A Case Weighting Scheme for Primary Outcome Analysis

Abstract

We introduce a case-weighting scheme, **P**ower-maximizing **W**eighting for **R**epeated-measurements with **D**elayed-effects, constructed with the aim of increasing the power of hypothesis tests for primary outcome analysis in randomized controlled trials (RCTs). The longitudinal nature of many RCTs means researchers often possess multiple observations on individuals but not necessarily the same number of observations on any given indivudal. As a result, the weighting scheme applied can substantially impact one's ability to find an effect. Our scheme, referred to as **PWRD** weighting, addresses these issues with repeated measurements by maximizing power for a broad class of interventions with just a few simple assumptions based off the theoretical underpinnings of these interventions. We illustrate this method on a large scale IES-funded RCT testing the efficacy of a reading intervention designed to help early elementary students at risk of falling below grade level proficiency. This intervention maintains that effects are both delayed and non-uniform and the treatment is anything but constant as a result. We show that for interventions of this type, **PWRD** weights provide greater power in expectation than commonly used competitors.

Keywords: randomized trials, pull-out intervention, power analysis, delayed effects

Contents

1 Introduction

In education research, randomized controlled trials (RCTs) generally provide the highest-quality possible evidence for determining the efficacy of interventions. This is largely due to the fact that well-balanced randomized trials tend to reduce bias, letting researchers enhance the accuracy of their analysis. Despite this desirable property, they do possess some drawbacks as well. In particular, RCTs are often underpowered for a variety of reasons. This lack of power typically arises due to both the increased tendency to pre-register analysis plans and the inherently expensive nature of RCT design. The first hurts the power of the experiment because the mode of analysis needs to be selected prior to examining the data from the experiment and the second places constraints on the sample size, thus limiting power as well. Many covariate methods have been proposed to help address this issue like targeted MLE (Moore and van der Laan, 2009; Balzer et al., 2016) and the LOOP estimator (Wu and Gagnon-Bartsch, 2018) but another strategy is to adjust the method of weighting.

In this paper, we choose to combat the lack of power in this manner by introducing a case-weighting scheme, Power-maximizing Weighting for Repeated-measurements with Delayed-effects. These weights are constructed with the aim of increasing the power of hypothesis tests for primary outcome analysis in randomized controlled trials. Researchers often possess multiple observations on individuals due to the longitudinal design of many randomized trials, but they will not necessarily possess the same number of observations on any given individual. As a result, the decision about which weighting scheme to apply can substantially influence one's ability to detect an effect. Our scheme, henceforth referred to as **PWRD** weighting, addresses these issues with repeated measurements by maximizing power for a broad class of education interventions.

We illustrate this method on a large IES-funded RCT testing the efficacy of a reading intervention designed to help early elementary students at risk of falling below grade level proficiency. This intervention, which we call RSEG (Reading Support in Early Grades), is a "pull-out" intervention providing targeted instruction to students who test below a certain benchmark rather than providing specific instruction to the entire classroom. The theory behind RSEG maintains that effects are delayed, i.e. students do not receive an immediate benefit upon joining the intervention. Instead, they must "test in" prior to receiving the treatment. RSEG additionally maintains that effects are non-uniform in that only students pulled out of the classroom are affected; as a result, the treatment is anything but constant when the intervention works in the intended manner.

The primary objective of **PWRD** weights is to construct a method of weighting that maximizes the likelihood of rejecting a null hypothesis of no effect if the theory behind the intervention is true. In particular, we provide a weighting scheme that increases power in expectation for interventions where students are pulled out to receive targeted instruction. Furthermore, our scheme allows us to define the exact causal parameter of interest rather than merely the estimate for a pre-determined parameter.

1.1 Literature Review

While there is extensive literature on both randomized trials and longitudinal data, much less has been written with respect to weighting methods for the combination of the two topics. The weighting scheme applied is typically of secondary importance and researchers in the education world often fail to explicitly refer to the weighting scheme they apply when handling repeated observations on students. Fortunately from the reader's perspective, most researchers apply one of the following three schemes: weights that do not vary (Meece and Miller, 1999), exit observation weighting (Simmons et al., 2008; Meece and Miller, 1999), or most commonly weighting through hierarchical linear models (HLM) (Lee, 2000; Ethington, 1997; Guo, 2005).

The first method, which we term "flat" weighting, can be viewed as the linear model with working independence structure (Laird, 2004; Fox, 2015) in the mixed model and general estimating equations literature. Due to this assumed independence, each observation receives the same weight or in other words, the weights are flat across all observations.

The second method, which we term "exit observation" weighting, lies on the opposite end of the spectrum from flat weighting. Rather than treating observations within students as independent, exit observation weighting treats them as completely dependent. In other words, each student rather than each student observation receives the same weight. The most common method of attaching exit observation weights ignores all student observations prior to the student's final record. For example, if we looked at a reading intervention for kindergarten through third grade students with an outcome of Grade 3 reading scores, exit observation weights would only analyze data from students in the third grade, i.e. when they exit the study. Observations on students in kindergarten through second grade would be ignored in outcome analysis.

Both the above methods are rather extreme. Flat weighting imposes unrealistic assumptions on our data whereas exit observation weighting discards valuable data. Instead of these options, the majority of researchers use hierarchical linear models (Raudenbush and Bryk, 2002; Bryk and Raudenbush, 1987) when conducting outcome analysis. This implicitly chooses a middle ground between flat and exit observation weighting. HLMs allow for some correlation between observations but also accept that they are not entirely dependent either.

A different and more extensive literature addresses the issue of causal effect estimation of time-varying exposures but largely through methods other than weighting. Instrumental variables (IV) (Bloom, 1984) are one prominent example. This technique, utilized broadly in economics literature, is gaining prominence in other fields as well. Briefly, instrumental variables are implemented in scenarios when the explanatory variable of interest is correlated with the error. 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 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 paricipants 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 in favor of an "as treated" analysis. Instead, under certain conditions the IV estimand will be identical to the complier-average causal effect (CACE) (Angrist et al., 1996; Baiocchi et al., 2014). The CACE 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). If formulated properly, 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. Instrumental variables additionally allow researchers to put bounds on the average treatment effect for the full population by bounding the difference between the CACE and average treatment effects for the never takers and always takers (Baiocchi et al., 2014).

Another area of research addressing the same issue 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 under certain conditions, will provide identical estimates of the treatment effect.

Much of this literature is constructed with sequentially randomized experiments with differing treatment regimes across time in mind, similar to how students in RSEG 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 $Z_{it} = 1$ denoting receiving the treatment. Then $\bar{Z}_i$ is the treatment regime throughout the length of the experiment so 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 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, $\sum_t Z_{it}$, through marginal structural models. More specifically, we can test this hypothesis by using ordinary least squares with IPTW (Robins et al., 2000). Note that with respect to RSEG, we do not work under an assumption of increasing effect as a function of exposure, but the assumption of increasing probability of exposure. The actual treatment effect is constant regardless of the exposure. Nonetheless, 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 treat a specific target parameter as given. Nevertheless, there may be no single weighting that is superior or more appropriate than the rest when aggregating repeated measures of the outcome in a comparative study. Rather than one ITT parameter,there are instead many possible ITT parameters each of which may be valid. **PWRD** weighting suggests the method of aggregating follow-up measurements into a single estimate prior to comparing averages across comparable treatment and control groups. In other words, **PWRD** weighting helps specify the target parameter rather than treating it as given.

1.2 Roadmap

In this paper, we first motivate this weighting method through a simple example using the 2006 health care reform in Massachusetts. We then discuss the connection of longitudindal data in education settings to pull-out interventions in Section 2.2. After, we use theoretical underpinnings behind pull-out interventions to define assumptions under which **PWRD** weighting will be power-maximizing. We then explicitly present the formulation for **PWRD** weights. In Section 3, we show how **PWRD** weights compares with other commonly used weighting schemes in a simulation study also presented therein. We make this comparison under both the standard and relaxed assumptions for pull-out interventions before showing how the methods perform when the assumptions fail as well. Finally, in Section 4, we recap how **PWRD** weights provide researchers with a tool that will best help them detect an effect for pull-out interventions. We then conclude by discussing how this method can be extended to other scenarios.

2 Method

2.1 Massachusetts Health Care Reform

To motivate the weighting method presented in this paper, we briefly look at the 2006 health care reform in Massachusetts. Legislators wrote the bill with the intention of providing universal health care coverage to Massachusetts residents through expansion of Medicaid, subsidized private insurance, and an individual mandate. There is little doubt that access to healthcare increased after its passage but it is less clear what benefits, if any, the legislation had on mortality.

We believe that one reason researchers have struggled to uncover an effect on mortality is because these studies are searching in the wrong area. To illustrate, examine two Massachusetts counties in Table 1.

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

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

Middlesex, home to Harvard and MIT, is one of the wealthiest counties in the United States. Prior to health care reform, only 7.2% of residents were living in poverty and under 15% of residents were uninsured. Since Medicaid expansion and subsidized private health insurance primarily target low income individuals, we would not expect to see many Middlesex County residents benefitting from this legislation. Suffolk County, on the other hand, includes much of Boston proper and had a poverty rate of 17.1% while over 18% of residents were uninsured. As a result, we would expect to find a greater benefit to mortality in Suffolk than Middlesex.

Based on this theory, i.e. that we will see the largest benefits to Medicaid expansion among counties with a large proportion of low-income residents, we divide counties into four brackets based on their 2006 poverty rate. Table 2 shows the average proportion of adults below the poverty limit in each bracket.

Bracket	Pov. Rate
1	5.7%
2	8.5%
3	10.9%
4	18.0%

Table 2: The average poverty rate in each bracket for 2006.

All else equal, we would expect an effect that is increasing across brackets where counties in Bracket 4 experience the largest benefit to Massachusetts healthcare reform. Nonetheless, it should be noted that poverty rate is a rough proxy that will underestimate the true proportion of individuals who stood to benefit from this health care reform. For example, adults with income between 150% and 300% of the federal poverty limit received subsidized health insurance and poverty rate does not capture those individuals.

2.1.1 Analysis of Mortality

To test this theory, we follow the same procedure as Sommers et al. (2014). We compare their standard method, i.e. a negative binomial regression of healthcare-amenable mortality counts on treatment controlling for baseline covariates, with a modified version of their model that interacts treatment with a county's poverty bracket. From this modified version, we obtain the effect estimates shown in Table 3.

Bracket	Coef.	S.E.
1	-0.002	0.024
2	-0.030	0.019
3	-0.032	0.022
4	-0.039	0.020

Table 3: The estimated effects on mortality for each of the four brackets.

The magnitude of the mortality benefit increases across the four brackets and the standard error largely decreases as well. The following question then arises: how do we best aggregate these coefficients into a single estimator? In other words, how do we weight these coefficients to best estimate the overall effect on mortality of the Massachusetts health care reform? Here, we choose to apply a variation of the **PWRD** weighting formulation described in Section 2.3. Simply, we weight observations based on the county's poverty bracket because greater numbers of low-income adults present a greater opportunity for individuals to benefit from this healthcare reform. We then calculate p-values adopting the permutation inference technique used by Kaestner (2016) for both the standard method and the weighted aggregate. Under both methods, we find significance at the 5% level when using a Wald statistic with a clustered standard error (Bell and McCaffrey, 2002). Neither the standard method nor the weighted aggregate uncovered a significant effect with the estimated coefficient as the test statistic although the weighted aggregate did provide a 5% reduction in the magnitude of the p-value.

Method	**Coef.**	**S.E.**	**tstat**	**Coef** p	**tstat** p
Standard	-0.028	0.007	-4.27	0.221	0.006
Weighted Agg. 4	-0.030	0.006	-5.25	0.210	0.006

Table 4: The estimated effect and test statistic under a standard analysis and under a weighted aggregation.

We speculate that gains in power for the p-value of the coefficient were smaller than expected for two reasons. The first relates to the size of the treatment group. Massachusetts is divided into 14 counties and we further divide those 14 counties into four poverty brackets, each bracket containing three or four counties. As a result, the coefficient of the estimated effect for any given bracket is susceptible to outliers when permuting treatment assignment, causing the permuted aggregated effects to vary greatly in magnitude and reducing power in the process. Note that this issue solely arises for the p-value of the coefficient and not the p-value of the Wald statistic. The larger standard error present due to these outliers reduces these permuted statistics and we have ample power to detect an effect under the Wald test statistic. We believe the second reason power was smaller than expected is merely that poverty rate is a weak proxy for the proportion of adults benefitting from health care reform. With a stronger tool at our disposal, we may have realized greater gains in power.

2.2 PWRD Weights Intuition

In educational settings assessing efficacy of interventions, students frequently enter and exit studies at different points. As a consequence, we often possess varying numbers of observations for any given student. For example, RSEG examined a reading intervention on early elementary students across four years and we possessed one to four observations for each student depending on their entry grade. Table 5 illustrates the design for the first cohort.

In RSEG and other longitudinal settings, the question of how to weight each observation or student is of great importance. The simplest outcome analysis might sidestep this question entirely by solely examining outcomes when students exit the study (e.g. 3rd grade observations in RSEG), but even here complications emerge. According to

	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 5: Progression of Cohort 1 through the four years of the RSEG study.

the theory behind RSEG, 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 hamper a researcher's ability to detect an effect.

Instead, researchers typically choose to make use of all their data. Perhaps the easiest way to handle *repeated* measurements is to fit a linear model predicting student-year observations from independent variables identifying the cohort and 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 (Laird, 2004; Fox, 2015). These analyses effectively attach equal weights to each student observation and thus we refer to them as "flat" weights. However, this weighting scheme fails to take into account which observations will help us best detect an effect and consequently some power is lost.

	Grade at Entry	Year 1	Year 2	Year 3	Year 4	Difference
Cohort 1	**3**	3.33	-	-	-	-
	2	-0.63	-0.38	-	-	0.25
	1	1.44	6.57	2.16	-	0.72
	0	-5.33	3.74	1.18	3.45	8.78

Table 6: Difference in means between treatment and control groups for the first cohort of students.

This is particularly the case in pull-out interventions where students only receive the intervention once they "test in", causing effects that are scattered and delayed rather than concentrated and instantaneous. Before students test in, they receive the same instruction they otherwise would have received if no intervention took place. Due to this, pull-out intervention theory maintains that students only receive an effect once they have been removed from the classroom to receive that targeted remediation. It naturally follows that the longer an individual participates in an intervention with targeted remediation, the more likely they will be to directly test into that intervention and thus beneft from it. Therefore, we expect to see greater effects from the intervention as students participate for a greater length of time.

Table 6 illustrates this phenomenon in RSEG. While there is some random variation, we generally see larger differences between unadjusted treatment and control means as students participate in the study for longer. **PWRD** weighting looks to address this issue with repeated measurements and non-instantaneous effects by building around the theory that the longer students participate in a study, the more likely they will be to experience a benefit. In other words, the expected size of the effect at time t will be proportional to the percentage of students who directly received the intervention by time t.

Years in RSEG	**Tested In**
1	38.3%
2	54.3%
3	61.1%
4	69.4%

Table 7: The proportion of students in the control group who have "tested in" to the pull-out intervention by how many years they have participated in the study.

Table 7 demonstrates how this works. Students who have been in the study for a greater length of time are more likely to have been "pulled out" for targeted remediation. As a result, we will then attach greater importance to observations from students who have participated for more time since those are the observations most likely to have benefited from the intervention.

2.3 Presentation of PWRD Weights

The following weighting scheme is constructed under the potential outcomes framework of Rubin (1974) and Splawa-Neyman et al. (1990) and for the class of intention-to-treat (ITT) estimators (Montori and Guyatt, 2001; Gupta, 2011). We let $Z = 1$ denote those who were assigned to the treatment and $Z = 0$ denote those assigned to the control. Y is the outcome of interest and we let F_t denote the group of students who are in year-of-follow-up t. For example, all observations of students who are in their second year of the intervention would belong to F_2. We let Y_{1t} denote the outcomes for students in time t who received the treatment.

We can then define Δ_t as the parameter representing the treatment effect in year-of-follow-up t, i.e.,

$$\Delta_t = \mathbb{E}(Y_{1t} - Y_{0t}|T = t).$$

This decomposes our overall average treatment effect into a collection of ITT estimates for each year-of-follow-up. We can then view Δ_t as the portion of that overall effect contributed by observations from year-of-follow-up t. In other words, our overall treatment effect can be interpreted as a linear combination of parameters Δ_t, i.e. $\sum_t \gamma_t \Delta_t$. **PWRD** weighting looks to discover which linear combination of parameters will maximize the power of tests based on the statistics $\sum_t \gamma_t \hat{\Delta}_t$ and $\sum_t \hat{\gamma}_t \hat{\Delta}_t$. To find that linear combination, we first make some assumptions about the treatment.

When the theory behind interventions with targeted remediation holds, we make the following two assumptions about the nature of the treatment effect:

Condition 2.1. *Individuals who receive the intervention at time j receive a homogenous effect $\tau_{ij+} \geq 0$ at some point between j and when they exit the study. Individuals who do not receive the intervention are unaffected.*

Condition 2.2. *Effect τ_{ij} received by individual i at time j is retained by individual i in full throughout the duration of the study, i.e. from $[j, t_i]$.*

We can view Condition 2.1 as an extension of the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980). Briefly, SUTVA states that 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 RSEG, this additionally means 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 remain in the classroom instead. We see this illustrated in Figure 1. From these two conditions, we now construct **PWRD** weights.

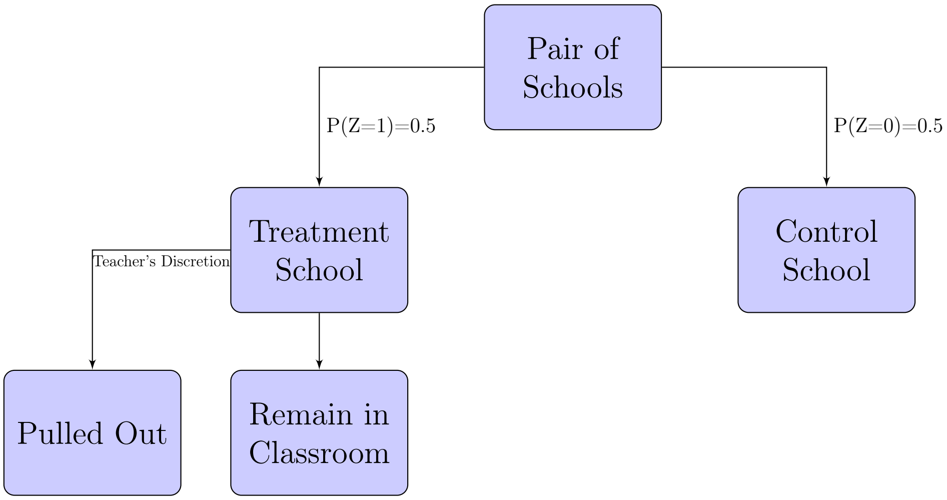

Figure 1: RSEG design for a pair of schools.

Proposition 2.3. *Take the no-effect hypothesis test that the linear combination of parameters $\sum_t \gamma_t \Delta_t$ is zero or negative against the alternative hypothesis that this linear combination is positive. Under Conditions 2.1 and 2.2, the following weights w maximize in expectation the power of the resulting test under the event that the null hypothesis is false:*

$$w = \alpha \cdot \Sigma^{-1} \boldsymbol{p}_0$$

where α is some constant, $\Sigma := \operatorname{Cov}\{(\hat{\Delta}_{(Z=1,t)} - \hat{\Delta}_{(Z=0,t)} : t)\}$, and $\boldsymbol{p}_{0t} := \mathbb{P}($student in control would 'test in' to the treatment by time $t|F_t)$.

In other words, let us examine test statistics of the following general form:

$$\frac{\mathbb{E}(w\Delta')}{\operatorname{Var}^{1/2}(w\Delta')}.$$

The signal-to-noise ratio will be maximized by weights of the following form:

$$w = \alpha \cdot \Sigma^{-1} \mathbf{p}_0,$$

which we call **PWRD** weights.

Very generally, we take the gradient of the above test statistic with respect to w. After setting this term equal to zero and simplifying through a grouping of scalar quantities, we obtain **PWRD** weights. We additionally add a constraint to ensure that weights are non-negative. For the complete derivation, see Appendix A. Neither $\mathbf{p}_{0t}$ nor Σ are directly observed, but both can be easily estimated: $\mathbf{p}_{0t}$ through the proportion $\hat{\mathbf{p}}_{0t}$ observed among students assigned to the control and Σ through a slightly more elaborate calculation centered around control-group residuals. Complications emerge in situations where the experimental design includes block random assignment with clustering of observations within each block. We handle this by scaling the "bread" component of Huber-White sandwich estimators of the variance (Huber, 1967; White, 1980) 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 presented above.

We do not currently apply any covariate adjustment under this formulation; instead we have focused on the simplest scenario where outcomes Y are unadjusted. This can be amended through the approach outlined in Lin et al. (2013) or through application of a Peters-Belson method (Peters, 1941; Belson, 1956) when estimating Σ. While not constructed with attributable effects (Rosenbaum, 2001) in mind, we can extend **PWRD** weights into that setting with minor adjustments.

2.3.1 Relaxing Pull-Out Intervention Assumptions

As stated in Section 2.3, one condition of pull-out intervention theory is that individuals who are pulled out of the classroom and into the intervention receive a non-negative effect and individuals who do not remain unaffected. However, Condition 2.1 weakens under additional scrutiny. Since schools have finite resources, it follows that pull-out interventions may transfer resources away from the classroom to allow for targeted remediation. In other words, this is a scenario where SUTVA fails to hold and students testing into the intervention itself adversely affects the potential outcomes of the remaining treatment students. Nonetheless, we can relax that assumption, i.e. relax that students who do not receive the intervention are unaffected, so long as the following is true:

Condition 2.4. *Individual i receiving the intervention at time j gains a homogeneous non-negative effect τ_{ij} at some point between j and when they exit the study. Individuals who do not receive the intervention may experience an effect, positive or negative, so long as the overall effect of all students is positive in aggregate.*

3 Results

While the theory behind the RSEG pull-out intervention was discussed in Section 2, the structure of the data merits additional discussion to understand simulation design. RSEG is a large-scale cluster randomized trial testing the efficacy of a reading intervention among early elementary students designed to assist students at risk of falling below grade-level proficiency. RSEG is block-randomized at the school level with 26 total blocks. 24 of these are pairs of schools, and the remaining blocks are a triplet of schools (in which two schools were assigned to treatment) and a singleton. The singleton originally belonged to a pair but the school assigned to the control dropped out of the study. Nearly all schools were matched within a district and the randomized trial is well-balanced as a result. Across these 52 schools, we observe 27000 unique students for somewhere between 1 and 4 years for a total of 52000 observations. As discussed in Section 2.2, the length of time for which each student participated in the study depends on the grade and year during which they entered the study. While we encounter some missing data, we possess demographic information (Race/Gender/DOB/Free Lunch Status/etc.) for the vast majority of students. We additionally observe DIBELS scores and end-of-year assessment scores for every student. DIBELS, a widely used reading assessment, serves as the tool by which students are pulled out of the classroom to receive RSEG targeted instruction and can additionally function as a pre-test. The end-of-year assessments are our primary outcome of interest.

3.1 Primary Simulation Design

In order to demonstrate how **PWRD** weighting performs in comparison to "flat" weighitng and hierarchical linear models on RSEG data, we need to construct a unique simulation design. This need primarily arises because all RSEG treatment observations actually received some unknown effect τ from the intervention. To address this need, we generate new student outcomes from a two-level model based on RSEG data as follows:

$$Y_{ijk} = \beta_0 + \beta_1 Grade_{ijk} + \mu_i + \epsilon_{ijk}$$

$$\mu_i = \gamma_0 + \omega_i$$

In other words, the outcome of student j in year-of-follow up k at school i is a function of the grade of the student and the random intercept of the school at which the student is enrolled. It should be noted that fixed effects like race, gender, socio-economic status and others could be added to this process but were excluded since we have presented **PWRD** weights without covariate adjustment. Once these outcomes are generated, the following two steps are taken. First, outcomes that fall below a given threshold are flagged 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. Next, we impose treatment effects on the students within treatment-schools and find the corresponding power across iterations of this data generation.

We look at three different versions 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 in to the intervention, receive a constant negative effect $-p\tau$ where $p \in (0, 1]$. The third version of treatment effect imposes $\tau \sim N(k, 2.5 * k)$ for some k to all treatment observations.

3.2 Analysis of Results

Figure 2 shows results of power for 1000 replications of the synthetic experiment across three weighting schemes: **PWRD**, flat, and HLM. It is immediately apparent that **PWRD** weights outperform the other two methods, especially for medium effect sizes where we observe a 30-40% increase in power. This is unsurprising since **PWRD** weights attach greater importance to student observations most likely to have received an effect from the intervention and down-weight the remaining observations. As the

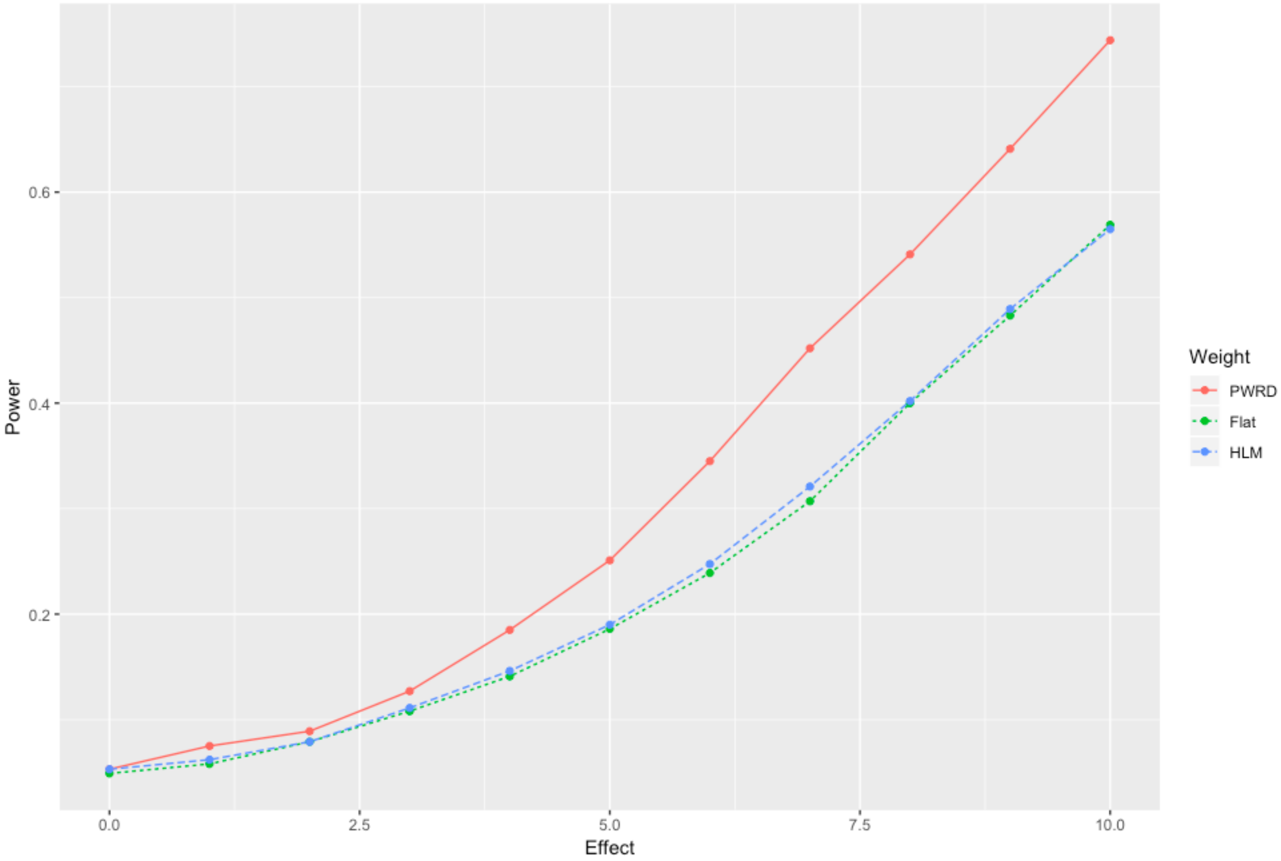

Figure 2: Power for the three weighting schemes across increasing effect sizes.

size of the effect increases, we begin to observe diminishing returns and each method provides researchers with ample opportunity to detect the effect. For smaller effect sizes, the improvement over flat weighting and HLM weighting may not seem large at first. That being said, we observe roughly a 10% improvement in power, which is not insubstantial. It should also be noted that despite the unconventional weighting, **PWRD** weights still maintain proper Type I error rates when no artificial effect is imposed on treatment observations.

A natural question, particularly in education settings, is whether this holds across different intraclass correlations (ICC). We now look at these same weighting schemes holding the imposed effect constant, but varying the ICC within schools. We see these results in Figure 3.

It is readily apparent that across ICCs that typically arise in these settings (Hedges et al., 2007), **PWRD** weighting consistently outperforms flat and HLM weighting. For intra-class correlations between 0.1 and 0.2, **PWRD** weights provide 30-40% more power than the alternatives. That gap does decrease for larger ICCs although this is at the upper range of reasonable ICC values and we still obtain a 10-20% improvement

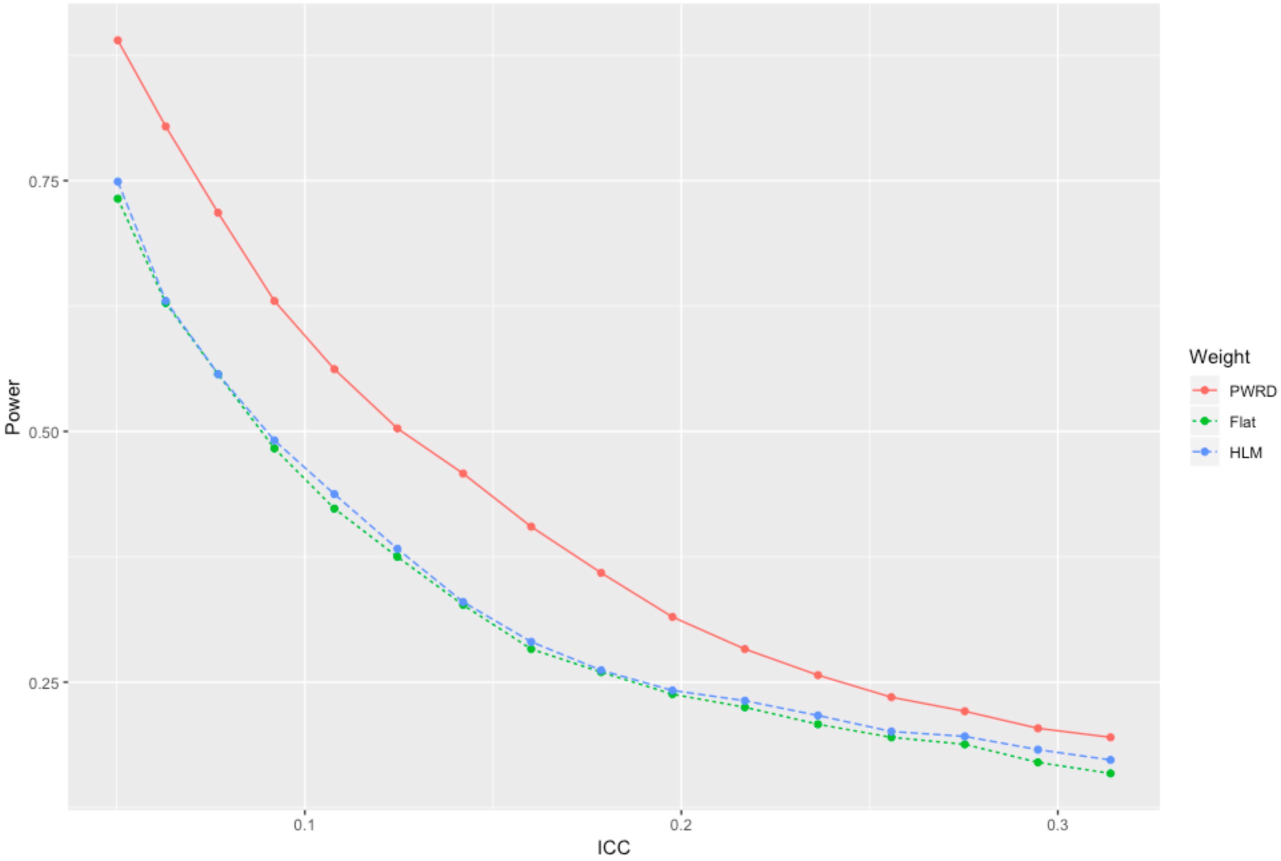

Figure 3: Power for the three weighting schemes across increasing intraclass correlations.

in power.

We now relax the assumption that students who are not pulled out and do not receive the targeted instruction are unaffected. Instead, we impose a negative effect that is 40% in magnitude of the positive effect imposed on students who are pulled-out. 40% was chosen such that the overall effect is positive in aggregate.

In Figure 4, we observe that **PWRD** weights actually perform better in comparison to the other two methods than they did without the relaxed assumption. This is to be expected as well. **PWRD** weights down-weight observations who are less likely to have benefitted from the intervention and up-weight observations who are likely to have experienced a positive effect. In other words, we weight down observations who are more likely to have received a *negative* effect, attaching greater importance to those more likely to have received a *positive* effect. Neither flat weighting nor HLM weighting are able to do this and their power to detect the effect is substantially reduced as a consequence. For small effect sizes, **PWRD** weights increase power by nearly 20% and this gap only widens as the effect size increases, over doubling the power that HLM weighting and flat weighting provide.

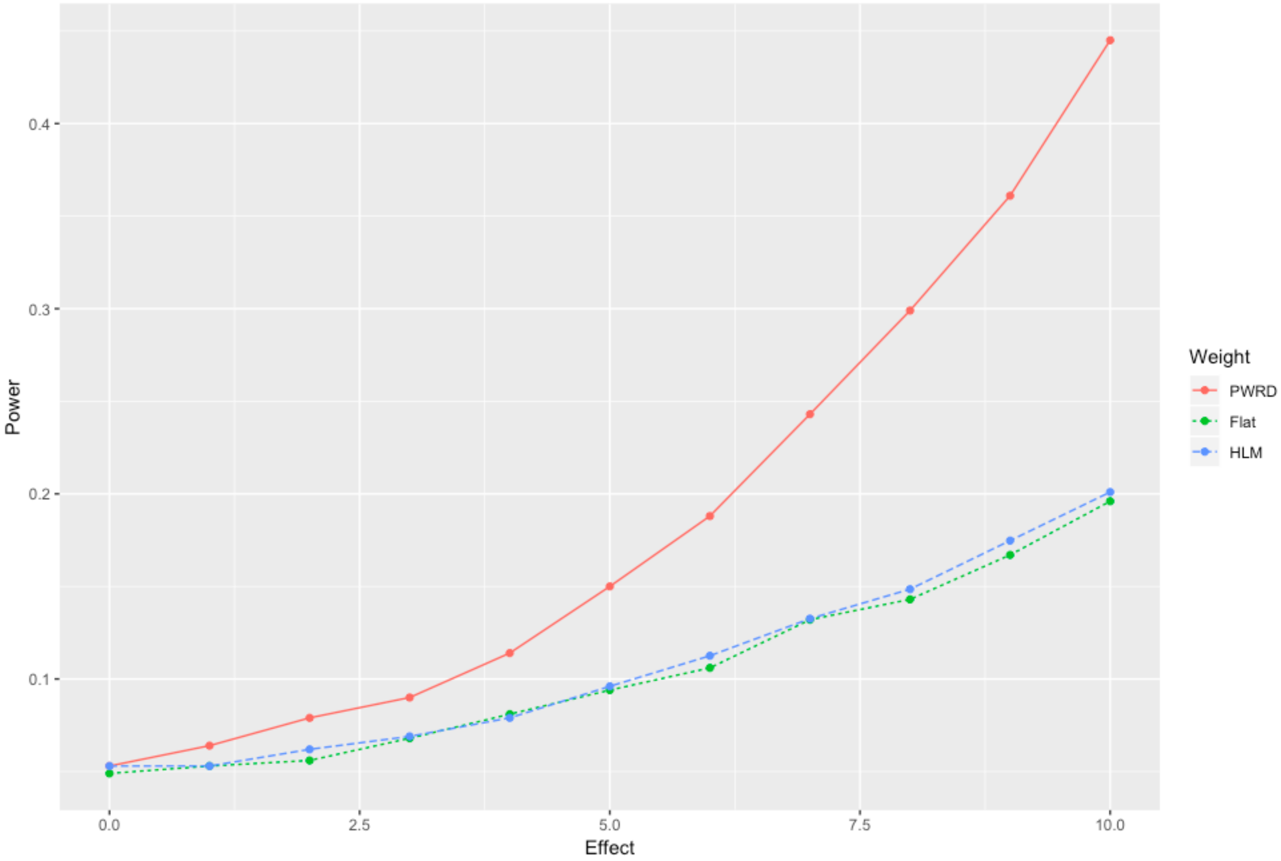

Figure 4: Power for the three weighting schemes across increasing effect sizes when Condition 2.1 does not hold.

The phenomenon we observe in Figure 4 holds when the magnitude of the negative effect varies as well. We observe this in Figure 5. Here, the adverse effect experienced by those in the treatment who do not test into the intervention varies from 0% of the benefit to 100% of the benefit. **PWRD** weighting provides a persistent 5-6 percentage point advantage in power for negative effects up to 60% of the positive effect before narrowing out. This corresponds to at least a 30% improvement in power at all levels of negative effect below 100%, capping out at an 85% improvement for a negative effect with 70% the magnitude of the positive effect.

We now examine what occurs in cases where the theory behind pull-out interventions entirely fails. This may arise either because the pull-out intervention does not work as intended or because the intervention does not fall within the broad class of pull-out interventions. We now impose an artificial treatment effect on all treatment observations such that $\tau_{it} \sim N(k, 2.5 * k)$ for $k = 1, \ldots, 10$. Note both that while the aggregate effect is still positive, any given student can be negatively affected and that effects are neither stacked nor persistent across time. We see the results in Figure 6.

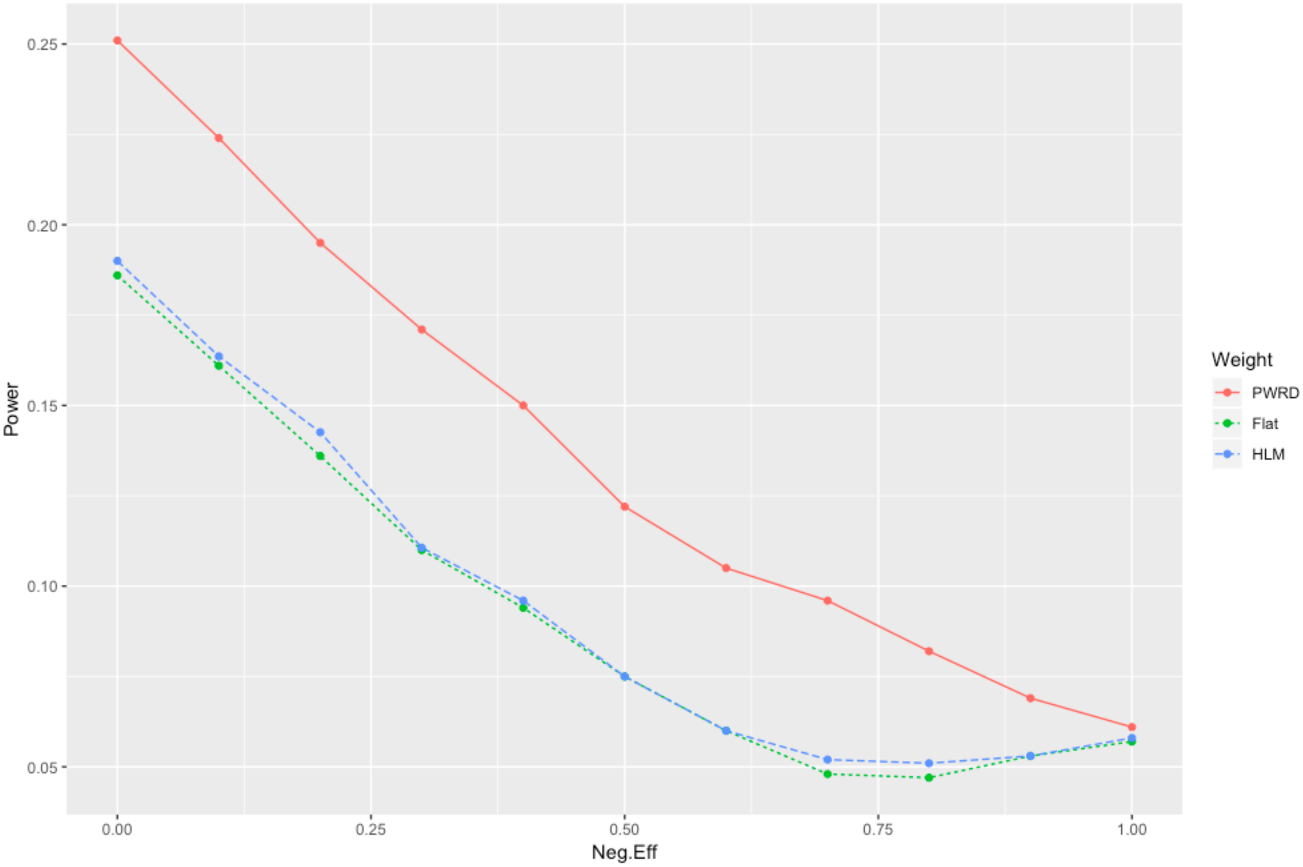

Figure 5: Power for the three weighting schemes across increasing negative effects. Here we add a positive effect of size 5 to students in the intervention and a negative effect that increases from 0 to 5.

We immediately observe that while HLM weighting slightly outperforms **PWRD** weighting, this improvement is minimal and never rises above a 3% improvement. By the time we reach larger effect sizes, we are able to reject under any of the weighting schemes. From these simulations, it is clear that **PWRD** weighting provides substantial gains in power in situations where the theory behind pull-out interventions holds and when the theory does not hold, we see only a marginal decrease in our ability to detect an effect.

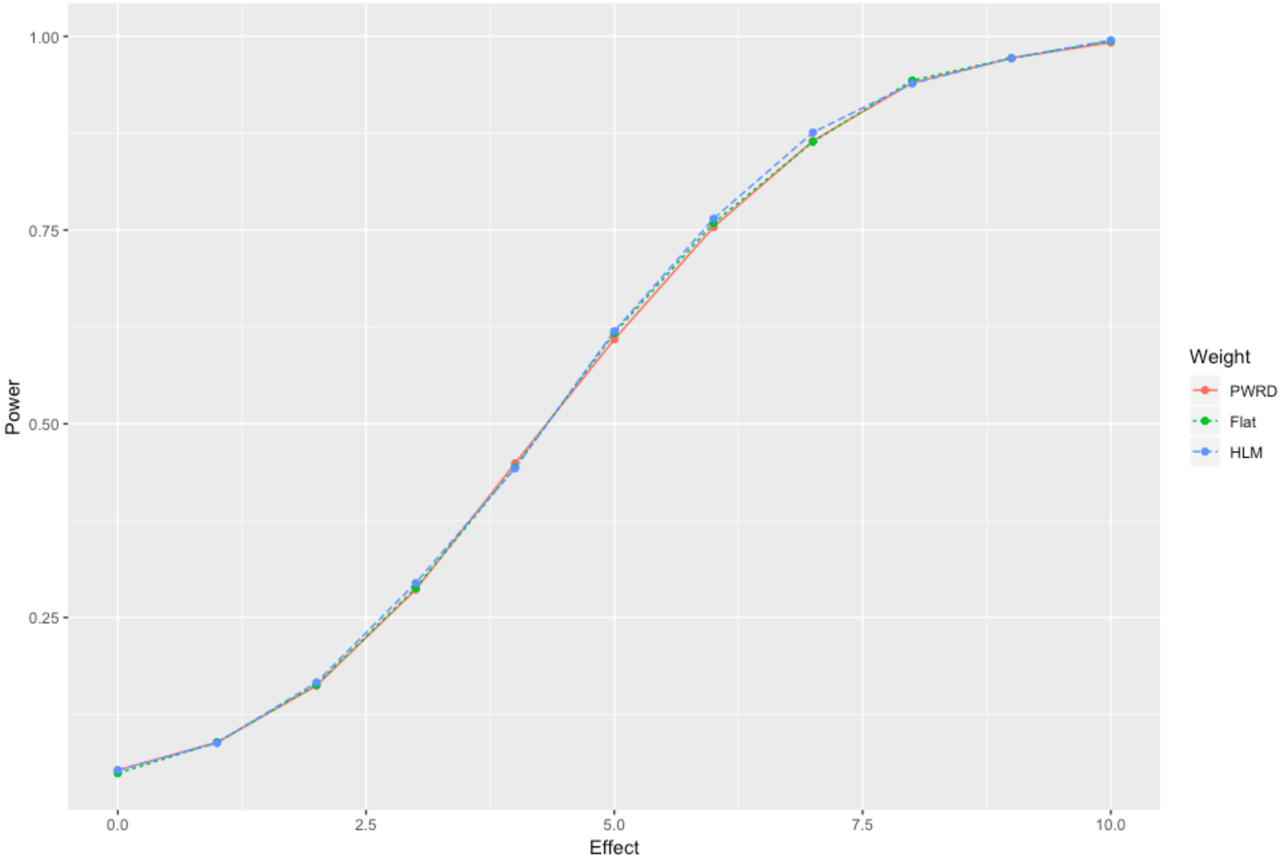

Figure 6: Power for the three weighting schemes across increasing effect sizes when none of the conditions hold.

4 Discussion

In this paper, we have presented a novel method of weighting that will maximize power in expectation for a broad class of interventions. More specifically, **PWRD** weighting gives researchers the best possible opportunity to detect an effect in pull-out interventions. This helps alleviate one of the main issues of randomized trials: namely that RCTs tend to be underpowered due to pre-registration and budgetary concerns. Furthermore, since the weights are derived solely from control data, this method is compatible with pre-registration of analysis plans in the first place. Additionally, our scheme allows us to define the exact causal parameter of interest rather than treat a pre-specified parameter as given.

It should be noted that while our weighting scheme is optimal when the theory behind pull-out interventions holds, no benefit is gained when all assumptions fail. Nonetheless, this weighting scheme does not hamper one's ability to detect an effect in this situation. In the future, we look to extend **PWRD** weighting to many other scenarios, both experimental and quasiexperimental, with longitudinal data and a treatment that accrues heterogeneously across observations. We believe that in each of these scenarios, similar weights can be formulated that will maximize power.

References

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.

Balzer, L. B., Petersen, M. L., van der Laan, M. J., and Collaboration, S. (2016). Targeted estimation and inference for the sample average treatment effect in trials with and without pair-matching. *Statistics in medicine*, 35(21):3717–3732.

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.

Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation review*, 8(2):225–246.

Bryk, A. S. and Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological bulletin*, 101(1):147.

Ethington, C. A. (1997). A hierarchical linear modeling approach to studying college effects. *HIGHER EDUCATION-NEW YORK-AGATHON PRESS INCORPORATED-*, 12:165–194.

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.

Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage Publications.

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.

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.

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.

Kaestner, R. (2016). Did massachusetts health care reform lower mortality? no according to randomization inference. *Statistics and Public Policy*, 3(1):1–6.

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

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.

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.

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

Moore, K. L. and van der Laan, M. J. (2009). Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. *Statistics in medicine*, 28(1):39–64.

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.

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*.

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

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

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.

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.

Sommers, B. D., Long, S. K., and Baicker, K. (2014). Changes in mortality after massachusetts health care reform: a quasi-experimental study. *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. *Bmj*, 340:c2073.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, pages 817–838.

Wu, E. and Gagnon-Bartsch, J. A. (2018). The loop estimator: Adjusting for covariates in randomized experiments. *Evaluation review*, 42(4):458–488.

Appendices

A PWRD Weights Derivation

Notation	Description
$\hat{\Delta}_t$	Estimated effect in year t
$\vec{\hat{\Delta}}$	Vector of estimated effects of dimension $(1 \times T)$
Δ_t	$\mathbb{E}\hat{\Delta}_t$, i.e. true effect in year t
w_t	Weight attached to year-of-follow-up t
$\vec{w}$	Vector of weights of dimension $(1 \times T)$
$\sum_t w_t \hat{\Delta}_t$	Pooled estimated effect
Σ_Δ	Covariance of effects across years t

Following the above notation, we write our target parameter, i.e. the unspecified linear combination of parameters Δ_t, as $\mathbb{E}(\sum_t w_t \hat{\Delta}_t) = \vec{w}\Delta'$. Similarly, the variance of this term can be written as follows: $\mathrm{Var}(\sum_t w_t \hat{\Delta}_t) = \vec{w}\Sigma_\Delta \vec{w}'$.

Our problem is then to select $\vec{w} = (w_1, \ldots, w_T) \geq 0$ to maximize the signal-to-noise ratio:

$$\max \frac{\mathbb{E}(\vec{w}\vec{\hat{\Delta}}')}{\mathrm{Var}^{1/2}(\vec{w}\vec{\hat{\Delta}}')}. \tag{A.1}$$

A.1 Determining the Optimum $\vec{w}_{opt}$

We first transform A.1 logarithmically which is equivalent to maximizing the following:

$$f(\vec{w}) = \log(\mathbb{E}(\vec{w}\vec{\hat{\Delta}}')) - \frac{1}{2}\log(\mathrm{Var}(\vec{w}\vec{\hat{\Delta}}')). \tag{A.2}$$

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

$$\nabla f(\vec{w}) : \frac{\mathbb{E}\vec{\hat{\Delta}}}{\mathbb{E}(\vec{w}\vec{\hat{\Delta}}')} - \frac{\vec{w}\Sigma_\Delta}{\vec{w}\Sigma_\Delta\vec{w}'} = \vec{0}.$$

Note that both $\mathbb{E}(\vec{w}\vec{\hat{\Delta}}')$ and $\vec{w}\Sigma_\Delta\vec{w}'$ are scalars, so we can rewrite this as follows:

$$(\mathbb{E}(\vec{w}\vec{\hat{\Delta}}'))^{-1}\mathbb{E}\vec{\hat{\Delta}} - (\vec{w}\Sigma_\Delta\vec{w}')^{-1}\vec{w}\Sigma_\Delta = \vec{0}.$$

We now rearrange the terms to solve for $\vec{w}_{opt}$:

$$\vec{w}_{opt} = \left(\frac{\vec{w}\Sigma_\Delta\vec{w}'}{\mathbb{E}(\vec{w}\vec{\hat{\Delta}}')}\right)\mathbb{E}\vec{\hat{\Delta}}\Sigma_\Delta^{-1}.$$

A.2 Estimation of $\vec{w}_{opt}$

From Slutsky's Theorem, we can then estimate $\vec{w}_{opt}$ as follows:

$$\vec{w}_{opt} = \left(\frac{\vec{w}\Sigma_\Delta \vec{w}'}{\vec{w}\hat{\vec{\Delta}}'}\right)\hat{\vec{\Delta}}\Sigma_\Delta^{-1}. \tag{A.3}$$

If we allow $\alpha = \left(\frac{\vec{w}\Sigma_\Delta \vec{w}'}{\vec{w}\hat{\vec{\Delta}}'}\right)$, we can then rewrite this as $\vec{w}_{opt} = \alpha\hat{\vec{\Delta}}\Sigma_\Delta^{-1}$. To check this simplifies, plug $\alpha \cdot \hat{\vec{\Delta}}\Sigma_\Delta^{-1}$ back into $\vec{w}$ in A.3. We have thus uniquely specified $\vec{w}_{opt}$. Furthermore, in principle we can define $\vec{w}_{opt}$ only up to a constant of proportionality such that $\vec{w}_{opt} = \hat{\vec{\Delta}}\Sigma_\Delta^{-1}$.

A.3 $\vec{w}_{opt}$ with a Non-Negativity Constraint

In Equation A.2, we wished to maximize $f(\vec{w}) = \log(\mathbb{E}(\vec{w}\hat{\vec{\Delta}}')) - \frac{1}{2}\log(\text{Var}(\vec{w}\hat{\vec{\Delta}}'))$. We now add in two constraints to prevent $w_t < 0$. In particular, we would now like to find $\max_w \log(\mathbb{E}(\vec{w}\hat{\vec{\Delta}}')) - \frac{1}{2}\log(\text{Var}(\vec{w}\hat{\vec{\Delta}}'))$ such that $w_t \geq 0 \; \forall \; t$ and $\mathbf{1}'w = 1$. In other words, we would like to maximize $\vec{w}$ such that each w_t is non-negative and $\sum_{t=1}^{T} w_t = 1$.

This is equivalent to:

$$\max_{\vec{w}} \log(\mathbb{E}(\vec{w}\hat{\vec{\Delta}}')) - \frac{1}{2}\log(\text{Var}(\vec{w}\hat{\vec{\Delta}}')) - u\vec{w}' + v\vec{w}'.$$

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

- **Stationarity**

$$(\mathbb{E}(\vec{w}\hat{\vec{\Delta}}'))^{-1}\mathbb{E}\hat{\vec{\Delta}} - (\vec{w}\Sigma_\Delta\vec{w}')^{-1}\vec{w}\Sigma_\Delta - u + v = \mathbf{0}.$$

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

- **Complementary Slackness**

$$u\vec{w}' = 0.$$

- **Primal Feasibility**

$$\vec{w} \geq 0, \mathbf{1}'\vec{w} = 1$$

- **Dual Feasibility**

$$u \geq 0$$

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

$$v - u = c_2\vec{w}\Sigma_\Delta - c_1\mathbb{E}\vec{\hat{\Delta}} \Rightarrow v \geq c_2\vec{w}\Sigma_\Delta - c_1\mathbb{E}\vec{\hat{\Delta}},$$

from stationarity, and

$$(c_1\mathbb{E}\vec{\hat{\Delta}} - c_2\vec{w}\Sigma_\Delta + v)\vec{w}' = 0,$$

from complementary slackness. After rearranging, we see that

$$\mathbf{0} \leq \vec{w} \leq \frac{v + c_1\mathbb{E}\vec{\hat{\Delta}}}{c_2}\Sigma_\Delta^{-1}.$$

From this, we then argue that $\vec{w}_{opt}$ is maximized by the following:

$$w_t = \begin{cases} \frac{v+c_1\mathbb{E}\vec{\hat{\Delta}}}{c_2}\Sigma_\Delta^{-1} & \text{if } v \geq -c_1\mathbb{E}\vec{\hat{\Delta}} \\ 0 & \text{if } v < -c_1\mathbb{E}\vec{\hat{\Delta}} \end{cases}.$$

In other words, $\vec{w}_{opt} = \max(0, \frac{v+c_1\mathbb{E}\vec{\hat{\Delta}}}{c_2}\Sigma_\Delta^{-1})$ where $\mathbf{1}'\vec{w} = 1$. We can then estimate $\vec{w}_{opt}$ following the same argument as in Appendix A.2.