. (1948). *Lecture Notes on Nonparametric Statistical Inference: Lectures Given for the University of North Carolina,[Chapel Hill], 1948*. University of North Carolina.

Pocock, S. J., Assmann, S. E., Enos, L. E., and Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine*, 21(19):2917–2930.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer Science & Business Media.

Powell, M. G., Hull, D. M., and Beaujean, A. A. (2020). Propensity Score Matching for Education Data: Worked Examples. *The Journal of Experimental Education*, 88(1):145–164.

Pustejovsky, J. (2017). clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections. R package version 0.2.3. *R Found. Stat. Comput., Vienna*.

Pustejovsky, J. E. and Tipton, E. (2016). Small Sample Methods for Cluster-Robust Variance Estimation and Hypothesis Testing in Fixed Effects Models. *Journal of Business & Economic Statistics*.

Quenouille, M. H. (1956). Notes on Bias in Estimation. *Biometrika*, 43(3/4):353–360.

Quenouille, M. H. et al. (1949). Problems in Plane Sampling. *The Annals of Mathematical Statistics*, 20(3):355–375.

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

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, volume 1. Sage.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512.

Robins, J. M., Blevins, D., Ritter, G., and Wulfsohn, M. (1992). G-Estimation of the Effect of Prophylaxis Therapy for Pneumocystis carinii Pneumonia on the Survival of AIDS Patients. *Epidemiology*, pages 319–336.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal Structural Models and Causal Inference in Epidemiology.

Rosenbaum, P. (2018). *Observation and Experiment*. Harvard University Press.

Rosenbaum, P. R. (1991). A Characterization of Optimal Designs for Observational Studies. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3):597–610.

Rosenbaum, P. R. (2001). Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot. *Biometrika*, 88(1):219–231.

Rosenbaum, P. R. (2002). Randomized Experiments. In *Observational studies*, pages 19–70. Springer.

Rosenbaum, P. R. (2007). Interference Between Units in Randomized Experiments. *Journal of the American Statistical Association*, 102(477):191–200.

Rosenbaum, P. R. et al. (2010). *Design of Observational Studies*, volume 10. Springer.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rowan, B., Hansen, B. B., White, M., Lycurgus, T., and Scott, L. J. (2019). A Summary of the BURST [R]: Reading Efficacy Trial. *Institute for Social Research*.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688.

Rubin, D. B. (1980). Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. *Journal of the American Statistical Association*, 75(371):591–593.

Rust, K. F. and Johnson, E. G. (1992). Chapter 2: Sampling and Weighting in the National Assessment. *Journal of Educational Statistics*, 17(2):111–129.

Sales, A. C. and Pane, J. F. (2019). The role of mastery learning in an intelligent tutoring system: Principal stratification on a latent variable. *The Annals of Applied Statistics*, 13(1):420–443.

Sales, A. C. and Pane, J. F. (2021). Student Log-Data from a Randomized Evaluation of Educational Technology: A Causal Case Study. *Journal of Research on Educational Effectiveness*, pages 241–69.

Schochet, P. Z. (2008). Statistical Power for Random Assignment Evaluations of Education Programs. *Journal of Educational and Behavioral Statistics*.

Schochet, P. Z. (2013). Student Mobility, Dosage, and Principal Stratification in School-Based RCTs. *Journal of Educational and Behavioral Statistics*, 38(4):323–354.

Simmons, D. C., Coyne, M. D., Kwok, O.-m., McDonagh, S., Harn, B. A., and Kame'enui, E. J. (2008). Indexing Response to Intervention: A Longitudinal Study of Reading Risk From Kindergarten Through Third Grade. *Journal of Learning Disabilities*, 41(2):158–173.

Smith, H. L. (1997). 6. Matching with Multiple Controls to Estimate Treatment Effects in Observational Studies. *Sociological Methodology*, 27(1):325–353.

Sobel, M. E. (2006). What do Randomized Studies of Housing Mobility Demonstrate? Causal Inference in the Face of Interference. *Journal of the American Statistical Association*, 101(476):1398–1407.

Sommers, B. D., Gawande, A. A., Baicker, K., et al. (2017). Health Insurance Coverage and Health - What the Recent Evidence Tells Us. *New England Journal of Medicine*, 377(6):586–593.

Sommers, B. D., Long, S. K., and Baicker, K. (2014). Changes in Mortality After Massachusetts Health Care Reform. *Annals of Internal Medicine*, 160(9):585–593.

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. (1990). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science*, pages 465–472.

Sussman, J. B. and Hayward, R. A. (2010). An IV for the RCT: using instrumental variables to adjust for treatment contamination in randomised controlled trials. *British Medical Journal*, 340:c2073.

Swaminathan, S., Sommers, B. D., Thorsness, R., Mehrotra, R., Lee, Y., and Trivedi, A. N. (2018). Association of Medicaid Expansion With 1-Year Mortality Among Patients With End-Stage Renal Disease. *JAMA*, 320(21):2242–2250.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*, volume 3, pages 72–89. Cambridge University Press.

Vanderweele, T. J., Hong, G., Jones, S. M., and Brown, J. L. (2013). Mediation and Spillover Effects in Group-Randomized Trials: A Case Study of the 4Rs Educational Intervention. *Journal of the American Statistical Association*, 108(502):469–482.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica: Journal of the Econometric Society*, pages 817–838.

White, M. C., Rowan, B., Hansen, B., and 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.

Wyss, R., Hansen, B. B., Ellis, A. R., Gagne, J. J., Desai, R. J., Glynn, R. J., and Stürmer, T. (2017). The “Dry-Run” Analysis: A Method for Evaluating Risk Scores for Confounding Control. *American Journal of Epidemiology*, 185(9):842–852.