Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

Using state data sets and meta-analysis of low-powered studies to evaluate a school-based dropout prevention program for students with disabilities

Tom Munk *, Ning Rui, William Zhu, Elaine Carlson

Westat, 1600 Research Boulevard Rockville, MD, 20850, USA

ARTICLE INFO

Keywords:
Meta-analysis
Dropout prevention
Impact evaluation
Difference-in-difference
Low-power studies

ABSTRACT

This study explores the use of state data sets and meta-analysis of low-powered studies to evaluate a school-based dropout prevention program for students with disabilities. The program was implemented in several states. A randomized controlled trial was infeasible because schools were not chosen at random; furthermore, pretest data were minimal. The use of extant state data allowed these obstacles to be overcome by providing valid pre- and post-intervention outcomes as well as a large selection of schools and variables to create reasonable matches for the treatment schools. Results from four states were synthesized meta-analytically to evaluate whether the program had a significant impact on any of seven proximal and distal outcome variables. No such impacts were demonstrated. More importantly, this paper demonstrates and explains the methodological steps and choices involved in a quasi-experimental evaluation approach that may be applied to cases for which large amounts of extant data are available.

1. The high school dropout problem

Dropping out of high school has serious negative outcomes for youth, including an increased likelihood of being unemployed, underemployed, dependent on welfare (Belfield & Levin, 2007a, 2007b), unhealthy (Hayes, Nelson, Tabin, Pearson, & Worthy, 2002), and incarcerated (Sanford et al., 2011; Stanard, 2003).[1] In recent years, high dropout rates and low graduation rates have generated attention at the national and state levels. The resulting increase in awareness about this crisis spawned a variety of federal efforts as well as several privately funded organizations dedicated to improving dropout and graduation rates in the United States. Concomitant with this attention and focused effort, the national 4-year adjusted cohort graduation rate increased from 79 percent in the 2010–11 school year[2] to 85 percent in the 2016–17 school year (McFarland et al., 2019). Additionally, the high-school status dropout rate in the United States decreased from 9.7 percent in 2006 to 5.3 percent in 2018 (McFarland et al., 2019).

Overall improvement has occurred; yet, some groups continue to fare worse than others in their school-completion outcomes. Specifically, although students with disabilities are also graduating more frequently, the adjusted cohort graduation rate was still only 67 percent in 2013–14, 18 percentage points less than the rate for all students, and the status dropout rate was 12.1 percent for youth with a disability versus 5.0 percent for youth without a disability in 2017 (National Center for Education Statistics, 2020).

One of the problems in addressing this gap is the relative dearth of evidence-based dropout-prevention interventions that focus on students with disabilities. In a May 2020 search, the What Works Clearinghouse (WWC) presented results for 46 interventions under the topic "Path to Graduation," of which 22 were "positive or potentially positive." However, when the additional filter of "children and youth with disabilities" was added, only one intervention was presented, "Check and Connect," and deeper investigation showed that it too, had zero studies "that fell within the scope of the Children Identified With Or At Risk For An Emotional Disturbance review protocol and met WWC evidence standards." More research is needed in this and many others areas of education, but conducting rigorous impact evaluations is challenging, even when considerable resources are committed to implementation of promising interventions.

To illustrate one approach to building more rigorous evaluations of existing initiatives, we report here on an evaluation of a national initiative that included intervention sites for implementing a dropout-

* Corresponding author at: 313 Sweet Bay Place, Carrboro, NC, 27510, USA.
E-mail address: tommunk@westat.com (T. Munk).

[1] Dropping out of school refers to students' departure from school prior to obtaining a high school credential

[2] The first school year that all states used a consistent, 4-year adjusted measure of school completion was 2010–11.

https://doi.org/10.1016/j.stueduc.2020.100969
Received 13 July 2018; Received in revised form 17 November 2020; Accepted 18 December 2020
Available online 31 December 2020

prevention intervention for students with and without disabilities. The intervention framework was based on cognitive behavioral intervention (CBI), and tailored to the needs of each site-based expert team in a data-driven process. This approach was backed by a research synthesis that supported the efficacy of CBI—across educational environments, types of disability, ages, and genders—in reducing dropout rates and addressing issues correlated with dropping out (Cobb, Sample, Alwell, & Johns, 2005). The intervention framework was used to support dropout-prevention efforts in high schools in several states. The purpose of this article is to describe an impact evaluation of the national initiative in the four states with mature programs and available data.

2. Strategies for evaluating school-level interventions and low-powered studies

The gold standard for evaluating the effectiveness of an intervention is the randomized controlled trial (RCT). Some examples are evident (Borman et al., 2005a, 2005b), but RCTs are often challenging and sometimes not feasible. Quasi-experiments can sometimes be successfully substituted for RCT. Recent publications about evaluation of whole-school interventions have supported or demonstrated the use of historical cohort control groups (Walser, 2014), comparative interrupted time series (St. Clair, Cook, & Hallberg, 2014), microanalytical simulation methods (Sondergeld, Beltyukova, Fox, & Stone, 2012), natural experiments with regression controls (Wong & Socha, 2008), and matching of treatment schools with nontreatment schools (Algozzine et al., 2012; Wong & Socha, 2008; Wong, Boben, Kim, & Socha, 2009).

Due to privacy concerns, logistics, and cost-related barriers to obtaining student-level data, researchers are increasingly turning to state-provided, school-level demographic and achievement data as an alternative source to estimate the effects of school-based interventions. These data are often easily accessible from state department of education websites. They typically provide a ready source for data on treatment schools and, most importantly, a wide selection of potential comparison schools. School-level data are often more than adequate to address key evaluation questions related to school-based interventions. Kaniuka, Vitale, and Romance (2013), for example, argued that the evidence on school-based reform will be strengthened through analysis of school-based pre- and post-intervention data in multisite and multi-year contexts. Although student-level data analyzed in a multilevel framework is preferred when the predictor variable is at the student level (Landeghem, Fraine, & Damme, 2005; Moerbeek, 2004), the concern is minimal when the predictor is at the group level. Even when individual- and school-level covariates are included in the model, the point estimates and standard errors from multilevel analyses of student- and school-level data are comparable to those from ordinary least squares regression models of school-level data (Jacob, Goddard, & Kim, 2014). This finding is consistent with both experimental and nonexperimental designs.

Table 1
Anticipated progression.

		Outcomes			
School year	Description	Attendance	Test scores	Dropout	Graduation
0	Training in late summer or early fall; school team develops implementation	0	0	0	0
1	Initial implementation	+	+	0	0
2	Mature implementation I	++	++	+	0
3	Mature implementation II	+++	+++	++	+
4	Equilibrating	++	++	+++	++
5	New equilibrium I	+	+	++	+++
6	New equilibrium II	+	+	+	++
7	New equilibrium III	+	+	+	+
8	Final equilibrium	+	+	+	+

Note: 0 = no effects; +=small positive effects; ++=moderate positive effects; +++=strong/maximal positive effects.

Low-powered studies are not ideal, but they are common. For example, the typical power of a psychological study does not exceed 0.5 (Cohen, 1962; Sedlmeier & Gigerenzer, 1989). Consequently, for such studies, researchers should expect to detect less than half of the real and meaningful effects that they are trying to detect. This high Type II error rate—false negatives— has led to a *replicability crisis* in psychology (Braver, Thoemmes, & Rosenthal, 2014).

There are many reasons for low-powered studies. In the competitive game of research publication, with a premium placed on significant results, collecting studies with small sample size and low power may be a more efficient research strategy than running one well-powered study (Bakker, van Dijk, & Wicherts, 2012). More relevant to evaluation is the fact that cases can be very expensive to obtain. In evaluation, budgets often allow for implementation of school reform efforts in only a limited number of schools. This limitation is true for the study represented by this paper; cases here are schools, and the intervention is implemented in each school, each of which represents a substantial investment.

If all of the schools in the current study were to be combined, the power of the study would be improved, but the fact that the schools were in different states means that the details of the program implementation and the data elements differed from state to state. It would have been inappropriate to combine the schools. There was, however, an alternative—meta-analysis.

Meta-analysis is most commonly thought of as a method of conducting a systematic review of literature and combining results from all relevant studies to identify patterns among the results (Glass, McGaw, & Smith, 1981). Effect sizes, such as Cohen's *d*, Hedges' *g*, or odds ratio, are computed for each study, thus standardizing the metric for the outcomes. After the effect size of each study is estimated, the overall mean weighted effect size is calculated. The weighting factor is usually the inverse of the variance of the study, so effect sizes from larger samples contribute more to the mean than those from smaller samples.

As Braver et al. (2014) pointed out, a meta-analytic approach can make use of low-powered studies; even studies deemed insignificant by traditional null-hypothesis significance testing can add strength to each other when combined. They propose a "continuously cumulating meta-analysis" approach to psychology's current replicability crisis, showing that, by combining results from replication attempts with original results via a fixed-effects meta-analysis, results that might be seen as failures to replicate "might nonetheless provide more, not less, evidence that the effect is real" (p. 333). They note that a similar approach can be used to "combine internal replications of multistudy articles" (p. 340). They suggest that when this approach is used, the combined studies should be tested for heterogeneity: Q-tests indicating significant variation in outcomes between studies or I^2 values indicating that more than half of the overall variance is between studies may bring into question the assumption that the true effect sizes are the same in the replications and force a search for the differences between the studies. These tests themselves, however, are extremely unreliable when the number of studies is small. (Borenstein, Hedges, & Rothstein, 2007).

In the years following Braver's article, internal meta-analyses of a single research team's similar studies has been advocated for researchers in psychology (Cumming, 2014; Goh, Hall, & Rosenthal, 2016) and consumer research (McShane & Böckenholt, 2017). Meta-analysis redirects attention toward effect sizes and away from individual studies' p-values. It allows similar studies to be joined, leveraging statistical power. It incorporates null findings instead of suppressing them, improving reliability, replicability, and transparency. Additionally, it allows for a simplified presentation of multiple studies (Goh et al.,

Table 2
Outcome variables used in the evaluation.

Variable	Description
Dropout Rate	The dropout rate calculation is the number of students (with disabilities) in Grades 9–12 with a withdrawal code corresponding to a dropout, divided by the number of students with disabilities in Grades 9–12 who attended the district. The number of students who attended the district is based on any student with disabilities who is reported in the student record.
Alternate Dropout Rate	Because one state provided one count of students (with disabilities) for the purposes of dropout calculations and another for the purposes of enrollment counts, and because these numbers do not always match, the researchers calculated an alternate dropout rate for students with disabilities using the enrollment count of students with disabilities as the denominator.
Graduation Rate	The graduation rate reflects the percentage of students (with disabilities) who entered Grade 9 in a given year and were in the graduating class 4 years later. The graduation rate is calculated by using information in the relevant student records. The graduation class size is the number of dropouts in Grades 9–12 from appropriate years, plus graduates, plus other completers.
Attendance Days	Attendance measures varied by state. They included the average number of attendance days by students with disabilities and students without disabilities in each school, the percentage of students absent for 15 days or more, and the attendance rate (days attended divided by total school days).
Reading Assessment Passing Rates	High school students take a variety of state-content area tests. Reading pass rates for students with disabilities and students without disabilities were used for this evaluation. The tests and passing thresholds differed from state to state as did the overall pass rates.

2016).

Other authors have raised cautions about internal meta-analysis (Ueno, Fastrich, & Murayama, 2016; Vosgerau, Simonsohn, Nelson, & Simmons, 2019). It is even more important for internal meta-analysis than for single studies that the studies be properly preregistered, that those preregistrations be followed in all essential aspects, and that the decision of whether to include a given study in an internal meta-analysis be made before any of those studies are run (Vosgerau et al., 2019). Violations of these principles increase false-positive rates beyond the nominal level (Ueno et al., 2016). Translating these concerns from the academic realm to the evaluation realm, it is even more essential with internal meta-analysis than with separate analyses for separate small studies that evaluation plans be thorough and followed assiduously. As evaluators know, this is just one more reason for the value of strong summative evaluation planning prior to extensive project activity.

3. Methods

This paper demonstrates the use of meta-analysis to combine several low-powered studies on the effects of the dropout prevention intervention for students with disabilities. Each study will be built on pre- and post-intervention data obtained from state data systems. Table 1 shows a theoretical framework representing the anticipated progression of outcomes based on the organization’s dropout prevention efforts. Proximal outcomes include increased attendance. Medial outcomes include improved test scores, and ultimate outcomes include decreased dropout rates and increased graduation rates. Outcomes are expected to build over time, taking as many as 5 years for maximum benefit in graduation rates. This study was designed to test whether the intervention demonstrated a desirable impact for both students with disabilities and students without disabilities (or the student body as a whole) on each of the outcomes. Outcome variables for the evaluation are described in Table 2. The availability and details of data for each outcome variable varied by state, which affected how each sub-study was designed.

Using extant state data to examine changes over time in dropout and graduation rates, as well as changes in more proximal outcomes (i.e., attendance and reading scores), in a set of quasi-experimental, matched-comparison-group difference-in-difference designs, we estimated the effect of services on all served schools in all four states that participated substantially in the intervention: State A, State B, State C, and State D. Importantly, the choices of variables and states were made prior to any analysis of data. We used all treated schools with the required data in each state and one-to-one matching with untreated schools in the state. We used meta-analysis to synthesize the results.

Comparison schools were selected using a variety of variables for students with and without disabilities, such as average school-level dropout rate, graduation rate, percentage of students passing state assessments, and attendance in the years before the start of treatment. Other matching variables used were school size, the percentage of students with disabilities, and the percentage of students who were Hispanic, Black, and eligible for free lunch[3] during the year that the treatment began. Matching variables differed slightly from state to state based on the availability of data. Before matching, treatment schools differed substantially from the rest of the schools in the state. Estimating treatment effects by comparing treatment groups with all of the schools in the state would have produced biased results because of the variation among schools. The matching procedures reduced, but did not eliminate, the standardized bias. More detail about matching procedures is provided in Appendix A in Supplementary material.

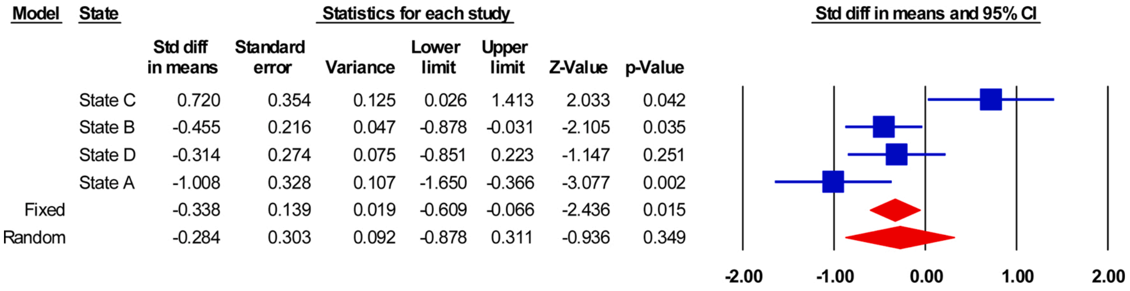
The meta-analysis was designed to achieve a summative evaluation purpose: Did the interventions, on average, have desirable impacts on seven key proximal and distal outcomes? In statistical terms, the meta-analysis had confirmatory, not exploratory goals. Combining results from multiple states increased sample size and therefore power. Separate meta-analyses were conducted for each of the outcome variables based on the availability of data, and results of the analyses were combined across states. The outcome variables were:

1 Dropout rates for students with disabilities in State A, State B, State C, and State D;
2 Dropout rates for all students in State A, State B, State C, and State D;
3 Graduation rates for students with disabilities in State A, State B, State C, and State D;
4 Graduation rates for all students in State A, State B, State C, and State D;
5 Attendance for students with disabilities in State A, State B, and State C;
6 State reading assessment passing rates for students with disabilities in State A and State B; and
7 State reading assessment passing rates for students without disabilities in State A and State B.

We calculated an effect size (Hedges’ g) for each outcome in each state, standardizing with the pooled pretest standard deviation as suggested by Morris (2008).[4] This represents 23 (State, Outcome) studies. Depending on the specific outcome variable, a positive or negative effect size may indicate that treated schools had better or worse outcomes. For example, a negative effect size for change in dropout rate indicated that treated schools did better than control schools in reducing dropout rates; however, a positive effect size for change in graduation rate indicated that treated schools did better than control schools in improving graduation rates.

[3] Free lunch eligibility is a powerful and commonly available predictor of educational outcomes (Domina et al., 2018).

[4] Morris’s (2008) study of the best estimates of effect size from pretest-posttest-control group designs “favored an effect size based on the mean pre-post change in the treatment group minus the mean pre-post change in the control group, divided by the pooled pretest standard deviation.”



Model	State	Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
	State C	0.720	0.354	0.125	0.026	1.413	2.033	0.042
	State B	-0.455	0.216	0.047	-0.878	-0.031	-2.105	0.035
	State D	-0.314	0.274	0.075	-0.851	0.223	-1.147	0.251
	State A	-1.008	0.328	0.107	-1.650	-0.366	-3.077	0.002
Fixed		-0.338	0.139	0.019	-0.609	-0.066	-2.436	0.015
Random		-0.284	0.303	0.092	-0.878	0.311	-0.936	0.349

Fig. 1. Forest plot of meta-analysis of program's effects on dropout rates for students with disabilities.

Means, standard deviations, and sample sizes for pre- and post-outcome data were used to estimate the standardized effect size, standard error, confidence interval, and p-value for each (State, Outcome) study. We report on and discuss significant results at this level for formative evaluation purposes: were there some state efforts that showed positive results for particular outcomes? Our summative results for each of the seven outcome variables were created meta-analytically. We combined results from up to four states for each outcome variable to create overall weighted mean effect sizes, standard errors, confidence intervals, and p-values. The weighting factor was the inverse variance for each (State, Outcome) pair so effect sizes from states incorporating more schools contributed more to the meta-analysis mean than did those from states with smaller samples. These steps are detailed in Appendix B in Supplementary material.

We used a Benjamini-Hochberg (B-H) adjustment to control the false discovery rate, restricting the expected share of significant findings where the true effect is zero (Type I errors) to five percent. We made these adjustments for the 23 formative (State, Outcome) studies, allowing us to confidently identify state efforts with positive outcomes for particular results. We made a separate set of adjustments for the seven outcomes at the summative meta-analytic level, allowing us to confidently identify the set of outcomes for which the school-based dropout prevention program had a consistently positive effect,

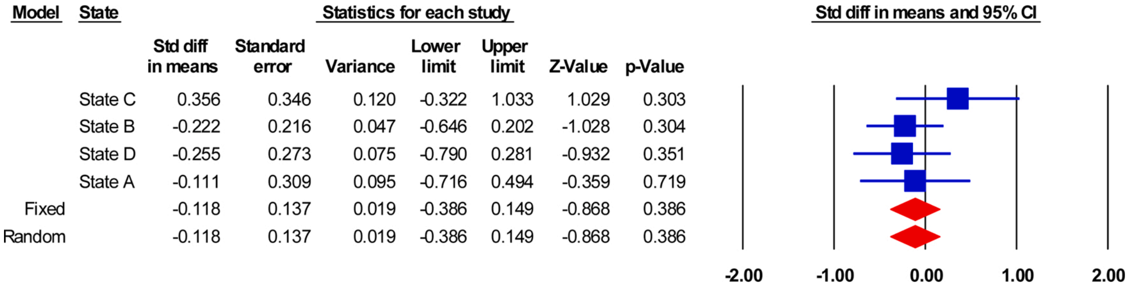
A key choice for all meta-analysts is whether the meta-analysis will use a fixed-effects or a random-effects model. The choice of models should be based on the inferences desired. A random-effects model is often the right choice for the most common use of meta-analysis: a review of a large number of very heterogeneous studies with the goal of making unconditional generalizations to any number of potential future studies or to a larger social-scientific phenomenon. We will present results of random-effects models, but will focus on fixed-effects models because our central question is not about variance across states, and our goal is not to generalize beyond the states studied, but to ascertain whether the dropout prevention work being evaluated was effective. Furthermore, the fact that we have at most, four studies per outcome variable dramatically limits the usefulness of a random-effects model (Goh et al., 2016; Hedges & Vevea, 1998). As Borenstein et al. (2007) note, "if the number of studies is very small, then it may be impossible to estimate the between-studies variance (tau-squared) with any precision. In this case, the fixed-effects model may be the only viable option. In effect, we would then be treating the included studies as the only studies of interest." In general, we suggest that evaluators carefully consider the kinds of inferences they wish to draw and the numbers of studies they are using, and choose their model accordingly. All calculations were performed using Borenstein et al.' Comprehensive Meta-Analysis Version 3 (Borenstein, Hedges, Higgins, & Rothstein, 2013).

4. School-level results

Appendix C in Supplementary material shows pre- and post-descriptive statistics for the treatment and comparison schools for the 23 (State, Outcome) studies. Treatment and control schools each demonstrated trends over time for each outcome; some of the trends were upward, some were downward; some were desirable and some were undesirable. For example, an upward trend is desirable for graduation rates, but undesirable for dropout rates. Because the sample sizes were small, we did not report significance tests for these sub-studies.

In the following seven figures, we present 95 % confidence intervals, forest plots, and p-values for each of the 23 difference-in-difference studies. We discuss, for formative evaluation purposes, the significant results. The 23 studies are grouped by seven outcome variables and meta-analyzed. We show both fixed-effects and random-effects meta-analytic results for each outcome variable. The fixed-effects results represent the answers to our seven summative research questions.

Fig. 1 summarizes and compares the effects on dropout rates for students with disabilities in the four states using a forest plot. Three of the four p-values are smaller than 0.05, but only State A's is smaller than its B-H-adjusted critical value. The fixed-effects meta-analysis shows an effect size that was significantly less than zero [g = -0.34, p = .02], suggesting that, on average, dropout rates declined more rapidly (or increased more slowly) in treated schools than in non-treated schools. However, this fixed-effects meta-analytic result is no longer significant after Benjamini-Hochberg adjustments for seven comparisons. A random-effects analysis also produces a null finding, as seen in the final row. This is primarily because of the State C study, which yields a positive standardized difference in means, while all other studies yield negative standardized differences in means. With the greatest variance of the studies, State C has a small weight in the fixed effects analysis. It is weighted more equally with the other states in the random-effects model, moving the standardized difference in means closer to zero.



Model	State	Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
	State C	0.356	0.346	0.120	-0.322	1.033	1.029	0.303
	State B	-0.222	0.216	0.047	-0.646	0.202	-1.028	0.304
	State D	-0.255	0.273	0.075	-0.790	0.281	-0.932	0.351
	State A	-0.111	0.309	0.095	-0.716	0.494	-0.359	0.719
Fixed		-0.118	0.137	0.019	-0.386	0.149	-0.868	0.386
Random		-0.118	0.137	0.019	-0.386	0.149	-0.868	0.386

Fig. 2. Forest plot of meta-analysis of program's effects on dropout rates for all students.

Model	State	Statistics for each study						
		Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
	State B	0.249	0.273	0.075	-0.286	0.785	0.912	0.362
	State A	-0.332	0.311	0.097	-0.941	0.277	-1.069	0.285
	State C	-0.101	0.343	0.118	-0.774	0.572	-0.294	0.768
	State D	0.015	0.272	0.074	-0.519	0.548	0.055	0.956
Fixed		-0.017	0.148	0.022	-0.306	0.273	-0.112	0.911
Random		-0.017	0.148	0.022	-0.306	0.273	-0.112	0.911

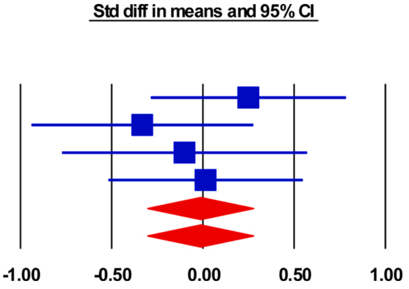


Fig. 3. Forest plot of meta-analysis of treatment effects on graduation rates for students with disabilities.

Model	State	Statistics for each study						
		Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
	State B	0.329	0.212	0.045	-0.087	0.745	1.548	0.122
	State D	0.233	0.273	0.075	-0.303	0.768	0.852	0.394
	State C	0.166	0.344	0.118	-0.508	0.839	0.482	0.630
	State A	-0.200	0.309	0.096	-0.807	0.406	-0.647	0.518
Fixed		0.178	0.135	0.018	-0.087	0.444	1.317	0.188
Random		0.178	0.135	0.018	-0.087	0.444	1.317	0.188

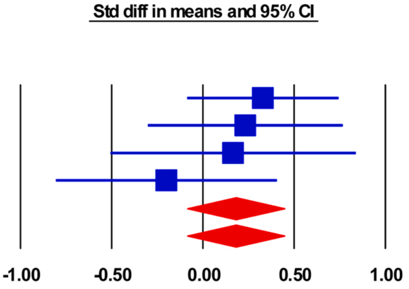


Fig. 4. Forest plot of meta-analysis of intervention effects on graduation rates for all students.

Model	State	Statistics for each study						
		Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
	State B (% absent >15 days)	-0.183	0.205	0.042	-0.583	0.218	-0.892	0.372
	State C (attendance rate)	0.259	0.344	0.119	-0.416	0.934	0.751	0.452
	State A (attendance days)	-0.231	0.310	0.096	-0.838	0.376	-0.747	0.455
Fixed		-0.107	0.153	0.023	-0.407	0.192	-0.702	0.482
Random		-0.107	0.153	0.023	-0.407	0.192	-0.702	0.482

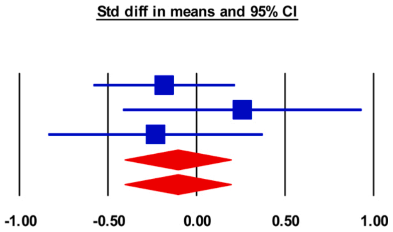


Fig. 5. Forest plot of meta-analysis of effects on attendance for students with disabilities.

Furthermore, the State C study dramatically increases the between-study variance, leading to a much larger standard error for the random effects model than the fixed effects model.

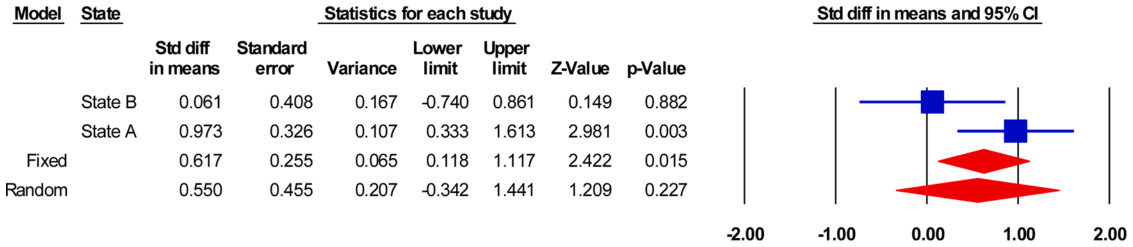
Fig. 2 summarizes and compares the program's effects on dropout rates for all students in the four studied states. Like the results described previously, the negative standard difference in means favored the treated group, suggesting that dropout rates declined more rapidly (or increased more slowly) for the treated group than for the control group. As shown in the figure, although the effect sizes for State B, State D, and State A were less than zero, their associated confidence intervals all crossed zero, suggesting the estimates were not statistically significant. The effect size for State C was positive, but also nonsignificant. As a result, the average effect size was nonsignificant (g= -.12, p = .39). No statistically valid conclusion could be drawn about the effect of the intervention on the dropout rates for all students in treated schools. The fixed effect and random effect results were identical because all observed variance could be attributed to within-study variance, leaving no between-study variance. The same agreement between fixed and random effects is found in Figs. 3–5, for the same reason.

The effects of the treatment on graduation rates for students with disabilities in State B, State A, State C, and State D are shown in Fig. 3. A positive standard difference in means favored the treated group for States B and D, suggesting that the graduation rate increased more rapidly (or declined more slowly) for the treated group than for the control group, but these results were not significant. For States C and A, the effect sizes were negative, but nonsignificant. As a result, the average effect size was nonsignificant (g= -.02, p = .91).

The effects of the intervention on graduation rates for all students in the four states are presented in Fig. 4. Although the effect estimates for State B, State C, and State D were positive, favoring the treatment group, the estimates were not statistically significant. The effect size for State A was negative, but it too, was nonsignificant. The overall average effect size was also nonsignificant (g = .18, p = .19). No statistically valid conclusion can be drawn about the effect of the intervention on the graduation rates for all students.

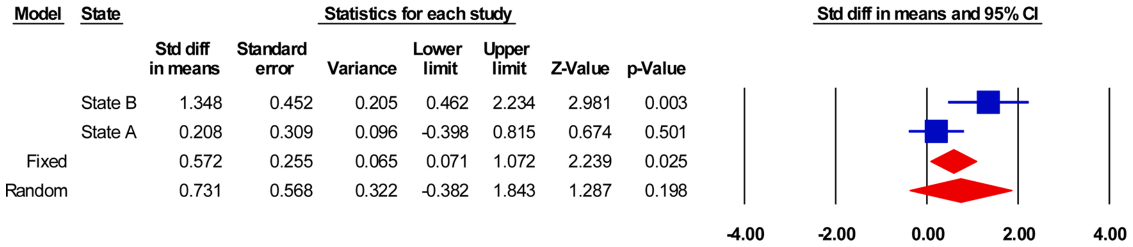
Fig. 5 summarizes and compares effects on attendance for students with disabilities in State C, State A, and State B. The attendance was measured by the percentage of students absent for 15 days or more in State B, by total attendance days in State A, and by the attendance rate in State C.[5] To ensure that the directions of effect sizes aligned correctly across the three states, the standardized difference in mean changes for State B was reverse coded (i.e., the positive was coded as negative and vice versa). In Fig. 5, a positive standard difference in means favored the treated group, suggesting that the attendance improved more rapidly (or degraded less rapidly) for the treated group than for the control group. As shown in the figure, only the effect estimate for State C was positive, but its associated confidence interval crossed zero, suggesting that the estimate was not statistically significant. The effect sizes for the other two states were negative, and not statistically significant. The resulting average effect size was nonsignificant (g= -.11, p = .48).

[5] These attendance measures are clearly different and are on different original scales, but there is confidence in combining them because they all measure a similar construct and all have been converted to an effect-size scale.



Model	State	Statistics for each study							Std diff in means and 95% CI
		Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value	
	State B	0.061	0.408	0.167	-0.740	0.861	0.149	0.882	
	State A	0.973	0.326	0.107	0.333	1.613	2.981	0.003	
Fixed		0.617	0.255	0.065	0.118	1.117	2.422	0.015	
Random		0.550	0.455	0.207	-0.342	1.441	1.209	0.227	

Fig. 6. Forest plot of meta-analysis of effects on state reading assessment passing rates for students with disabilities.



Model	State	Statistics for each study							Std diff in means and 95% CI
		Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value	
	State B	1.348	0.452	0.205	0.462	2.234	2.981	0.003	
	State A	0.208	0.309	0.096	-0.398	0.815	0.674	0.501	
Fixed		0.572	0.255	0.065	0.071	1.072	2.239	0.025	
Random		0.731	0.568	0.322	-0.382	1.843	1.287	0.198	

Fig. 7. Forest plot of meta-analysis of effects on state reading assessment passing rates for students without disabilities.

Passing rates in reading assessments for students with disabilities in State B and State A and reading assessment rates for students without disabilities in these states are presented in Figs. 6 and 7, respectively. A positive standard difference in means favored the treated group, suggesting that the passing rate increased more rapidly (or decreased more slowly) for the treatment group than for the control group. As shown in the figures, the effect estimates for both states for both outcome measures were positive. The estimate for students with disabilities was significant in State A; the estimate for students without disabilities was significant in State B. These two findings remained after the B-H adjustment to the critical p-value for 23 studies.

The average effect sizes for students with disabilities and students without disabilities were also significant in the fixed-effects analyses (SWD: $g = .62$, $p = .02$; SWOD: $g = .57$, $p = .03$), but these two findings were not robust to the Benjamini-Hochberg adjustment for seven comparisons at the outcome level. Furthermore, the wide difference between the effect sizes in the two states created sizable between-study variance, which led to non-significant results in the random-effects analyses (SWD: $g = .55$, $p = .23$; SWOD: $g = .73$, p = .20).

5. Conclusion

Twenty-three school level difference-in-difference analyses using carefully chosen comparison schools yielded five significant results, three of which were robust to Benjamini-Hochberg adjustments. These results suggested that reading scores and dropout rates were significantly improved in State A by the program, as were reading scores for students without disabilities in State B. All other results were nonsignificant, possibly because most of the analyses were based on a fairly small numbers of schools. Based on these formative results, an evaluator might investigate the work in States A and B to see what worked well.

The summative meta-analytic results, designed to combine studies and evaluate the overall quality of the intervention with greater power, found no results that were significant at the 0.05 level after Benjamini-Hochberg adjustment for multiple comparisons.

More importantly, this study demonstrated an impact evaluation method that may be useful in some cases for which RCT is infeasible, but large amounts of extant data are available. The method, which we recommend to evaluators in similar situations, is as follows:

1 Before implementing the program to be studied, determine the desired outcomes for which extant pre-post data are available.
2 Also prior to implementation, use a careful matching method to locate comparison units. Match on key variables that may predict the outcomes, especially the pretest values of the outcomes.
3 After completion of the treatment, compare control and treatment units with changes in the outcome variables.
4 If, as in our case, different data sets must be used for different groups of treatment units, and if the outcome variables differ substantially between data sets, perform separate analyses and use meta-analysis to combine the results.

Recent literature suggests that the benefits of using state-provided data outweigh its drawbacks, and this data may be used as an inferentially effective and cost-effective approach to assessing effects of school-based interventions. When the data must spread across similar but differing data sets, meta-analytic methods can be used to combine results.

6. Limitations

One limitation of this study applies to matching procedures. Because states did not provide this data, we were unable to match schools on the severity of the disabilities represented in each school. This would have been useful; instead, we presume that the distributions were similar since sites were not selected for the program based on the severity of students' disabilities.

Secondly, we speak to the central issue of power. We have demonstrated here that when an evaluation situation requires a low-powered data analysis, power and focus on summative outcomes can be improved via meta-analysis. We note that some type II errors can also be avoided by increasing the significance level from 0.05 to 0.10, at the cost of increasing the likelihood of type I errors. For example, in this study, such a loosened significance threshold would have allowed the evaluator to find positive effects on three of the seven outcomes. This could be an appropriate choice for some summative evaluations when scientific advance is not the primary purpose of the efforts. In such cases, incorrect null findings may be almost as bad as incorrect positive findings. We do not report these results here because such decisions should be made before an analysis is completed.

In this example, several results remained that had insufficient precision to generate a solid inference. We expect this will be the case in many evaluation situations. The meta-analytic combining of several low-powered studies is not a silver bullet, and sometimes results in a study that is still low-powered, just less so. We believe that by providing control groups and increasing power, the uses of extant data and meta-analysis demonstrated here can improve, if not revolutionize, some evaluation practice.

Acknowledgments

This article is based upon work done for the Center to Improve Project Performance (ED-OSE-13-C-0049). The Center was funded by the U.S. Department of Education, Office of Special Education Programs. We wish to thank Debra Rog for her comments on an early draft of the article and Rob Olsen for his comments on the final draft. We also wish to thank Wendy LaRue and Audra Barrett for preparing the article for submission.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.stueduc.2020.100969.

References

Algozzine, B., Wang, C., White, R., Cooke, N., Marr, M. B., Algozzine, K., et al. (2012). Effects of multi-tier academic and behavior instruction on difficult-to-teach students. *Exceptional Children, 79*(1), 45–64. https://doi.org/10.1177/0014402912079001, 03.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543–554.

Belfield, C. R., & Levin, H. M. (2007a). *The return on investment for improving California's high school graduation rate*. Santa Barbara, California: California Dropout Research Project, University of California at Santa Barbara.

Belfield, C. R., & Levin, H. M., eds. (2007). The price we pay: Economic and social consequences of inadequate education. Washington, DC: Brookings Institution Press.

Borenstein, M., Hedges, L., & Rothstein, H. (2007). *Meta-analysis: Fixed effect vs. Random effects*. https://www.meta-analysis.com/downloads/M-a_f_e_v_r_e_sv.pdf.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2013). *Comprehensive meta-analysis version 3*. Englewood, NJ: Biostat.

Borman, G. D., Slavin, R. E., Cheung, A. C. K., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2005a). Success for all: First-year results from the national randomized field trial. *Educational Evaluation and Policy Analysis, 27*(1), 1–22. https://doi.org/10.3102/01623737027001001.

Borman, G. D., Slavin, R. E., Cheung, A. C. K., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2005b). The national randomized field trial of Success for all: Second-year outcomes. *American Educational Research Journal, 42*(4), 673–696. https://doi.org/10.3102/00028312042004673.

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science, 9*(3), 333–342. https://doi.org/10.1177/1745691614529796.

Cobb, B., Sample, P., Alwell, M., & Johns, N. (2005). *The effects of cognitive-behavioral interventions on dropout for youth with disabilities*. Clemson, SC: National Dropout Prevention Center for Students with Disabilities, Clemson University.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*, 145–153. https://doi.org/10.1037/h0045186.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7–29.

Domina, T., Pharris-Ciurej, N., Penner, A. M., Penner, E. K., Brummet, Q., Porter, S. R., et al. (2018). Is free and reduced-price lunch a valid measure of educational disadvantage? *Educational Researcher, 47*(9), 539–555. https://doi.org/10.3102/0013189X18797609.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: SAGE.

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass, 10*(10), 535–549.

Hayes, R. L., Nelson, J., Tabin, M., Pearson, G., & Worthy, C. (2002). Using school-wide data to advocate for student success. *Professional School Counseling, 6*(2), 86–95. http://www.jstor.org/stable/42732397.

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods, 3*(4), 486.

Jacob, R. T., Goddard, R. D., & Kim, E. S. (2014). Assessing the use of aggregate data in the evaluation of school-based interventions: Implications for evaluation research and state policy regarding public-use data. *Educational Evaluation and Policy Analysis, 36*(1), 44–66. https://doi.org/10.3102/0162373713485814.

Kaniuka, T. S., Vitale, M. R., & Romance, N. R. (2013). Aggregating school based findings to support decision making: Implications for educational leadership. *Issues in Educational Research, 23*(1), 69–82.

Landeghem, G. V., Fraine, B. D., & Damme, J. V. (2005). The consequence of ignoring a level of nesting in multilevel analysis: A comment. *Multivariate Behavioral Research, 40*(4), 423–434. https://doi.org/10.1207/s15327906mbr4004_2.

McFarland, J., Hussar, B., Zhang, J., Wang, X., Wang, K., Hein, S., et al. (2019). *The condition of education 2019 (NCES 2019-144). U.S. Department of Education*. Washington, DC: National Center for Education Statistics. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2019144.

McShane, B. B., & Böckenholt, U. (2017). Single-paper meta-analysis: Benefits for study summary, theory testing, and replicability. *The Journal of Consumer Research*, 43.

Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research, 39*, 129–149. https://doi.org/10.1207/s15327906mbr3901_5.

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods, 11*(2), 364–386. https://doi.org/10.1177/1094428106291059.

National Center for Education Statistics. (2020). *Trends in high school dropout and completion rates in the United States*. https://nces.ed.gov/programs/dropout/.

Sanford, C., Newman, L., Wagner, M., Cameto, R., Knokey, A.-M., & Shaver, D. (2011). *The posthigh school outcomes of young adults with disabilities up to 6 years after high school. Key findings from the National Longitudinal Transition Study-2 (NLTS2) (NCSER 2011-3004)*. Menlo Park, CA: SRI International.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*(2), 309–316. https://doi.org/10.1037/0033-2909.105.2.309.

Sondergeld, T. A., Beltyukova, S. A., Fox, C. M., & Stone, G. E. (2012). Using microanalytical simulation methods in educational evaluation: An exploratory study. *Mid-western Educational Researcher, 25*(1), 24.

St. Clair, T., Cook, T. D., & Hallberg, K. (2014). Examining the internal validity and statistical precision of the comparative interrupted time series design by comparison with a randomized experiment. *The American Journal of Evaluation, 35*(3), 311–327. https://doi.org/10.1177/1098214014527337.

Stanard, R. P. (2003). High school graduation rates in the United States: Implications for the counseling profession. *Journal of Counseling & Development, 81*(2), 217–221.

Ueno, T., Fastrich, G. M., & Murayama, K. (2016). Meta-analysis to integrate effect sizes within an article: Possible misuse and Type I error inflation. *Journal of Experimental Psychology General, 145*(5), 643.

Vosgerau, J., Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2019). 99% impossible: A valid, or falsifiable, internal meta-analysis. *Journal of Experimental Psychology General, 148*(9), 1628.

Walser, T. M. (2014). Quasi-experiments in schools: The case for historical cohort control groups. *Practical Assessment, Research & Evaluation, 19*(6), 8.

Wong, K. K., & Socha, T. (2008). A pilot study to identify comparison schools for math and science partnership participating schools: Preliminary findings on one math/science partnership. *Peabody Journal of Education, 83*(4), 654–673. https://doi.org/10.1080/01619560802418677.

Wong, K. K., Boben, M., Kim, C., & Socha, T. (2009). Comparison of MSP and non-MSP schools in six states. *Journal of Educational Research & Policy Studies, 9*(2), 73–95.