ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 20-274

Improving Low-Performing Schools: A Meta-Analysis of Impact Evaluation Studies

Beth E. Schueler
University of Virginia

Catherine Armstrong Asher
Harvard University

Katherine E. Larned
Harvard University

Sarah Mehrotra
Education Trust

Cynthia Pollard
Harvard University

The public narrative surrounding efforts to improve low-performing K-12 schools in the U.S. has been notably gloomy. Observers argue that either nothing works or we don't know what works. At the same time, the federal government is asking localities to implement evidence-based interventions. But what is known empirically about whether school improvement works, how long it takes, which policies are most effective, and which contexts respond best to intervention? We meta-analyze 141 estimates from 67 studies of turnaround policies implemented post-NCLB. On average, these policies have had a moderate positive effect on math but no effect on ELA achievement as measured by high-stakes exams. We find evidence of positive impacts on low-stakes exams in STEM and humanities subjects and no evidence of harm on non-test outcomes. Some elements of reform, namely extended learning time and teacher replacements, predict greater effects. Contexts serving majority-Latinx populations have seen the largest improvements.

VERSION: August 2020

Suggested citation: Schueler, Beth E., Catherine Armstrong Asher, Katherine E. Larned, Sarah Mehrotra, and Cynthia Pollard. (2020). Improving Low-Performing Schools: A Meta-Analysis of Impact Evaluation Studies. (EdWorkingPaper: 20-274). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/qxjk-yq91

Improving Low-Performing Schools: A Meta-Analysis of Impact Evaluation Studies

Beth E. Schueler, University of Virginia
Catherine Armstrong Asher, Harvard University
Katherine E. Larned, Harvard University
Sarah Mehrotra, Education Trust
Cynthia Pollard, Harvard University

Abstract: The public narrative surrounding efforts to improve low-performing K-12 schools in the U.S. has been notably gloomy. Observers argue that either nothing works or we don't know what works. At the same time, the federal government is asking localities to implement evidence-based interventions. But what is known empirically about whether school improvement works, how long it takes, which policies are most effective, and which contexts respond best to intervention? We meta-analyze 141 estimates from 67 studies of turnaround policies implemented post-NCLB. On average, these policies have had a moderate positive effect on math but no effect on ELA achievement as measured by high-stakes exams. We find evidence of positive impacts on low-stakes exams in STEM and humanities subjects and no evidence of harm on non-test outcomes. Some elements of reform, namely extended learning time and teacher replacements, predict greater effects. Contexts serving majority-Latinx populations have seen the largest improvements.

Keywords: school improvement, turnaround, accountability policy

Note: The authors thank Empower Schools and the University of Virginia for financial support, the many authors of turnaround studies who responded to our survey and other queries, Shuang Geng and Jacqueline Hammaker for excellent research assistance, and Luke Miratrix, Katie Gonzalez and Daphna Bassok for helpful advice. Please send correspondence to Beth Schueler at beth_schueler@virginia.edu.

Why Review School Improvement Evaluations?

Americans tend to view their public school system as the government's primary program for combating social and economic inequality (Hochschild & Scovronick, 2004). However, both the income gap between wealthy and poor families and academic outcome gaps between high- and low-income children appear to have grown over the past three decades (Duncan & Murnane, 2011; Carter & Reardon, 2014). This is true for both more immediate indicators of K-12 academic performance such as test scores (Reardon, 2011) as well as longer-term outcomes including years of schooling (Duncan & Murnane, 2011) and college going (Bailey & Dynarski, 2011; Gamoran, 2013; Gamoran, 2015). Even those studies finding that income-based gaps have remained stable or gotten smaller, still conclude that there is significant work to be done to close these substantial inequalities (Hanushek, Peterson, Talpey & Woessmann, 2019; Hashim, Kane, Kelley-Kemple, Laski & Staiger, 2020). Race-based gaps have narrowed but remain meaningful in size (Reardon, 2011). One potential strategy for ensuring that schools play a greater role in narrowing these gaps is the rapid and dramatic improvement, sometimes called the "turnaround," of the country's lowest performing public K-12 schools and districts which serve high concentrations of low-income students and disproportionate numbers of students of color.

Unfortunately, the public narrative surrounding the effectiveness of school turnaround during the current era of test-based accountability has been notably pessimistic. Many observers assert that when it comes to efforts to improve persistently low-performing schools, either nothing works or—at best—the research literature is without quality evidence to guide the field. Many education policy pundits have suggested that we simply do not have examples of effective school turnaround from which to learn. Michael Petrilli, President of the Thomas B. Fordham Institute, has argued, "We don't know what to do about chronically low-performing schools.

Nothing has worked consistently and at scale" (Washington Post, 2017). Andy Smarick (2010), a former White House aide and U.S. Department of Education official, has observed, "overall, school turnaround efforts have consistently fallen far short of hopes and expectations. Quite simply, turnarounds are not a scalable strategy for fixing America's troubled urban school systems" (p. 21). Stuit (2010) argues that given how rare successful turnaround is empirically, it is "easier to close a low-performing school than to turn one around" (p. 10).

Many of these remarks have come in the wake of the Obama administration's unprecedented investments in school turnaround—which included some of the largest federal education grant programs in history (Dragoset et al., 2017). Criticisms of Obama-era policy efforts to turn around chronically underperforming schools have come from both sides of the political spectrum. Those on the right often argue that turnaround efforts have been either too prescriptive, too expensive, or both. Those on the left often criticize the focus on test-based accountability-driven sanctions, the disruptive nature of turnaround reforms, and the limited degree to which these policies have provided educators with the needed supports to address the out of school challenges that many students in turnaround contexts confront. Still others point to supply-side considerations, arguing that the staff replacements that have been central to many turnaround efforts can only be effective insofar as there is some untapped supply of highly effective educators who are currently not employed at low-performing schools and are ready to replace the teachers and administrators who are working in these challenging settings.

Perhaps most illustrative of the seeming unanimity, assessments of the Obama-era school turnaround efforts represent a rare point of agreement between U.S. Secretary of Education Betsy DeVos and American Federation of Teachers President Randi Weingarten. DeVos was blunt in her 2017 address before the Conservative Political Action Conference, criticizing the

results of the Obama-era School Improvement Grants, “They tested their model, and it failed… miserably.” Weingarten similarly referred to the program as a “terrible investment,” calling instead for community-based and career and technical education programs to address students’ wholistic needs (Washington Post, 2015).

Scholars of educational leadership and policy have also been skeptical of turnaround efforts, particularly in the post-No Child Left Behind Act (NCLB) era of universal test-based accountability. In fact, several have argued that many of the reforms required by the Obama administration’s school improvement policies were likely to actually reinforce the very conditions that promote persistent low performance in the first place, including high staff turnover, the use of inexperienced staff, limited financial investments, lack of focus on curriculum and instruction, limited parental engagement, and continued racial and socioeconomic segregation (e.g., Trujillo, 2012; Trujillo & Renee, 2015; Murphy & Bleiberg, 2019). Several scholars have focused specifically on reconstitution and mass staff replacements—a key component of multiple common turnaround models in this period—as particularly likely to generate unintended consequences (Malen et al., 2002; Rice & Malen, 2003; Malen & Rice, 2004; Rice & Croninger, 2005; Trujillo & Renee, 2015). Other criticisms have focused on the limited evidence of effectiveness among the partner organizations that collaborate with schools, districts, and states to implement turnaround efforts (Peurach & Neumerski, 2015; Meyers & VanGronigen, 2018).

Researchers and leaders have also disagreed on the durational aspect of turnaround. In their seminal 2007 report, “The Turnaround Challenge,” leaders at Mass Insight Education and Research Institute define school turnaround as, “a dramatic and comprehensive intervention in a low-performing school that produces significant gains in student achievement within two

academic years" (p. 2). However, there is an active debate regarding the amount of time necessary to achieve "turnaround"; indeed, one lesson emerging from the research on "comprehensive school reform" prior to the passage of NCLB was that such school improvement reforms tend to need more than three years of implementation to demonstrate substantial results (e.g., Berends et al., 2002; Borman et al., 2003; Gross et al., 2009; Hochbein, 2012). More recently, Sun, Loeb and Kennedy (2020) find that effects of School Improvement Grants tend to increase after the first year of intervention. Peurach and Neumerski (2015) go even further as to argue that although gains can be realized within three years, the establishment of school-level improvement infrastructure is more of a seven-year (or even longer) process and system-level infrastructural improvement requires ongoing effort over decades. These authors therefore call for, "a strong need to balance the rhetorical urgency of "turnaround" with the understanding that building educational infrastructure to improve large numbers of underperforming schools will likely require massive, sustained technical, financial, policy, and political support" (p. 1).

But what is known empirically about the effectiveness of efforts to improve low-performing schools and districts? Does rigorous quantitative research confirm the pessimistic public narrative surrounding turnaround reforms? Have the school improvement programs enacted in the period between the passage of No Child Left Behind Act (NCLB) and the passage of the Every Student Succeeds Act (ESSA) generated improvements for the children served by these schools? If so, how long has it taken for gains to materialize, which programs have been most successful, and does the effect of these programs depend on features of the contexts in which they are implemented (e.g., the demographics of the student populations)? In other words, what guidance does the literature provide for those leaders seeking to improve low-achieving schools and systems? In this paper, we comprehensively review the existing impact evaluation

studies of school and district improvement interventions implemented between NCLB and ESSA passage and conduct a meta-analysis to tackle the following research questions:

(1) What has been the effect of efforts to improve low-performing U.S. K-12 school and districts on student academic achievement?

(2) How many years has it taken for these improvement interventions to demonstrate results?

(3) Are particular contexts associated with larger effects?

(4) Have particular intervention features been associated with larger effects than others?

Our study would not have been possible ten years ago. The authors of a 2008 federal Institute of Education Sciences-sponsored report, "Turning Around Chronically Low-Performing Schools," after reviewing the existing research evidence on school turnaround, wrote that they were unable to uncover a single study that met What Works Clearinghouse standards for high-quality experimental or quasi-experimental evidence (Herman et al., 2008). In more recent years, a series of studies of school improvement efforts have emerged relying on rigorous methods designed for making credible causal inferences. Collectively, the findings from these various studies have been described as "mixed" (Barnum, 2017). Complicating matters further, each of these studies evaluates different policy approaches implemented in different settings. Therefore, despite the proliferation of empirical work in this area, there is still confusion about what—if anything—works overall or in specific contexts. Efforts to examine the mechanisms behind successful turnaround have been rare (e.g., Sun, Liu, Zhu & LeClair, 2019). Therefore, this review aims to look across this body of research for patterns that might help the field draw broader conclusions about how to improve low-performing schools.

We find that there is more room for optimism than the public narrative might suggest. On average, the existing impact evaluations reveal that turnaround reforms have generated

moderate—although hardly transformational—positive gains in math achievement as measured by high-stakes standardized tests, as well as several low-stakes achievement measures. We find no strong evidence of average turnaround effects on high-stakes English Language Arts (ELA) achievement. The average reforms do not seem to require three years of implementation to demonstrate results, though gains do appear to increase the longer an intervention is implemented. We find no strong evidence of differential effects between the four major types of policy response to low-performance (turnaround, labeling, charter conversion, and closure), though we uncover suggestive evidence that charter conversion generates greater gains while school closure is less effective for the directly affected cohorts. Our results also highlight more granular intervention features and populations associated with more effective improvement efforts. These findings do not seem to be driven by characteristics of the studies or methods used to evaluate the interventions. Before presenting our methods and results, we next briefly describe the key policy activity related to improving low-performing schools and districts that occurred leading up to, during, and at the culmination of the time period that is the focus of our study.

School Improvement Policy from NCLB to ESSA

No Child Left Behind. The 2002 passage of NCLB in the early years of the George W. Bush administration began an era of universal test-based educational accountability that brought with it a new approach to improving low-performing schools. To receive newly increased Title I funding, states were required to establish specific, grade-level standards, test most students annually, set criteria to determine "proficiency" on those standards, and ultimately achieve universal proficiency in math and reading by 2014 (McGuinn, 2006). Schools that did not demonstrate "adequate yearly progress" (AYP) toward the universal proficiency goal were subject to sanctions that increased in severity each year a school failed to make AYP. The

ultimate sanction involved school restructuring, including state takeover of the school, closure and reopening as a charter school, contracting with a private school operator, replacing staff deemed relevant to the failure (often called "reconstitution"), or any other major reorganization. Not surprisingly, a large majority of schools chose the least disruptive option ("any other major reorganization") and very few schools that entered the restructuring phase were ever able to exit (Scott, 2008; Scott, 2009; Smarick, 2010; Murphy & Bleiberg, 2019). Initially, there was limited causal evidence on effects of these policies, though a review of successful turnaround case studies highlights the importance of signaling a genuine commitment to improvement, leadership, committed staff, and instructional focus (Herman et al., 2008). In more recent years, a series of studies with higher internal validity have revealed mixed results on the effects of the school improvement interventions of this era. Our meta-analysis looks across these studies to reconcile those findings and draw broader lessons for the field.

American Recovery and Reinvestment Act. The 2009 American Recovery and Reinvestment Act (ARRA), passed at the start of the Barack Obama administration, dedicated unprecedented federal resources for school turnaround, establishing the School Improvement Grant (SIG) program targeting the lowest-performing five percent of schools within each state. The new $3 billion allocated to fund the program represented a considerably more substantial investment than the similar School Improvement Fund (SIF) program begun in 2007 (Dee, 2012; Trujillo & Renee, 2015). SIG investments were also accompanied by $4 billion in Race to the Top competitive grant funding that was awarded in part based on states' plans for turning around their lowest achieving schools (Smarick, 2010; McGuinn, 2012; Murphy & Bleiberg, 2019).

Schools accepting SIG funds were required to either implement one of three federally prescribed improvement models. The "transformation" model required replacing the school's

principal and implementing a series of additional whole-school reform efforts, including those focused on performance-based teacher evaluations and human resource decisions, data-driven instruction, extended learning time, professional development, technical assistance, and wraparound supports. The "turnaround" model required replacing the principal and at least half of the teachers, providing the new school leader with operational flexibilities, and implementing "comprehensive instructional reforms." Third, the "restart" model involved converting a traditional public school into one run by an outside management organization such as a charter school operator (Dee, 2012; Dragoset et al., 2017; Murphy & Bleiberg, 2019). The fourth and final model was school closure. A large majority of SIG schools relied on the least extreme "transformation" model, and second most popular was the "turnaround" model. Charter conversions and closures were least common (Hurlburt et al. 2011). While a handful of rigorous studies have shown SIG grants had positive effects on achievement (e.g., Dee, 2012; Carlson & Lavertu, 2018), others have generated more disappointing results (e.g., Dickey-Griffith, 2013). Studies of average effects nationwide have not shown the kind of dramatic improvements that would constitute the kind of successful "turnaround" at scale that the program was designed to produce (Dragoset et al., 2017).

No Child Left Behind Waivers. Although NCLB had been due for reauthorization as of 2007, political gridlock prevented this from occurring until 2015. In the meantime, the U.S. Department of Education encouraged state education agencies to apply for waivers from some NCLB requirements, contingent on the adopting a particular system of interventions for low-performing schools. Thirty-nine states received waivers in 2012 (Derthick & Rotherham, 2012; Dougherty & Weiner, 2017) and were required to classify schools based on performance and implement reforms targeting the lowest performing five percent of schools (called "priority

schools") and the ten percent of schools with the largest achievement gaps or lowest performance among subgroups of students (called "focus schools"). While focus schools were given substantial flexibility to implement reforms as they saw fit, the reforms required of the priority schools largely mirrored those embraced by Race to the Top and the SIG program. These included restart and closure, but most schools ended up implementing something most similar to what had been labelled "transformation" under SIG. Most evaluations of waiver policies in individual states have generated null or negative results (e.g., Dougherty & Weiner, 2017; Dee & Dizon-Ross, 2017; Hemelt & Jacob, 2018). However, Bonilla and Dee (2018) reveal more optimistic evidence from the focus school reforms as implemented in Kentucky. Again, these mixed results highlight the need for a synthesis of the empirical work in this area.

Every Student Succeeds Act. Toward the end of the Obama administration, Congress formally revisited NCLB, reauthorizing the federal K-12 education law as the Every Student Succeeds Act (ESSA). Partly in response to criticism that SIGs and NCLB waivers represented an overly prescriptive overreach of federal authority, ESSA devolved greater responsibility for turnaround to states and localities (Mann, 2016). States are required to identify their lowest performing five percent of schools—and this identification can now be based in part on non-test measures of student achievement—and must determine how best to remedy the low performance, using evidence-based reforms. The federal government no longer mandates particular interventions and many states delegated further turnaround authority to local districts (Barone, 2017). Given this new freedom to select interventions, it is all the more important that scholars identify effective turnaround practices for policymakers.

Methods

Literature Search. We began with an electronic search of abstracts using the databases Academic Search Premier, ERIC, Ed Abstracts, and ProQuest Dissertations & Theses, for the period between January 2000 and January 2018. Search terms included both turnaround content- (e.g., "school improvement grant") and methodology- (e.g., "randomized controlled trial") related keywords to capture experimental and quasi-experimental designs. Appendix A1 provides a complete list of search terms. This search uncovered 11,120 studies, as we show in the Appendix A2 flowchart of our search, screening, and inclusion process.

Though our database search only covered studies published prior to January 2018, we continued to monitor publications during the coding phase (e.g., Journal of Policy Analysis and Management, Education Finance and Policy, AERA-sponsored journals) to ensure we did not miss recently published research (through summer 2019). Next, we searched within websites of research firms (Mathematica, MDRC, SRI International, RAND) as well as the National Bureau of Economic Research, What Works Clearinghouse, and working paper series from Stanford University's Center for Education Policy Analysis and Brown University's Annenberg Institute. We searched the programs for education and policy academic conferences (2016 to 2019) to capture unpublished working papers. These steps allowed us to identify 2,090 sources.

After removing duplicates, we netted a total of 4,041 records. We screened these records in two phases. First, we reviewed the titles and abstracts of all studies to determine whether they were potentially relevant for our sample. For each of the studies that seemed potentially relevant, we scanned the reference list to identify additional studies for review. Through this process, we identified 207 studies that appeared to be relevant impact evaluations of school or district turnaround interventions based on the titles and abstracts alone and excluded 3,834 records.

Inclusion Review. The 207 studies were assessed against six inclusion criteria. First, the study needed to constitute an impact evaluation of a policy effort to substantially improve outcomes for students in low-performing schools or districts. We chose to define school improvement broadly—beyond the four federal SIG models—to ensure we could capture the most effective policy options for addressing underperformance of struggling schools (regardless of whether they had previously been federally endorsed). Therefore, we included studies as long as the intervention targeted low-performing schools or districts for improvement. Interventions that met our definition include reconstitution, transformation, restart, charter conversion, labeling (e.g, A-F grades, "warning" designation), accountability pressure (e.g., not making AYP under NCLB), comprehensive school reform, conversion to a community school model, technical assistance, state takeover, districtwide turnaround, and school closure. Although school closure may seem distinct from the rest of the interventions, we included these interventions because closure was one of the four federally defined models under the SIG program and is a common possible response to persistent low-performance. We excluded 33 studies based on this criterion.

Second, the intervention under study must have been implemented after the passage of NCLB. We focus on this era because universal test-based accountability could influence the effectiveness of turnaround interventions and because there have already been comprehensive reviews of research prior to this period (including Borman, Hewes, Overman, & Brown; 2003; Carallo & McDonald, 2001; Figlio & Loeb, 2011; Lee, 2008; Wong & Meyer, 1998). We ultimately excluded 17 of the 207 studies that examined interventions prior to NCLB.

Third, the study authors must have used a comparison group to estimate the impact of the intervention. We included both experimental and quasi-experimental research methods, including randomized controlled trials, regression discontinuity, difference-in-differences,

instrumental variables, and matching or multivariate regression. We excluded pre- and post-intervention comparisons without a counterfactual. We excluded 23 studies on this criterion.

Fourth, studies must have examined at least one student academic outcome such as test scores, suspensions, attendance, graduation, or grade progression. Fifth, the studies needed to report standardized effect sizes or provide enough information for our team to calculate standardized effect sizes on at least one outcome. Sixteen studies did not meet these criteria.

Finally, given differences in turnaround approaches across sectors and nations, the intervention must have targeted K-12 schools or districts in the U.S. Three studies did not meet this criterion. We also excluded 33 duplicative records (e.g., an interim report of an eventually published paper), and nine studies we were unable to accces, even after emailing study authors.

Expert Review. Next, we contacted at least one author of each study we had identified, as well as a handful of school turnaround experts who had not authored studies on our list, following Cooper (2010). Via online survey, we asked them to review our list of studies, confirm we had the most up-to-date versions of their papers, suggest additional studies for our sample, and provide feedback. In total, we heard back from 51 people, representing authors from 87 percent of the studies. Experts generally confirmed that our tracking was comprehensive and provided information for six additional studies that met the inclusion criteria described above. Our ultimate analysis sample includes 67 studies.

Study Coding. Our codebook included 328 codes. These were most commonly binary or categorical indicators that we grouped into four categories. First were codes related to the study itself, such as publication type and year, author affiliations, funder, identification strategy, covariates included, and sensitivity checks. The second category related to the study's sample,

including sample sizes and demographic characteristics, the geographic region, and grade level of the intervention.

The third category of codes covered features of the interventions. At the highest level, this meant distinguishing between major categories of school improvement reforms: (1) turnaround, (2) labeling, (3) charter conversion, or (4) closure, and coding for the length of the intervention. Additionally, we coded for a wide-ranging set of intervention components based on the intervention descriptions in the study or public record. We consolidated our codes into 14 common intervention features and ensured each study was represented by at least one feature:

1. New funding: there was a documented source of additional funding.
2. Governance change: governance of schools was transferred from the traditional locally elected school board to another party, such as a state takeover of a district.
3. Change in school manager: day-to-day management of treated units was transferred to a new group, such as a charter management organization.
4. Human Resource changes: there were changes to how teaching staff were managed, paid, or evaluated, including flexibility from existing collective bargaining agreements, changes to compensation systems, or the inclusion of a new performance-based teacher evaluation.
5. Teacher Professional Development: teachers were provided with professional development.
6. Administrator Technical Assistance: school or district administrators received supports in the form of professional development or technical assistance.
7. Teacher replacements: at least 35 percent of teachers were replaced or the study authors otherwise called out teacher replacements as a major part of the intervention.
8. Principal replacements: at least 50 percent of school principals were replaced or leadership replacement was explicitly described as a major part of the intervention.

9. Extended learning time: students were provided with extended learning time in the form of longer school days, or more days of instruction in the school year.
10. Tutoring: at least some students in the treated schools were identified for small group or individualized tutoring to support learning beyond typical instruction.
11. Curricular change: schools changed their curricula as part of the intervention.
12. Data use: school staff used student data to inform instruction.
13. Wraparound services: schools or districts provided additional non-instructional services to students and families, such as counseling, health services, or food support.
14. School choice: parents could send select schools other than that assigned by neighborhood.

Our list of consolidated intervention feature codes does not include separate codes for the federal SIG models (e.g., "transformation," "restart," "reconstitution"). Instead, we opted to capture the core substantive features of these models (e.g., principal replacements, change in school manager) so that evaluations of SIGs could be examined with interventions with similar features but that were not part of the SIG program. We consider this a strength over simply analyzing the SIG interventions alone.

The fourth and final category of codes were effect estimates (effect sizes, standard errors, and p-values) by outcome domain. Our primary domains of interest were academic achievement on state math and ELA exams used for accountability. However, we also examined effects on a number of low-stakes exams grouped into two categories: (1) science, technology, and math (STEM) exams, and (2) humanities (English and social studies). We also analyzed effects on three non-test outcomes: attendance, discipline, and graduation. For the small number of effect sizes for which authors did not report standard errors, we imputed standard errors using the reported p-values from a two-tailed t-test with the appropriate degrees of freedom. If an outcome

domain did not contain sufficient information to calculate a standardized effect for a specific outcome domain, that outcome domain was not coded.

In some cases, individual studies contained evaluations of multiple distinct interventions (e.g. light touch versus more intensive interventions in the same state) or effects among multiple samples (e.g. the same intervention but in two different states). These studies were coded as multiple observations (one for each evaluation/estimate). This resulted in a final sample of 141 estimates from our 67 studies, each containing an effect size for at least one outcome domain.

Our team included six coders, five of whom held graduate degrees in either public policy or education policy and one with a bachelor's degree in statistics and public policy. Initially, all coders read and coded the same three studies to produce a consistent application of codes. We reconciled our coding as a group, resolving differences and tracking decision rules. We then assigned each remaining study to two coders who coded the study independently and then reconciled as a pair. We convened the larger group regularly to resolve outstanding questions and to ensure consistency across coders.

Meta-Analytic Methods. The majority of studies in our sample reported impact estimates using model-based, standardized mean differences between the treated and control units. We used these model-based point estimates to calculate Hedge's *g* effect sizes where possible. Where outcomes had not already been standardized, we used reported sample standard deviations to standardize effects and their associated standard errors. For all estimates, we applied the following correction to minimize the slight upward bias found in Hedges' *g* effects with small sample sizes (Borenstein, Hedges, Higgins, and Rothstein, 2009):

$$g^* = \left(1 - \frac{3}{4(n_T + n_C) - 9}\right) g$$

where n_T is the treatment group sample size, n_C is the comparison group sample size, and g is the original effect size estimate. Because many studies had student-level data but reported on the effects of school- or district-level interventions, we based the correction on the number of treated and control units. Additional corrections were applied to standard errors that did not account for the clustering of observations within treatment units, following Littel, Corcoran, & Pillai (2008).

Because multiple evaluations within the same study are likely to be correlated, traditional meta-analysis models that treat each evaluation as independent would not have been appropriate. Historically, researchers have chosen to randomly or systematically select a single impact estimate per study or to aggregate across multiple estimates and arrive at a single average effect per study. However, to avoid losing information or masking potential heterogeneity within these studies, we use Robust Variance Estimation (RVE, Hedges, Tipton, & Johnson, 2010), a flexible meta-analytic approach that accounts for the nesting of impact estimates within clusters (in our case, evaluations nested in studies).

RVE typically adjusts for one of two common types of dependencies that might occur between effect sizes (Tanner-Smith, Tipton, & Polanin, 2016): (1) "correlated effects", or multiple effect size measures for the same sample or multiple outcome measures for the same treatment-control contrast and (2) "hierarchical effects", multiple evaluations with distinct samples within studies. While our sample contains occasional examples of correlated effects, we follow the advice of Tanner-Smith and Tipton (2014) and model the most prevalent type of dependency observed in our sample, the "hierarchical effects". We specify the following model for an impact estimate from evaluation *i* in study *j* in each of our outcome domains *k*:

$$Y_{ij}^{k} = \beta_0^k + u_j^k + \varepsilon_{ij}^k$$

where β_0 is the overall average impact on outcome domain *k*, u_j is a study-level random effect assumed to have $Var(u_j) = \tau^2$, and ε_{ij} is the residual of a specific effect size from its study-specific average with $Var(\varepsilon_{ij}) = \omega^2 + v_{ij}$, where $\omega^2$is the within-study variation in effect sizes, and v_{ij} is the observed variance of the effect size (i.e. the squared standard error). Like in traditional metanalytic models, β_0 is estimated as a weighted average of each effect size. The weight for effect size *i* in study *j* in our hierarchical model is drawn from the parameters above:

$$w_{ij} = \frac{1}{v_{ij} + \hat{\tau}^2 + \hat{\omega}^2}$$

where τ^2 and ω^2 are both estimated via method of moments (Hedges, Tipton, & Johnson, 2010).

We implement RVE in Stata using the Robumeta package (Tanner-Smith & Tipton, 2014) parameterized for hierarchical effects. This approach includes a small-sample correction designed to correct for the anti-conservative bias in standard error estimation in a meta-analysis with a small number of impact estimate (Tipton, 2015). For outcome domains where we have at least 20 effect sizes, we assess the robustness of our findings to researcher decisions made by the original authors or in the process of coding the meta-analysis. For our two primary outcome domains (high stakes math and ELA test scores), we expand our meta-regression model to include evaluation- and study-level moderators represented below as *Xi/j* :

$$Y_{ij}^k = \beta_0^k + \Gamma' X_{i/j} + u_j^k + \varepsilon_{ij}^k$$

Results

Describing Included Studies. As we describe in Table 1, we included 67 studies in our analytic sample. These studies provide a total of 141 estimates across seven available outcome domains. The majority of studies were published in peer-reviewed journals (51 percent), and the remainder were divided evenly among university-based center or think tank reports, working

papers, and research firm reports. By far the most common type of author was a university-based researcher. A large share did not report a funder (43 percent), but others reported receiving funds from federal or foundation sources or state or local agencies. Methods sometimes varied within studies, but the most common research design was difference-in-differences (47 percent). Our sample included a fair number of regression discontinuity analyses, as well as a smaller set of randomized controlled trials, instrumental variables, and matching or regression-based designs.

Table 2 describes the contexts in in which the interventions our studies evaluated took place. Each of the four major regions of the U.S. were represented. Among studies that provided demographic information, 93 percent of estimates come from contexts where a majority of students qualify for subsidized lunch. Nearly half of the estimates were from studies of contexts serving majority-African American student populations while 21 percent were from majority-Latinx communities. The interventions targeted a range of school grade levels as well.

Finally, we characterize the interventions that the studies in our sample evaluated in Table 3. The large majority examined school-level interventions while only 12 percent examined districtwide efforts. Approximately half of estimates examined interventions after one year of implementation, eighteen percent after two years, and 35 percent after three or more.

Ultimately, we examined four broad categories of policy efforts, representing related but distinct approaches to addressing low-performance: (1) turnaround, (2) labeling, (3) charter conversion, and (4) closure. The majority of estimates (54 percent) evaluated what we refer to as "turnaround" interventions designed to rapidly and dramatically improve existing schools. Twenty-three percent of estimates studied "labeling" interventions that aimed at encouraging improvement by generating public labels reflecting a school's performance (e.g., F-grades on A-F school report card rating scales or failing AYP). Third most common (17 percent) were

estimates examining policies to address underperformance by closing low performing schools. Importantly, our estimates in this category include estimates of the effect of being displaced by a school closure (i.e., being in a school that closes) as well as the effect of being in a school that receives new students from closed schools. Finally, six percent evaluated efforts to convert low-performing traditional public schools to charter schools.

We found substantial variation in the features of the interventions that were tested. The most common intervention feature highlighted by authors of our studies was principal replacements. Other common features included new funding, teacher replacements, professional development or technical assistance for administrators, human resource changes, and teacher professional development. Less common were data use, curricular changes, changes in school managers, extended learning time, governance changes, tutoring, and wraparound services.

At the bottom of Table 3, we show the variation in treatment among school closure estimates. A large majority estimated the average effect of being in a school that is closed and being displaced from that school (83 percent). Among these, some focus on the effect of closure for a subset of displaced students who transferred to a school that was higher performing than the closed school (21 percent), and 8 percent estimate the effects of closure in a slow-phaseout model (eliminating one grade at a time). Finally, 17 percent of closure estimates study the effect of attending a school that receives students from a closed school.

Effects on Academic Achievement. We present the pooled effect size estimates for our full sample in Table 4. We find that school improvement efforts have had a significant average effect of 0.062 standard deviations (SD) on high-stakes math achievement ($p<0.05$) and a smaller positive effect of 0.016 SD on high-stakes ELA achievement, though the ELA estimate is not statistically significant. We also examine pooled effects on low-stakes exams, finding

statistically significant effects of 0.068 SD in STEM subjects and 0.088 SD in humanities. This is notable given the smaller sample we have to estimate these effects (25 estimates from 15 studies in STEM and 20 estimates from 11 studies in ELA). We take these results as suggestive that—at worst—school improvement efforts have not superficially improved high stakes math performance at the expense of true learning and likely increased achievement.

Robustness Checks. We conduct a number of checks to test the sensitivity of our results to coding decisions and present results in Table 4. First, there were a handful of studies that did not explicitly report precise sample sizes at the level of the intervention, which were needed to implement the Hedges correction. Rather than exclude these studies entirely, we inferred sample size based on publicly available information, but also we estimated pooled effect sizes after limiting our sample to those studies for which the sample size was not inferred. Our results are similar for this subsample (0.056 SD in high stakes math). In other cases, authors reported p-values and we inferred standard errors using the maximum possible t-statistic (Cooper, 2010). Again, we find limiting to those studies for which we did not infer standard errors did not meaningfully change our findings.

To ensure outliers were not driving findings, we estimate the pooled effect sizes after excluding the top and bottom five percent of effects in our sample and again find our conclusions unchanged (though for ELA, the effect achieves statistical significance). Additionally, some studies examined similar interventions implemented with similar populations but with slightly different samples or different methods. To test whether this potential "double counting" of interventions could be influencing our findings, we estimated pooled effect sizes after limiting our sample to those studies that had no overlap in the samples under investigation. Our findings were also robust to this sample restriction.

Finally, we considered methodological decisions made by our study samples' original authors. We tested whether results were consistent after limiting our sample to those studies using student-level (rather than school-level) data, to studies that were able to capitalize on statewide data (versus data limited to a smaller sample), and to those studies that correctly clustered their standard errors at the level of treatment. Again, none of these restrictions shift our broad conclusions. The estimates are either very similar or in some cases larger in magnitude in both math and ELA. In math, the estimates are always statistically significant and in ELA the estimates sometimes achieve statistical significance with the sample restrictions. We conduct the same set of sample restriction tests for the low-stakes outcomes and again find the magnitude of our point estimates are generally consistent and marginally significant. These results suggest a consistent but modest positive effect on low-stakes assessments.

We also test whether other research characteristics of the studies in our sample are driving our main findings (see Table 5). First, we test whether results vary depending on the research methodology deployed. Using randomized controlled trials as the reference group, we find no significant differences between the effects using quasi-experimental designs. Studies using matching or regression methods only show differences in effects that are relatively large in magnitude in both subjects (-0.214 SD in math and -0.120 SD in ELA), and marginally significant in math. Due to these differences and because we consider these methods less credible for generating causal impact estimates, we control for the use of this method in future results. Table 5 also shows that we find no differences in effect size estimates based on author(s) affiliations or the source of study funding.

Publication Bias. We next examine the extent to which our findings could reflect publication bias. For example, researchers may be less likely to report or submit null or non-

significant results for publication and journals may be less likely to publish them. If these phenomena were occurring, we would expect the pooled effects in our sample coming from peer-reviewed journal articles to be larger than those from working papers or other non-peer-reviewed outlets. In the right panel of Table 5, we display the estimated effect size for peer-reviewed journal articles (0.058 SD in math, 0.010 SD in ELA) and find that estimates from these journals are not statistically different from any of the other publication venues.

In the absence of publication bias and conditional on the uniformity of interventions under study, we should expect to see a specific distribution of effect sizes and standard errors, with the most precise studies estimating effects near our average, and less precise studies shower a symmetric, wider spread of effect estimates (Duval & Tweedie, 2000). We plot this graphically in a series of funnel plots (see Appendix Figure A1) for math and ELA effect sizes by major type of intervention (turnaround, labeling, closure, and charter conversion). We find the symmetrical pattern is generally true of the estimates included in our sample although more so for the labeling and closure interventions than for the turnaround and charter conversion. We formally test for the asymmetry of these results using Egger's test (Egger, Smith, Schneider, & Minder, 1997). The results, presented in the first column and panel of Appendix Table A1, indicate the potential presence of publication bias for both Math and ELA. Because the Egger's test aggregates multiple effects within studies, we conduct a modified test following Lynch, Hill, Gonzalez, & Pollard (2019) using our RVE model. Again, we find evidence of potential publication bias, though methodologists have found that Egger's test can have inflated Type I error rates when using standardized mean difference outcomes (Pustejovsky & Rodgers, 2019).

There are two reasons our studies' estimates might differ from the expected pattern even in the absence of publication bias. First, publication bias tests assume that there is a single "true"

effect and variation across studies is due to sampling. This is plausible for labeling or closure interventions, but school "turnaround" is a broad category of intervention. We would thus anticipate these studies to have a wider range of "true" effects than many other subjects of meta-analysis, which is what we see in Appendix Figure A1. These differences are also reflected in the formal Egger's test, which does not find evidence of bias when we limit to studies of labeling and closure interventions. To account for this, we later disaggregate our estimates based on the type of turnaround policy. Second, one typical motivation for meta-analysis has been to pool estimates from a series of small sample studies to improve statistical power. Many of our studies already have high statistical power because they rely on quasi-experimental methods and student-level administrative data. Thus we see a larger congregation of precisely estimated effects than would be expected in a meta-analysis of small RCTs. For charter conversion, we observe a handful of studies with effects that are both large and precise, suggesting the possibility that interventions studied in this category represent outliers. In sum, we are not entirely able to rule out the possibility of publication bias. However, we are also unable to rule out the possibility that the patterns we observe are due to the diversity of interventions under study and resulting true variation in effects. Indeed, scholars of meta-analytic methods have found that it can be difficult, if not impossible, to disentangle publication bias from true heterogeneity of effects (Peters et al., 2010).

Effects on Non-Test Outcomes. A subset of the studies in our sample reported effects on non-test based academic attainment and behavioral outcomes. As shown in Table 6, we find suggestive evidence that school improvement efforts have had positive effects on non-test outcomes and no evidence they have had a negative impact on these measures. More specifically, on average, school improvement efforts have had a 0.108 SD positive effect on school

attendance outcomes, have somewhat reduced disciplinary infractions by 0.006 SD and have increased graduation rates by 0.044 SD. However, none of the effects on any of the non-test outcomes achieve statistical significance. This is not particularly surprising given these estimates were calculated with much smaller samples than the test score estimates (between four and nine effect sizes from between three and six studies). The graduation effect is robust to the inclusion of a control for whether the authors relied on matching or regression research methods alone, as are the high stakes math and ELA estimates. None of the attendance, discipline, or low-stakes exam estimates came from studies using matching or regression methods and therefore including controls for method could not have changed our estimates.

Features of Effective Intervention Contexts. In Table 7, we display the results of our exploration regarding whether school and district improvement efforts were more effective in some contexts than others. In the top panel, we test whether reforms produced greater impacts depending on the socioeconomic or racial makeup of a student population. We find efforts were more effective at improving math achievement in schools or districts serving majority-Latinx student populations, even after controlling for whether a majority of the population qualified for subsidized lunch, producing effects 0.219 SD larger than those interventions implemented in majority white, non-majority subsidized lunch. The ELA effects are also larger in majority-Latinx contexts by 0.081 SD, though this difference is not statistically significant. We do not find that improvement interventions produced different effects for populations where a majority of students qualify for subsidized lunch or majority-African American student populations. Across subjects, these findings are robust to including a control for the use of matching or regression methods only, suggesting these differences are not driven by features of the research methodologies. In the bottom panel, we show that effects do not vary dramatically by region.

Length of Effective Interventions. Another important debate in the literature on school improvement is related to the length of intervention necessary to achieve results. We test whether effects are different for those interventions that were implemented for two years (versus one) or for three or more years (versus one) and report the results in Table 8. We find positive effects appear, on average, even among those interventions that have been implemented for only a single year in both math and ELA, however, none of these coefficients are statistically significant. In both subjects, the effects appear to be slightly larger for those interventions implemented for longer periods of time (two or three or more years), though again these differences do not achieve statistical significance. Results are not sensitive to controls for research methodology.

When we restrict the sample to those studies examining interventions we classified as "turnaround" (excluding labeling, closure, and charter conversion studies), we find that ELA gains do not emerge until year two of these interventions and even evidence of small negative effects in year one. However, again, none of these differences are statistically significant so we are limited in our ability to draw strong conclusions on this topic. Taken as a whole we interpret these results to provide suggestive evidence that it is not impossible for school improvement efforts to demonstrate results within the first year of implementation, however, greater duration does appear to be associated with greater results.

Features of Effective Interventions. We further explore whether various features of the interventions that our studies evaluated meaningfully predict variation in effect size estimates. First, we compare effect size estimates for our four major intervention categories: turnaround, labeling, charter conversion, and closure, and display results in Table 9. Given it is the most common type in our data, turnaround studies serve as the comparison group. We find no major statistically significant differences in effectiveness between the four categories. Charter

conversion appears more effective in both subjects than turnaround while closure appears somewhat less effective in math, though these differences do not achieve statistical significance. One thing to keep in mind when interpreting the closure results is that closure estimates here include both effects for students in closed schools and in schools receiving students from closed schools. Additionally, lthough our sample is not large enough to formally test for differences, the qualitative pattern of results suggest that closure effects are more positive on displaced students when they are able to transfer to higher value-added schools (relative to their closed schools).

We also test whether effects vary for school- versus district-level interventions. While none of the differences are statistically significant, that is not entirely surprising given that a small number of studies in our sample evaluated districtwide interventions. Our results are suggestive that districtwide interventions tend to generate larger positive results on the order of 0.11-0.12 standard deviations in math and ELA than school-level interventions, consistent with some of the theory regarding the important role districts can play in building capacity for school-level improvement. Results are robust to the inclusion of a control for research methodology.

Next, we turn to the specific features of the interventions that our studies evaluated. We coded 14 key intervention features that authors in the studies associated with the treatments under study. First, we test whether "less is more" and find the opposite—the larger the number of intervention features highlighted for a particular treatment, the larger the effects. In math, each additional intervention feature is associated with 0.029 SD larger effects. In ELA, the difference is 0.018 SD. Results are robust to controlling for research methods. The relationship between number of features and effects does take on a linear functional form, as shown in Figure 1.

To analyze the effect of each intervention feature, we conduct a series of bivariate meta-regression models testing for a difference in the impact of studies with and without each feature.

We display these findings visually in Figure 2 where we plot the "overall" average effect of school improvement interventions, as well as the implied coefficients in the absence and presence of each intervention feature. Estimates for evaluations missing a particular feature (the reference) are colored in grey. The bars colored black represent differences where $p<0.10$ when comparing the presence of that feature to evaluations without that feature. In both the math and ELA panels, the features are sorted based on the magnitude of the differential feature effect in math. We find two intervention features that appear to be particularly associated with greater reform effectiveness: extended learning time and teacher replacements. These findings are consistent across subjects. Two additional features—tutoring and wraparound services—in some cases achieve statistical significance. We find no significant differences in effects for reforms that included new funding, governance change, change in school manager, human resources (HR) changes (e.g., freedom from collective bargaining agreements, merit-based pay), professional development or technical assistance for teachers or administrators, principal replacements, curricular changes, data use, and school choice. Accompanying regression results with standard errors can be found in Appendix Table A2.

To further explore the potential of the various intervention features, we test whether any of the promising features remained significant predictors when including five features in a single model (see Table 10). For this comparison, we also included teacher professional development in this model because the coefficients were non-trivial in size and relevant to the ongoing debate about the relative importance of teacher development versus replacement in turnaround contexts. We find that the most promising features of school improvement efforts include extended learning time and teacher replacements. More specifically, as shown in Column 1, extended learning time is associated with larger effects in math (by 0.094 SD) and in ELA (by 0.099 SD).

Interventions that included teacher replacements as a major reform component, on average, produced larger positive effects on the order of 0.086 SD in math and 0.088 SD in ELA than those that did not involve significant teacher replacements. Results for extended learning time and teacher replacements are similar when limiting the sample to turnaround studies (Panel 2) and are robust to including a control for whether a study relied on matching or regression methods alone (Columns 2 and 5) as well as controlling for the number of intervention features.

Discussion

This comprehensive meta-analytic review of post-NCLB efforts to improve low-performing U.S. schools establishes that observers can not credibly claim that nothing works for improving struggling schools. Though there is still much to learn, the literature using causal methods to evaluate turnaround reforms has grown significantly over the past fifteen years. Furthermore, we find that, on average, school improvement policies that have been subjected to evaluation demonstrated positive effects on high-stakes math achievement and on low-stakes exam performance in both STEM and humanities subjects. These effects are neither radically transformation nor trivial—they can be characterized as medium in size (Kraft, 2020). We find no evidence that these reforms hurt ELA achievement on high-stakes tests or non-test-based outcomes, although more work should be done to examine impacts on long-term outcomes. Our results are not driven by study characteristics, including research methods deployed, author affiliations, type of publication, and type of funder. We also find suggestive evidence that it is indeed possible for school improvement efforts to generate gains after a single year of implementation, at least in the post-NCLB environment. However, gains do appear to increase the longer interventions are in effect and it remains unclear what is needed to sustain gains.

Additionally, the district-level interventions in our sample are more effective than school-level ones, though the difference is not significant. This finding is consistent with theory identifying districts as important for supporting capacity building in schools (e.g., Honig & Hatch, 2004; Carnoy, Elmore, & Siskin, 2003; Cohen & Hill, 2008; Johnson et al., 2015). Alternatively, it is possible the districts selected for districtwide intervention were more ripe for the reforms under study than the individual schools. We also find no evidence of the "less is more" theory; greater numbers of intervention features are associated with larger positive effects. However, we cannot rule out the possibility that authors highlighted more intervention features or reformers kept better track of those features in contexts where efforts were successful.

In terms of describing the policy features of effective school improvement efforts, our findings show that extended learning time and teacher replacements seem particularly promising—including for generating gains in reading. It is not surprising that one of these features relates to instructional quality given the well-established importance of teacher quality in explaining student short and longer-term success (e.g., Chetty et al., 2011; Hanushek & Rivkin, 2010). However, we cannot speak to the manner in which teacher replacements were instantiated, and the value of such a reform may be dependent on who is replaced and the existing supply of available higher-performing teachers. Our findings are also consistent with a literature on the value of additional learning time, particular for low-performing students.

It is important to keep in mind that the studies in our sample often estimate the effect of intervention relative to the threat of intervention. Comparison groups are typically (and intentionally) also very low-performing. Therefore, our results should not be taken as a summary of the effects of broader accountability policies. For example, we do not find that governance

changes are associated with greater effectiveness. That does not mean, however, that the threat of a state takeover or a charter conversion could not meaningfully improve school performance.

An additional caveat is that we are limited to studying the types of interventions that have been rigorously studied specifically as a vehicle for improving low-performing schools. For example, race-based integration or the redrawing of attendance zones within districts could be effective methods for improving low-performing schools or systems even though these are not typically described as "turnaround" policies. However, we did not come across studies evaluating these approaches that met our inclusion criteria. Furthermore, we cannot fully rule out the possibility that our estimates overstate the true effect of the full universe of interventions that have targeted low-performing schools in the event that researchers are more likely to study effective policies.

An important area for future research would be to use these impact estimates to explore which intervention features are most cost effective and scalable. Although school improvement efforts have on average generated gains, those who argue "nothing works" could still be correct in practice if the most effective interventions are cost prohibitive. However, our findings suggest the opposite is true given that interventions involving large amounts of new funding were not more effective than those without it – but this issue should be explored more thoroughly.

A few recommendations for researchers emerged from our coding and analysis. First, whenever possible scholars reporting on impact evaluations should be careful to include enough information for others to calculate effect sizes for all outcomes examined in a given study (and reviewers and editors should encourage them to do so) to facilitate future meta-analytic work. We found that this was sometimes overlooked for non-test outcomes (attendance, discipline, graduation) where papers would report a non-standardized effect (on, for example, days of

attendance or number of suspensions) without a measure of the variability in the broader sample or in cases where authors used transformed variables in analysis but reported descriptive statistics for untransformed versions. This will be increasingly important going forward now that the federal Every Student Succeeds Act allows states and districts to include non-test indicators of student performance in their accountability systems that are used for identifying low-performing schools to target for intervention. Not only are these outcomes policy relevant but also of substantive importance for students' long-term well-being. Therefore, more work is needed to better unpack the effects of school improvement efforts on non-test outcomes.

Finally, our finding on the variation in effectiveness of school improvement efforts based on the racial/ethnic makeup of the student population raises several important questions. Why is it that turnaround efforts with positive effects—at least those that have undergone evaluation—have been concentrated in predominantly Latinx communities as opposed to majority-African American or low-income school systems? We were also surprised by the small number of studies that reported effects for specific sub-populations of students, which hindered our ability to deeply investigate the implications of turnaround interventions for racial and socioeconomic equity. This review reveals an urgent need to better understand how policymakers can more successfully improve low-performing schools and systems serving African American communities. The next generation of school improvement research should much more carefully examine what works for whom so leaders can effectively tailor their policy efforts to the different contexts in which the lowest performing schools are operating. Understanding these issues is critical to improving the educational institutions serving some of our nation's most vulnerable children and ultimately to narrowing persistent race- and class-based opportunity and outcome gaps.

References

* indicates studies that contributed estimates to the meta-analysis

*Abdulkadiroğlu, A., Angrist, J. D., Hull, P. D., & Pathak, P. A. (2016). Charters without lotteries. *The American Economic Review*, 106(7), 1878-1920.

*Ahn & Vigdor (2014). The Impact of No Child Left Behind's Accountability Sanctions on School performance: Regression Discontinuity Evidence from North Carolina. NBER Working Paper 20511.

Bailey, M. & Dynarski, S. (2011). "Inequality in postsecondary education" in Greg Duncan and Richard Murnane eds., *Whither opportunity: Rising inequality, schools, and children's life chances.* New York, NY: Russell Safe Foundation.

Barnum, M. (2017). New study deepens nation's school turnaround mystery, finding little success in Rhode Island. *Chalkbeat*. https://chalkbeat.org/posts/us/2017/08/17/new-study-deepens-nations-school-turnaround-mystery-finding-little-success-in-rhode-island/

Barone, C. (2017). "What ESSA says: Continuities and departures" in Frederick Hess and Max Eden, eds., *The Every Student Succeeds Act: What it means for schools, systems, and states* (pp. 59-74). Cambridge, MA: Harvard Education Press.

Berends, M., Bodilly, S. J., & Kirby, S. N. (2002). *Facing the challenges of whole-school reform: New American Schools after a decade*. Rand Corporation.

*Bifulco, Duncombe & Yinger (2005). Does whole-school reform boost student performance? The case of New York City. *Journal of Policy Analysis and Management, 24(1),* 47-72.

*Bifulco, R., & Schwegman, D. (2018). Who Benefits from Accountability-Driven School Closure? Evidence from New York City. Center for Policy Research, The Maxwell School: Working Paper Series Paper 212.

*Bonilla, S. & Dee, T. (2020). The effects of school reform under NCLB waivers: Evidence from focus schools in Kentucky. *Education Finance and Policy*, 15(1), 75-103.

Borenstein, M., Hedges, L., Higgins, H. and Rothstein, H. (2009). *Introduction to Meta-Analysis*. John Wiley and Sons.

Borman, G., Hewes, G., Overman, L. & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research*. 73(2), 125-230.

*Bross, Harris & Liu (2016). The Effects of Performance-Based School Closure and Charter Takeover on Student Performance. Education Research Alliance for New Orleans.

*Brummet, Q. (2014). The effect of school closings on student achievement. *Journal of Public Economics*, 119, 108-124

Calkins, A., Guenther, W., Belfiore, G. & Lash, D. (2007). *The turnaround challenge: Why America's best opportunity to dramatically improve student achievement lies in our worst-performing schools*. Mass Insight Education & Research Institute.

*Carlson, Borman & Robinson (2011). A Multistate District-Level Cluster Randomized Trial of the Impact of Data-Driven Reform on Reading and Mathematics Achievement. *Educational Evaluation and Policy Analysis*, 33(3), 378-398.

*Carlson, D. & Lavertu, S. (2018). School improvement grants in Ohio: Effects on student achievement and school administration. *Educational Evaluation and Policy Analysis*.

*Carlson, D., & Lavertu, S. (2016). Charter school closure and student achievement: Evidence from Ohio. *Journal of Urban Economics*, 95, 31-48.

Carnoy, M., Elmore, R. F., & Siskin, L. S. (Eds.). (2003). *The new accountability: High schools and high-stakes testing*. New York, NY: Routledge Falmer.

Carter, P.L., & Reardon, S.F. (2014). Inequality Matters: Framing a Strategic Inequality Research Agenda. *The William T. Grant Foundation*.

Chalkbeat (2017). New study deepens nation's school turnaround mystery.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *Quarterly Journal of Economics*, *126*(4), 1593-1660.

*Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics*, 93(9-10), 1045-1057.

*Chin, M., Kane, T. J., Kozakowski, W., Schueler, B. E., & Staiger, D. O. (2018). School District Reform in Newark: Within-and Between-School Changes in Achievement Growth. *Industrial and Labor Relations Review*, 72(9), 1-32.

Cohen, D. K., & Hill, H. C. (2008). *Learning policy: When state education reform works*. Yale University Press.

Cooper, H. (2010). *Research synthesis and meta-analysis: A step-by-step approach*. Thousand Oaks, CA: SAGE Publications.

Corallo, C. & McDonald, D. (2001). What works with low-performing schools: A review of research literature on low-performing schools. Washington, D.C.: Office of Educational Research and Improvement.

*de la Torre, M., Allensworth, E., Jagesic, S., Sebastian, J., Salmonowicz, M., Meyers, C., & Gerdeman, R.D. (2012). Turning around low-performing schools in Chicago. Research

Report. Chicago, IL: The University of Chicago Consortium on Chicago School Research.

*Dee, T. (2012). School turnarounds: Evidence from the 2009 stimulus. NBER Working Paper 17990.

*Dee, T. & Dizon-Ross, E. (2019). School performance, accountability, and waiver reforms: Evidence from Louisiana. *Educational Evaluation and Policy Analysis.*

Derthick & Rotherham (2012). Obama's Education Waivers. *Education Next*, 12(2), 56-61.

Desimone, L. (2002). How can comprehensive school reform models be successfully implemented? *Review of Educational Research, 72(3),* 433-479.

DeVos, B. (2017). U.S. Secretary of Education Betsy DeVos' Prepared Remarks at the 2017 Conservative Political Action Conference.

*Dickey-Griffith, D. (2013). Preliminary effects of the school improvement grant program on student achievement in Texas. *Georgetown Public Policy Review*, 21-39.

*Dougherty, S. & Weiner, J. (2017). The Rhode to Turnaround: The Impact of Waivers to No Child Left Behind on School Performance. *Educational Policy*, 1-32.

Duke, D. (2015). *Leadership for low-performing schools: A step-by-step guide to the school turnaround process*. Lanham, Maryland: Rowman & Littlefield.

*Dragoset, Thomas, Herrmann, Deke, James-Burdumy, Graczewski, Boyle, Upton, Tanenbaum, & Giffin (2017). School Improvement Grants: Implementation and Effectiveness. NCEE 2017-4013. *National Center for Education Evaluation and Regional Assistance*.

Duncan & Murnane (2011). "Introduction: The American Dream, Then and Now" in Greg Duncan and Richard Murnane eds., *Whither opportunity: Rising inequality, schools, and children's life chances.* New York, NY: Russell Sage Foundation.

Duval, S. & Tweedie, R. (2000). Trim and fill: A simple funnel plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56(2), 455-463.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

*Engberg, Gill, Zamarro & Zimmer (2012). Closing schools in a shrinking district: Do student outcomes depend on which schools are closed? *Journal of Urban Economics*, 71, 189-203.

Figlio, D. & Loeb, S. (2011). School accountability. *Handbook of the Economics of Education*, vol. 3, p. 383-417.

*Figlio, D. N., & Rouse, C. E. (2006). Do accountability and voucher threats improve low-performing schools?. *Journal of Public Economics*, 90(1-2), 239-255.

*Fruehwirth & Traczynski (2013). Spare the Rod? The Dynamic Effects of Failing Accountability on Schools. Working Paper.

*Fryer, Roland. 2014. Injecting Charter School Best Practices Into Traditional Public Schools: Evidence From Field Experiments. *Quarterly Journal of Economics*, 129, p. 1355-1407.

Gamoran, A. (2013). *Inequality is the problem: Prioritizing research on reducing inequality.* William T. Grant Foundation Annual Report.

Gamoran, A. (2015). *The future of educational inequality in the United States: What went wrong, and how can we fix it?* William T. Grant Foundation.

*Gandhi, A., Slama, R., Park, S., Russo, P., Winner, K., Bzura, R., Jones, W. & Williamson, S. (2018) Focusing on the Whole Student: An Evaluation of Massachusetts's Wraparound Zone Initiative, *Journal of Research on Educational Effectiveness*, 11:2, 240-266.

*Gigliotti, P. (2019). Leveraging Managerial Autonomy to Turn Around Low-Performing Schools: Evidence from the Innovation Schools Program in Denver. Working Paper.

*Gill, B., Zimmer, R. Christman, J.B., & Blanc, S. (2007). State Takeover, Restructuring, Private Management, and Student Achievement in Philadelphia Santa Monica, CA: RAND Corporation and Research for Action.

*Gold, E., Norton, M., Good, D., Levin, S. (2012). Philadelphia's Renaissance Schools Initiative: 18 Month Interim Report. Research for Action.

*Goulas, Raymond, Bierbaum, Bell, Mazzola & Snow (2017). The Impact of Scaling the New Orleans Charter Restart Model on Student Performance. Center for Research on Education Outcomes (CREDO).

*Gordon, de la Torre, Cowhy, Moore, Sartain & Knight (2018). School Closings in Chicago: Staff and Student Experiences and Academic Outcomes. University of Chicago Consortium on School Research.

Gross, B., Booker, T. K., & Goldhaber, D. (2009). Boosting student achievement: The effect of comprehensive school reform on student achievement. *Educational Evaluation and Policy Analysis*, *31*(2), 111–126.

*Guthrie, J.E. & Henry, G.T. (2016). When the LATE Ain't ATE: Comparing Alternative Methods for Evaluating Reform Impacts in Low-Achieving Schools. Working Paper.

*Hallberg, Williams, Swanlund & Eno (2018). Short Comparative Interrupted Time Series Using Aggregate School-Level Data in Education Research. *Educational Researcher.* 47(5), 295-306.

*Hallgren, Gonzalez, Kelly, Demers & Gill (2019). Year 2 Report of the Atlanta Public Schools Turnaround Strategy. Mathematica Policy Research Report.

*Han, C., Raymond, M., Woodworth, J., Negassi, Y., Richardson, W., & Snow, W. (2017). Lights Off: Practice and Impact of Closing Low-Performing Schools. CREDO.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, *100*(2), 267-71.

Hanushek, Peterson, Talpey & Woessmann (2019). The unwavering SES achievement gap: Trends in U.S. student performance. NBER Working Paper No. 25648.

*Harris, D. & Larsen, M. (2016). The Effects of the New Orleans Post-Katrina School Reforms on Student Academic Outcomes. Education Research Alliance for New Orleans.

*Harris, D. & Larsen, M. (2018). The Effects of the New Orleans Post-Katrina Market-Based School Reforms on Student Achievement, High School Graduation, and College Outcomes. Tulane University, Education Research Alliance for New Orleans.

Hashim, S., Kane, T., Kelley-Kemple, T., Laski, M., & Staiger, D. (2020). Have income-based achievement gaps widened or narrowed? Working Paper.

Hedges L.V., Tipton E., Johnson M.C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods,* 1, 39–65.

*Heissel, J. & Ladd, H. (2018). School turnaround in North Carolina: A regression discontinuity analysis. *Economics of Education Review*, 62, 302-320.

*Hemelt, S. W. (2011). Performance effects of failure to make Adequate Yearly Progress (AYP): Evidence from a regression discontinuity framework. *Economics of Education Review*, 30(4), 702-723.

*Hemelt, S. & Jacob, B. (2020). How does an accountability program that targets achievement gaps affect student performance? *Education Finance and Policy*, 15(1), 45-74.

*Henry, G., Guthrie, E. & Townsend, L. (2015). Outcomes and Impacts of North Carolina's Initiative to Turn Around Its Lowest-Achieving Schools. Consortium for Educational Research and Evaluation - North Carolina

*Henry, G.T., McNeill, S.M., & Harbatkin, E. (2019). Effects of school turnaround on K-3 student achievement (EdWorkingPaper No.19-66). Retrieved from Annenberg Institute at Brown University: http://edworkingpapers.com/ai19-66

*Henry, Gary T., and Erica Harbatkin. (2019). The Next Generation of State Reforms to Improve their Lowest Performing Schools: An Evaluation of North Carolina's School Transformation Initiative. (EdWorkingPaper: 19-103). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/evs5-nc27

Herman, R., Aladjem, D., McMahon, P., Masem, E., Mulligan, I., O'Malley, A.S., et al. (1999). *An educator's guide to school- wide reform.* Arlington, VA: Educational Research Service.

Herman, R., Dawson, P., Dee, T., Greene, J., Maynard, R., Redding, S., and Darwin, M. (2008). *Turning Around Chronically Low-Performing Schools: A practice guide* (NCEE #2008-4020). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

*Hernandez, M. (2019). Is There No Excuse? The Effects of the New Orleans School Reforms on School Discipline. Education Research Alliance for New Orleans.

Hochbein, C. (2012). Relegation and reversion: Longitudinal analysis of school turnaround and decline. Journal of Education for Students Placed at Risk, 17(1-2), 92-107.

Hochschild & Scovronick (2004). *The American dream and the public schools*. New York, NY: Oxford University Press.

*Holbein & Ladd (2017). Accountability pressure: Regression discontinuity estimates of how No Child Left Behind influenced student behavior. *Economics of Education Review*, 58, 55-67.

Honig, M. & Hatch, T. (2004). Crafting coherence: How schools strategically manage multiple, external demands. *Educational Researcher*, 33(8), 16-30.

Hurlburt, S., Le Floch, K. C., Therriault, S. B., & Cole, S. (2011). *Baseline analyses of SIG applications and SIG-eligible and SIG-awarded schools* (NCEE 20114019). Washington, DC: U.S. Department of Education.

Johnson, S. M., Marietta, G., Higgins, M., Mapp, K., & Grossman, A. (2015). Achieving coherence in district improvement. Cambridge, MA: Harvard Education Press.

*Kemple, J. (2015). High school closures in New York City: Impacts on students' academic outcomes, attendance, and mobility. The Research Alliance for New York City Schools.

*Kirshner, Gaertner & Pozzoboni (2010). Tracing Transitions: The Effect of High School Closure on Displaced Students. *Educational Evaluation and Policy Analysis*. 32(3), 407-429.

Kraft, M. (2020). Interpreting effect sizes of education interventions. Educational Researcher. https://scholar.harvard.edu/files/mkraft/files/kraft_2019_effect_sizes.pdf

*Larsen, M. (2018). Does Closing Schools Close Doors? The Effect of High School Closings on Achievement and Attainment. Working Paper.

Lee, J. (2008). Is test-driven external accountability effective? Synthesizing the evidence from cross-state causal-comparative and correlational studies. *Review of Educational Research*, 78(3), 608-644.

*LiCalsi, C., Citkowicz, M., Friedman, L. & Brown, M. (2015). Evaluation of Massachusetts Office of District and School Turnaround Assistance to Commissioner's Districts and Schools. American Institutes for Research.

*LiCalsi, C., Garcia Piriz, D. (2016). Evaluation of Level 4 school turnaround efforts in Massachusetts—Impact of School Redesign Grants. Washington, DC: American Institutes for Research.

Littell, J. H., Corcoran, J., & Pillai, V. (2008). *Systematic reviews and meta-analysis*. New York, NY: Oxford University Press.

Lynch, K., Hill, H. C., Gonzalez, K. E., & Pollard, C. (2019). Strengthening the research base that informs STEM instructional improvement efforts: A meta-analysis. *Educational Evaluation and Policy Analysis*, 41(3), 260–293. https://doi.org/10.3102/0162373719849044

Malen, B., Croninger, R., Muncey, D., & Redmond-Jones, D. (2002). Reconstituting schools: "Testing" the "theory of action." *Educational Evaluation and Policy Analysis*, *24*, 113–132.

Malen, B. & Rice, J.K. (2016). School reconstitution as a turnaround strategy: An analysis of the evidence. In W. Mathis & T. Trujillo (Eds.), *Learning from the federal market-based reforms: Lessons for the Every Student Succeeds Act (ESSA).* Charlotte, NC: Information Age Publishing.

Mann, E. (2016). School turnaround under ESSA: Progress, but not a silver bullet. Washington, DC: Brookings Institution.

McGuinn, P. (2006). *No Child Left Behind and the transformation of federal education policy 1965-2005*. University Press of Kansas.

McGuinn, P. (2012). Stimulating Reform: Race to the Top, Competitive Grants, and the Obama Education Agenda. *Educational Policy,* 26(1), 136-159.

Meyers, C. & VanGronigen, B. (2018). So many educational service providers, so little evidence. *American Journal of Education*, 125(1), 109-139.

Moher, D. Liberati, A., Tetzlaff, J. & Altman, D. (2009). Preferred reporting items for systemic reviews and meta-analyses: The PRISMA statement. *PLoS Med.* 6(7).

Murphy, J. & Datnow, A. (Eds.0 (2003). *Leadership lessons from the comprehensive school reform.* Thousand Oaks, CA: Corwin Press.

Murphy, J. & Meyers, C. (2008). *Turnaround Around Failing Schools: Leadership Lessons from the Organizational Sciences*. Thousand Oaks, California: Sage Publications.

Murphy, J. & Bleiberg, J. (2019). *School turnaround politics and practices in the U.S.: Learning from failed school reform.* Switzerland: Springer.

*Opper, Johnston, Engberg & Xenakis (2019). Assessing the Short-term Impact of the New York City Renewal Schools Program. RAND Corporation Working Paper.

*Orland, M., Hoffman, A.E., Vaughn, S. (2010). Evaluation of the comprehensive school reform program implementation and outcomes: fifth-year report. Washington DC: Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service.

*Osek, U., Hansen, M., & Gonzalez, T. (2012). A leg up or boot out? Student achievement and mobility under school restructuring. National Center for Analysis of Longitudinal Data in Education Research.

*Papay J., & Hannon, M. (2018). The Effects of School Turnaround Strategies in Massachusetts. Brown University Working Paper.

Peurach, D. & Neumerski, C. (2015). Mixing Metaphors: Building Infrastructure for Large Scale School Turnaround. *Journal of Educational Change*, 16(4), 379-429.

Peters, Sutton, Jones, Abrams, Rushton & Moreno (2010). Assessing publication bias in meta-analyses in the presence of between-study heterogeneity. *Statistics in Society*, 173(3), 575-591.

*Peterson, P. & Chingos, M. (2008). Impact of For-Profit and Non-Profit Management on Student Achievement: The Philadelphia Experiment. Working paper.

*Pham, L., Henry, G.T., Kho, A., & Zimmer, R. School turnaround over the long haul: An extended evaluation of Tennessee's achievement school district and local innovation zones. Working Paper.

*Player, D. & Katz, V. (2016). Assessing School Turnaround: Evidence from Ohio. *The Elementary School Journal*. 116(4), 675-698.

Pustejovsky, J. & Rodgers, M. (2019). Testing for funnel plot asymmetry of standardized mean differences. *Research Synthesis Methods*, 10(1), 57-71.

Reardon, S. (2011). "The widening academic achievement gap between the rich and the poor" in Greg Duncan and Richard Murnane eds., *Whither opportunity: Rising inequality, schools, and children's life chances.* New York, NY: Russell Sage Foundation.

*Rice, J., Bojorquez, J.C., Diaz, M., Wendt, S., Nakamoto, J., (2014) Evaluation of Michigan's School Improvement Grant: Outcomes after three years. Report submitted to the Michigan Department of Education, Office of Education Improvement & Innovation.

Rice, J. K., & Croninger, R. G. (2005). Resource generation, reallocation, or depletion: An analysis of the impact of reconstitution on school capacity. *Leadership and Policy in Schools*, 4, 73–103.

Rice, J. K., & Malen, B. (2003). The human costs of education reform: The case of school reconstitu- tion. *Educational Administration Quarterly*, *39*, 635–666.

Rice, J. K., & Malen, B. (2010). *School reconstitu- tion as an education reform strategy: A synopsis of the evidence*. Atlanta, GA: National Education Association.

*Rockoff, J., & Turner, L. J. (2010). Short-run impacts of accountability on school quality. American Economic Journal: Economic Policy, 2(4), 119-47.

*Rouse, Hannaway, Goldhaber & Figlio. (2013). "Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure." *American Economic Journal: Economic Policy*, 5(2):251-81.

*Saw, Schneider, Frank, Chen, Keesler, Martineau (2017). The Impact of Being Labeled as a Persistently Lowest Achieving School. *American Journal of Education,* 123, 585-613.

*Schueler, B., Goodman, J., & Deming, D. (2017). Can states take over and turn around school districts? Evidence from Lawrence, Massachusetts. *Educational Evaluation and Policy Analysis*, *39*, 311-332.

Scott, C. (2008). *A call to restructure restructuring: Lessons from the No Child Left Behind Act in five states.* Washington, DC: Center on Education Policy.

Scott, C. (2009). *Improving low-performing schools: Lessons from five years of studying school restructuring Under No Child Left Behind*. Washington, DC: Center on Education Policy.

*Slavin, Cheung, Holmes, Madden & Chamberlain (2013). Effects of a Data-Driven District Reform Model on State Assessment Outcomes. *American Educational Research Journal*, 50(2), 371-396.

Smarick, A. (2010). The turnaround fallacy: Stop trying to fix failing schools. Close them and start fresh. *Education Next*. 10(1), 20-26.

*Springer, M. (2008). The influence of an NCLB accountability plan on the distribution of student test score gains. Economics of Education Review, 27, 556-563.

*Steinberg, M.P. & MacDonald, J.M. (2019). The effects of closing urban schools on students' academic and behavioral outcomes: Evidence from Philadelphia. *Economics of Education Review*, 69: 25-60.

*Strunk, K. O., Marsh, J. A., Hashim, A. K., Bush-Mecenas, S., & Weinstein, T. (2016). The impact of turnaround reform on student outcomes: Evidence and insights from the Los Angeles Unified School District. *Education Finance and Policy*, 11(3), 251-282.

*Strunk, McEachin & Westover (2014). The Use and Efficacy of Capacity-Building Assistance for Low-Performing Districts. *Journal of Policy Analysis and Management*, 33(3), 719-751.

*Strunk & McEachin (2014). More Than Sanctions: Closing Achievement Gaps Through California's Use of Intensive Technical Assistance. *Educational Evaluation and Policy Analysis*, 36(3), 281-306.

*Sun, M., Penner, E., & Loeb, S. (2017). Resource- and Approach-Driven Multi-Dimensional Change. *American Education Research Journal*. 54(4), 607-643.

Sun, M., Liu, J., Zhu, J. & LeClair, Z. (2019). Using a text-as-data approach to understand reform processes. *Educational Evaluation and Policy Analysis*.

Sun, M., Kennedy, A. & Loeb, S. (2020). The Longitudinal Effects of School Improvement Grants. (EdWorkingPaper: 20-177). Retrieved from: https://doi.org/10.26300/cyyk-4r44

Tanner-Smith, E. & Tipton, E. (2014). Robust variance estimation with dependent effect sizes. *Research Synthesis Methods*, 5(1), 13-30.

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. Psychological Methods, 20(3), 375–393. https://doi.org/10.1037/met0000011

Trujillo, T. (2012). The Paradoxical Logic of School Turnarounds: A Catch-22. Teachers College Record. 16797.

Trujillo, T. & Renee, M. (2015). Irrational Exuberance for Market-Based Reform: How Federal Turnaround Policies Thwart Democratic Schooling. Teachers College Record, 117(6).

U.S. Department of Education (2010). *Evaluation of the comprehensive school reform program implementation and outcomes: fifth-year report.* Washington DC: Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service.

Washington Post (2015). Feds spent $7 billion to fix failing schools, with mixed results.

Washington Post (2017). What should America do about its worst public schools? States still don't seem to know.

*West, M. R., & Peterson, P. E. (2006). The efficacy of choice threats within school accountability systems. *The Economic Journal*, 116(510), C46-C62.

*Winters, M. (2017). Costly progress: DeBlasio's renewal school program. Manhattan Institute.

*Winters & Cowen (2012). Grading New York: Accountability and student proficiency in America's largest school district. *Educational Evaluation and Policy Analysis*. 34(3), 313-327.

Wong, K. & Meyer, S. (1998). Title I schoolwide programs: A synthesis of findings from recent evaluation. *Educational Evaluation and Policy Analysis*, 20(2), 115-136.

Yatsko, S., Lake, R., Bowen, M., & Cooley Nelson, E. (2015). Federal School Improvement Grants (SIGs). *Peabody Journal of Education, 90*(1), 27-52.

Table 1. Characteristics of Studies in the Sample

	Studies		Estimates	
	Number	Percent	Number	Percent
Total	67	100	141	100
Publication Type				
Peer- or Board-Reviewed Journal Article	34	51	81	57
Research Firm Report	10	15	18	13
Think Tank Report	2	3	5	4
University-Based Center Report	10	15	18	13
Working Paper	11	16	19	13
Author Affiliation				
Contract Research Firm	10	15	18	13
Government Agency-based Researcher(s)	1	1	1	1
University-based Researcher(s)	50	75	112	79
Multiple Types	6	9	10	7
Funder Type				
Federal Government	16	24	37	26
Foundation(s)	11	16	22	16
Foundation(s) + Federal Government	4	6	10	7
Local Education Agency	1	1	3	2
State Education Agency	6	9	8	6
Unknown	29	43	61	43
Methodology				
Randomized controlled trial	-	-	5	4
Regression discontinuity	-	-	50	35
Difference in differences	-	-	66	47
Instrumental variables	-	-	8	6
Matching or regression only	-	-	12	9

Note: Estimates refer to the total effect sizes available across all outcomes.

Table 2. Characteristics of Contexts Where Evaluated Interventions Were Implemented

	Studies		Estimates	
	Number	Percent	Number	Percent
Total	67	100	141	100
Region				
Midwest	12	18	23	16
Northeast	16	24	30	21
South	23	34	61	43
West	6	9	14	10
Multiple	7	10	8	6
Anonymous	3	4	5	4
Demographics				
Majority Receive Subsidized Lunch	-	-	129	93
Majority African American	-	-	44	48
Majority Latinx	-	-	19	21
Majority White	-	-	10	12
Grade Level				
All Grade Levels	28	42	36	26
Elementary	4	6	16	11
Elementary/Middle	27	40	62	44
Middle	2	3	11	8
Middle/Secondary	3	4	4	3
Secondary	3	4	12	9

Notes: Estimates refer to the total effect sizes available across all outcomes. Demographic percentages represent the fraction of estimates for whom information on that demographic characteristic is not missing.

Table 3. Characteristcs of Interventions Evaluated by Studies in the Sample

	Studies		Estimates	
	Number	Percent	Number	Percent
Total	67	100	141	100
Intervention Level				
District	8	12	10	7
School	59	88	131	93
Years of Treatment				
1 Year	-	-	66	47
2 Years	-	-	26	18
3+ Years	-	-	49	35
Intervention Type				
Charter Conversion	-	-	9	6
Closure	-	-	24	17
Labeling	-	-	32	23
Turnaround	-	-	76	54
Key Intervention Features[a]				
Principal Replacements	-	-	77	55
New Funding	-	-	68	48
Teacher Replacements	-	-	54	38
Administration PD/TA	-	-	51	36
Human Resources Changes	-	-	42	30
Teacher Professional Development	-	-	41	29
Data Use	-	-	35	25
Curricular Change	-	-	28	20
Change School Manager	-	-	26	18
Extended Learning Time	-	-	24	17
Governance Change	-	-	16	11
Tutoring	-	-	13	9
School Choice	-	-	13	9
Wraparound Services	-	-	11	8
Closure Interventions[b]				
Effect of Leaving School	-	-	20	83
Move to Higher Performing School	-	-	5	21
Slow Phaseout	-	-	2	8
Receive from Closed Schools	-	-	4	17

Note: Estimates refer to the total effect sizes available across all outcomes. [a]Key intervention features are not mutually exclusive, so the percentage column does not sum to 100. [b]Closure intervention subtypes are not mutually exclusive and percentages refer to the percent of closure studies.

Table 4. Pooled Effect Size Estimates of School or District Improvement Interventions and Robustness Checks

	Math	ELA	Low-stakes STEM	Low-stakes Humanities
Overall	0.062*	0.016	0.068*	0.088*
	(0.018)	(0.015)	(0.030)	(0.036)
k[n]	112[54]	103[51]	25[15]	20[11]
Sample Size Not Inferred	0.056**	0.01	0.051+	0.067+
	(0.018)	(0.014)	(0.026)	0.028
k[n]	104[48]	95[45]	23[13]	19[10]
Standard Errors Not Inferred	0.093**	0.03	0.073+	0.093+
	(0.026)	(0.022)	(0.036)	(0.041)
k[n]	85[41]	82[40]	20[12]	18[10]
No Top/Bottom 5% of Effects	0.051**	0.035**	0.046	0.091
	(0.015)	(0.011)	(0.032)	(0.047)
k[n]	100[51]	91[48]	17[12]	15[10]
No Sample Overlap Across Studies	0.078***	0.028+	0.072	0.089+
	(0.019)	(0.014)	(0.039)	(0.040)
k[n]	75[44]	66[41]	17[10]	14[9]
Student Level Data	0.101***	0.035*	0.071+	0.099+
	(0.022)	(0.017)	(0.038)	(0.048)
k[n]	70[40]	61[37]	19[12]	13[8]
Statewide Data Coverage	0.065*	0.019	0.072	0.12
	(0.027)	(0.024)	(0.042)	(0.065)
k[n]	70[33]	61[30]	16[11]	11[7]
Standard Errors Clustered Correctly	0.064**	0.016	0.076+	0.098+
	(0.020)	(0.016)	(0.034)	(0.040)
k[n]	89[44]	83[42]	20[13]	16[10]

Notes: + p<0.10, * p<.05, ** p<.01, *** p<.001. Pooled effect size estimates with robust variance estimated standard errors reported in parentheses. For sample size, k is the number of effect sizes and n is the number of studies.

Table 5. Pooled Effect Size Estimates by Study Characteristics

	Research Method	
	Math	ELA
Randomized Controlled Trial	0.181	0.109
	(0.085)	(0.069)
Regression Discontinuity (Difference)	-0.129	-0.097
	(0.093)	(0.070)
Difference-in-Differences (Difference)	-0.101	-0.090
	(0.093)	(0.076)
Instrumental Variables (Difference)	0.029	0.124
	(0.085)	(0.108)
Matching or Regression (Difference)	-0.214+	-0.120
	(0.090)	(0.071)
k[n]	112[54]	103[51]

	Author Affiliation	
	Math	ELA
University	0.065**	0.020
	(0.024)	(0.018)
Research Firm (Difference)	0.008	0.027
	(0.050)	(0.048)
Gov't Agency or Multiple (Difference)	-0.022	-0.017
	(0.041)	(0.030)
k[n]	112[54]	103[51]

	Publication Type	
	Math	ELA
Peer-Reviewed Journal Article	0.058*	0.010
	(0.024)	(0.020)
Research Firm or Think Tank Report (Difference)	0.018	0.046
	(0.046)	(0.046)
University Center Report (Difference)	0.043	0.046
	(0.061)	(0.050)
Working Paper (Difference)	0.006	0.018
	(0.060)	(0.046)
k[n]	112[54]	103[51]

	Funder Type	
	Math	ELA
Unknown	0.050+	-0.009
	(0.027)	(0.026)
Federal Government (Difference)	-0.011	0.032
	(0.035)	(0.030)
Foundation(s) (Difference)	0.044	0.068
	(0.057)	(0.046)
Foundation & Federal (Difference)	0.170	0.153
	(0.058)	(0.099)
Local or State Education Agency (Difference)	0.099	0.097
	(0.125)	(0.101)
k[n]	112[54]	103[51]

Note: + p<.10, * p<.05, ** p<.01, *** p<.001. Pooled effect size estimates with robust variance estimated standard errors reported in parentheses. For sample size, k is the number of effect sizes and n is the number of studies.

Table 6. Pooled Effect Size Estimates of Turnaround Interventions on Test and Non-Test Outcomes

	Unconstrained	Controlling for Method
Math	0.062**	0.071**
	(0.018)	(0.022)
k[n]	112[54]	
ELA	0.016	0.019
	(0.015)	(0.016)
k[n]	103[51]	
Low-stakes STEM	0.068*	-
	(0.030)	-
k[n]	25[15]	
Low-stakes Humanities	0.088*	-
	(0.036)	-
k[n]	20[11]	
Attendance	0.108	-
	(0.144)	-
k[n]	12[8]	
Discipline	-0.006	-
	(0.038)	-
	4[3]	
Graduation	0.044	0.034
	(0.054)	(0.079)
k[n]	7[6]	
Control for Matching or Regression Only Methods	no	yes

Notes: * p<.05, ** p<.01, *** p<.001. In the second column, we do not report estimates on outcomes for which there were no studies that relied on matching or regression only methods given controlling for the use of this type of method could not affect our estimates. Pooled effect size estimates with robust variance estimated standard errors reported in parentheses. For sample size, k is the number of effect sizes and n is the number of studies.

Table 7. Effect Size Estimates by Demographics of Student Population

	Race & Socioeconomic Status			
	Math		ELA	
Non-Majority Subsidized Lunch, Majority White	0.021	0.021	0.013+	0.013+
	(0.009)	(0.009)	(0.002)	(0.002)
Majority Subsidized Lunch (Difference)	0.034	0.035	0.032	0.035
	(0.032)	(0.033)	(0.032)	(0.033)
Majority African American (Difference)	0.022	0.064	0.009	0.026
	(0.069)	(0.057)	(0.042)	(0.048)
Majority Latinx (Difference)	0.219**	0.219**	0.081	0.082
	(0.069)	(0.069)	(0.074)	(0.073)
Control for Matching or Regression Only Methods	no	yes	no	yes
k[n]	72[39]		65[36]	
	Region			
	Math		ELA	
Multiple or Anonymous	0.030	0.094*	0.031	0.063
	(0.031)	(0.047)	(0.050)	(0.062)
Midwest (Difference)	-0.002	-0.064	-0.018	-0.048
	(0.038)	(0.052)	(0.053)	(0.063)
Northeast (Difference)	0.088	0.026	0.030	-0.002
	(0.052)	(0.063)	(0.063)	(0.073)
South (Difference)	0.033	-0.020	-0.024	-0.053
	(0.045)	(0.046)	(0.056)	(0.063)
West (Difference)	0.062	0.000	0.026	-0.006
	(0.085)	(0.093)	(0.062)	(0.072)
Control for Matching or Regression Only Methods	no	yes	no	yes
k[n]	112[54]		103[51]	

Note: + p<0.10, * p<.05, ** p<.01, *** p<.001. Pooled effect size estimates with robust variance estimated standard errors reported in parentheses. For sample size, k is the number of effect sizes and n is the number of studies.

Table 8. Effect Size Estimates by Years of Treatment

	All Intervention Types			
	Math		ELA	
1 Year of Treatment	0.044	0.051	0.005	0.008
	(0.024)	(0.027)	(0.023)	(0.025)
2 Years of Treatment (Difference)	0.054	0.052	0.006	0.003
	(0.043)	(0.046)	(0.039)	(0.040)
3+ Years of Treatment (Difference)	0.032	0.046	0.047	0.053
	(0.038)	(0.040)	(0.034)	(0.034)
Control for Matching or Regression Methods	no	yes	no	yes
k[n]	112[54]		103[51]	
	Turnaround Interventions Only			
	Math		ELA	
1 Year of Treatment	0.027	0.028	-0.022	-0.020
	(0.053)	(0.054)	(0.066)	(0.067)
2 Years of Treatment (Difference)	0.052	0.052	0.022	0.021
	(0.070)	(0.070)	(0.080)	(0.081)
3+ Years of Treatment (Difference)	0.064	0.077	0.078	0.091
	(0.065)	(0.067)	(0.071)	(0.074)
Control for Matching or Regression Methods	no	yes	no	yes
k[n]	61[33]		61[33]	

Notes: Pooled effect size estimates with robust variance estimated standard errors reported in parentheses. For sample size, k is the number of effect sizes and n is the number of studies. The bottom panel limits the sample to estimates of turnaround interventions only (excluding labeling, charter conversion, and closure studies).

Table 9. Effect Size Estimates by Treatment Type

	Policy Type			
	Math		ELA	
Turnaround	0.063*	0.071*	0.012	0.016
	(0.028)	(0.030)	(0.026)	(0.028)
Labeling (Difference)	-0.005	-0.012	0.003	-0.001
	(0.046)	(0.047)	(0.027)	(0.029)
Charter Conversion (Difference)	0.141	0.182+	0.116	0.155
	(0.105)	(0.079)	(0.118)	(0.087)
Closure (Difference)	-0.057	-0.023	-0.008	0.011
	(0.032)	(0.045)	(0.032)	(0.036)
Control for Matching or Regression Methods	no	yes	no	yes
k[n]	112[54]		103[51]	
	Unit of Intervention			
	Math		ELA	
School-Level	0.055**	0.064*	0.008	0.01
	(0.018)	(0.022)	(0.014)	(0.016)
District-Level (Difference)	0.118	0.109	0.122	0.120
	(0.066)	(0.067)	(0.062)	(0.062)
Control for Matching or Regression Methods	no	yes	no	yes
k[n]	112[54]		103[51]	
	Number of Key Intervention Features			
	Math		ELA	
Constant	-0.038	-0.030	-0.043	-0.041
	(0.025)	(0.025)	(0.022)	(0.023)
Effect of Each Additional Intervention Feature	0.029***	0.029***	0.018**	0.019**
	(0.006)	(0.006)	(0.005)	(0.005)
Control for Matching or Regression Methods	no	yes	no	yes
k[n]	112[54]		103[51]	

Note: + p<0.10, * p<.05, ** p<.01, *** p<.001. Pooled effect size estimates with robust variance estimated standard errors reported in parentheses. For sample size, k is the number of effect sizes and n is the number of studies.

Table 10. Exploring Robustness of Effects by Turnaround Intervention Features

	(1)	(2)	(3)	(4)	(5)	(6)
	All Intervention Types					
	Math			ELA		
Extended Learning Time	0.094+	0.082	0.066	0.099*	0.092*	0.098*
	(0.049)	(0.050)	(0.054)	(0.032)	(0.033)	(0.045)
Tutoring	0.042	0.033	0.026	-0.020	-0.022	-0.022
	(0.064)	(0.063)	(0.052)	(0.035)	(0.034)	(0.040)
Wraparound Services	0.031	0.026	0.023	0.029	0.026	0.027
	(0.048)	(0.045)	(0.042)	(0.040)	(0.041)	(0.041)
Teacher PD	0.054	0.062+	0.025	0.009	0.013	0.009
	(0.036)	(0.034)	(0.054)	(0.032)	(0.033)	(0.046)
Teacher Replacements	0.086*	0.109**	0.058	0.088*	0.100**	0.087*
	(0.033)	(0.038)	(0.042)	(0.032)	(0.033)	(0.039)
Constant	0.011	0.017	-0.017	-0.016	-0.014	-0.018
	(0.015)	(0.016)	(0.030)	(0.017)	(0.017)	(0.024)
Control for Matching or Regression Methods	no	yes	no	no	yes	no
Control for N of Intervention Features	no	no	yes	no	no	yes
k[n]		112[54]			103[51]	
	Turnaround Interventions Only					
	Math			ELA		
Extended Learning Time	0.134	0.128	0.132	0.128*	0.125*	0.148*
	(0.082)	(0.080)	(0.083)	(0.051)	(0.052)	(0.053)
Tutoring	0.036	0.040	0.031	-0.093	-0.089	-0.070
	(0.094)	(0.091)	(0.101)	(0.079)	(0.078)	(0.105)
Wraparound Services	0.017	0.012	0.015	0.011	0.009	0.011
	(0.063)	(0.063)	(0.065)	(0.064)	(0.063)	(0.063)
Teacher PD	0.077	0.088+	0.074	0.017	0.020	0.041
	(0.045)	(0.048)	(0.056)	(0.039)	(0.042)	(0.051)
Teacher Replacements	0.140*	0.133*	0.136+	0.133*	0.131*	0.163*
	(0.057)	(0.056)	(0.074)	(0.054)	(0.054)	(0.062)
Constant	-0.009	-0.006	-0.016	-0.034	-0.033	0.013
	(0.025)	(0.025)	(0.075)	(0.029)	(0.030)	(0.069)
Control for Matching or Regression Methods	no	yes	no	no	yes	no
Control for N of Intervention Features	no	no	yes	no	no	yes
k[n]		61[33]			61[33]	

Notes: ~p<.10, * p<.05, ** p<.01, *** p<.001. Pooled effect size estimates with robust variance estimated standard errors reported in parentheses. For sample size, k is the number of effect sizes and n is the number of studies.

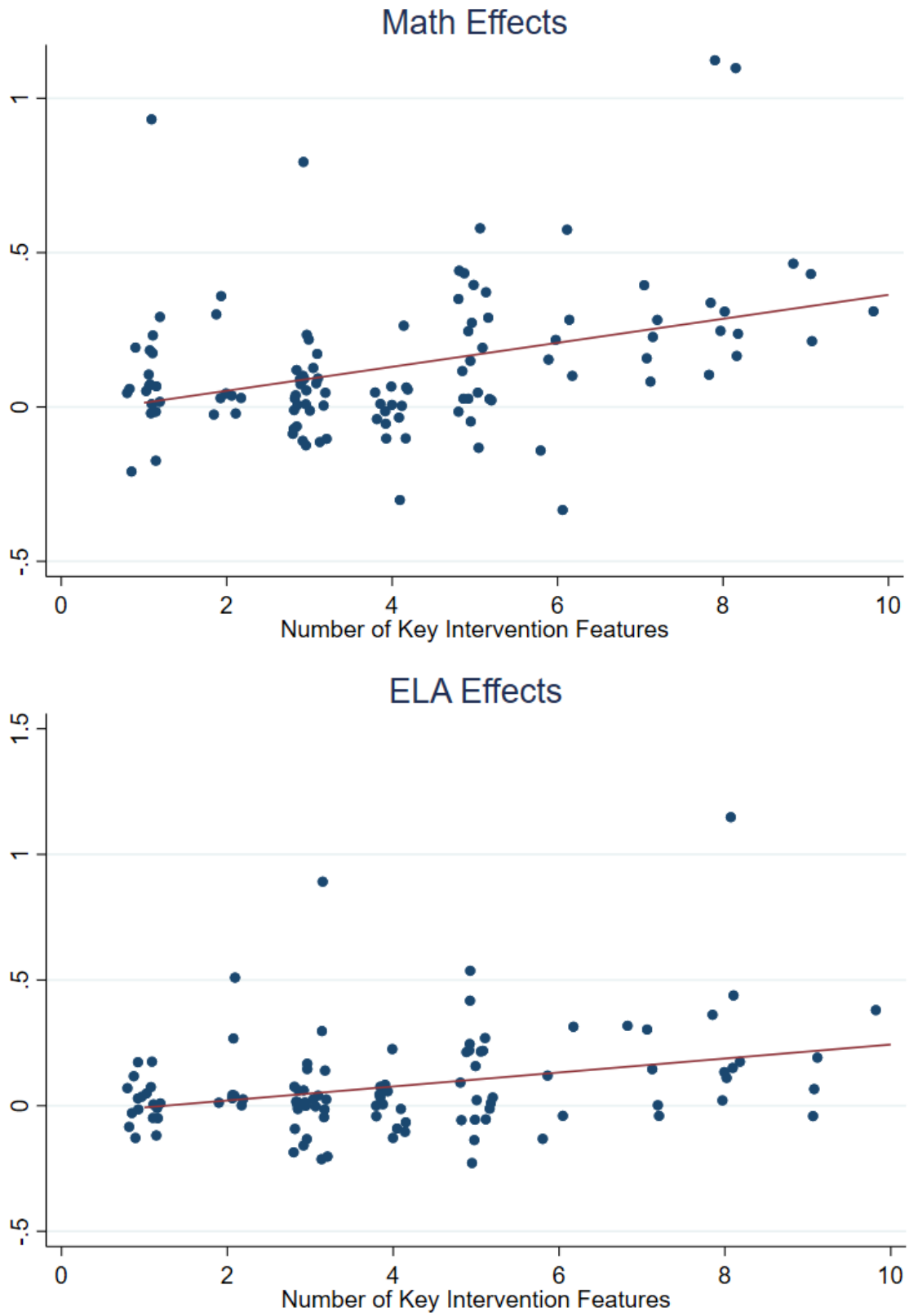
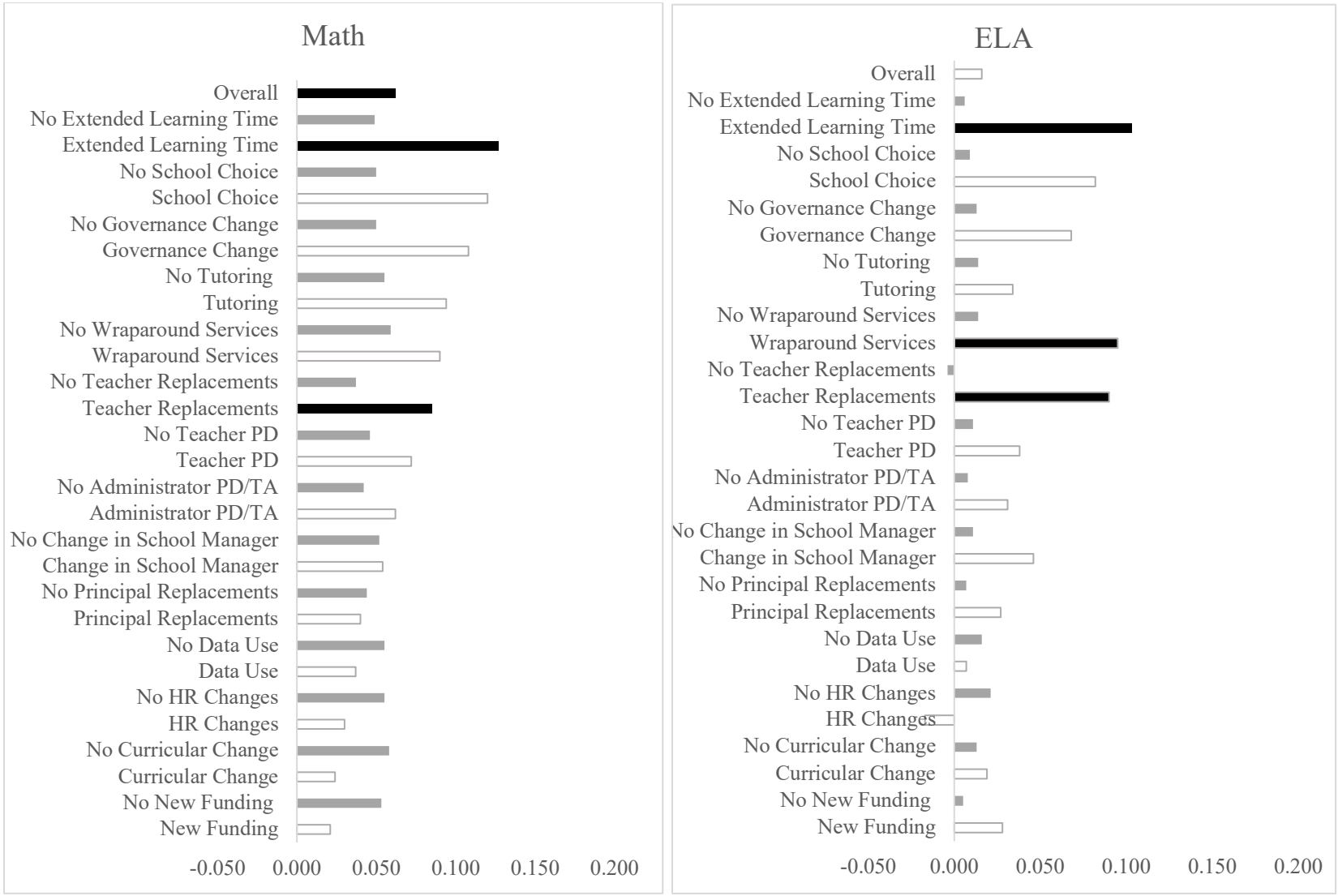


Figure 1. Turnaround Effects by Number of Intervention Features



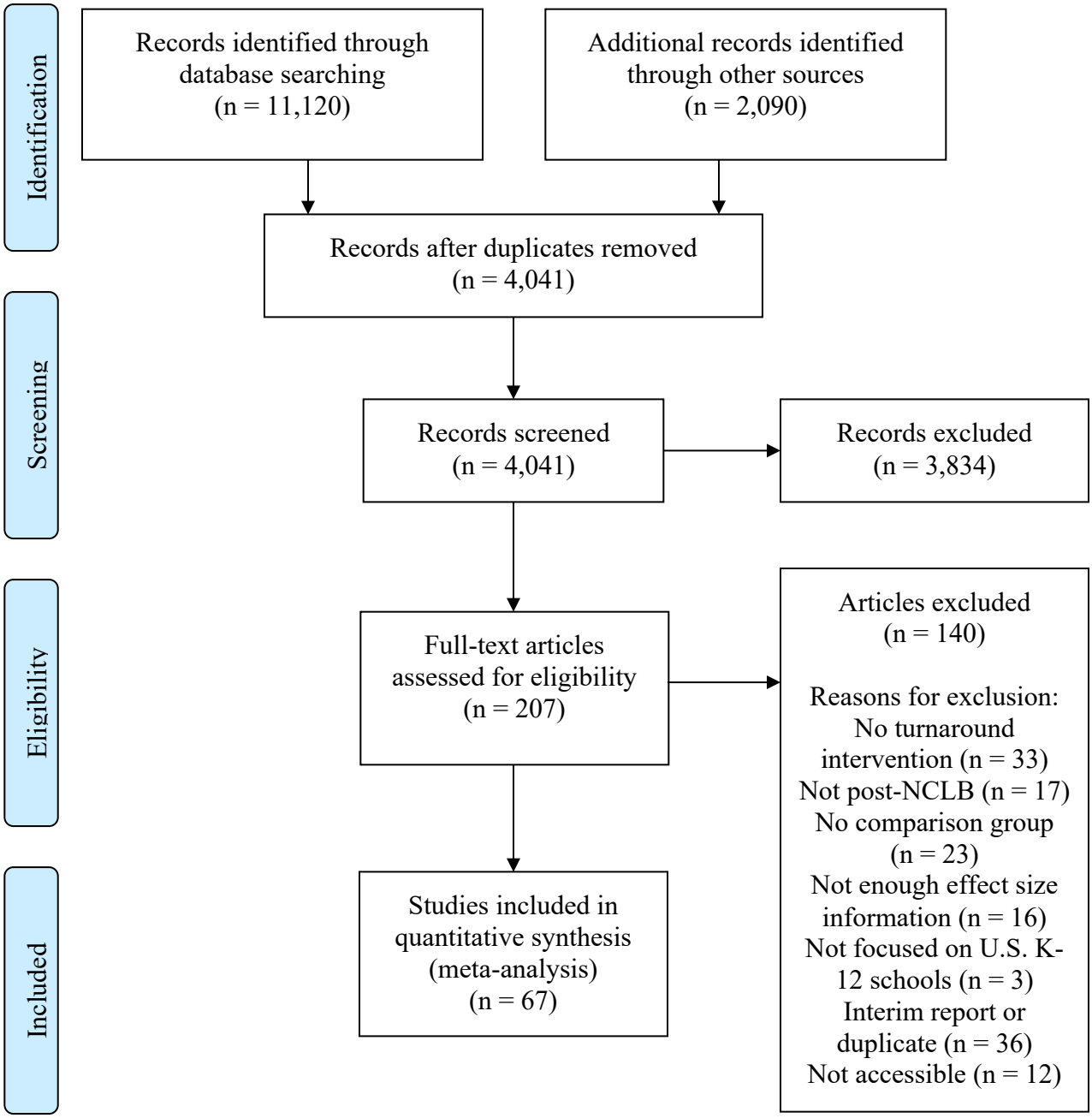
Note: Each pair (with and without a particular intervention feature) represent a separate regression. In other words, we include one intervention feature in each model. Bars in black represent differences where p<.10 (comparing with and without the feature).

Figure 2. Effect Size Estimates by Intervention Features

Appendix A1: Database Search Terms

school
AND
turnaround or "turn around" or closure or close or takeover or "take over" or transformation or restart or “charter conversion” or “focus school” or “school improvement grant” or “SIG grant” or “ESEA waiver” or “NCLB waiver” or “low performing school” or “low-performing school” or failing or “whole school reform” or “comprehensive school reform” or “success for all” or “priority school” or “failing grade” or “bottom 5%” or “bottom five percent” or “bottom 5 percent” or “lowest 5%” or “lowest five percent” or “lowest 5 percent” or “school report card” or “adequate yearly progress”
AND
impact or effect or quasi-experiment or RCT or “field experiment” or intervention or evaluation or “difference in difference” or “differences in differences” or “regression discontinuity” or causal or random or matching or “comparison group” or “comparison schools” or “comparison district” or “standard deviation” or “matched sample” or “instrumental variable”

Appendix A2: Flow Diagram of Search and Inclusion Process



Appendix Tables

Table A1. Tests of Publication Bias

	Egger's Test of Publication Bias			
	Math		ELA	
	Full-Sample	Labeling and Closure	Full-Sample	Labeling and Closure
Precision	0.007	0.014	0.004	0.010
	(0.010)	(0.018)	(0.005)	(0.011)
Intercept	1.272***	0.901	0.983**	0.390
	(0.323)	(0.594)	(0.0356)	(0.437)
[n]	[54]	[18]	[51]	[16]
	Modified Egger's Test Using RVE & Standard Error			
	Math		ELA	
	Full-Sample	Labeling and Closure	Full-Sample	Labeling and Closure
Standard Error	1.200***	0.841*	1.245***	0.482
	(0.222)	(0.287)	(0.295)	(0.380)
Intercept	-0.001	0.007	-0.027	0.000
	(0.016)	(0.014)	(0.017)	(0.009)
k[n]	112[54]	42[19]	103[51]	36[16]

Notes: + p<.10, * p<.05, ** p<.01, *** p<.001. The top panel regresses the study-level average standard normal deviation (the average effect size divided by the average standard error) on the inverse of the average standard error. The bottom panel uses RVE to regress each effect size on its standard error. For sample size, k is the number of effect sizes and n is the number of studies. Slight discrepancies in n between the two panels reflect the fact that some studies contained effect sizes for multiple types of interventions; these studies are not included in the Egger's test results.

Appendix Table A2. Effect Size Estimates by Turnaround Intervention Features

	All Studies		Turnaround Studies Only	
	Math	ELA	Math	ELA
No Extended Learning Time	0.049* (0.017)	0.006 (0.014)	0.046 (0.028)	-0.006 (0.026)
Extended Learning Time (Difference)	0.127* (0.044)	0.103** (0.026)	0.148* (0.054)	0.114** (0.035)
No Tutoring	0.055** (0.017)	0.014 (0.015)	0.075* (0.031)	0.009 (0.027)
Tutoring (Difference)	0.094 (0.101)	0.034 (0.046)	0.207* (0.035)	0.068 (0.032)
No Wraparound Services	0.059** (0.018)	0.014 (0.015)	0.063+ (0.029)	0.007 (0.026)
Wraparound Services (Difference)	0.090 (0.065)	0.095* (0.021)	0.093 (0.072)	0.102+ (0.030)
No School Choice	0.050** (0.016)	0.009 (0.014)	-	-
School Choice (Difference)	0.120 (0.092)	0.082 (0.065)	-	-
No Governance Change	0.050** (0.017)	0.013 (0.016)	0.059+ (0.029)	0.019 *0.031)
Governance Change (Difference)	0.108 (0.063)	0.068 (0.049)	0.060 (0.080)	0.034 (0.064)
No Teacher PD	0.046* (0.018)	0.011 (0.017)	0.043 (0.032)	0.002 (0.039)
Teacher PD (Difference)	0.072 (0.042)	0.038 (0.037)	0.094 (0.055)	0.046 (0.053)
No Data Use	0.055* (0.020)	0.016 (0.018)	0.062 (0.038)	0.018 (0.042)
Data Use (Difference)	0.037 (0.044)	0.007 (0.034)	0.038 (0.061)	0.001 (0.052)
No Change in School Manager	0.052** (0.018)	0.011 (0.017)	0.065+ (0.033)	0.021 (0.034)
Change in School Manager (Difference)	0.054 (0.049)	0.046 (0.034)	0.007 (0.067)	0.010 (0.048)
No Teacher Replacements	0.037* (0.016)	-0.004 (0.015)	0.033 (0.023)	-0.008 (0.026)
Teacher Replacements (Difference)	0.085* (0.036)	0.090* (0.032)	0.157* (0.059)	0.132* (0.053)
No Curricular Change	0.058** (0.019)	0.013 (0.016)	0.060 (0.033)	0.011 (0.033)
Curricular Change (Difference)	0.024 (0.044)	0.019 (0.038)	0.019 (0.064)	0.012 (0.052)
No Administrator PD/TA	0.042* (0.019)	0.008 (0.019)	0.029 (0.035)	-0.017 (0.048)
Administrator PD/TA (Difference)	0.062 (0.037)	0.031 (0.030)	0.085 (0.051)	0.058 (0.055)
No New Funding	0.053* (0.023)	0.005 (0.011)	0.068 (0.061)	-0.013 (0.044)
New Funding (Difference)	0.021 (0.038)	0.028 (0.037)	0.027 (0.069)	0.034 (0.057)
No HR Changes	0.055** (0.018)	0.021+ (0.010)	0.076* (0.033)	0.036 (0.024)
HR Changes (Difference)	0.030 (0.061)	-0.017 (0.056)	-0.020 (0.063)	-0.049 (0.054)
No Principal Replacements	0.044* (0.018)	0.007 (0.012)	0.036 (0.026)	0.011 (0.029)
Principal Replacements (Difference)	0.040 (0.028)	0.027 (0.038)	0.050 (0.058)	0.017 (0.057)
k[n]	112[54]	103[51]	61[33]	61[33]

Notes: + p<.10, * p<.05, ** p<.01, *** p<.001. Pooled effect size estimates with robust variance estimated standard errors reported in parentheses. Each panel represents a distinct meta-regression with a single covariate. In other words, we include one intervention feature in each model. In the last two columns, the sample is limited to estimates of turnaround interventions (excluding labeling, charter conversion, and closure studies). For the turnaround only sample, we did not have enough studies that included school choice to test for this difference. For sample size, k is the number of effect sizes and n is the number of studies.

Appendix Figures

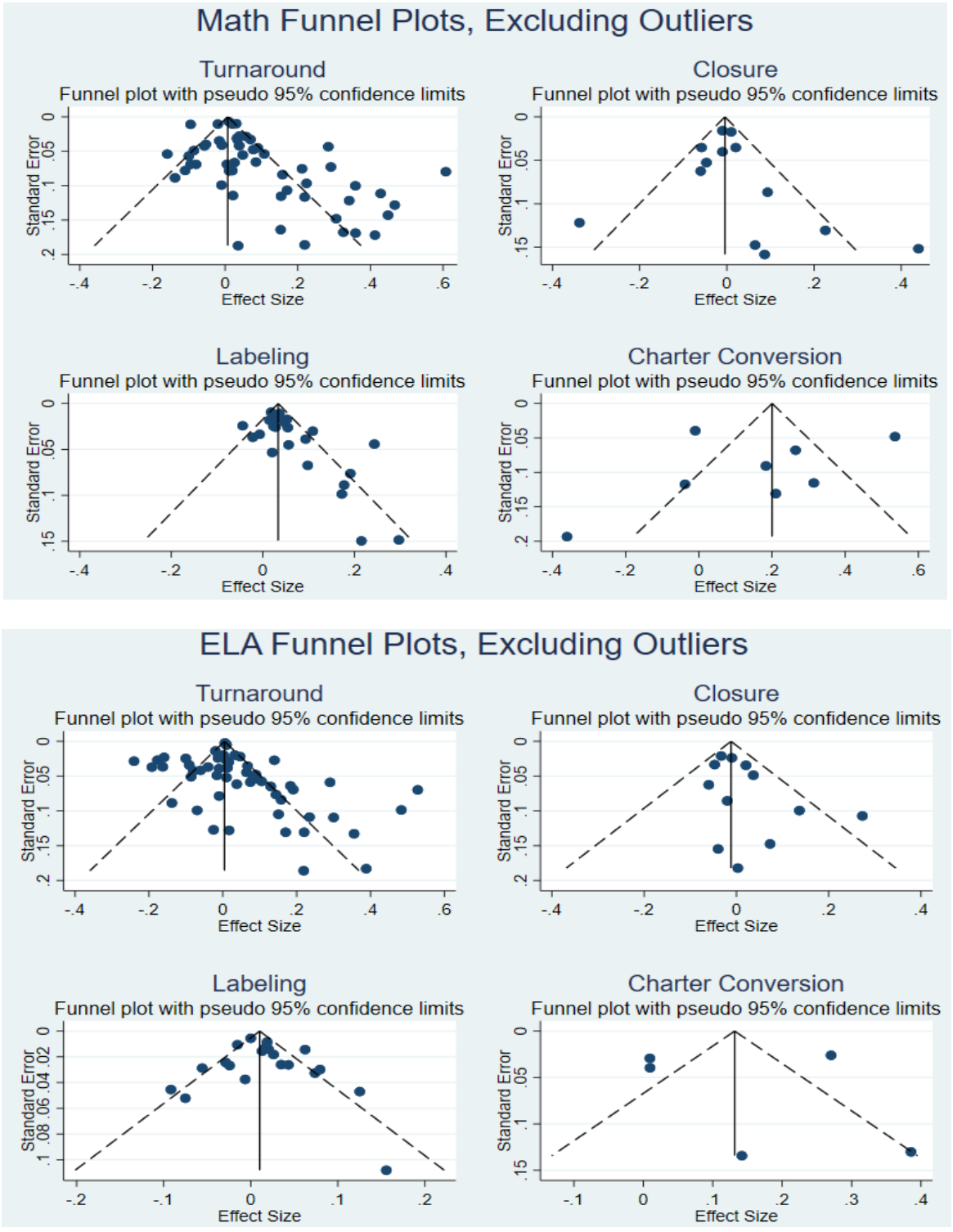


Figure A1. Funnel Plots of Math and ELA Effect Sizes by Policy Type with Full Sample