An Aggregation Scheme for Increased Power in Primary Outcome Analysis

Timothy Lycurgus & Ben B. Hansen*

Abstract

A novel aggregation scheme increases power in randomized controlled trials and quasi-experiments when the intervention possesses a robust and well-articulated theory of change. Longitudinal data analyzing interventions often include multiple observations on individuals, some of which may be more likely to manifest a treatment effect than others. An intervention's theory of change provides guidance as to which of those observations are best situated to exhibit that treatment effect. Our *p*ower-maximizing *w*eighting for *r*epeated-measurements with *d*elayed-effects scheme, PWRD aggregation, converts the theory of change into a test statistic with improved asymptotic relative efficiency, delivering tests with greater statistical power. We illustrate this method on an IES-funded cluster randomized trial testing the efficacy of a reading intervention designed to assist early elementary students at risk of falling behind their peers. The salient theory of change holds program benefits to be delayed and non-uniform, experienced after a student's performance stalls. In this instance, the PWRD technique's effect on power is found to be comparable to that of a doubling of (cluster-level) sample size.

Keywords: delayed effects, effect aggregation, partial compliance, asymptotic relative efficiency, repeated measurements.

*The authors wish to acknowledge valuable comments and suggestions from Brian Rowan and Mark White, as well as research support from the National Science Foundation (DMS1646108) and the Institute for Education Sciences (R305A120811, R305D210029).

Contents

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 for randomized controlled trials and quasi-experiments. 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 increases 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. PWRD aggregation is primarily designed to assist with hypothesis testing rather than with estimation yet it may be implemented in tandem with standard estimation techniques.

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 the only students who are affected are those whose progress in reading has slowed. 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 follow-up times subsequent to points where student learning would otherwise have stalled. 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).

1.1 Roadmap

In this paper, 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 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 4, we then illustrate how PWRD aggregation compares with those same methods for BURST itself. Finally, in Section 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 Method

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 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 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 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. BURST Cohort 1's experience seems to have been of this type: as seen in Table 2, mean treatment-control differences in the exit observation year as compared to the entry year increase steadily from Cohort 1.2, with just 2 years of BURST, to Cohorts 1.1 and to Cohort 1.0, which enjoyed up to 4 years of BURST's supports.

In addition, exit observation analysis lacks appeal to researchers who prefer to use all of the 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

	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: Differences in mean reading scores between treatment and control groups for the first of four cohorts of students. The final column (δ) gives the difference in these differences as calculated for the final year of participation versus the first year of participation.

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 Model 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, 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, Model 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 Model 2.1 and Model 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 utilizing more than one posttest per student. Exit observation analysis allowed us to articulate a well-defined overall average treatment effect: the expected difference in outcomes among third grade students. Using repeated measures of the outcome 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, or 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 PWRD Aggregation

The three estimation methods presented in Section 2.1 all possess certain benefits. For example, exit observation analysis allows for a well-articulated overall ATE 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 kindergarten during the first year of the study (i.e. Cohort 1.0 in Table 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 4), but rather we use this subset as an example that best-serves to illustrate the intuition and process behind PWRD aggregation.

As with the principal stratification method of Sales and Pane (2021), PWRD aggregation requires estimation of separate treatment effects for each subgroup of interest. In Sales and Pane (2021), these are latent subgroups determined through dosage levels. For our method in the context of BURST, subgroup refers to the cohort year of follow-up. This too relates to dosage levels as the theory of change suggests that students were more likely to have received targeted remediation when they had participated in the intervention for longer, i.e. during later years of follow-up. Because schools may implement the intervention differently over time, the method calls for separate estimates of the treatment effect for each combination of cohort and year of follow-up. These covariate-adjusted treatment effect estimates for Cohort 1.0 during each year of follow-up are presented in Table 3. For an explanation of how the covariate adjustment was performed, see Section 4.

As a departure from Sales and Pane (2021) however, PWRD aggregation then serves as the tool by which we aggregate the four estimated effects in the course of hypothesis testing. This aggregate need not correspond to any meaningful average of individual effects in order to heighten our power to detect the presence of an effect. This formulation simultaneously allows us to sidestep the debates reviewed in Section 2.1 as to how the treatment effect is best parameterized, while making use of the full, longitudinal data in a fashion best suited to detect that effect.

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

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 would have received if no intervention took place. As a consequence, the theory of change entails the exclusion restriction (Angrist et al., 1996) that students only obtain an effect once they have received the supplemental instruction. The longer an individual has participated in an intervention of this nature, the greater the likelihood of their having become eligible to benefit from it.

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

Table 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 4 shows that for Cohort 1.0, student eligibility for the BURST intervention indeed increased in step with longer participation in the study. Accordingly, the theory of change posits that the expected size of the effect in cohort g during year of follow-up t will be proportional to the percentage of students in cohort g who were eligible for supplemental instruction by t, i.e. proportional to $\mathbf{p}_0 := \big(p_{0gt} : g, t\big)$, where $p_{0gt} :=$ $\mathbb{P}$(An individual in cohort g is eligible to receive the supplemental instruction by year of follow-up t). Thus $\mathbf{p}_0$ represents the proportion of students who were not *excluded* from having been affected, in virtue of the assumed exclusion restriction.

Nonetheless, the expected size of the effect as estimated through $\hat{\mathbf{p}}_0$ is not the only consideration of PWRD aggregation. Define Δ_{gt} as the parameter representing the ITT effect for cohort g during year of follow-up t, i.e.,

$$\Delta_{gt} := \mathbb{E}(Y_{gt}^{(Z=1)} - Y_{gt}^{(Z=0)} | G = g, T = t),$$

where := denotes "defined as", $Z = 1$ denotes assignment to the treatment and $Z = 0$ denotes assignment to control. Suppose corresponding ITT estimators (Gupta, 2011; Montori and Guyatt, 2001) $\{\hat{\Delta}_{gt} : g, t\}$ to have been designated. (PWRD aggregation is constructed under the potential outcomes framework of Rubin (1974), Holland (1986), and Splawa-Neyman et al. (1990). Note that our unit of observation is at the student-year level rather than at the student-level.) The estimated relative covariances among $\{\hat{\Delta}_{gt} : g, t\}$, denoted $\hat{\Sigma}$, also factor into our method, with those effect estimates that are relatively uncorrelated with the others receiving greater weight.

PWRD aggregation calculates a power-optimizing weighted combination of cohort/year of follow-up ITT estimates — an aggregate

$$\hat{\Delta}_{agg} := \sum_{g,t} \omega_{gt} \hat{\Delta}_{gt}, \tag{2.3}$$

with specially chosen weights ω ($\omega_{gt} \geq 0$, all g, t; $\sum_{g,t} \omega_{gt} = 1$). To find the specific ω that maximizes 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 2.1. *Individuals who receive supplemental instruction as a result of the intervention at time j receive an effect $\tau \geq 0$ at some point 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. *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 2.1 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 in the study. With respect to BURST, we argue this implies individuals testing into the intervention to receive targeted remediation will not affect the potential outcomes of individuals in the treatment who remain in the classroom without any supplemental instruction. This corresponds to a situation where there is no interference across individuals (Sobel, 2006).

Effectively, Conditions 2.1 and 2.2 amount to assuming that the effect for cohort g during year of follow-up t is proportional to the share of the cohort non-excluded by time t. A technical condition simplifies the development by excluding pathological cases.

Condition 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 2.4. *Consider test statistics of the forms: $\sum_{g,t} \omega_{gt} \hat{\Delta}_{gt}$, with g and t ranging over cohorts and times of follow-up respectively; $\sum_{g,t} \omega_{gt} \hat{\Delta}_{gt} - \sum_{g,t} \omega_{gt} \delta_{0gt}$, where δ_0 is a vector of hypothesized values of Δ, $\Delta := (\Delta_{gt} : g, t)$; and $\hat{v}^{-1/2}(\sum_{g,t} \omega_{gt} \hat{\Delta}_{gt} - \sum_{g,t} \omega_{gt} \delta_{0gt})$, where $\hat{v}$, perhaps an estimate of $\mathrm{Var}(\sum_{g,t} \omega_{gt} \hat{\Delta}_{gt})$, satisfies $n\hat{v} \to_p c > 0$. Consider the family of statistical hypotheses $\{K_\eta : \Delta = \eta \mathbf{p}_0,\ \eta \geq 0\}$. Under Conditions 2.1, 2.2, and 2.3, and for tests of $H_0 = K_0$ against alternatives K_η, $\eta > 0$,* *asymptotic relative efficiency is maximized by*

$$\omega = (\Sigma^{-1}\mathbf{p}_0)_+ \Big/ \sum_j (\Sigma^{-1}\mathbf{p}_0)_{+j}. \tag{2.4}$$

In (2.4), $(\Sigma^{-1}\mathbf{p}_0)_+$ *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 words, so long as the effect is proportional to the share of non-excluded, the slope (Pitman, 1948) of test statistics described in Proposition 2.4 may be maximized by weights proportional both to expected sizes of cohort-year effects, $\mathbf{p}_0$, and also the relative precisions of estimated cohort-year effects, Σ: $\omega \propto \Sigma^{-1}\mathbf{p}_0$. Note that any test statistic of the form:

$$\frac{\sum_{g,t} \omega_{gt}\hat{\Delta}_{gt} - \sum_{g,t} \omega_{gt}\delta_{0gt}}{\hat{v}^{1/2}}, \tag{2.5}$$

such as the t-statistic combining estimates $\hat{\Delta}_{gt}$ with fixed weights ω_{gt}, will be covered by Proposition 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 level of non-excluded students. Our method maximizes the "signal-to-noise" ratio for test statistics of the form presented in Equation 2.5 (which includes t-statistics). Thus, test statistics incorporating PWRD aggregation weights will provide researchers with a greater opportunity to detect an effect of the intervention when the theory of change holds. While designed to assist with hypothesis testing, our method may be used in tandem with a different approach to ITT estimation like flat weighting or exit observation analysis. Alternatively, the researcher may forgo ITT estimation entirely and instead present an instrumental variables estimate of a local ATE that examines average effects across non-excluded cohorts.

In general terms, we derive these weights by taking the gradient of the test slope of Equation 2.3 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 A.

2.2.1 PWRD Aggregation in the BURST Evaluation

In order to implement PWRD aggregation, researchers first require estimates of $\mathbf{p}_0$ and Σ to formulate the aggregation weights $\hat{\omega}$. In addition to contributing to $\hat{\omega}$, $\hat{\Sigma}$ assists in calculation of the standard error for $\hat{\Delta}_{agg}$.

Neither $\mathbf{p}_0$ nor Σ is directly observed, but both can be estimated easily. We estimate p_{0gt} through the proportion $\hat{p}_{0gt}$ observed among students assigned to the control. In the BURST example, this is the probability in cohort g of *ever* having tested in by time t, rather than the probability of testing in to supplemental instruction during year t: once a student becomes eligible for the first time, each subsequent observation for that student is deemed eligible as well. Thus, treatment received by a student in year t does not affect their weight in year $t+1$ or afterward; assuming the exclusion restriction, $\hat{\mathbf{p}}_0$ is pre-treatment in the sense that treatment assignment does not affect it. In theory, testing in to receive supplemental instruction from BURST solely occurred through *Dynamic Indicators of Basic Early Literacy Skills* (DIBELS), a reading assessment administered as a part of this intervention. If a student's DIBELS score fell within a certain range, they were eligible for the intervention. In practice, teachers may have used their own discretion when determining who received the supplemental instruction. Nonetheless, we estimate $\mathbf{p}_0$ solely using DIBELS, as PWRD aggregation is consistent with ITT analysis. Thus, we construct PWRD aggregation weights using the proportions of students who *should* have received the intervention if it was implemented with fidelity. That is, the level of non-excluded students within a given year of follow-up t is the expected proportion of students who were eligible for supplemental instruction by t as determined through DIBELS.

Often $(\hat{\Delta}_{g,t} : g, t)$ will be estimates from a common regression fit, in which case an accompanying estimate of the covariance of coefficient estimates can be used to estimate Σ in Equation 2.4. Our analysis of BURST used the Peters-Belson (1941; 1956) method, and called for a somewhat more elaborate calculation centered around control-group residuals (Hansen and Bowers, 2009). 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 served as $\hat{\Sigma}$.

To calculate the standard error, PWRD aggregation combines with standard techniques addressing 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. Simply, 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_{gt} 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.

2.3 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 share of non-excluded observations, PWRD aggregation maximizes power. But will PWRD aggregation have adverse effects on outcome analysis when the theory of change does not hold and the proportionality assumption fails? In particular:

- When there is no effect of the intervention, will PWRD aggregation lead to incorrect Type I errors?
- When the effect accrues in a different fashion than the theory of change hypothesizes, will PWRD aggregation yield less power than alternative methods?

To answer the first question, Appendix B proves from weak technical conditions that PWRD aggregation maintains proper Type I error rates (rather than over or under-rejecting a null hypothesis of no effect).

Simulations to be presented in Section 3 serve to address the second question, comparing power of t-tests based on Equation 2.3's $\hat{\Delta}_{agg}$, with weights ω as given by Equation 2.4, to t-tests based on flat-weighted, exit observation weighted or random effect-adjusted treatment ITT estimates. Our simulation study considers treatment effects of forms more and less favorable to PWRD aggregation, as well as a base scenario in which Δ_{gt} is proportional to p_{0gt}.

Our simulation study additionally provides an alternative to the above competitors that suggests analysts simply implement PWRD aggregation together with a standard method like flat weighting or exit observation analysis. Given the theory of change holds, PWRD aggregation will yield substantially more power than extant alternatives; if instead the intervention works in a manner different than hypothesized by the theory of change, the standard method will protect against a large loss of power. The max t procedure of Hothorn et al. (2008) allows these two methods to be implemented simultaneously while maintaining valid Type I error rates. Crucially, it avoids adding assumptions other than requiring a consistent estimate of the statistics' covariance. Furthermore, in the scenario that the two test statistics are highly correlated with one another, the max t procedure will provide power close to the power of any one of the correlated statistics.

Separately, failures of the $\Delta \propto \mathbf{p}_0$ model stemming from within-cluster interference can be studied analytically, without need for simulations.

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 2.1 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 2.1 in favor of the following:

Condition 2.5. *Individual i receiving supplemental instruction due to the intervention at time j gains non-negative effect τ at some point 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 2.5, we now present Proposition 2.6, a corollary to Proposition 2.4:

Proposition 2.6. *Under Conditions 2.2, 2.3, and 2.5, the following aggregation weights ω will maximize the slope of test statistics discussed in Proposition 2.4 for the family of hypothesis tests and alternative hypotheses also elaborated in Proposition 2.4:*

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

According to Proposition 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 2.6 follows the same structure as the derivation of Proposition 2.4 found in Appendix A.

3 Simulations

In order to demonstrate how PWRD aggregation performs in comparison to exit observation analysis, flat weighting, mixed effects models, and a combination of PWRD aggregation with flat weighting through the max t procedure, 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:

$$\begin{aligned} Y_{ijk} &= \beta_0 + \beta_1 \text{Grade}_{ijk} + \mu_k + \epsilon_{ijk} \\ \mu_k &= \gamma_0 + \nu_k \end{aligned}, \tag{3.1}$$

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 and Tipton, 2016).

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 points. Following guidance from Kraft (2020), we will refer to effect sizes less than 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 was 0.17σ (4 points) and the 90th percentile was 0.5σ (11.8 points) (Kraft, 2020). Thus, our simulation study examines these methods on effect sizes that frequently appear in reading interventions.

3.1.1 Effect 1

Figure 1 shows the power from 1,000 replications of the synthetic experiment for each effect size across the analytical schemes mentioned above. The mixed effects model is specified according to Equation 3.1, but with an independent variable representing the treatment. It is immediately apparent that PWRD aggregation outperforms the standard 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. In addition, note that while the max t combination of PWRD aggregation and flat weighting offers less power than PWRD aggregation, it still yields far greater power than the standard methods alone. Thus, this method should prove attractive both when the analyst wishes to be protected against a loss of power if the theory of change fails or when the analyst wishes to both test the null hypothesis and estimate the treatment effect using a standard approach.

It is natural to ask whether the gains in power present in Table 1 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

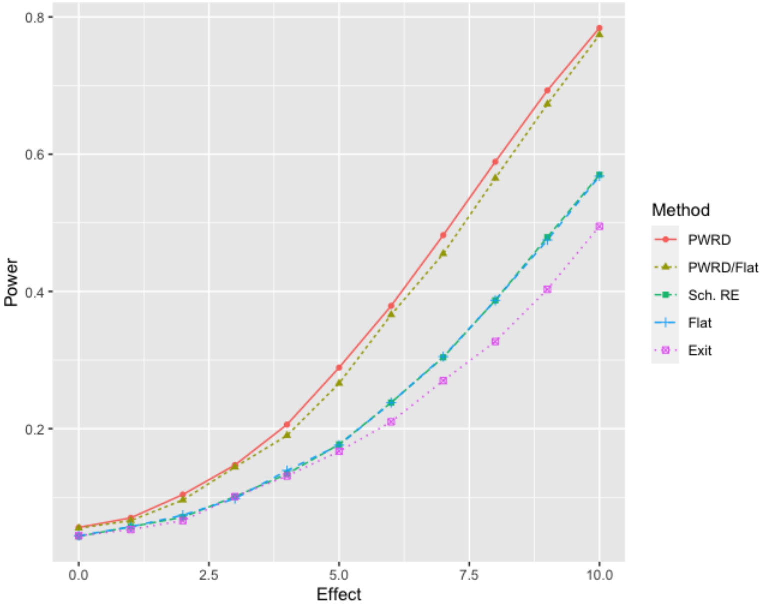

Figure 1: Power for the three methods under Effect 1 across increasing effect sizes when the theory of change holds. PWRD/Flat denotes the combination of the methods through the max t procedure.

correlation (ICC). We present these results in Figure 2.

In Figure 2, PWRD aggregation consistently outperforms the standard methods 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 35% improvement in power. Lastly, the max t combination of PWRD aggregation and flat weighting offers substantially greater power than the standard methods do on their own.

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 2.1 with Condition 2.5 and thus Proposition 2.4 with Proposition 2.6. We chose 40% to ensure the overall effect is positive in aggregate.

In Figure 3, we observe that under the relaxed assumption, PWRD aggregation performs even better in comparison to the traditional 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

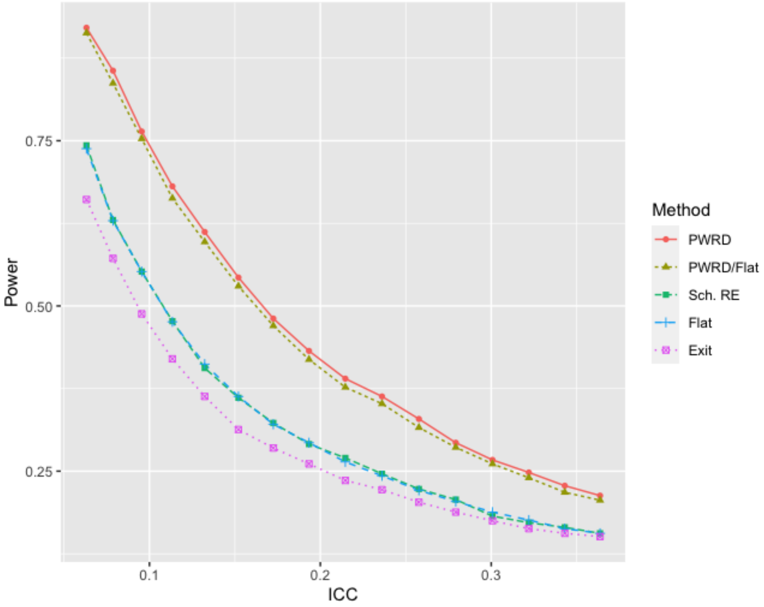

Figure 2: Power for the three methods under Effect 1 with increasing intraclass correlations. PWRD/Flat denotes the combination of the methods through the max t procedure.

greater importance to those more likely to have received a *positive* effect. None of the other models 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 roughly 30% 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. Once again, the max t combination of PWRD aggregation and flat weighting greatly outperforms the traditional alternatives although not to the extent of standard PWRD aggregation.

The phenomenon present in Figure 3 holds when the magnitude of the negative effect varies as well. We observe this in Figure 4. 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.

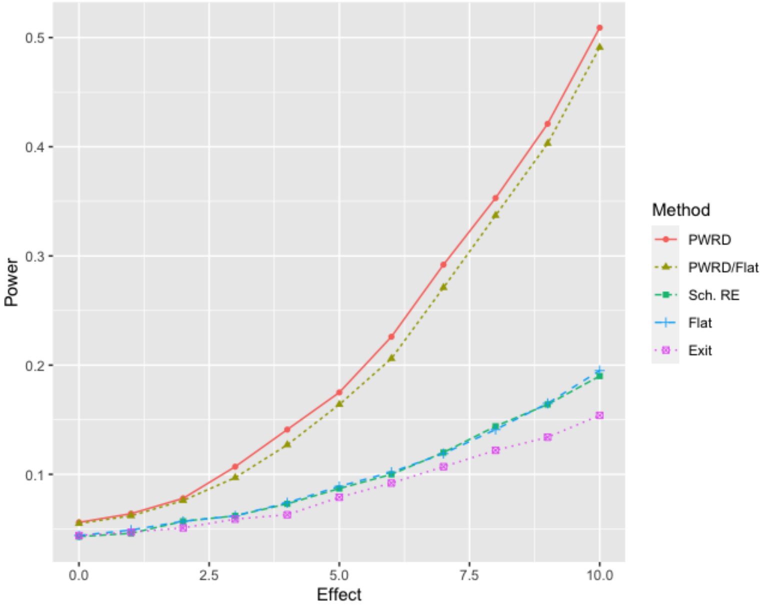

Figure 3: Power for the three methods under Effect 2 across increasing effect sizes when Condition 2.1 does not hold and is replaced with Condition 2.5. PWRD/Flat denotes the combination of the methods through the max t procedure.

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 5.

We observe that while the standard methods 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. In this scenario, we did not require the additional protection against the theory of change failing offered by simultaneously implementing PWRD aggregation and flat weighting through the max t procedure.

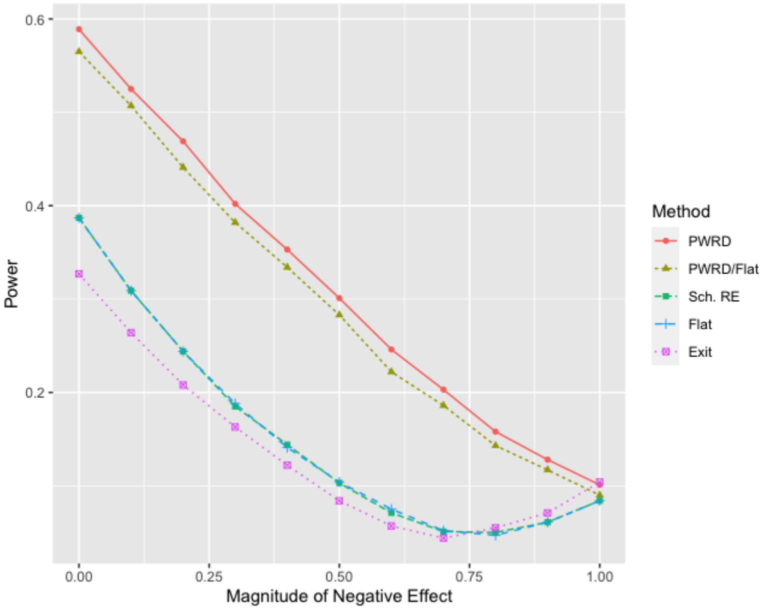

Figure 4: Power for the three methods under Effect 2 with increasingly negative effects. Here we add a positive effect of size 8 to students in the intervention and a negative effect that increases from 0% to 100% of the positive effect. PWRD/Flat denotes the combination of the methods through the max t procedure.

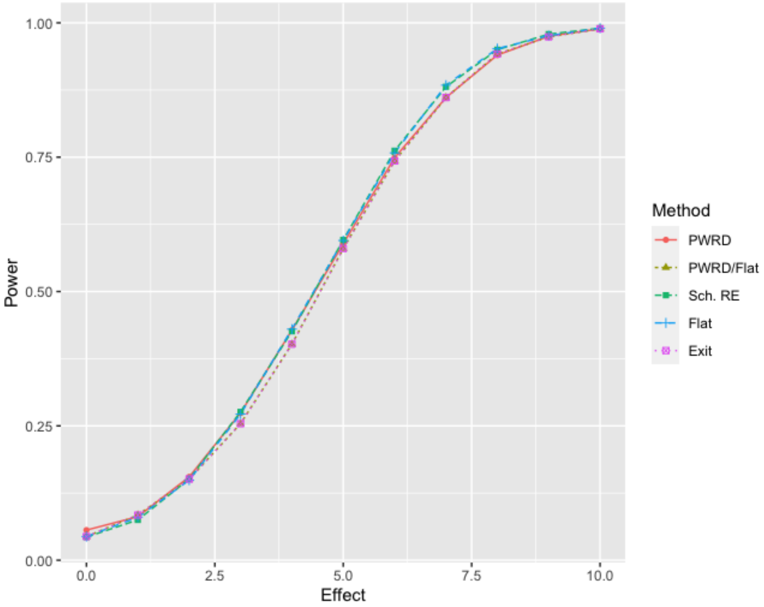

Figure 5: Power for the three methods under Effect 3, i.e. across increasing effect sizes when none of Conditions 2.1, 2.2, or 2.5 hold. PWRD/Flat denotes the combination of the methods through the max t procedure.

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. 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 2.1, 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 DIBELS scores and end-of-year assessment scores for each student. DIBELS 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.

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 2.2.1 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 4. We then calculate Σ with the grouping of control-group residuals described in greater detail in Section 2.2. 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 5.

Method	**Est.**	**S.E.**	**t value**	**Sig.**	**Test Slope**
Exit	9.88	9.72	1.02	-	0.082
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 5: BURST results on a subset of Cohort 1.0 for various methods, including PWRD aggregation.

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.

4.2 BURST[R]: Reading

We now conduct the same analysis described 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 covariate adjustment. Results are presented in Table 6.

Method	**Est.**	**S.E.**	**t value**	**Sig.**	**Test Slope**
Exit	-1.10	3.25	-0.34	-	0.189
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 6: BURST results for various methods, including PWRD aggregation.

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 asymptotic relative efficiency of PWRD aggregation versus exit observation analysis, flat weighting, and mixed effects modeling was 1.30, 2.02, and 1.78 respectively. This suggests that we would have required over 15, 52, and 40 additional schools in BURST in order to achieve the same power we possessed under PWRD aggregation with these alternatives.

5 Discussion

The strategy of using a regression coefficient to conduct a hypothesis test is standard in settings across the social sciences. This approach assists in implementation of commonly used methods like exit observation analysis, flat weighting, and mixed effects models. Nonetheless, these conventional regressions may prove to be suboptimal in any given scenario because they fail to account for which observations are most likely to benefit from the treatment. In this paper, we have presented a novel method of aggregation that takes advantage of that structure by converting an intervention's theory of change into statistical power for a broad class of interventions. We have shown both mathematically and through a simulation study that when the theory of change holds, PWRD aggregation provides far greater power than extant alternatives.

This method is broadly applicable in education settings, where suitable theories of change are expected, for independent reasons, in competitions for desirable research funding. We demonstrated extraction from a theory of change of the weights needed for PWRD aggregation, with a theory of change likely to be typical of interventions providing supplemental instruction. In it and similar circumstances, the subgroups on which we apply PWRD aggregation weights (e.g. years of follow-up) are determined independently of the proportion of students eligible for supplemental instruction; the estimated proportions only factor into the aggregation weights themselves. Thus, the method solely requires researchers possess some measure of the proportion of students who are non-excluded in each subgroup on either the treatment group or the control group, rather than both. Only in settings where the level of non-excluded (i.e. the proportion who stand to benefit) serves to delineate the subgroups on which we apply

PWRD aggregation do we require measurement of non-exclusion rates for both the treatment and the control groups.

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. To further protect against any potential loss of power in settings where the theory of change fails, PWRD aggregation may be used in tandem with standard estimation techniques like exit observation analysis or flat weighting through a max t procedure, as described in Section 2.3. This additionally provides a standard ITT estimate of the treatment effect that PWRD aggregation on its own does not provide, while yielding far greater power than the traditional method yields when the theory of change holds.

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.

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.

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.

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.

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

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

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.

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.

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.

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.

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

Hothorn, T., Bretz, F., and Westfall, P. (2008). Simultaneous Inference in General Parametric Models. Biometrical Journal: Journal of Mathematical Methods in Biosciences, 50(3):346–363.

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.

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.

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.

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.

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.

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.

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

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.

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. (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.

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.

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.

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.

Appendices

A Proof of Proposition 2.4

Consider the parameter $\Delta_{agg} = \mathbb{E}(\sum_{g,t} \omega_{gt}\hat{\Delta}_{gt}) = \omega'\Delta$ where Δ_{gt}, and thus Δ_{agg}, follow a proportionality assumption, i.e. $\Delta_{gt} \propto \eta p_{0gt}$. The variance of $\omega'\Delta$ satisfies $\mathrm{Var}(\sum_{g,t} \omega_{gt}\hat{\Delta}_{gt}) = \omega'\Sigma_\Delta\omega$, where Σ_Δ denotes the covariance of effects across cohort-years $\{g, t\}$, and is assumed fixed at a common value across hypotheses K_η, $-\infty < \eta < \infty$.

Now examine the test statistic $\sum_{g,t} \omega_{gt}\hat{\Delta}_{gt}$, the argument for the other forms being similar. Our problem is to select $\omega = (\omega_{1,1}, \ldots, \omega_{G,T}) \geq 0$ that maximizes the test slope of $\sum_{g,t} \omega_{gt}\hat{\Delta}_{gt}$ which in turn will maximize the asymptotic relative 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):

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

where $\Delta'_{agg}(0)$ denotes the derivative at zero of a function of the form $d \mapsto \Delta(d)$. The corresponding asymptotic relative 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 asymptotic relative efficiency $\big(h(\omega_{opt})/h(\omega_{alt})\big)^2$. Under flat weighting, $\omega_{alt_{gt}} := n_{gt}/N$, where n_{gt} denotes the number of observations in cohort g during year of follow-up t and N denotes the total number of observations.

A.1 Determining the Optimum ω_{opt}

We would like to determine which ω maximizes the test slope in (A.1). Under the assumption that $\Delta_{g,t} \propto \eta p_{0gt}$, then $\Delta'_{gt}(0) \propto p_{0gt}$ as well. Thus, to determine which ω maximizes the test slope in (A.1), we maximize the following:

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

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

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

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}:

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

A.2 Estimation of ω_{opt}

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

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

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 A.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 $\hat{\omega}_{opt} = \Sigma_\Delta^{-1}\hat{\mathbf{p}}_0$.

A.3 ω_{opt} With a Non-Negativity Constraint

In Equation A.3, we wished to maximize $f(\omega) = \log(\omega'\mathbf{p}_0) - \frac{1}{2}\log(\text{Var}(\omega'\hat{\Delta}))$. We now add in two constraints to prevent $\omega_g < 0$. In particular, we would now like to find $\max_\omega \log(\omega'\mathbf{p}_0) - \frac{1}{2}\log(\text{Var}(\omega'\hat{\Delta}))$ such that $\omega_{gt} \geq 0 \ \forall \ \{g, t\}$ and $\mathbf{1}'\omega = 1$. In other words, we would like to maximize ω such that each ω_{gt} is non-negative and $\sum_{g=1}^{G}\sum_{t=1}^{T}\omega_{gt} = 1$. This is equivalent to solving: $max_\omega \log(\omega'\mathbf{p}_0) - \frac{1}{2}\log(\text{Var}(\omega'\hat{\Delta})) - u'\omega + v'\omega$.

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:

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

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 A.2.

B PWRD Aggregation and Type I Errors

In Section 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 share of non-excluded observations, 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 B.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 2.3.*

Condition B.2. $\sqrt{n}(\hat{\Delta} - \Delta) \to_d N\big(\mathbf{0}, \mathrm{Cov}(\Delta)\big)$.

With Conditions 2.3, B.1 and B.2, we formulate a simple proposition about the distribution of the test statistic specified in Equation 2.5.

Proposition B.3. *Take fixed aggregation weights* w*. Under the null hypothesis* H_0 *and when Conditions 2.3, B.1, and B.2 hold,*

$$\frac{\sum_{g,t} w_{gt}\hat{\Delta}_{gt} - \sum_{g,t} w_{gt}\delta_{0gt}}{(w'\mathrm{Cov}(\hat{\Delta})w)^{1/2}} \to_d N(0,1).$$

Proposition B.3 states that with a consistent estimator of the covariance and an estimator that is asymptotically multivariate normal, the test statistic specified in Equation 2.5 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, $\hat{v}^{1/2}$, present in Equation 2.5 and Section 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 B.1.

While Proposition B.3 allows us to determine the asymptotic distribution of test statistics with the form in Equation 2.5 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 $\sum_{g,t} \hat{\omega}_{gt}\hat{\Delta}_{gt}$ includes an auxiliary statistic: $\hat{\omega}_{gt}$. 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 $\hat{\Delta}_{gt}$ by aggregation weights $\hat{\omega}$, yet does not incorporate the covariance between each $\hat{\omega}_{gt}$. To address this issue, we first present a mild condition on $\hat{\mathbf{p}}_0$.

Condition B.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 B.4 is immediate from the Weak Law of Large Numbers. Conditions 2.3, B.1, and B.4 allow us to circumvent our standard error not incorporating additional variation from $\hat{\omega}$ through Proposition B.5.

Proposition B.5. *Consider t-statistics of the form*

$$\frac{(\sum_{g,t} \hat{\omega}_{gt} \hat{\Delta}_{gt} - \sum_{g,t} \hat{\omega}_{gt} \delta_{0gt})}{(\hat{\omega}' \widehat{\text{Cov}}(\hat{\Delta}) \hat{\omega})^{1/2}}, \tag{B.1}$$

where $\hat{\omega} = (\widehat{\text{Cov}}[\hat{\Delta}]^{-1} \hat{\mathbf{p}}_0)_+ \Big/ \sum_j (\widehat{\text{Cov}}[\hat{\Delta}]^{-1} \hat{\mathbf{p}}_0)_{+_j} \in [0,1]$ *represents weights for PWRD aggregation. Under Conditions 2.3, B.1, and B.4, the difference between* (B.1) *and*

$$\frac{(\sum_{g,t} \omega_{gt} \hat{\Delta}_{gt} - \sum_{g,t} \omega_{gt} \delta_{0gt})}{(\omega' \text{Cov}(\hat{\Delta}) \omega)^{1/2}},$$

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

$$\left[\frac{\sum_{g,t} \hat{\omega}_{gt} \hat{\Delta}_{gt} - \sum_{g,t} \hat{\omega}_{gt} \delta_{0gt}}{(\hat{\omega}' \widehat{\text{Cov}}(\hat{\Delta}) \hat{\omega})^{1/2}} - \frac{\sum_{g,t} \omega_{gt} \hat{\Delta}_{gt} - \sum_{g,t} \omega_{gt} \delta_{0gt}}{(\omega' \text{Cov}(\hat{\Delta}) \omega)^{1/2}} \right] \to_P 0. \tag{B.2}$$

Simply, Proposition B.5 states that the t-statistic centered around $\sum_{g,t} \hat{\omega}_{gt} \delta_{0gt}$, 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. B.3 and covered by Prop. 2.4, which is centered around the parameter $\sum_{g,t} \omega_{gt} \delta_{0gt}$ and scaled by the sampling s.d. of $\sum_{g,t} \omega_{gt} \hat{\Delta}_{gt}$. 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 B.5 can be found in Appendix B.1, the following subsection.

B.1 Proof of Proposition B.5

To show Proposition B.5, we begin by showing that $\|\hat{\omega} - \omega\|_2 \to_P 0$. Writing $\hat{\Sigma} := n\widehat{\text{Cov}}(\hat{\Delta})$, Condition B.1 says $\|\hat{\Sigma} - \Sigma\|_2 = o_P(1)$. Because Σ is positive-definite (Condition 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 B.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.3) below the left denominator converges to the denominator at the right, a positive constant:

$$\frac{\sqrt{n}(\sum_{g,t}\hat{\omega}_{gt}\hat{\Delta}_{gt} - \sum_{g,t}\hat{\omega}_{gt}\delta_{0gt})}{(\hat{\omega}'n\widehat{\text{Cov}}(\hat{\Delta})\hat{\omega})^{1/2}} - \frac{\sqrt{n}(\sum_{g,t}\omega_{gt}\hat{\Delta}_{gt} - \sum_{g,t}\omega_{gt}\delta_{0gt})}{(\omega'n\text{Cov}(\hat{\Delta})\omega)^{1/2}}. \tag{B.3}$$

Noting that (B.3) is equivalent to the left-hand side of (B.2) in the statement of the Proposition, we just need to show that $\sqrt{n}\big[\sum_{g,t}\hat{\omega}_{gt}\hat{\Delta}_{gt} - \sum_{g,t}\hat{\omega}_{gt}\delta_{0gt} - \sum_{g,t}\omega_{gt}\hat{\Delta}_{gt} + \sum_{g,t}\omega_{gt}\delta_{0gt}\big] \to_P 0$, which is equivalent to showing $\sqrt{n}\big[\sum_{g,t}(\hat{\omega}_{gt} - \omega_{gt})(\hat{\Delta}_{gt} - \delta_{0gt})\big] \to_P 0$. 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, $n^{1/2}[(\hat{\omega} - \omega)(\hat{\Delta} - \delta_0)] = o_P(1)$.