



## Conjuring Power from a Theory of Change: The PWRD Method for Trials with Anticipated Variation in Effects

Timothy Lycurgus, Ben B. Hansen & Mark White

To cite this article: Timothy Lycurgus, Ben B. Hansen & Mark White (2022): Conjuring Power from a Theory of Change: The PWRD Method for Trials with Anticipated Variation in Effects, Journal of Research on Educational Effectiveness, DOI: [10.1080/19345747.2022.2142178](https://doi.org/10.1080/19345747.2022.2142178)

To link to this article: <https://doi.org/10.1080/19345747.2022.2142178>



View supplementary material [↗](#)



Published online: 12 Dec 2022.



Submit your article to this journal [↗](#)



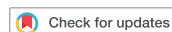
View related articles [↗](#)



View Crossmark data [↗](#)



This article has been awarded the Centre for Open Science 'Preregistered' badge.



# Conjuring Power from a Theory of Change: The PWRD Method for Trials with Anticipated Variation in Effects

Timothy Lycurgus<sup>a</sup>, Ben B. Hansen<sup>b</sup>, and Mark White<sup>c</sup>

<sup>a</sup>d3center, Institute for Social Research, University of Michigan, Ann Arbor, MI, USA; <sup>b</sup>Department of Statistics, University of Michigan, Ann Arbor, MI, USA; <sup>c</sup>Department of Teacher Education and School Research, Quality in Nordic Teaching (QUINT) Center, Universitetet i Oslo, Oslo, Norway

## ABSTRACT

We present an aggregation scheme that increases power in randomized controlled trials and quasi-experiments when the intervention possesses a robust and well-articulated theory of change. Intervention studies using longitudinal data 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 *power-maximizing weighting for repeated-measurements with delayed-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 doubling the number of clusters in the experiment.

## ARTICLE HISTORY

Received 26 July 2021  
Revised 19 September 2022  
Accepted 5 October 2022


## KEYWORDS

Enhanced statistical precision; logic model; repeated measures design; varying length of follow-up

## 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" (IES, 2020, p. 48): How and why does a desired improvement in outcomes occur as a consequence of the intervention? That is, what is the logic model of how effects accumulate and which students are expected to benefit?

**CONTACT** Timothy Lycurgus  [tlycugu@umich.edu](mailto:tlycugu@umich.edu)  d3center, Institute for Social Research, University of Michigan, Ann Arbor, MI, USA.

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/19345747.2022.2142178>

© 2022 Taylor & Francis Group, LLC

This paper introduces a scheme, PWRD aggregation of effects, for converting theories of change into statistical power for randomized controlled trials and quasi-experiments. Given an efficacious program, a theory of change that correctly identifies where effects are likely to concentrate, and measurements indicating which students stand to benefit, this power-maximizing weighting for repeated measurements with *delayed* effects method increases the probability of detecting program benefits. It maintains the canonical intention-to-treat (ITT) perspective on program benefits. The method is primarily designed to assist with hypothesis testing rather than with estimation, yet it may be implemented in tandem with standard estimation techniques. PWRD aggregation is applicable when there are baseline or post-treatment measures of intervention delivery or availability, in combination with primary outcomes measured on varying numbers of occasions. It offers dramatic improvements of power over both fixed effects and hierarchical linear models (HLMs)—at least, over conventional uses of these regression techniques that do not make use of side information mapping expected accrual of program benefits. It is easy to combine with these and other regression methods that are commonly used for analysis of education RCTs.

We illustrate PWRD aggregation 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, 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 cohorts of students 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 envionring theory of the intervention.

In underlying concept if not in its goals, the method relates to instrumental variables estimation (Angrist et al., 1996; Baiocchi et al., 2014; Bloom, 1984) and principal stratification (Frangakis & Rubin, 2002; Page, 2012; Sales & 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). For recent methodology proposing different weighting schemes to aggregate average treatment effect on the

treated (ATT) estimates into an overall effect estimate, see Callaway and Sant'Anna (2021) and Sun and Abraham (2021).

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 supporting 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 “Simulations,” we present a simulation study mirroring BURST design to show PWRD aggregation performance in comparison with commonly used methods under various assumptions. In Section “PWRD Analysis Findings,” we then illustrate how PWRD aggregation compares with those same methods for BURST itself. Finally, in Section “Discussion,” 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.

## Method

### *Review: Comparative Studies with Repeated Measurements of the Outcome*

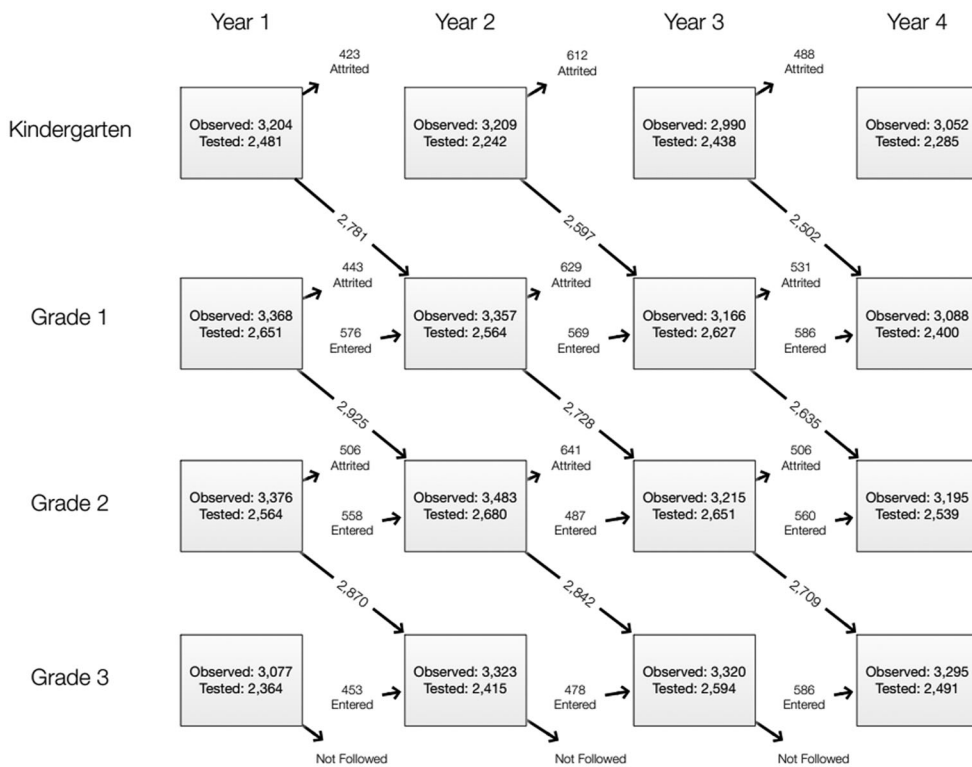
In educational settings assessing the efficacy of interventions, students frequently enter and exit studies at different points. For example in BURST, we examined a reading intervention on early elementary students across four years. Depending on their grade at the study's outset, the number of observations on each student varied from one to four. Figure 1 illustrates this phenomenon for BURST's various cohorts. In this paper, we refer to subcohorts as “X.Y,” where X and Y denote, respectively, the year and grade in which the cohort entered the study. For example, Cohort 1.3 is the set of students who entered the study in the first year as third graders.

Data sources for similarly structured efficacy trials will incorporate an analogous design as BURST, with varying numbers of observations on any given participant. The method chosen to handle multiple observations in these longitudinal studies is of great importance. 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 BURST in Figure 1, this entails using data from the bottom row and discarding the remaining observations. This method, herein termed “exit observation” analysis, treats the student rather than the student-year as the unit of analysis. Exit observation analysis typically uses models such as

$$Y_{ij3} = \beta_0 + \tau Z_{ij3} + \beta X_{ij3} + \epsilon_{ij3} \quad (\mathbb{E}(\epsilon_{ij3}) = 0; \text{Var}(\epsilon_{ij3}) = \sigma^2), \quad (1)$$

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 logic model, 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.K.



**Figure 1.** The path of students and cohorts throughout the four years of the BURST intervention.

**Table 1.** Differences in mean reading scores between treatment and control groups for the first of four cohorts of students. The final column ( $\delta$ ) gives the difference in these differences as calculated for the final year of participation versus the first year of participation.

	Entry Grade	Entry Year	Exit Year	$\delta$
Cohort 1	3	5.2	5.2	–
	2	– 1.3	0.4	0.9
	1	0.3	3.2	2.9
	K	– 7.0	2.1	9.1

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 1, 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 Cohort 1.K, 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 and then estimating standard errors of these coefficients with appropriate attention to “clustering” by student or by school; in mixed modeling and general estimating equations literature, this is known as the linear model with “working independence structure” (Fox, 2015; Laird, 2004). These analyses effectively attach equal

weight to each student-year observation and thus we refer to them as “flat” weights. In combination with least squares, flat weighting delivers minimum-variance unbiased coefficient estimates under the model that

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

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, while clustering ensures consistency of standard errors by taking into account heterogeneity across groups. Model 2 differs from the exit-observations-only model, Model 1, in allowing multiple values of  $k$  for each student  $i$ ; in BURST,  $k$  ranges from one to four under flat weighting. While Model 2 is general, other flat weighting identification strategies may differ. For example, one specification of flat weighting may include fixed effects for years and interactions between the treatment and year. An example of flat weighting may be found in Meece and Miller (1999).

With multiple observations per student, Model 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 (Raudenbush & Bryk, 2002) during 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 1 and Model 2, we may represent the two-level mixed effects model appropriate to analysis of BURST within the single regression equation

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

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  $\mu_j \sim N(0, v)$ . 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 (ATE): the expected difference in outcomes among third grade students. Using repeated measures of the outcome would seem to remove that possibility. The overall ATE 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 possesses 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 (2018), who in turn extended the work of Bell and McCaffrey (2002).

### **PWRD Aggregation**

The three estimation methods presented in Section “Review: Comparative Studies with Repeated Measurements of the Outcome” 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 those methods, leverages the intervention’s logic model 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. the cohort beginning in the top left box of [Figure 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 will not restrict the sample to high-implementation schools when estimating treatment effects. Rather we use this subset as an example that best illustrates 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, the subgroups are directly observable and refer to the cohort year of follow-up. Yet this too relates to dosage levels as the theory of change suggests that different subgroups have varying levels of exposure to the treatment: those students who participate for longer are more likely to receive a greater dose of supplemental instruction. 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 the subset of Cohort 1.K during each year of follow-up are presented in [Table 2](#). Note that since this is an ITT analysis, all student observations are used, not merely those exposed to the treatment.

As a departure from Sales and Pane (2021) however, PWRD aggregation serves as the tool by which we aggregate the four estimated effects in the course of hypothesis testing. This aggregate, similar to the two-way fixed effects difference-in-differences estimator (TWFEDD) (Goodman-Bacon, 2021), need not correspond to an independently



**Table 2.** Estimated change in outcome in each year of follow-up for a subset of Cohort 1 that entered the study in Grade K.

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.** The proportion of students in Cohort 1 who entered the study in grade K who have “tested in” to BURST to receive supplemental instruction by how long they have participated in the study.

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

meaningful average of individual effects. However, unlike TWFEEDD, PWRD aggregation does estimate its target estimand even if that estimand itself is not easily interpretable. Therefore, this formulation simultaneously allows us to sidestep the debates reviewed in Section “Review: Comparative Studies with Repeated Measurements of the Outcome” 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.

PWRD aggregation is particularly beneficial in terms of power versus extant methods for the analysis of 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 receive the supplemental instruction. The longer an individual participates in an intervention of this nature, the greater the likelihood that they become eligible to benefit from it, but prior to that occurrence, they are “excluded” from benefitting from the intervention.

Table 3 shows that for Cohort 1.K, student eligibility for the BURST intervention indeed increased in step with longer participation in the study. This holds both for those students belonging to treatment schools and for those students attending control schools. Accordingly, the working model describing how effects accumulate 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  $\mathbf{p} := (p_{gt} : g, t)$ , where  $p_{gt} := \mathbb{P}(\text{An individual in cohort } g \text{ is eligible to receive the supplemental instruction by year of follow-up } t)$ . Thus  $\mathbf{p}$  represents the proportion of students who were not *excluded* from having been affected, in virtue of the assumed exclusion restriction.

The expected size of the effect as estimated through  $\hat{\mathbf{p}}$  is not the only consideration of PWRD aggregation. Define  $\Delta_{gt}$  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 treatment and  $Z=0$  denotes assignment to control. Suppose corresponding ITT estimators  $\{\hat{\Delta}_{gt} : g, t\}$  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.) Now let us define:

$$\Sigma := \text{Cov}(\hat{\Delta}_{g,t} : g, t),$$

the relative covariances among  $\{\hat{\Delta}_{gt} : g, t\}$ . This factor, the estimated relative covariances  $\hat{\Sigma}$ , will contribute to our method as well, with those effect estimates that are relatively precise and uncorrelated with the other estimates receiving greater weight. This relates to precision weighting where, for example, estimates with smaller variances also receive greater weight (Raudenbush & Bryk, 2002). Unlike precision weighting, however, PWRD aggregation accounts for the correlations among cohort-year ATE estimates as well, which are often substantial.

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

$$\hat{\Delta}_{agg} := \sum_{g,t} w_{gt} \hat{\Delta}_{gt}, \quad (3)$$

with specially chosen weights  $w$  ( $w_{gt} \geq 0$ , all  $g, t$ ;  $\sum_{g,t} w_{gt} = 1$ ). To find the specific  $w$  that maximizes power to detect an effect, we first make multiple assumptions about the nature of the treatment, given the theory of the intervention is correct:

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

The second portion of 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  $t$ .

From these conditions, we now construct PWRD aggregation (formally presented in Proposition A.2 in [Appendix A](#)). Take the null hypothesis that there is no effect of the intervention compared to an alternative hypothesis that there is an effect and that effect is proportional to the dosage received:

$$\begin{aligned} H_0 : \Delta &= \mathbf{0} \\ H_a : \Delta &= \eta \mathbf{p}, \eta > 0 \end{aligned} \quad (4)$$

Now take test statistics centered around the aggregation of separate cohort-year ATEs, i.e.  $\hat{\Delta}_{agg} = \sum_{g,t} w_{gt} \hat{\Delta}_{gt}$ . Given the above conditions hold, the asymptotic relative efficiency and thus, the power of these test statistics will be maximized using weights of the following form:

$$\mathbf{w} = (\Sigma^{-1} \mathbf{p})_+ / \sum_j (\Sigma^{-1} \mathbf{p})_{+j}, \quad (5)$$

where  $(\Sigma^{-1} \mathbf{p})_+$  denotes the element-wise maximum of  $(\Sigma^{-1} \mathbf{p})$  and  $\mathbf{0}$ , and  $(\cdot)_{+j}$  denotes the  $j$ th element of  $(\cdot)_+$  such that  $\mathbf{w}' \mathbf{1} = 1$ . Note that Conditions 2.1 and 2.2 are not required for the validity of the hypothesis test, but are necessary for finding an efficient estimator of  $\Delta_{agg}$ .

In sum, so long as the effect is proportional to the share of non-excluded observations, the “signal-to-noise” ratio of test statistics centered around  $\hat{\Delta}_{agg}$  will be maximized by weights proportional both to the expected sizes of cohort-year effects,  $\mathbf{p}$ , and also to the relative precisions of estimated cohort-year effects,  $\Sigma$ :  $\mathbf{w} \propto \Sigma^{-1} \mathbf{p}$ . Any test statistic of the form

$$\frac{\sum_{g,t} w_{gt} \hat{\Delta}_{gt} - \sum_{g,t} w_{gt} \delta_{0gt}}{\hat{v}^{1/2}}, \quad (6)$$

such as the  $t$ -statistic combining estimates  $\hat{\Delta}_{gt}$  with fixed weights  $w_{gt}$ , will be covered by Proposition A.2.

Using PWRD aggregation weights in place of some other set of weights provides test statistics with greater asymptotic relative efficiency. Improving relative efficiency by 20% corresponds to a 20% reduction in the sample size required to achieve the same level of power (Van der Vaart, 2000). Thus, test statistics incorporating PWRD aggregation weights provide researchers with a greater opportunity to detect an effect of the intervention when the working model of how effects accumulate holds. For a formal description of asymptotic relative efficiency with respect to PWRD aggregation, see [Appendix A](#). While designed to assist with hypothesis testing, the 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 selecting  $\mathbf{w}$  to maximize an approximation known as “test slope” of the expected value of  $\hat{\Delta}_{agg}$ , as given in [Equation 3](#). 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](#).

While we have presented PWRD aggregation using a one-sided test, there is nothing to prevent researchers from applying a two-sided test. In fact, if the magnitude of the negative effect increases with greater exposure, then PWRD aggregation with a two-sided test would once again increase power to detect that negative effect.

### PWRD Aggregation in the BURST Evaluation

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

Neither  $\mathbf{p}$  nor  $\Sigma$  is directly observed, but both can be estimated easily. We estimate  $p_{gt}$  through the proportion  $\hat{p}_{gt}$  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,  $\hat{\mathbf{p}}$  is pretreatment in the sense that treatment assignment does not affect it. That is,  $\mathbf{p}$  is defined in terms of potential outcomes under the control.

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  $\mathbf{p}$  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.

Note that if we expected that treatment eligibility and thus, exposure differed between treatment and control groups (perhaps because treatment schools had greater incentive to provide DIBELS), we instead could have calculated PWRD aggregation weights using the proportion  $\hat{p}_{1gt}$  observed among students assigned to the treatment. Nonetheless, we had evidence to suggest this was not the case (Rowan et al., 2019).

Often  $(\hat{\Delta}_{g,t} : g, t)$  are estimates from a common regression fit, in which case an accompanying estimate of the covariance of coefficient estimates can be used to estimate  $\Sigma$  in Equation 5. Our analysis of BURST used the Peters-Belson (Peters, 1941; Belson, 1956) method and called for a somewhat more elaborate calculation centered around control-group residuals (Hansen & Bowers, 2009). For a more thorough explanation, see Rowan et al. (2019).

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 (2018). 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  $\Delta_{gt}$  either through design-based approaches outlined in Lin (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.

### ***Considerations When the Logic Model Fails***

When the theory of change correctly identifies which students will benefit from the intervention and at what point that will occur, PWRD aggregation maximizes the asymptotic relative efficiency of this method versus extant alternatives for the family of hypotheses  $K_\eta : \Delta = \eta \mathbf{p}$ . That is in BURST, if 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 the intervention 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 hypothesized by the theory of change, 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).

We address the second question both conceptually and through simulations presented in Section “Simulations.” The constant effects assumption behind base PWRD aggregation is rather strong, yet this is merely our “working model” of the treatment effect. The working model is informed by the intervention’s logic model. In BURST, for example, it is posited that effects only accrue to students who receive the supplemental instruction. The working model, however, goes beyond the logic model with the simplifying assumption of a constant effect. So long as the working model is roughly accurate (i.e. the effect is proportional to the exposure, an assumption of the underlying theory of the intervention) we believe that PWRD aggregation will provide a benefit.

For example, effects may increase in magnitude with greater exposure. In this scenario, PWRD aggregation will still provide a gain to power because all else equal, the method emphasizes cohort-years with greater exposure which in turn have a greater opportunity to benefit from the increasing effect. This occurs despite a violation of Condition 2.1.

Instead, effects may decrease over time. Here, the working model is incorrect. Yet in the context of BURST, PWRD aggregation may still provide a benefit if the intervention is successfully implemented. To illustrate, take the set of students who are eligible and benefit immediately after receiving the supplemental instruction. If that effect diminishes, then they would once again require supplemental instruction and thus, once again receive the benefit (assuming the intervention does, in fact, work). Therefore, cohort-years with greater cumulative exposure would likely have larger effects and PWRD aggregation would yield more power. On the other hand, if students who initially

receive a benefit gradually lose that boost to their reading performance and do not benefit a second time, a scenario where PWRD aggregation may be harmful, we would argue that the intervention does not work and we should not detect an effect. The goal of BURST is to improve stalled reading abilities and if BURST is only a temporary palliative, then it has failed at achieving its aims.

Our simulation study empirically examines the benefits and drawbacks of PWRD aggregation by comparing the power of  $t$ -tests based on Equation 2.3's  $\hat{\Delta}_{agg}$ , with weights  $\mathbf{w}$  as given by Equation 5, to  $t$ -tests based on flat-weighted, exit observation weighted or random effect-adjusted 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 the working model is correct and  $\Delta_{gt}$  is proportional to  $p_{gt}$ .

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 the intervention holds, PWRD aggregation yields substantially more power than extant alternatives; if instead effects accumulate in a manner different than that hypothesized by the theory of the intervention, the standard method protects against a large loss of power. The step-down Dunnett procedure (Dunnett & Tamhane, 1991; 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 step-down Dunnett procedure will provide power close to the power of any one of the correlated statistics.

In fact, this step-down procedure need not solely use test statistics from standard PWRD aggregation and an analysis method typical to these settings. Researchers could instead include a third test statistic using a variant of PWRD aggregation that takes into account a different working model of the treatment effect, e.g. a working model that posits effects diminish or increase over time.

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

### 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; Gottfried, 2013; Vanderweele et al., 2013), often calling for specialized methods or other accommodations (Bowers et al., 2018; Rosenbaum, 2007; Sobel, 2006; Vanderweele et al., 2013). 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 2.3.** *Individual  $i$  receiving supplemental instruction due to the intervention at time  $j$  gains non-negative effect  $\tau$  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.*

This allows for a corollary to standard PWRD aggregation (i.e. Proposition A.2). The formal presentation of the corollary may be found in [Appendix A.3](#). Simply, the corollary argues that 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 A.3 follows the same structure as the derivation of Proposition A.2 found in [Appendix A](#).

## Simulations

In order to demonstrate how PWRD aggregation performs in comparison to exit observation analysis, flat weighting, mixed effects models, and a PWRD aggregation/flat weighting combination using the step-down Dunnett 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:

$$\begin{aligned} Y_{ijk} &= \beta_0 + \beta_1 \text{Grade}_{ijk} + \mu_k + \epsilon_{ijk}, \\ \mu_k &= \gamma_0 + \nu_k \end{aligned} \quad (7)$$

with  $\nu_k \sim N(0, \xi)$ . The outcome of student  $i$  in year of follow-up  $j$  at school  $k$  is a function of the grade of the student and the random intercept of the school at which the student is enrolled,  $\mu_k$ . 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 eligible for supplemental instruction. Once a student tests in to receive the treatment, 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 the methods under three different treatment effects. Under the first, all treatment observations flagged as eligible for supplemental instruction receive some constant, positive effect  $\tau$ . Under the second, flagged treatment observations receive a constant, positive effect  $\tau$  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 1000 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 1000 iterations. We use cluster-robust standard errors with clusters at the school level from the `clubSandwich` package in R (Pustejovsky & Tipton, 2018).

## Simulation Results

We 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  $0.05\sigma$  (1.2 points in our simulation study) as small, those between  $0.05\sigma$  and  $0.2\sigma$  (4.7 points) as moderate, and those greater than  $0.2\sigma$  as large. Across 1,260 effect sizes on reading outcomes from 495 RCTs, the mean effect size was  $0.17\sigma$  (4 points) and the 90th percentile was  $0.5\sigma$  (11.8 points) (Kraft, 2020). Thus, our simulation study examines these methods on effect sizes that frequently appear in reading interventions.

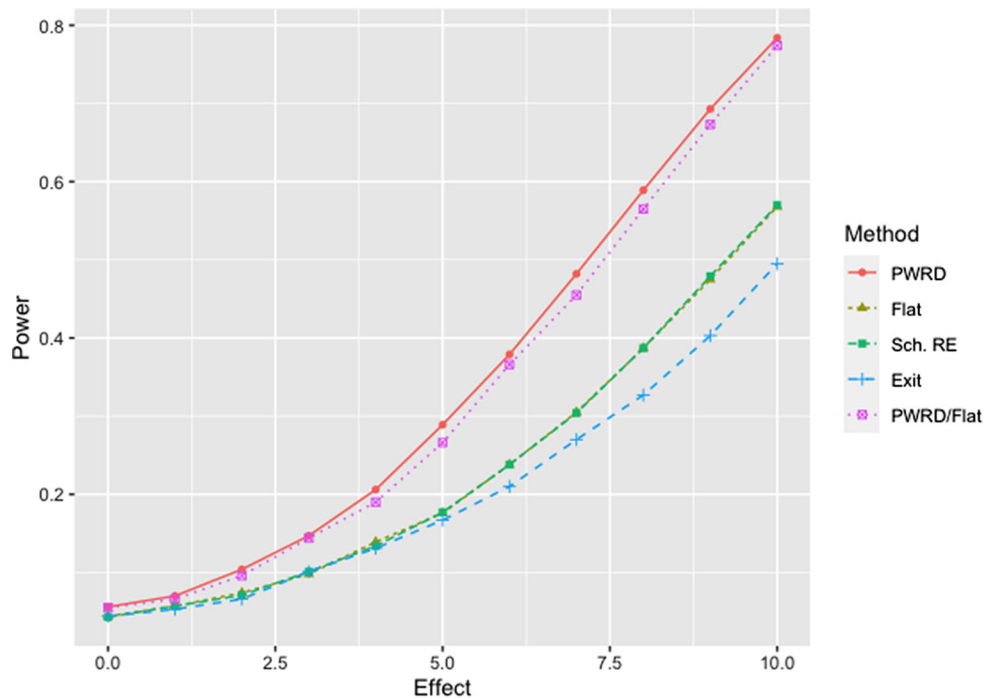
### Effect 1

Figure 2 shows the power from 1000 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 step-down Dunnett combination of PWRD aggregation and flat weighting offers less power than PWRD aggregation, it still yields far greater power than the standard methods yield. Thus, this method should prove attractive both when the analyst wishes to be protected against a loss of power if the theory of change is incorrect or when the analyst wishes to both test the null hypothesis and estimate the treatment effect using a standard approach.

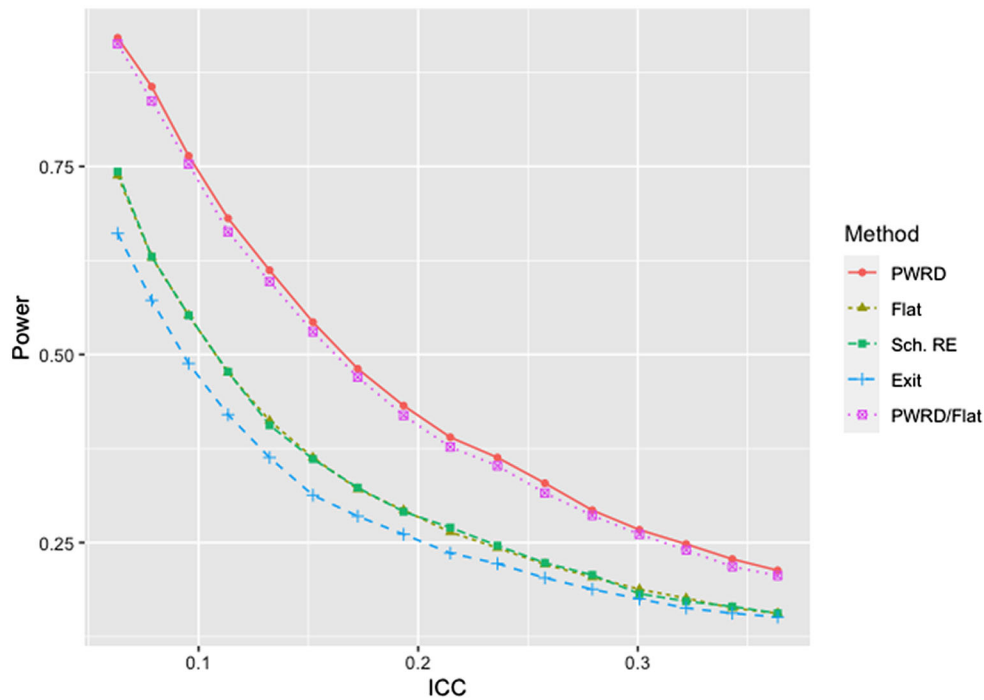
It is natural to ask whether the gains in power present in Figure 2 hold across different levels of correlation of observations within a school. To examine this we conducted additional simulations holding the imposed effect constant, but varying the intraclass correlation (ICC). We present these results in Figure 3.

In Figure 3, PWRD aggregation consistently outperforms the standard methods across ICCs that typically arise in educational settings (Hedges et al., 2007). For intraclass

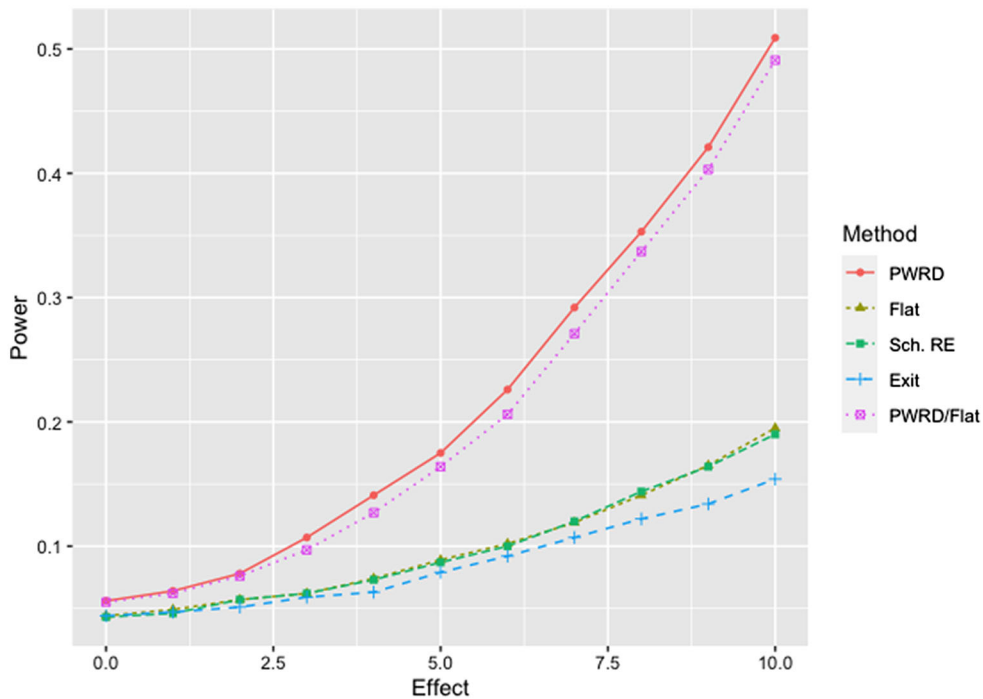




**Figure 2.** Power for the four methods under Effect 1 across increasing effect sizes when the theory of change holds. PWRD/Flat denotes the combination of the methods through the step-down Dunnett procedure.



**Figure 3.** Power for the four methods under Effect 1 with increasing intraclass correlations. PWRD/Flat denotes the combination of the methods through the step-down Dunnett procedure.



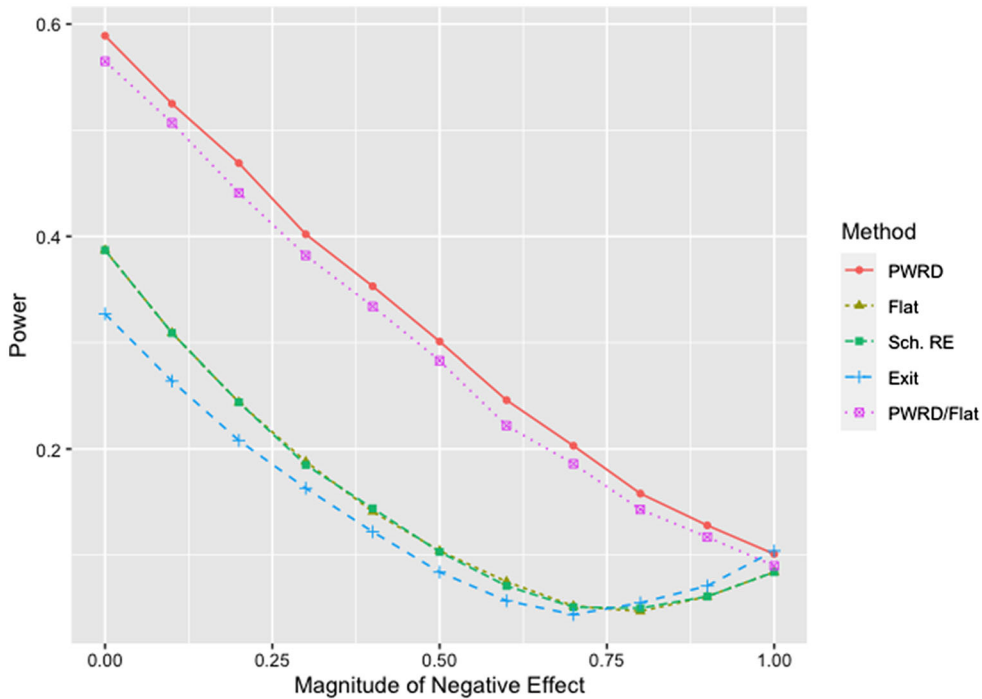
**Figure 4.** Power for the four methods under Effect 2 across increasing effect sizes when Condition 2.1 does not hold and is replaced with Condition 2.3. PWRD/Flat denotes the combination of the methods through the step-down Dunnett procedure.

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 step-down Dunnett combination of PWRD aggregation and flat weighting offers substantially greater power than the standard methods do on their own.

### 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, which corresponds to replacing Condition 2.1 with Condition 2.3 and thus Proposition A.2 with Proposition A.3. We chose 40% to ensure the overall effect is positive in aggregate.

In Figure 4, 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 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 widens as the effect size increases. For example, our



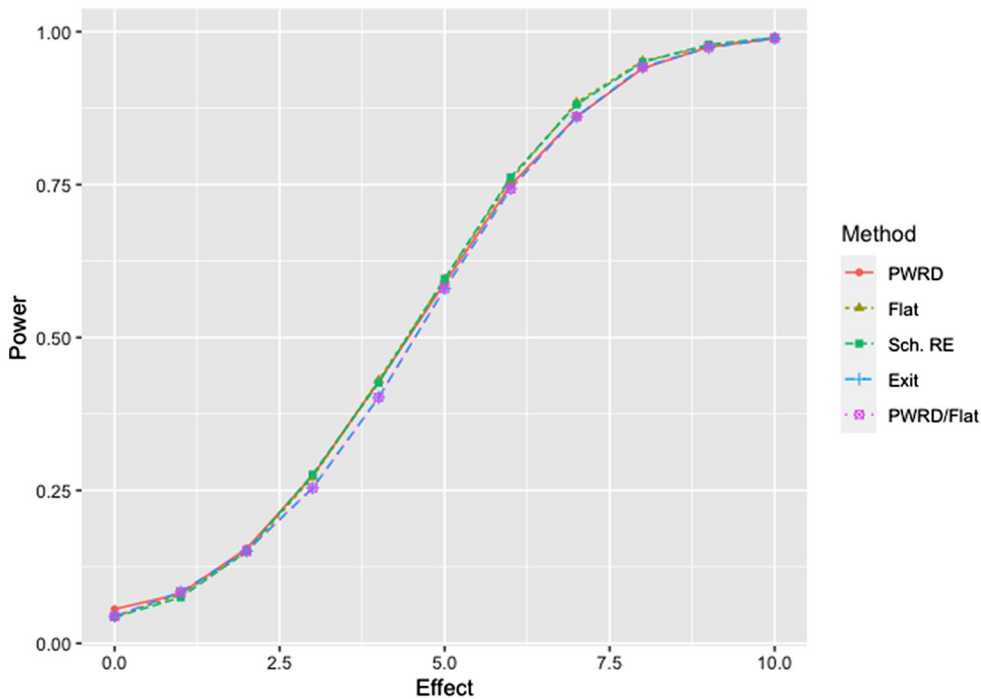
**Figure 5.** Power for the four methods under Effect 2 with increasingly negative effects. Here we add a positive effect of size 8 to students in the intervention and a negative effect that increases from 0% to 100% of the positive effect. PWRD/Flat denotes the combination of the methods through the step-down Dunnett procedure.

method more than doubles the power of mixed effects models and flat weighting for large effect sizes. Once again, the step-down Dunnett 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 4 holds when the magnitude of the negative effect varies as well. We observe this in Figure 5. Under this scenario, the size of the benefit remains constant. Instead, the adverse effect for those treatment students who do not test into the intervention varies from 0% of the benefit to 100% of the benefit. PWRD aggregation provides a persistent 15-20 percentage point advantage in power for negative effects up to 60% of the positive effect before narrowing out. This corresponds to at least a 40% improvement in power for all magnitudes of the negative effect; under certain circumstances, the method provides double the power. When the negative effect is equal in magnitude to the positive effect, PWRD aggregation no longer provides a benefit.

### 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



**Figure 6.** Power for the four methods under Effect 3, i.e. across increasing effect sizes when none of Conditions 2.1, 2.2, or 2.3 hold. PWRD/Flat denotes the combination of the methods through the step-down Dunnett procedure.

just that it does not work as hypothesized by the theory of the intervention. Instead, effects may accumulate in a different fashion. Here, we impose an artificial treatment effect on all treatment observations such that  $\tau_{ijk} \sim N(l, 2.5 * l)$  for  $l = 1, \dots, 10$ . 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 6.

We observe that while the standard methods outperform PWRD aggregation, this improvement is minimal and never exceeds 3%. For example, with an imposed effect of size 5 (on the border between a moderate and large effect), the standard methods accurately reject a null hypothesis of no effect 59.6% of the time. PWRD aggregation, on the other hand, rejects the null hypothesis 58.9% of the time. For effect sizes greater than 6 (roughly  $0.25\sigma$ ), we are able to reject frequently under any of the schemes. From these simulations, it is clear that PWRD aggregation provides substantial gains in power in situations when the theory of the intervention holds. While an effect accumulating in a manner similar to Effect 3 may be more common than either Effect 1 or Effect 2, our simulations show only marginal decreases in its ability to reject the null hypothesis when the assumptions fail. In this scenario, we did not require the additional protection against a failure in the intervention's theory offered by simultaneously implementing PWRD aggregation and flat weighting through the step-down Dunnett procedure.

## PWRD Analysis Findings

This section presents results for BURST, both on Cohort 1.K and on the overall randomized trial, using PWRD aggregation and commonly applied alternative methods. The theory behind BURST was presented in Section “Method.” 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 27000 unique students on 1–4 occasions each, for a total of 52,000 student-year observations. The length of time for which each student participated in the RCT depended on the grade and year at 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 pretest. The end-of-year assessments were our primary outcome of interest. The BURST reading intervention study was conducted under a University of Michigan IRB exemption.

### *Burst Cohort 1.K*

In this section, we first show how the aggregation weights  $\hat{\mathbf{w}}$  were generated before we present the results themselves. In order to calculate  $\hat{\mathbf{w}}$ , we estimate  $\mathbf{p}$  and  $\Sigma$ . We know from Section “PWRD Aggregation in the BURST Evaluation” that we estimate  $\mathbf{p}$  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 3. We then calculate  $\Sigma$  through a grouping of control-group residuals described in greater detail in Rowan et al. (2019). We then formulate:

$$\hat{\mathbf{w}} = (\Sigma^{-1}\mathbf{p})_{+} / \sum_j (\Sigma^{-1}\mathbf{p})_{+j} = (0.25, 0, 0.32, 0.43),$$

where the weights correspond to the first through fourth years of follow-up respectively. Note that while more students were eligible for supplemental instruction by the second year than by 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 suggests that outcome analysis would be best served by attaching no weight to those observations.

We then employ a Peters–Belson (Belson, 1956; Peters, 1941) 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  $\hat{\mathbf{w}}$  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 used to predict treatment outcomes. The differences between the fitted and observed values serve to estimate the average treatment effect. Results appear in [Table 4](#).

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 gives a relatively large  $t$  statistic, conceivably because students in their fourth year of follow-up, i.e. in third grade, were best situated to benefit from BURST.

**Burst[R]: Reading**

We now conduct the same analysis described previously, yet using the complete data from BURST. For PWRD aggregation, we calculate separate effect estimates and aggregation weights for each cohort-year. As with analysis on Cohort 1.K, we employ a Peters-Belson approach to covariate adjustment. Results are presented in [Table 5](#).

None of these methods detect an effect of BURST on student achievement: unfortunately, this program would appear not to have provided a benefit. A possible explanation for the apparent lack of an effect is that schools possessed limited resources; more students required supplemental instruction than schools had the ability to serve at levels recommended by the theory of the intervention (Rowan et al., 2019). Despite the theory of change not appearing to have held, PWRD aggregation still provides valid standard errors and a valid hypothesis test. This continues to hold if the intervention’s effect is negative.

Nonetheless, if the theory of change were correct, 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 using these alternatives.

**Table 4.** BURST results on a subset of Cohort 1 students who entered the study in grade K for various methods, including PWRD aggregation.

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 for various methods, including PWRD aggregation.

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

## 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 identifying relevant logic models and converting them into statistical power. In the motivating example, an intervention providing supplemental reading instruction, the salient logic model entailed that benefits would accrue only after the student's learning trajectory had stalled. We have shown both mathematically and through a simulation study that when the working model of how effects accumulate is accurate, PWRD aggregation provides far greater power than extant alternatives.

This method is applicable in education settings, where suitable theories of change are expected in research funding competitions. We demonstrated how to extract the weights needed for PWRD aggregation from a theory of change that is likely to be typical of interventions providing supplemental instruction. In it and similar circumstances, the method requires measurements of intervention delivery or exposure by cohort year. PWRD aggregation is constructed around theories of an intervention that should be determined *a priori*, so the method is consistent with pre-registration of analysis plans for increased transparency in outcome analysis.

While PWRD aggregation is optimal when its supporting theory of change holds, no benefit is gained when that theory is incorrect. Nonetheless, the 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 when the working model fails, PWRD aggregation may be used in tandem with standard estimation techniques like exit observation analysis or flat weighting through a step-down Dunnett procedure, as described in Section "Considerations When the Logic Model Fails." In this variation of the method, flat weighting or exit observation analysis contributes a standard ITT estimate of the treatment effect for increased interpretability, PWRD aggregation contributes a more efficient estimator that yields greater power than the traditional method when effects accumulate in the hypothesized manner, and the Dunnett procedure preserves type 1 error rates despite multiple testing of a single null hypothesis. PWRD aggregation can be adapted to other scenarios, both experimental and quasi-experimental, with longitudinal data and a suitably theorized treatment that accrues heterogeneously across observations.

## Acknowledgements

The authors wish to acknowledge valuable comments and suggestions from Johann Gagnon-Bartsch, Sophia Rabe-Hesketh, Brian Rowan, Adam Sales, Anders Skrondal and three anonymous reviewers. Responsibility rests with the authors for any shortcomings that may remain.



## Open Research Statements

### *Study and Analysis Plan Registration*

The study and analysis are registered on the Registry of Efficacy and Effectiveness Studies (Study #473: <https://sreereg.icpsr.umich.edu/sreereg/subEntry/2127/pdf?section=all&action=download>).

### *Data, Code, and Materials Transparency*

The data, code, and materials that support the findings of this study are not publicly available.

### *Design and Analysis Reporting Guidelines*

This manuscript was not required to disclose use of reporting guidelines, as it was initially submitted prior to JREE mandating open research statements in April 2022.

### *Transparency Declaration*

The lead author (the manuscript's guarantor) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

### *Replication Statement*

This manuscript reports an original study.

## Open Scholarship



This article has earned the [Center for Open Science](#) badge for Preregistered through Open Practices Disclosure. The materials are openly accessible at <https://sreereg.icpsr.umich.edu/sreereg/subEntry/2127/pdf?section=all&action=download>.

## Funding

This work was supported by the National Science Foundation (DMS1646108) and the Institute for Education Sciences (R305A120811, R305D210029). The opinions expressed are those of the authors and do not represent views of the Foundation, the Institute or the U.S. Department of Education.

## References

- Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434), 444–455. <https://doi.org/10.1080/01621459.1996.10476902>
- Baiocchi, M., Cheng, J., & Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in Medicine*, 33(13), 2297–2340. <https://doi.org/10.1002/sim.6128>
- Bell, R. M., & 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*, 5(3), 195–202. <https://doi.org/10.2307/2985420>
- Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review*, 8(2), 225–246. <https://doi.org/10.1177/0193841X8400800205>
- Bowers, J., Desmarais, B. A., Frederickson, M., Ichino, N., Lee, H.-W., & Wang, S. (2018). Models, methods and network topology: Experimental design for the study of interference. *Social Networks*, 54, 196–208. <https://doi.org/10.1016/j.socnet.2018.01.010>
- Callaway, B., & Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2), 200–230. <https://doi.org/10.1016/j.jeconom.2020.12.001>
- Cox, D. (1958). *The planning of experiments*. John Wiley.
- Dunnett, C. W., & Tamhane, A. C. (1991). Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts. *Statistics in Medicine*, 10(6), 939–947. <https://doi.org/10.1002/sim.4780100614>
- Ethington, C. A. (1997). *A hierarchical linear modelling approach to studying college effects*. Higher Education (vol. 12, pp. 165–194). Agathon Press Incorporated.
- 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. <https://doi.org/10.1002/pam.20479>
- Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage Publications.
- Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58(1), 21–29. <https://doi.org/10.1111/j.0006-341X.2002.00021.x>
- Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2), 254–277. <https://doi.org/10.1016/j.jeconom.2021.03.014>
- Gottfried, M. A. (2013). The spillover effects of grade-retained classmates: Evidence from urban elementary schools. *American Journal of Education*, 119(3), 405–444. <https://doi.org/10.1086/669851>
- Guo, S. (2005). Analyzing grouped data with hierarchical linear modelling. *Children and Youth Services Review*, 27(6), 637–652. <https://doi.org/10.1016/j.childyouth.2004.11.017>
- Hansen, B. B., & Bowers, J. (2009). Attributing effects to a cluster-randomized get-out-the-vote campaign. *Journal of the American Statistical Association*, 104(487), 873–885. <https://doi.org/10.1198/jasa.2009.ap06589>
- 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. <https://doi.org/10.1080/01621459.1986.10478354>
- Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal. Biometrische Zeitschrift*, 50(3), 346–363. <https://doi.org/10.1002/bimj.200810425>
- 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.
- Institute of Education Sciences (IES). (2020). Education Research Grant Program Request for Applications, CFDA number: 84.305 A. [https://ies.ed.gov/funding/pdf/2021\\_84305A.pdf](https://ies.ed.gov/funding/pdf/2021_84305A.pdf)
- Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4), 241–253. <https://doi.org/10.3102/0013189X20912798>
- Laird, N. (2004). Analysis of longitudinal and cluster-correlated data. In *NSF-CBMS Regional Conference Series in Probability and Statistics* (pp. 1–155). Institute for Mathematical Sciences.
- Lee, V. E. (2000). Using hierarchical linear modelling to study social contexts: The case of school effects. *Educational Psychologist*, 35(2), 125–141. [https://doi.org/10.1207/S15326985EP3502\\_6](https://doi.org/10.1207/S15326985EP3502_6)
- Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, 7(1), 295–318. <https://doi.org/10.1214/12-AOAS583>

- Meece, J. L., & 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. [https://doi.org/10.1207/s1532799xssr0303\\_2](https://doi.org/10.1207/s1532799xssr0303_2)
- Middleton, J. A., & Aronow, P. M. (2015). Unbiased estimation of the average treatment effect in cluster-randomized experiments. *Statistics, Politics and Policy*, 6(1–2), 39–75. <https://doi.org/10.1515/spp-2013-0002>
- 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. <https://doi.org/10.1080/19345747.2012.688410>
- 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. <https://doi.org/10.1080/00220671.1941.10881036>
- Pustejovsky, J. E., & Tipton, E. (2018). Small sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics*, 36(4), 672–683. <https://doi.org/10.1080/07350015.2016.1247004>
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (vol. 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. <https://doi.org/10.1093/biomet/88.1.219>
- Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association*, 102(477), 191–200. <https://doi.org/10.1198/016214506000001112>
- Rowan, B., Hansen, B. B., White, M., Lycurgus, T., & Scott, L. J. (2019). *A summary of the BURST [R]: Reading efficacy trial* (ED593875). ERIC.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688–701. <https://doi.org/10.1037/h0037350>
- Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75(371), 591–593. <https://doi.org/10.2307/2287653>
- Sales, A. C., & 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. <https://doi.org/10.1214/18-AOAS1196>
- Sales, A. C., & Pane, J. F. (2021). Student log-data from a randomized evaluation of educational technology: A causal case study. *Journal of Research on Educational Effectiveness*, 14(1), 241–269. <https://doi.org/10.1080/19345747.2020.1823538>
- Schochet, P. Z. (2013). Student mobility, dosage, and principal stratification in school-based RCTs. *Journal of Educational and Behavioral Statistics*, 38(4), 323–354. <https://doi.org/10.3102/1076998612458322>
- Simmons, D. C., Coyne, M. D., Kwok, O-m., McDonagh, S., Ham, B. A., & 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. <https://doi.org/10.1177/0022219407313587>
- 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. <https://doi.org/10.1198/016214506000000636>
- Splawa-Neyman, J., Dabrowska, D. M., & Speed, T. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4), 465–472.
- Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2), 175–199. <https://doi.org/10.1016/j.jeconom.2020.09.006>
- Van der Vaart, A. W. (2000). *Asymptotic statistics* (vol. 3, pp. 72–89). Cambridge University Press.

- Vanderweele, T. J., Hong, G., Jones, S. M., & 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. <https://doi.org/10.1080/01621459.2013.779832>
- White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4), 817–838. <https://doi.org/10.2307/1912934>
- White, M. C., Rowan, B., Hansen, B., & Lycurgus, T. (2019). Combining archival data and program-generated electronic records to improve the usefulness of efficacy trials in education: General considerations and an empirical example. *Journal of Research on Educational Effectiveness*, 12(4), 659–684. <https://doi.org/10.1080/19345747.2019.1636438>

## Appendices

### A. Presentation of Proposition A.2 and Its Proof

First take the following technical condition that simplifies the development by excluding pathological cases.

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

**Proposition A.2.** Consider test statistics of the forms:  $\sum_{g,t} w_{gt} \hat{\Delta}_{gt}$ , with  $g$  and  $t$  ranging over cohorts and times of follow-up respectively;  $\sum_{g,t} w_{gt} \hat{\Delta}_{gt} - \sum_{g,t} w_{gt} \delta_{0gt}$ , where  $\delta_0$  is a vector of hypothesized values of  $\Delta$ ,  $\Delta := (\Delta_{gt} : g, t)$ ; and  $\hat{v}^{-1/2}(\sum_{g,t} w_{gt} \hat{\Delta}_{gt} - \sum_{g,t} w_{gt} \delta_{0gt})$ , where  $\hat{v}$ , perhaps an estimate of  $\text{Var}(\sum_{g,t} w_{gt} \hat{\Delta}_{gt})$ , satisfies  $n\hat{v} \rightarrow_p c > 0$ . Consider the family of statistical hypotheses  $\{K_\eta : \Delta = \eta \mathbf{p}, \eta \geq 0\}$ . Under Conditions 2.1, 2.2, and A.1, and for tests of  $H_0 = K_0$  against alternatives  $K_\eta$ ,  $\eta > 0$ , asymptotic relative efficiency is maximized by

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

(In this display,  $(\Sigma^{-1} \mathbf{p})_+$  denotes the element-wise maximum of  $(\Sigma^{-1} \mathbf{p})$  and  $\mathbf{0}$ , and  $(\cdot)_{+j}$  denotes the  $j$ th element of  $(\cdot)_+$  such that  $\mathbf{w}' \mathbf{1} = 1$ .)

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

Now examine the test statistic  $\sum_{g,t} w_{gt} \hat{\Delta}_{gt}$ , the argument for the other forms being similar. Our problem is to select  $\mathbf{w} = (w_{1,1}, \dots, w_{G,T}) \geq 0$  that maximizes the test slope of  $\sum_{g,t} w_{gt} \hat{\Delta}_{gt}$  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):

$$h(\mathbf{w}) = \frac{\Delta'_{agg}(0)}{\text{Cov}_0^{1/2}(\mathbf{w}' \hat{\Delta})} = \frac{\Delta'_{agg}(0)}{[\mathbf{w}' \Sigma_\Delta \mathbf{w}]^{1/2}}, \quad (\text{A.1})$$

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  $\mathbf{w}$  may be represented by  $(h(\mathbf{w}_1)/h(\mathbf{w}_2))^2$ . The form of the two test statistics is identical; they merely incorporate different aggregation weights  $\mathbf{w}$ . Thus, it follows that finding  $\mathbf{w}_{opt}$ , where  $\mathbf{w}_{opt}$  maximizes the test slope, will also maximize the asymptotic relative efficiency  $(h(\mathbf{w}_{opt})/h(\mathbf{w}_{alt}))^2$ . Under flat weighting,  $w_{alt_{gt}} := n_{gt}/N$ , 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 $W_{opt}$

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

$$\max_{\mathbf{w}} \frac{\mathbf{w}'\mathbf{p}}{\text{Var}^{1/2}(\mathbf{w}'\hat{\Delta})}. \quad (\text{A.2})$$

We first transform A.2 logarithmically which is equivalent to maximizing  $f(\mathbf{w}) = \log(\mathbf{w}'\mathbf{p}) - \frac{1}{2} \log(\text{Var}(\mathbf{w}'\hat{\Delta}))$ .

To maximize, we take the gradient of  $f(\mathbf{w})$  and set the gradient equal to the zero-vector,  $\mathbf{0}$ , i.e.  $\nabla f(\mathbf{w}) : \frac{\mathbf{p}'}{\mathbf{w}'\mathbf{p}} - \frac{\mathbf{w}'\Sigma_{\Delta}\mathbf{w}}{\mathbf{w}'\Sigma_{\Delta}\mathbf{w}} = \mathbf{0}$ . Note that both  $\mathbf{w}'\mathbf{p}$  and  $\mathbf{w}'\Sigma_{\Delta}\mathbf{w}$  are scalars, so we can rewrite this as  $(\mathbf{w}'\mathbf{p})^{-1}\mathbf{p}' - (\mathbf{w}'\Sigma_{\Delta}\mathbf{w})^{-1}\mathbf{w}'\Sigma_{\Delta} = \mathbf{0}$ . We now rearrange the terms to solve for  $\mathbf{w}_{opt}$ :

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

For the proof with the non-negativity constraint, see the [online Supplemental Appendix A](#).

### A.2. Estimation of $W_{opt}$

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

$$\hat{\mathbf{w}}_{opt} = \left( \frac{\mathbf{w}'\Sigma_{\Delta}\mathbf{w}}{\mathbf{w}'\hat{\mathbf{p}}} \right) \hat{\mathbf{p}}'\Sigma_{\Delta}^{-1}. \quad (\text{A.3})$$

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

### A.3. PWRD Aggregation with Interference

**Proposition A.3.** *Under Conditions 2.2, 2.3, and A.1, the following aggregation weights  $\mathbf{w}$  will maximize the slope of test statistics discussed in Proposition A.2 for the family of hypothesis tests and alternative hypotheses also elaborated in Proposition A.2:*

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

## B. PWRD Aggregation and Type I Errors

In Section ‘‘PWRD Aggregation,’’ we demonstrated how PWRD aggregation maximizes the test slope and thus, the corresponding power for the family of hypotheses  $K_{\eta} : \Delta = \eta\mathbf{p}$ . 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  $\widehat{\text{Cov}}(\hat{\Delta})$  is consistent for  $\text{Cov}(\hat{\Delta})$ , in the sense that  $|\widehat{\text{Cov}}(\hat{\Delta}) - \Sigma|_2 \rightarrow_p 0$ , where  $\Sigma$  is as in Condition A.1.*

**Condition B.2.**  $\sqrt{n}(\hat{\Delta} - \Delta) \rightarrow_d N(\mathbf{0}, \text{Cov}(\Delta)).$

With Conditions A.1, 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  $\omega$ . Under the null hypothesis  $H_0$  and when Conditions A.1, B.1, and B.2 hold,

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

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  $\omega$  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  $\omega$ , PWRD aggregation does not incorporate fixed weights. Rather, two components of PWRD aggregation,  $\hat{\mathbf{p}}$  and  $\hat{\Sigma}$ , are random variables. Consequently, the aggregated statistic  $\sum_{g,t} \hat{w}_{gt} \hat{\Delta}_{gt}$  includes an auxiliary statistic:  $\hat{w}_{gt}$ . 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 (2018). Thus, our standard error scales the covariance between each  $\hat{\Delta}_{gt}$  by aggregation weights  $\hat{\mathbf{w}}$ , yet does not incorporate the covariance between each  $\hat{w}_{gt}$ . To address this issue, we first present a mild condition on  $\hat{\mathbf{p}}$ .

**Condition B.4.**  $\hat{\mathbf{p}}$  is root- $n$  consistent, i.e.  $\|\hat{\mathbf{p}} - \mathbf{p}\|_2 = O_P(n^{-1/2})$ .

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

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

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

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

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

where  $\mathbf{w} = (\Sigma^{-1} \mathbf{p})_{+} / \sum_j (\Sigma^{-1} \mathbf{p})_{+j}$ , is asymptotically negligible:

$$\left[ \frac{\sum_{g,t} \hat{w}_{gt} \hat{\Delta}_{gt} - \sum_{g,t} \hat{w}_{gt} \delta_{0gt}}{(\hat{\mathbf{w}}' \widehat{\text{Cov}}(\hat{\Delta}) \hat{\mathbf{w}})^{1/2}} - \frac{\sum_{g,t} w_{gt} \hat{\Delta}_{gt} - \sum_{g,t} w_{gt} \delta_{0gt}}{(\mathbf{w}' \text{Cov}(\hat{\Delta}) \mathbf{w})^{1/2}} \right] \rightarrow_P 0. \quad (\text{B.2})$$

Simply, Proposition B.5 states that the  $t$ -statistic centered around  $\sum_{g,t} \hat{w}_{gt} \delta_{0gt}$ , where  $\hat{\mathbf{w}} = (\hat{\Sigma}^{-1} \hat{\mathbf{p}})_{+} / \sum_j (\hat{\Sigma}^{-1} \hat{\mathbf{p}})_{+j}$ , 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. A.2, which is centered around the parameter  $\sum_{g,t} w_{gt} \delta_{0gt}$  and scaled by the sampling s.d. of  $\sum_{g,t} w_{gt} \hat{\Delta}_{gt}$ . 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 the [online Supplementary Appendix B](#).