Drivers of Student Success in K-12 Education

A Systematic Synthesis of Causal Evidence

Avi Turetsky

Manus AI

(human-AI collaborative working paper)

Additional AI tools: Perplexity, Claude (Anthropic), Gemini (Google), Semantic Scholar, Elicit

May 2026 — Working Draft

Abstract

This document synthesizes the causal evidence across ten core domains of K-12 education research. The synthesis directly cites 63 high-impact papers, drawn from a systematically reviewed registry of 119 empirical studies, meta-analyses, and critical re-evaluations (all of which are included in the bibliography for further reading). We assess the magnitude and reliability of interventions including teacher quality, early childhood education, class size reduction, school funding, charter schools, reading instruction, high-dosage tutoring, social-emotional learning, out-of-school factors, and international system design. The synthesis highlights areas of strong consensus, areas of intense methodological debate, and domains where observational findings are often confounded by selection bias. We conclude by identifying 17 high-priority replication candidates with publicly available data, proposing five cross-cutting themes that unify the evidence base, and cataloging instances where policy claims have exceeded empir-

ical findings.

Declaration of AI Assistance: In accordance with emerging transparency norms for AI-assisted academic publishing, we disclose the following use of artificial intelligence in the preparation of this manuscript. **Manus AI** acted as the primary computational research assistant, executing literature retrieval, fact-checking, and drafting under human direction. **Perplexity** was employed for iterative fact-checking of effect sizes and verification of paper existence. **Claude** (Anthropic) and **Gemini** (Google) were used for independent editorial review, structural critique, and internal consistency checks. Specifically, Claude was provided with direct API access to the full text of source PDFs to conduct grounded verification of quantitative claims against the original papers. **Semantic Scholar** (via programmatic API) and **Elicit** were integrated for citation network analysis and automated literature screening. A programmatic API approach was preferred over interactive visual tools to ensure a rigorous, reproducible, and transparent methodology for identifying high-centrality missing nodes.

Contents

1 Introduction

The empirical literature on K-12 education policy is vast, politically salient, and methodologically heterogeneous. Over the past three decades, the field has undergone a "credibility revolution," moving away from simple observational correlations toward rigorous causal identification strategies that exploit natural experiments, randomized controlled trials, and quasi-experimental variation in policy rollouts. However, despite this methodological progress, significant debates remain regarding the magnitude, persistence, and scalability of key interventions.

This document provides a structured synthesis of the causal evidence across ten core research clusters. It directly cites 63 seminal papers, meta-analyses, and critical re-evaluations, drawn from a broader verified registry of 119 studies that inform the analysis. For each cluster, we review the foundational findings, examine the prevailing methodological disputes, and synthesize the current state of the evidence, with explicit attention to effect sizes and their policy implications. Effect sizes throughout this document are reported as Cohen's d (standard deviation units) unless otherwise noted. Where effect sizes are reported in other units (e.g., percentage point changes in graduation rates, earnings impacts), we note the original metric explicitly.

The synthesis is organized to serve multiple audiences. For academic researchers, we emphasize methodological debates and replication priorities. For policymakers, we flag where evidence is sufficiently robust to justify large-scale investment and where it is not. For the general public, we translate technical findings into accessible language while maintaining fidelity to the underlying research.

A note on the scope of this review: we focus primarily on the United States context, with Cluster 10 providing international comparative perspective. We prioritize studies published after 2000, though foundational older work (e.g., the Tennessee STAR experiment, the Perry

Preschool Program) is discussed where it remains the primary causal evidence. We do not attempt to be exhaustive; instead, we focus on the studies that have most influenced policy debates and that represent the strongest causal evidence available.

2 Methodological Framework

Before examining specific substantive clusters, it is necessary to establish the methodological hierarchy used to evaluate evidence in this review. Education research is particularly susceptible to selection bias: students are not randomly assigned to schools, teachers, or neighborhoods, meaning that observational correlations rarely represent true causal effects. A student who attends a high-performing charter school may have done so because of highly motivated parents who would have found other ways to support their child's education regardless of school assignment. A student assigned to a high-quality teacher may have been placed there by a counselor who recognized their potential. These selection processes, if unaddressed, produce inflated estimates of program effectiveness.

2.1 The Hierarchy of Causal Evidence

This synthesis prioritizes research designs capable of isolating causal mechanisms, ranked roughly as follows:

1. **Randomized Controlled Trials (RCTs):** The gold standard for causal inference. Students or schools are randomly assigned to treatment and control conditions, ensuring that any differences in outcomes are attributable to the intervention rather than pre-existing differences. The Tennessee Project STAR experiment (Krueger, 1999) and the Head Start Impact Study (Puma et al., 2010) are the most prominent examples in this literature. RCTs are often limited in scale or duration, and their results may not generalize to different populations or contexts.

2. **Natural Experiments and Lotteries:** Over-subscribed charter school lotteries

(Abdulkadiroğlu et al., 2016) provide RCT-equivalent evidence within specific subpopulations, as students who apply to a school are randomly assigned to receive or not receive an offer. The key assumption—that lottery losers represent a valid counterfactual for lottery winners—is generally well-supported in the urban charter school literature.

3. **Regression Discontinuity (RD):** Exploiting arbitrary administrative cutoffs to compare near-identical students on either side of a threshold. A canonical example is Maimonides' Rule (Angrist and Lavy, 1999): Israeli law requires a class to be split once enrollment exceeds 40 students, so a cohort of 41 students generates two classes of ∼20, while a cohort of 40 has one class of 40. Students on either side of this threshold are otherwise comparable, making the enrollment cutoff an instrument for class size. Students just above and just below the cutoff are assumed to be similar in all respects except their treatment status, providing a locally valid causal estimate.

4. **Difference-in-Differences (DiD):** Analyzing the differential impact of policy rollouts across states or districts over time (e.g., school finance reforms, Jackson et al. 2016). The key assumption is that treated and control units would have followed parallel trends in the absence of the policy change.

5. **Instrumental Variables (IV) and Value-Added Models (VAMs):** Using exogenous shocks or controlling for prior test scores to isolate teacher or school effects (Chetty et al., 2014a). These methods require strong and often untestable assumptions about the exclusion restriction (IV) or the absence of non-random student sorting (VAMs).

2.2 Common Threats to Validity

Even rigorous designs face significant threats to validity when translated to policy. **Scale-up effects** (general equilibrium effects) occur when an intervention works in a small trial but fails when implemented broadly due to supply constraints or behavioral responses. The most

dramatic example in this literature is California's class size reduction initiative: the policy was motivated by strong experimental evidence from Project STAR, but when implemented statewide, it required hiring so many new teachers so quickly that it diluted average teacher quality, particularly in high-poverty districts (Jepsen and Rivkin, 2009).

Fadeout is another pervasive issue, particularly in early childhood interventions (Puma et al., 2010), where initial cognitive gains dissipate within a few years, complicating cost-benefit analyses. The fadeout phenomenon does not necessarily imply that interventions are ineffective in the long run, as some research suggests that non-cognitive skills developed early may produce adult benefits even when test scores converge.

Publication bias poses a significant threat to meta-analytic conclusions. Studies with null or negative results are less likely to be published, meaning that the published literature systematically overstates effect sizes. This is particularly concerning in the SEL and growth mindset literatures, where large-scale pre-registered trials have consistently produced smaller effects than the earlier published literature suggested.

Heterogeneous treatment effects mean that an intervention's average effect may mask substantial variation across subgroups. Urban charter schools produce large positive effects for low-income minority students in cities, but near-zero or negative effects in rural and suburban settings (CREDO at Stanford University). Policymakers who apply urban charter findings to rural contexts are committing an ecological fallacy.

3 Cluster 1: Teacher Quality and Value-Added Models

The consensus in education economics is that teacher quality is the most important school-based determinant of student achievement. Before proceeding, it is necessary to define what the economic literature means by "teacher quality." Unlike traditional education research, which often measures quality through observable credentials (e.g., master's degrees, certification status, or years of experience), the modern causal literature defines teacher quality

almost exclusively as *value-added*—the marginal contribution a specific teacher makes to a student's academic growth, controlling for the student's prior achievement and demographic characteristics. As Rivkin et al. (2005) and others have repeatedly demonstrated, observable credentials explain very little (typically less than 5%) of the true variance in teacher effectiveness.

This finding is robust across multiple methodological approaches and has been replicated in dozens of countries. Foundational work by Rivkin et al. (2005) and Rockoff (2004) demonstrated that the standard deviation of teacher value-added effects on student test scores is approximately $d = 0.10$–0.15. Subsequent research using Project STAR data suggested even larger effects: Nye et al. (2004) estimated teacher variance components corresponding to an interquartile range of approximately $d = 0.34$–0.48, though this represents an upper bound; the modal estimate across the broader literature is closer to $d = 0.10$–0.20.

The magnitude of these effects is profound. To put $d = 0.15$ in perspective, a student assigned to a 75th percentile teacher rather than a 25th percentile teacher for three consecutive years would likely close a substantial portion of the average black-white achievement gap. This realization shifted the focus of education reform in the early 2000s heavily toward teacher evaluation and accountability, culminating in the Race to the Top initiative and widespread adoption of VAM-based teacher evaluation systems.

3.1 The Value-Added Debate

The central methodological debate in this cluster concerns the validity of Value-Added Models (VAMs) used to estimate teacher effectiveness. Chetty et al. (2014a,b) provided the most influential defense of VAMs, using data from more than one million students to argue that VAM estimates are unbiased predictors of teacher causal effects. They found that replacing a bottom-5% teacher with an average teacher raises the lifetime earnings (in present discounted value) of a single classroom by approximately $250,000 per year of teaching. This

figure became the most-cited number in education policy debates of the 2010s.

However, this conclusion has been rigorously contested. Rothstein (2010) demonstrated that standard VAMs fail falsification tests—teachers appear to affect students' *prior* test scores, indicating significant non-random sorting of students to teachers based on unobservable characteristics. This critique has been contested: Chetty et al. (2014a) and Goldhaber and Chaplin argue that the bias is small in magnitude (approximately 3% of the estimated VAM effect), and that Rothstein's test itself has limitations. Nevertheless, the instability of teacher rankings across different model specifications remains a major concern for high-stakes policy use. A teacher ranked in the top quartile by one VAM specification may be ranked in the bottom quartile by a slightly different specification, raising serious questions about the reliability of these measures for personnel decisions.

The American Statistical Association issued a formal statement in 2014 expressing significant concerns about the reliability and validity of VAMs, advising against their use as the sole or primary basis for high-stakes teacher evaluation decisions. Despite this, many states continued to use VAMs as a significant component of teacher evaluation systems throughout the 2010s.

3.2 Non-Cognitive Teacher Effects

Beyond the methodological debates over VAM validity, recent research has questioned whether test-score-based metrics capture the full scope of teacher influence. Jackson (2018a) found that teachers have substantial effects on non-cognitive outcomes—such as attendance, behavior, and course grades—and that these non-test-score effects predict high school graduation and college-going even when a teacher's test-score value-added is low. Crucially, Jackson found that a teacher's non-test-score value-added is weakly correlated with their test-score value-added (correlations of approximately $r = 0.15$–0.30), meaning that schools that evaluate teachers solely on test scores are missing a substantial portion of what teachers contribute

to student development.

Blazar (2018) confirmed through random assignment that teachers have causal impacts on student self-efficacy and behavior, though he cautioned against using these measures for formal evaluation given their susceptibility to social desirability bias and the difficulty of measuring them reliably at scale. Together, these findings suggest that effective teaching is multidimensional and that narrow test-score-based accountability systems may inadvertently incentivize teachers to focus on test preparation at the expense of broader developmental goals.

3.3 Teacher Labor Market and Retention

The teacher quality literature has increasingly focused on the supply side of the teacher labor market. Hanushek (2011) and subsequent work have documented that teacher salaries in the US are compressed relative to other professions requiring similar levels of education, and that the salary structure does not reward effectiveness. High-performing teachers earn roughly the same as low-performing teachers with similar experience and credentials, providing weak incentives for high-ability individuals to enter or remain in teaching.

Recent research has examined whether performance pay can improve teacher effectiveness. The evidence is mixed: some studies find positive effects of performance pay on student outcomes (Fryer, 2014), while others find that poorly designed incentive structures can crowd out intrinsic motivation and lead to gaming of evaluation metrics. The consensus is that teacher compensation reform is necessary but must be carefully designed to avoid perverse incentives.

3.4 Policy Implications and Exceeded Claims

The policy translation of VAM research frequently exceeds the empirical findings. The Chetty et al. (2014b) $250,000 lifetime earnings figure is routinely cited by policymakers

to justify aggressive termination of low-performing teachers. However, the authors themselves explicitly cautioned against this interpretation, noting that high-stakes use of VAMs would likely induce teaching-to-the-test and alter the reliability of the metric (Goodhart's Law). The empirical reality is that while teacher quality matters immensely, our ability to measure it precisely enough for high-stakes personnel decisions remains highly contested. Some researchers argue that the appropriate policy response is to invest heavily in teacher recruitment, preparation, and mentoring, while others contend that better-designed evaluation systems—even if imprecise—can improve incentives and selection. The empirical literature does not resolve this policy debate definitively.

4 Cluster 2: Early Childhood Education

Early childhood education (ECE) interventions are widely cited as having the highest return on investment in the education sector, though the evidence base reveals complex dynamics regarding the persistence of these effects. The theoretical foundation for ECE investment rests on the neuroscience of brain development, which shows that the early years (birth to age 5) are characterized by extraordinary neural plasticity, and on Heckman's skill-begets-skill model, which posits that early investments in human capital have compounding returns over the life course.

4.1 Long-Run Returns from Intensive Programs

The foundational evidence for ECE relies heavily on small-scale, intensive interventions from the 1960s and 1970s. Heckman et al. (2010) demonstrated that the HighScope Perry Preschool Program produced a 7–12% annual return on investment, with significant effects on crime reduction, employment, and earnings persisting to age 40. The Perry program was remarkable in its intensity: it provided two years of high-quality preschool education to 58 deeply disadvantaged African American children in Ypsilanti, Michigan, combined with

weekly home visits by trained teachers. Similarly, the Abecedarian Project, which provided full-time, high-quality childcare from infancy through age 5, showed large improvements in adult health and metabolic outcomes (Campbell et al., 2014), including significantly lower rates of hypertension and diabetes at age 35.

These findings have been enormously influential in policy debates, but they must be interpreted with significant caution. Both programs were tiny, targeted at the most disadvantaged children, and implemented with exceptional fidelity by highly trained staff. The external validity of these findings to modern, scaled-up public pre-K programs is far from guaranteed.

4.2 The Fadeout Problem in Modern Programs

Modern, scaled-up public pre-K programs show a strikingly different pattern. The Head Start Impact Study (Puma et al., 2010), a large-scale RCT, found that initial cognitive gains largely faded by 3rd grade, with no statistically significant differences between Head Start participants and the control group on most measures. More concerningly, the Tennessee Voluntary Prekindergarten RCT (Lipsey et al., 2018) found that by 3rd grade, pre-K participants actually scored *lower* than the control group on academic measures and had higher rates of disciplinary infractions. This finding suggests that scaling up early childhood programs without maintaining rigorous quality controls and specialized pedagogical approaches can be actively detrimental. The Tennessee program relied heavily on didactic, whole-group instruction rather than the play-based, individualized approaches characteristic of successful boutique programs.

The reconciliation of short-term test score fadeout with long-term life outcome improvements remains a central theoretical challenge. Heckman et al. (2006) has argued extensively that the primary mechanism is the development of non-cognitive skills—such as self-regulation, conscientiousness, and executive function—which are highly malleable in early childhood and predict adult success better than raw cognitive scores. If early childhood programs

successfully build these non-cognitive traits, the long-term benefits may manifest in reduced criminality and higher employment even if the initial cognitive boost dissipates. This hypothesis is supported by the long-run quasi-experimental evaluations of historical Head Start rollouts (Bailey et al., 2021) and universal pre-K programs (Cascio and Schanzenbach, 2013), which find significant long-term benefits for educational attainment and earnings, particularly for disadvantaged students, even when short-term test score effects are modest.

4.3 Policy Implications and Exceeded Claims

The rhetoric surrounding universal pre-K frequently conflates distinct programmatic models. Policymakers routinely cite the massive ROI of the Perry Preschool project to justify modern universal pre-K expansions. This is a severe extrapolation: Perry was a highly intensive, multi-year intervention targeting 58 deeply disadvantaged children, featuring weekly home visits and highly trained staff. Assuming that a modern, scaled-up state pre-K program (which often struggles with staff retention and quality control) will yield Perry-like returns is an analytical error. The Tennessee Pre-K findings (Lipsey et al., 2018) serve as a stark reminder that poor-quality ECE can be worse than no ECE at all.

The appropriate policy conclusion from this literature is not that universal pre-K is a bad investment, but that program quality is the decisive variable. Well-designed, adequately funded programs with trained staff and evidence-based curricula can produce meaningful long-term benefits. Poorly designed programs that simply provide childcare without a coherent pedagogical approach are unlikely to replicate the Perry or Abecedarian results.

5 Cluster 3: Class Size Reduction

The effect of class size on student achievement is one of the most heavily studied, yet persistently debated, topics in education policy. The intuitive appeal of smaller classes—more individualized attention, better classroom management, more time for each student—has

made class size reduction a perennially popular policy proposal, despite mixed evidence on its cost-effectiveness.

5.1 Experimental and Quasi-Experimental Evidence

The strongest causal evidence comes from the Tennessee STAR experiment. Krueger (1999) reanalysis found that students randomly assigned to small classes (13–17 students) outperformed those in regular classes (22–25 students) by approximately $d = 0.22$ in reading and $d = 0.27$ in math in kindergarten, with effects persisting through 4th grade. Importantly, the effects were significantly larger for minority and low-income students, suggesting that class size reduction may be a particularly powerful tool for reducing achievement gaps.

Outside of experimental settings, Angrist and Lavy (1999) exploited Maimonides' Rule as a natural experiment to identify the causal effect of class size in Israel. Maimonides' Rule—derived from the 12th-century Talmudic scholar Rabbi Moses Maimonides—stipulates that a class must be split once enrollment exceeds 40 students. Under this rule, a school with 40 students has one class of 40, but a school with 41 students must form two classes of approximately 20–21 students each. This creates sharp, discontinuous variation in class size that is plausibly unrelated to student characteristics: a student enrolling in a cohort of 41 is effectively randomly assigned to a class half the size of a student in a cohort of 40, purely by virtue of one additional student registering. Because families cannot precisely manipulate enrollment to land just above or below the threshold, students on either side of each cutoff (40, 80, 120, etc.) are comparable in all observable and unobservable respects, providing a clean regression discontinuity design. Angrist and Lavy (1999) found that class size reductions generated by this rule significantly improved reading and math scores, with effects comparable in magnitude to those found in the Tennessee STAR experiment. Similarly, Fredriksson et al. (2013) used a regression discontinuity design in Sweden to show that one fewer student per class raised adult earnings by approximately 3%, suggesting that class size effects persist into adulthood. These international findings provide important external

validation of the STAR results, though the magnitude of effects varies across contexts.

5.2 Cost-Effectiveness and General Equilibrium Effects

The skepticism toward class size reduction has deep roots in the empirical literature. Hanushek (1986), in an influential meta-analysis of 147 educational production function studies, found no systematic positive relationship between smaller classes and student achievement. Of 112 estimates of teacher/pupil ratio effects, only 9 showed a statistically significant positive relationship, while 14 showed a statistically significant *negative* relationship, and 89 were statistically insignificant. This "vote-counting" result—drawing on studies from across the country, covering different grade levels and outcome measures—provided the primary empirical basis for skepticism toward class size reduction as a cost-effective policy tool for over two decades.

Despite positive findings from the STAR experiment, widespread implementation of class size reduction has proven problematic. Jepsen and Rivkin (2009) showed that California's massive class size reduction initiative, which reduced class sizes from 28–30 to 20 students in grades K–3 beginning in 1996, yielded near-zero net benefits for disadvantaged students. The sudden demand for new teachers led wealthy districts to poach experienced teachers from poorer districts, forcing the latter to hire uncertified instructors. This general equilibrium effect—the labor market response to a large-scale policy change—entirely negated the structural benefit of smaller classes for the students who needed it most. The California experience is a cautionary tale about the dangers of extrapolating from small-scale experiments to large-scale policy implementation without considering supply-side constraints.

Furthermore, cross-national analyses (Wößmann and West, 2006) find inconsistent and generally small effects of class size across different education systems. Countries with large average class sizes (e.g., South Korea, Japan) consistently outperform countries with smaller classes (e.g., many European nations) on international assessments, suggesting that class size

is a weak predictor of system-level performance compared to teacher quality and curriculum rigor.

5.3 Policy Implications

Class size reduction is popular with parents and teachers but ranks poorly on cost-effectiveness metrics. Given the massive expense of hiring additional teachers and building new classrooms—estimated at $15,000–$20,000 per student per year for a reduction from 25 to 15 students—interventions like high-dosage tutoring (discussed in Cluster 7) are increasingly viewed as more efficient mechanisms for delivering individualized attention. The evidence suggests that class size reduction is most beneficial when implemented in the early grades, for disadvantaged students, and in contexts where teacher quality can be maintained during the expansion. Blanket mandates that reduce class sizes across all grades and districts without attention to teacher quality are unlikely to be cost-effective.

6 Cluster 4: School Funding and Resources

The question of whether "money matters" in education has been central to education economics for over half a century. For decades, the prevailing consensus was that it did not—a view built on two distinct pillars of research separated by thirty years: the Coleman Report in the 1960s and Hanushek's meta-analyses in the 1990s.

6.1 The Origins of the Null-Effects Consensus

The original skepticism regarding school funding stems from the Coleman Report (1966), a massive federally mandated study which famously concluded that measurable school resources had little independent effect on student achievement compared to a student's family background. This conclusion was enormously influential, providing intellectual cover for persistent funding inequities between wealthy and poor districts.

This skepticism was later reinforced and formalized in the economics literature by Hanushek's influential meta-analyses (Hanushek, 1997, 2003). Reviewing decades of production-function studies, Hanushek argued there was no consistent, statistically significant relationship between per-pupil expenditure and student achievement. Together, Coleman and Hanushek cemented a powerful narrative: throwing money at schools is an inefficient way to improve outcomes.

6.2 The Methodological Flaw in Early Research

The historical context of school finance explains why these early studies failed to find an effect. Prior to the 1990s, US school funding was overwhelmingly tied to local property taxes, resulting in massive disparities. However, early attempts to measure the impact of these disparities relied on cross-sectional observational data, pooling all districts within a state into a single regression. This approach is heavily confounded by selection bias operating in two opposite directions simultaneously. First, wealthy suburban districts have high property wealth (high spending) and advantaged student populations (high test scores), creating a positive correlation driven by family background rather than school resources. Second, state compensatory funding often directs *more* money to poor urban districts precisely because they have the *lowest* achievement, creating a negative correlation between spending and test scores. When these two opposing mechanisms are pooled together in a single cross-sectional regression, the true positive causal effect of funding is masked, often resulting in an aggregate estimate near zero.

6.3 The Quasi-Experimental Revolution

This null-effects consensus was initially challenged by Greenwald et al. (1996), whose competing meta-analysis found significant positive effects, but the debate only shifted decisively with the advent of modern quasi-experimental designs evaluating court-ordered school finance reforms.

Jackson et al. (2016) provided the most compelling evidence, using variation in school finance reforms across states and over time to estimate the causal effect of funding increases. They found that a 10% increase in per-pupil spending across all 12 school years led to approximately 7.25% higher adult wages, a 9.5 percentage point higher probability of graduating high school, and a 3.67 percentage point reduction in adult poverty (published *QJE* estimates). Crucially, effects for children from low-income families were substantially larger—approximately 9.5% wage gains, an 11.6 percentage point increase in high school graduation, and a 6.8 percentage point poverty reduction—compared to near-zero effects for children from higher-income families, suggesting that targeted funding increases for disadvantaged schools can be a powerful tool for reducing inequality. The Jackson et al. findings have been widely replicated and are now considered the strongest causal evidence on the long-run effects of school funding.

6.4 Mechanisms and Distributional Impacts

Recent studies consistently confirm that targeted funding increases improve outcomes. Lafortune et al. (2018) showed that post-1990 finance reforms significantly increased per-pupil spending in low-income districts by approximately $1,200 per year and raised test scores in those districts by approximately 0.10 standard deviations over ten years, effectively closing a portion of the achievement gap. Hyman (2017) found that Michigan's finance reform raised college enrollment and graduation rates. The mechanisms driving these improvements typically include reductions in class size, increases in teacher salaries (which attract better candidates), and capital improvements. Jackson et al. (2016) document the input changes within their own sample; Neilson and Zimmerman (2014) provide complementary evidence specifically on the effects of school capital construction.

The debate is no longer *whether* money matters, but *how* it is spent. Unrestricted block grants often yield lower returns than funds explicitly targeted toward instructional quality or high-need populations. This reconciles the apparent contradiction between the positive

findings of Jackson et al. (2016) regarding general funding increases and the null findings of Jepsen and Rivkin (2009) regarding California's Class Size Reduction initiative. When new funding is deployed in ways that inadvertently dilute teacher quality (as in California), the structural benefits are negated. Conversely, when funding is used to attract and retain high-quality educators or improve capital facilities, the returns are substantial.

6.5 Policy Implications

This literature supports a causal role for school resources in long-run outcomes, particularly for low-income children. The current system of local property-tax-based school finance is associated with substantial across-district variation in per-pupil resources. Translating this evidence into specific finance-system reforms involves additional design questions—including funding allocation, accountability for spending, and local-control tradeoffs—on which the empirical literature is more limited. Simply increasing spending without attention to how it is used is insufficient. The most effective funding reforms are those that (1) target resources toward high-need students and districts, (2) provide flexibility for districts to allocate funds based on local needs, and (3) include accountability mechanisms to ensure that funds are used for evidence-based interventions.

7 Cluster 5: Charter Schools and Vouchers

Research on school choice—encompassing charter schools and voucher programs—reveals highly heterogeneous effects that depend heavily on the specific model, regulatory environment, and student population served. The school choice literature is also among the most politically contested in education research, with advocates and critics often selectively citing evidence to support their prior positions.

7.1 The "No Excuses" Charter Model

The most robust positive findings in the charter literature come from urban "No Excuses" charter schools. Using lottery-based designs, Angrist et al. (2013a) found that Boston charter schools generated substantial test score gains. Per year of attendance, middle school lottery estimates were approximately $d = 0.20$ in English Language Arts and $d = 0.40$ in math; high school estimates were approximately $d = 0.22$ in ELA and $d = 0.29$ in math. Middle school math effects were larger than high school math effects, while ELA effects were comparable across levels. Abdulkadiroğlu et al. (2016) confirmed similar effects in New York City charter schools. Importantly, Cohodes et al. (2021) demonstrated that these Boston charters maintained their large positive effects even as the sector scaled up, suggesting the model is replicable within urban contexts.

The "No Excuses" model is characterized by extended school days and years, strict behavioral expectations, frequent formative assessments, intensive teacher coaching, and a strong college-preparatory culture. These structural features are thought to be the active ingredients driving the model's success, though it remains unclear which elements are most critical and whether the model can be successfully transplanted to different demographic contexts.

Conversely, non-urban charters and virtual charter schools generally show null or negative effects on student achievement. The CREDO at Stanford University (2015) national report highlighted this variance, showing that while urban charters often outperform traditional public schools, the average national effect is near zero due to severe underperformance in the virtual and suburban sectors. Virtual charter schools, in particular, have consistently shown large negative effects, with students losing the equivalent of 180 days of learning in math and 72 days in reading compared to traditional public school students (CREDO at Stanford University, 2015).

7.2 Voucher Programs and the Supply-Side Mechanism

Evidence on private school vouchers is mixed and increasingly negative in recent large-scale programs. While early evaluations of the Milwaukee voucher program (Rouse, 1998) and the New York scholarship program (Howell et al., 2002) showed modest positive effects for some subgroups, recent statewide programs have yielded concerning results.

Abdulkadiroğlu et al. (2018) found that the Louisiana Scholarship Program caused severe test score declines—approaching $d = -0.40$ in math—for voucher recipients in the first year, a finding consistent with a supply-side interpretation. The divergence between early small-scale voucher results and recent statewide programs likely reflects these supply-side dynamics: when programs scale beyond the capacity of high-quality private schools willing to accept regulation, average effects decline precipitously. In mature, highly regulated sectors (like Boston charters), students transfer to demonstrably higher-quality schools. In rapidly expanding, lightly regulated voucher programs, the participating private schools often experienced declining enrollment prior to the program and may offer lower instructional quality than the public schools students left behind.

7.3 Policy Implications and Exceeded Claims

The school choice literature illustrates the dangers of generalizing from the best examples to the average. Urban "No Excuses" charters produce genuine, large, and replicable gains for disadvantaged students in cities. But these results cannot be extrapolated to voucher programs, virtual charters, or rural charter schools. The policy implication is not that school choice is uniformly good or bad, but that the regulatory environment and quality control mechanisms are decisive. Choice without quality assurance is likely to harm the students it is intended to help.

8 Cluster 6: Reading Instruction

The debate over how to teach reading—often termed the "Reading Wars"—is one of the few areas in education research where a strong scientific consensus has emerged on a specific question: the value of systematic phonics instruction for early word recognition. Debates persist regarding broader instructional design, implementation fidelity, and the role of comprehension and oral-language instruction. The scientific consensus on reading instruction has been building for decades, but has only recently begun to translate into widespread policy change.

8.1 The Scientific Consensus on Phonics

The National Reading Panel (2000) established that effective reading instruction requires five components: phonemic awareness, phonics, fluency, vocabulary, and comprehension. Meta-analyses consistently show that systematic phonics instruction produces significantly better outcomes than non-systematic or whole-language approaches, with effect sizes around $d = 0.41$ (Ehri et al., 2001). Recent reviews (Castles et al., 2018) confirm that the alphabetic principle is essential for reading acquisition: children must learn to decode the relationship between letters and sounds before they can read fluently and comprehend complex texts. A 2025 exploratory quantitative analysis by Hansford et al. (2025) found that structured literacy outperformed balanced literacy across a sample of studies (unweighted means: $d = 0.43$ vs. $d = 0.19$; weighted means: $d = 0.46$ vs. $d = 0.29$). The authors describe their work as exploratory rather than a formal meta-analysis, and the balanced literacy estimate's confidence interval includes zero in the weighted analysis; readers should interpret these figures accordingly.

The policy implications of this consensus are profound. Despite the clear evidence favoring systematic phonics, surveys of teacher preparation programs frequently find that balanced literacy and whole-language approaches remain dominant in curricula. This disconnect between

the evidence base and classroom practice represents one of the most significant translational failures in education policy. The "Science of Reading" movement, which gained significant momentum in the early 2020s, has begun to change this, with many states passing legislation requiring phonics-based reading instruction and updating teacher preparation standards.

8.2 Intervention Fadeout: The Case of Reading Recovery

While early intervention is critical, sustaining gains remains challenging. Reading Recovery, a widely used early intervention, provides intensive, one-on-one daily tutoring for 12 to 20 weeks to first-grade students struggling with beginning reading. The program initially showed strong short-term gains of $d = 0.30$–0.42 in an i3 scale-up RCT (May et al., 2016). However, a long-term regression discontinuity follow-up of the same cohort (May et al., 2023) found that by fourth grade, Reading Recovery students actually had *lower* reading scores than the control group. This finding has been contested on methodological grounds, including concerns about differential attrition and the possibility that the comparison group received more subsequent intervention; a neutral reading of the evidence is that the long-term benefits of Reading Recovery are uncertain and that the fadeout finding warrants replication. This fadeout underscores the necessity of continuous, high-quality core instruction (Tier 1) rather than relying solely on short-term pull-out interventions.

The Reading Recovery finding is particularly instructive because the program is widely considered a model of evidence-based practice in reading intervention. The fact that even a well-designed, evidence-based intervention can produce negative long-term effects when not paired with strong Tier 1 instruction highlights the importance of viewing interventions as supplements to, rather than replacements for, high-quality core instruction. Tier 1 instruction that is grounded in the Science of Reading—systematic phonics, explicit vocabulary instruction, and structured comprehension strategies—is the foundation on which all other reading interventions must build.

8.3 Policy Implications

The reading instruction literature offers one of the clearest policy mandates in all of education research: systematic, explicit phonics instruction should be the foundation of all early reading programs. States and districts that have made this transition—including Mississippi, which has seen dramatic improvements in National Assessment of Educational Progress (NAEP) reading scores over the past decade—provide compelling evidence that evidence-based reading instruction can produce meaningful gains at scale.

9 Cluster 7: High-Dosage Tutoring

High-dosage tutoring has emerged as one of the most effective and reliably replicable interventions in K-12 education, standing in contrast to the mixed results often seen in broader school reform efforts. The COVID-19 pandemic, which produced historic learning losses particularly concentrated among disadvantaged students, has dramatically increased interest in tutoring as a recovery mechanism.

9.1 Efficacy and Effect Sizes

The foundational meta-analysis by Cohen et al. (1982) established that tutoring programs generally produce positive effects, but recent rigorous evaluations have clarified the specific parameters required for success. Nickow et al. (2020) synthesized recent experimental evidence across 96 tutoring studies, finding that tutoring programs produce a pooled average effect size of $d = 0.37$ on learning outcomes. These are large effects by the standards of education interventions, comparable to the effects of attending a high-quality urban charter school.

Cook et al. (2015) provided a striking demonstration of these effects in a randomized controlled trial in Chicago public schools, where daily individualized math tutoring for male high

school students generated a $d = 0.65$ increase in math scores and significantly reduced course failures. These gains are particularly notable given the difficulty of moving test scores for older adolescents, who are typically less responsive to educational interventions than younger children. The success of the Chicago program demonstrated that high-dosage tutoring can serve as a powerful remedial tool even for students who are multiple grade levels behind. Crucially, the intervention was tightly structured, using a specific curriculum and frequent formative assessments to guide instruction, rather than relying on unstructured homework help.

9.2 Implementation Parameters

The literature indicates that tutoring is most effective when conducted during the school day rather than after school, and when delivered by teachers or paraprofessionals rather than volunteers (Kraft and Falken, 2021). The during-school-day requirement is significant: after-school tutoring programs typically suffer from low attendance rates, particularly among the most disadvantaged students, who face the greatest barriers to participation (transportation, family obligations, competing priorities). Embedding tutoring within the school day eliminates these barriers and ensures that the students who need it most actually receive it.

Furthermore, while 1-on-1 tutoring is the gold standard, small group tutoring (up to 1-to-4 ratio) can maintain much of the efficacy at a significantly lower cost, making it a viable policy lever for scale-up. Nickow et al. (2020) found that the effect size for small-group tutoring ($d = 0.30$) was only modestly smaller than for 1-on-1 tutoring ($d = 0.40$), while the cost per student was substantially lower. This finding is critical for policymakers considering large-scale tutoring programs, as it suggests that high-dosage tutoring can be scaled without proportional increases in cost.

9.3 Cost-Effectiveness and Scale-Up

High-dosage tutoring is not cheap: well-designed programs typically cost $2,000–$4,000 per student per year. However, when compared to the cost of class size reduction ($15,000–$20,000 per student per year for a reduction from 25 to 15 students) or the long-run costs of educational failure (lower earnings, higher incarceration rates, reduced civic participation), the cost-effectiveness of tutoring is highly favorable. The post-COVID policy landscape has seen significant federal investment in tutoring programs through the American Rescue Plan, and early evidence from these programs is promising.

10 Cluster 8: Social-Emotional Learning and Non-Cognitive Skills

The role of non-cognitive skills—such as grit, growth mindset, self-control, and social-emotional competencies—in driving student success has garnered substantial attention, though the causal evidence for interventions targeting these skills is heterogeneous. The literature in this cluster is characterized by a sharp divergence between the strong evidence for universal SEL programs (comprehensive curricula delivered to all students in a classroom to build broad behavioral and social skills) and the weak evidence for targeted psychological interventions (brief, specific exercises designed to alter individual students' internal mindsets, such as grit or growth mindset).

10.1 Universal SEL Programs

Durlak et al. (2011) conducted a highly influential meta-analysis of 213 school-based universal SEL programs, showing that they improved academic achievement by $d = 0.27$, reduced conduct problems, and improved social skills. These programs, which focus on concrete behavioral skills, classroom climate, and explicit prosocial routines, show consistent positive

effects across diverse populations and settings. The CASEL (Collaborative for Academic, Social, and Emotional Learning) framework, which underpins most evidence-based SEL programs, emphasizes five core competencies: self-awareness, self-management, social awareness, relationship skills, and responsible decision-making.

The mechanisms through which SEL programs improve academic achievement are not fully understood, but likely include reductions in disruptive behavior (which creates a better learning environment for all students), improvements in students' ability to regulate their emotions and focus on academic tasks, and the development of positive relationships with teachers and peers that support learning.

10.2 Targeted Psychological Interventions: Grit and Growth Mindset

Conversely, targeted psychological interventions aimed at shifting specific internal beliefs (grit, growth mindset) have faced intense scrutiny. Credé et al. (2017) conducted a meta-analysis of the grit literature, concluding that grit is largely redundant with conscientiousness and that its incremental validity for predicting performance is near zero after controlling for other personality traits. Similarly, Sisk et al. (2018) found that growth mindset interventions produce very small overall effects ($d = 0.08$ for intervention studies). While effects were larger for high-risk student populations, the authors noted these results should be interpreted with caution due to the small number of effect sizes and non-significant moderator comparisons. Yeager et al. (2019) demonstrated that growth mindset interventions only improve achievement in schools where peer norms support challenge-seeking, suggesting that the social context is a critical moderator.

These divergent findings require careful reconciliation. Universal SEL programs (such as those analyzed by Durlak et al. (2011)) typically focus on concrete behavioral skills, classroom climate, and explicit prosocial routines, yielding reliable positive effects. Targeted

psychological interventions aimed at shifting specific internal beliefs (grit, growth mindset) are highly context-dependent and generally weak at scale. The synthesis suggests that building concrete behavioral routines is effective, whereas brief psychological "nudges" to alter internal mindsets are unlikely to produce sustained academic gains without supportive environmental conditions.

10.3 Policy Implications and Exceeded Claims

The enthusiasm for "grit" (Duckworth et al., 2007) rapidly outpaced the evidence base. Policymakers and schools began grading students on grit, despite the creator's explicit warnings against using self-report surveys for high-stakes evaluation (Duckworth, 2016). The empirical reality is that while non-cognitive skills matter immensely (Heckman et al., 2006), brief psychological "nudges" to alter them rarely produce sustained academic gains at scale. The appropriate policy response is to invest in comprehensive, evidence-based SEL programs that build concrete skills and positive classroom climates, rather than in brief, targeted mindset interventions that promise transformative results from minimal inputs.

11 Cluster 9: Out-of-School Factors

The Coleman Report (1966) famously highlighted the dominant role of family background in determining student achievement, concluding that "schools bring little influence to bear on a child's achievement that is independent of his background and general social context." Modern research continues to validate the core finding that out-of-school factors are powerful determinants of educational outcomes, while also demonstrating that high-quality schools can significantly mitigate, though not eliminate, the effects of disadvantage.

11.1 The Income-Achievement Gap and Parental Investment

Recent decades have seen a massive widening of the income-achievement gap. Reardon (2011) demonstrated that the achievement gap between children from high- and low-income families is roughly 30 to 40 percent larger among children born in 2001 than among those born twenty-five years earlier. This divergence is driven largely by disparities in parental investment, particularly early childhood enrichment activities and summer learning opportunities, rather than purely school-based factors. Reardon's analysis showed that while the black-white achievement gap has narrowed significantly since the 1970s, the income-achievement gap has widened, suggesting that economic inequality is now the primary driver of educational inequality in the United States.

The mechanisms through which income affects achievement are multiple and mutually reinforcing. Higher-income families invest more in children's cognitive development through books, educational toys, enrichment activities, and private tutoring. They live in neighborhoods with better schools, lower crime rates, and more social capital. They experience less chronic stress, which affects parental sensitivity and children's cortisol regulation. And they have greater access to healthcare, nutrition, and stable housing, all of which affect cognitive development.

11.2 COVID-19 Learning Loss

The COVID-19 pandemic provided a stark, involuntary natural experiment in the effects of out-of-school time and school disruption. Goldhaber et al. (2022) and others have documented historic declines in NAEP scores, with the severest impacts concentrated among low-income students and those in districts that maintained remote instruction for longer periods. The 2022 NAEP results showed the largest average score declines since the assessment began in the 1970s, with 4th grade reading scores falling by 3 points and 8th grade math scores falling by 8 points nationally. These declines were not evenly distributed: stu-

dents in the lowest-performing quartile experienced the largest losses, widening already-large achievement gaps.

This catastrophic learning loss underscores the critical, equalizing function that physical schools perform, despite their inability to fully close out-of-school gaps. Schools provide not only instruction but also nutrition, social interaction, mental health support, and a structured environment that supports learning. The pandemic demonstrated that these non-instructional functions are critical components of what schools contribute to child development.

11.3 Neighborhood and Family Poverty

Neighborhood poverty and family socioeconomic status exert profound effects on cognitive development and school readiness (Wolf et al., 2017). Borman and Dowling (2010) reanalyzed Coleman's data using modern multilevel models, confirming that family background remains the primary driver of inequality. These findings are reinforced by Chetty et al.'s Moving to Opportunity research, which demonstrates that moving to lower-poverty neighborhoods during childhood produces substantial long-run improvements in earnings, college attendance, and other outcomes. The Moving to Opportunity findings are particularly striking because they suggest that neighborhood effects operate through mechanisms beyond school quality—including peer effects, local labor market connections, and exposure to violence—that cannot be addressed through school reform alone.

11.4 The Summer Learning Gap

A significant portion of the achievement gap forms during the summer months. Alexander et al. (2007) demonstrated that the "summer slide" accounts for approximately two-thirds of the 9th-grade reading achievement gap between high- and low-income students, with low-income students losing ground during the summer while higher-income students maintain

or improve their skills. Interventions to address this, such as voluntary summer reading programs, show promise but yield modest effects ($d \approx 0.14$) (Kim, 2006). More intensive summer programs, including summer school and summer learning camps, produce larger effects but face significant challenges with attendance and engagement.

12 Cluster 10: International Systems

Comparative international research provides a macro-level perspective on education system design, often using data from PISA (Programme for International Student Assessment) and TIMSS (Trends in International Mathematics and Science Study). While the US context is unique, international comparisons provide crucial benchmarks for what is possible when education systems are designed coherently at the national level.

12.1 System-Level Drivers of Performance

Hanushek and Woessmann (2015) argue that the "knowledge capital" of a nation—as measured by cognitive test scores—is a powerful determinant of long-run economic growth. They estimate that a one standard deviation increase in a nation's cognitive skills (as measured by PISA/TIMSS) is associated with approximately 2% higher annual long-run economic growth, an effect that compounds dramatically over decades. This finding provides a powerful macroeconomic rationale for investing in education quality, beyond the individual-level returns to schooling.

High-performing systems often share common characteristics: highly selective teacher recruitment (Finland recruits teachers from the top third of university graduates; South Korea from the top 5%), rigorous national curricula that emphasize deep conceptual understanding over procedural fluency, significant autonomy for teachers and schools coupled with strong accountability mechanisms, and high social status for the teaching profession. Mourshed et al. (2010) found that improving school systems adopt interventions appropriate to their

specific performance stage (e.g., basic literacy and numeracy for poor-to-fair systems, versus professionalizing teaching for good-to-great systems), while sharing a common reliance on the continuity of new leadership to sustain reform. Mourshed et al. (2010) found that improving school systems adopt interventions appropriate to their specific performance stage (e.g., basic literacy and numeracy for poor-to-fair systems, versus professionalizing teaching for good-to-great systems), while sharing a common reliance on the continuity of new leadership to sustain reform. These features stand in sharp contrast to the US system, which recruits teachers from a wide range of academic backgrounds, lacks a coherent national curriculum, and has struggled to elevate the status and compensation of teaching as a profession.

12.2 Lessons from High-Performing Systems

The international evidence suggests several lessons for US education policy. First, teacher quality is the most consistent predictor of system-level performance, and systems that invest heavily in teacher recruitment, preparation, and professional development consistently outperform those that do not. Second, curriculum coherence matters: systems with clear, rigorous, and well-sequenced curricula produce better outcomes than those with fragmented or incoherent curricula. Third, equity and excellence are not in tension: the highest-performing systems (Finland, Singapore, South Korea) also have among the smallest achievement gaps, suggesting that a focus on equity does not come at the expense of overall performance.

Translating these system-level features to the decentralized US context remains challenging. Porter et al. (2022) note that federal efforts result in highly variable implementation due to local control, and that the US system's structural features—including local property-tax-based funding, decentralized curriculum decisions, and weak national standards—create significant barriers to the kind of coherent, system-level reform that has characterized high-performing nations. The Common Core State Standards initiative represented an attempt to address the curriculum coherence problem, but its political backlash illustrates the difficulty of implementing national-level reforms in the US context.

12.3 Integration with the Cross-Cutting Synthesis

The international evidence operates at a different level of analysis than the other clusters in this review, which focus primarily on specific interventions within the US context. However, it provides important context for interpreting the US findings. The fact that high-performing international systems consistently prioritize teacher quality and curriculum coherence reinforces the findings from Clusters 1 and 6 that these are the most important levers for improving student outcomes. The international evidence also suggests that the US system's structural features—particularly its reliance on local property taxes for school funding—are a significant barrier to achieving the kind of equity that characterizes high-performing systems, reinforcing the findings from Cluster 4.

13 Publicly Available Data Sources

For researchers seeking to replicate or extend the findings in this review, several high-quality, publicly available datasets are essential. The following table summarizes the most important datasets, their coverage, and their primary uses in education research.

Table 1: Key Publicly Available Datasets for K-12 Education Research

Dataset	**Coverage**	**Primary Uses**
Stanford Education Data Archive (SEDA)	All US public school districts, 2009–present	Studying funding equity, systemic reforms, and achievement gaps across districts

Continued on next page

Dataset	**Coverage**	**Primary Uses**
National Assessment of Educational Progress (NAEP)	State-level, 1969–present	Long-run trends in achievement, state policy comparisons
Early Childhood Longitudinal Study (ECLS-K)	Birth cohorts, K–8th grade	Early childhood fadeout, non-cognitive skill development, family background effects
Common Core of Data (CCD)	All public schools, 1986–present	Demographic and fiscal data, school characteristics
PISA and TIMSS	International, 3-year cycles	Cross-country comparisons of achievement and system features
National Longitudinal Survey of Youth (NLSY)	Birth cohorts, 1979–present	Long-run adult outcomes, intergenerational mobility
Current Population Survey (CPS)	Monthly, 1940–present	Educational attainment, labor market outcomes
American Community Survey (ACS)	Annual, 2005–present	Local demographic and economic conditions, school district characteristics

Several of these datasets require restricted-use data agreements for access to individual-level records. The ECLS-K and NLSY restricted-use files, in particular, contain sensitive information that requires approval from the National Center for Education Statistics (NCES) or the Bureau of Labor Statistics (BLS). Researchers should plan for a 3–6 month approval process when designing studies that require these data.

14 Citation Distortion and the Policy-Research Gap

A recurring pattern in the K-12 education literature is the systematic exaggeration of research findings as they propagate from academic journals into policy advocacy. This phenomenon is not merely anecdotal; it has been formally documented across multiple scientific disciplines and is particularly acute in education research, where the policy stakes are high and the demand for actionable interventions often outpaces the evidence base.

14.1 Mechanisms of Distortion

The formal literature on knowledge propagation identifies several distinct mechanisms through which research findings are overstated or distorted over time:

1. **Citation Distortion:** Greenberg (2009) analyzed citation networks and identified three specific pathways of distortion: *citation bias* (preferentially citing papers that support a claim while ignoring refutations), *amplification* (expanding a belief system through papers that present no new data but merely repeat the claim), and *invention* (the gradual conversion of a hypothesis into established fact through citation alone, with no new empirical support).

2. **Promising Trials Bias:** In the specific context of education research, Sims et al. (2023) quantified "promising trials bias"—the tendency for early, small-scale randomized controlled trials to systematically overestimate the true effect size of an intervention. In a retrospective analysis of 22 such trials, Sims found that the average "promising" trial exaggerated the true effect size by 52 percent. When these inflated early estimates are used to justify large-scale policy adoption, subsequent scale-up failures are virtually guaranteed.

3. **Citation Amnesia:** A related phenomenon where contradicting or null findings are systematically forgotten or uncited, even when they are methodologically superior to or

post-date the positive findings. This creates a "telescoping effect" where the evidence base appears much stronger to policymakers than it actually is, because the negative results have been effectively erased from the active citation network.

14.2 Documented Cases in K-12 Education

This review has identified several prominent instances where these mechanisms have actively shaped K-12 education policy. In each case, the original research contained genuine, albeit carefully caveated, findings that were subsequently stripped of their context and amplified into universal policy mandates.

Table 2: Documented Cases of Citation Distortion in Education Policy

Original Finding	**Original Caveats**	**Policy Distortion (Amplification & Invention)**
Teacher VAMs (Chetty et al., 2014b)	High-stakes use of VAMs will likely induce teaching-to-the-test and corrupt the metric (Goodhart's Law).	The $250,000 lifetime earnings figure was stripped of caveats and used to justify aggressive, high-stakes teacher termination policies nationwide.

Continued on next page

Original Finding	**Original Caveats**	**Policy Distortion (Amplification & Invention)**
Early Childhood (Heckman et al., 2010)	The Perry Preschool Program was a highly intensive, boutique intervention targeting 58 deeply disadvantaged children.	The 7–12% ROI figure was cited routinely to justify modern, scaled-up universal pre-K expansions, which often lack the quality controls of the original program.
Grit (Duckworth et al., 2007)	Grit is a specific psychological construct; self-report surveys should not be used for high-stakes evaluation.	Schools began formally grading students on grit; subsequent meta-analyses (Credé et al., 2017) showed the construct was largely redundant with conscientiousness.

The common thread across these cases is the vulnerability of education research to epistemic drift. Because effect sizes in education are typically small (often $d < 0.20$), the 52% exaggeration factor identified by Sims et al. (2023) can mean the difference between an intervention appearing transformative versus marginal. Policymakers must actively guard against citation distortion by demanding systematic reviews rather than relying on highly cited "superstar" papers, and by explicitly discounting effect sizes from early-stage, small-scale trials before committing to systemic scale-up.

15 Cross-Cutting Synthesis

Across the ten clusters reviewed, five cross-cutting themes emerge that define the current state of K-12 education research and that have significant implications for policy and practice.

15.1 The Persistence of Selection Bias

In almost every domain, observational estimates are routinely found to be biased upward when subjected to rigorous quasi-experimental or experimental stress tests. The credibility revolution has revealed that many widely-cited effect sizes in the education literature are inflated. Early estimates of the effects of attending private schools, for example, were largely explained by the selection of more motivated and advantaged students into private schools. Early estimates of the effects of attending high-spending schools were confounded by the correlation between school spending and family income. The consistent finding that rigorous designs produce smaller effect sizes than observational studies should make policymakers skeptical of claims based on non-experimental evidence, particularly when those claims are used to justify large-scale policy changes.

15.2 The Fadeout Phenomenon and Non-Cognitive Mediation

Interventions that produce large short-term gains (e.g., Reading Recovery, early Head Start) frequently see those gains fade out over time. However, this fadeout in test scores does not always preclude long-term benefits in adult outcomes. The most plausible reconciliation of this apparent paradox is that non-cognitive skills—self-regulation, conscientiousness, executive function—act as a crucial mediator. Early childhood programs that successfully develop these traits may produce adult benefits (reduced criminality, higher employment, better health) even when the initial cognitive boost dissipates. This hypothesis is supported by Heckman et al. (2006)'s extensive work on the economics of human development, which shows that non-cognitive skills are highly malleable in early childhood and are strong pre-

dictors of adult success. Future research should prioritize measuring non-cognitive outcomes directly, rather than relying solely on test scores as proxies for program effectiveness.

15.3 Implementation Fidelity Trumps Intervention Design

The efficacy of many interventions is highly dependent on implementation fidelity and context. Interventions that scale well—like high-dosage tutoring and structured phonics—typically have highly structured, standardized delivery mechanisms that can be replicated with high fidelity across diverse settings. Interventions that fail to scale—like California's class size reduction and Tennessee's pre-K expansion—typically require a level of human capital (highly trained teachers) or quality control that cannot be maintained when programs expand rapidly. This finding has profound implications for how policymakers should think about scaling evidence-based interventions: the question is not just "does this work?" but "can this be implemented with fidelity at scale?"

15.4 The Centrality of Cost-Effectiveness

While effect sizes (d) are the standard currency of academic research, they are insufficient for policy decisions without cost data. Class size reduction produces reliable gains but is immensely expensive ($15,000–$20,000 per student per year for a reduction from 25 to 15 students). High-dosage tutoring produces comparable or larger gains at a fraction of the cost ($2,000–$4,000 per student per year). Similarly, systematic phonics instruction requires significant initial professional development but relatively low ongoing marginal costs compared to structural interventions. Policymakers should demand cost-effectiveness analyses alongside effect size estimates, and should prioritize interventions that produce the largest gains per dollar spent.

15.5 The Structural Limits of School Reform

The out-of-school factors literature (Cluster 9) and the international comparisons literature (Cluster 10) both point to a fundamental structural limit on what school reform alone can achieve. Schools cannot fully compensate for the effects of poverty, neighborhood disadvantage, and family instability. The income-achievement gap has widened over the past four decades even as school quality has improved, suggesting that economic inequality is the primary driver of educational inequality. This does not mean that school reform is futile—the evidence from high-dosage tutoring, structured phonics, and equitable funding clearly shows that schools can make a significant difference. But it does mean that school reform must be accompanied by broader social and economic policies that address the root causes of inequality.

16 Replication Agenda (Data-Available Candidates)

Based on the verified registry, we flag 17 high-priority candidates for replication. We restrict this list to studies where the underlying data (or a highly comparable proxy) is publicly available or accessible via standard application procedures (e.g., NCES restricted-use data).

Selection Criteria: Candidates were flagged where (i) policy citation count is exceptionally high and the study drives current funding decisions, (ii) at least one credible methodological challenge has been published, and (iii) the underlying data is publicly available or accessible via application. *Note: The candidates below are grouped chronologically by thematic cluster to mirror the structure of this review, rather than being ranked by priority; all 17 are considered equally high-priority for the field.*

Table 3: 17 High-Priority Replication Candidates (Data Available)

Citation	**Cluster**	**Data Source**	**Rationale for Replication**
Chetty et al. (2014a,b)	Teacher Quality	IRS/School Data	VAM stability across different state testing regimes requires verification; $250K figure drives major policy.
Rothstein (2010)	Teacher Quality	NCES Data	Falsification tests for VAMs need updating with newer cohort data and modern VAM specifications.
Krueger (1999)	Class Size	Project STAR (Public)	Randomization integrity and attrition concerns remain unresolved; long-run effects contested.
Puma et al. (2012)	Early Childhood	Head Start Impact (ICPSR)	Critical to verify fade-out mechanisms vs. non-cognitive mediation with longer follow-up.
Lipsey et al. (2018)	Early Childhood	TN Pre-K (State Data)	Urgent need to replicate negative findings in other state pre-K expansions.

Continued on next page

Citation	**Cluster**	**Data Source**	**Rationale for Replication**
Heckman et al. (2010)	Early Childhood	Perry/Abecedarian (Public)	Small sample sizes ($n = 58$) warrant replication with modern intensive programs.
Jackson et al. (2016)	School Funding	PSID / CCD	Long-run adult outcomes rely on historical rollout variations; needs post-2000 replication.
Hanushek (1997)	School Funding	Various (CCD/SEDA)	Needs updating with modern quasi-experimental funding data and SEDA achievement measures.
Angrist et al. (2010,12)	Charter Schools	State DOE Data	Boston charter lottery effects need replication in post-pandemic context and other cities.
Abdulkadiroğlu et al. (2018)	Charter Schools	State DOE Data	Louisiana voucher negative effects require multi-state verification.
Ehri et al. (2001)	Reading	Meta-analysis data	Phonics effect sizes need updating with modern structured literacy trials and preregistered studies.

Continued on next page

Citation	**Cluster**	**Data Source**	**Rationale for Replication**
May et al. (2023)	Reading	i3 Grant Data	Reading Recovery long-term fadeout needs replication in other pull-out intervention programs.
Cook et al. (2015)	Tutoring	Chicago Public Schools	High-school math tutoring effects require replication in non-urban settings and with female students.
Nickow et al. (2020)	Tutoring	Meta-analysis data	Requires replication focusing on cost-effectiveness and scaling parameters post-COVID.
Durlak et al. (2011)	SEL	Meta-analysis data	Needs replication distinguishing universal vs. targeted interventions with pre-registered designs.
Borman & Dowling (2010)	Out-of-School	EEO Data (ICPSR)	Needs replication with modern, post-COVID inequality data and updated multilevel methods.
Reardon (2011)	Out-of-School	SEDA / NAEP	Income-achievement gap trajectory needs post-2020 updating with pandemic-era data.

17 Conclusion

The "credibility revolution" has transformed K-12 education research, yielding robust causal evidence that targeted interventions—such as high-dosage tutoring, systematic phonics, and equitable school funding—can significantly improve student trajectories. However, the literature also demands humility: many popular interventions yield small or null effects at scale, and translating successful programs from controlled settings to broad implementation remains a profound challenge.

The failed scale-ups documented in this review—California's class size reduction initiative, Tennessee's pre-K expansion, the Louisiana voucher program—are instructive. In each case, the intervention that worked in a controlled or small-scale setting produced null or negative effects when rapidly expanded. The California case illustrates the danger of general equilibrium effects: a policy that works when only a few districts adopt it may fail when all districts adopt it simultaneously, because the labor market response (in this case, the dilution of teacher quality) changes the conditions under which the intervention operates. The Tennessee case illustrates the danger of assuming that program quality can be maintained at scale: the boutique programs that produced the Perry and Abecedarian results relied on exceptional staff and intensive oversight that cannot be replicated in a statewide rollout without massive investment in quality control.

The COVID-19 pandemic has created both a crisis and an opportunity for K-12 education. The crisis is the historic learning loss documented in NAEP scores and other assessments, which has disproportionately affected the most disadvantaged students. The opportunity is the unprecedented federal investment in education recovery through the American Rescue Plan, which has provided billions of dollars for evidence-based interventions. The evidence reviewed in this synthesis suggests that high-dosage tutoring and structured literacy instruction are among the most cost-effective uses of these recovery funds.

As the field moves forward, the priority must shift from identifying "what works" in isolated settings to understanding the mechanisms of scalability and the long-term persistence of educational impacts. The failed scale-ups of California's class size reduction initiative, Tennessee's pre-K expansion, and the Louisiana voucher program explicitly demonstrate that structural interventions without pedagogical fidelity and human capital maintenance are insufficient. Future policy must prioritize interventions that combine strong theoretical foundations with robust, scalable implementation mechanisms—and must invest in the long-term follow-up studies needed to distinguish short-term test score gains from lasting improvements in human development.

References

Abdulkadiroğlu, Angrist, Hull, and Pathak (2016). Charters without Lotteries: Testing Takeovers in New Orleans and Boston. *American Economic Review*. Registry ID: 38.

Abdulkadiroğlu, Pathak, and Walters (2018). Free to Choose: Can School Choice Reduce Student Achievement? *American Economic Journal: Applied Economics*. Registry ID: 40.

Alexander, Entwisle, and Olson (2007). Lasting Consequences of the Summer Learning Gap. *American Sociological Review*. Registry ID: 64.

Angrist and Lavy (1999). Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. *Quarterly Journal of Economics*. Registry ID: 22.

Angrist and Lavy (2019). Maimonides' Rule Redux. *American Economic Review*. Registry ID: 23.

Angrist, Pathak, and Walters (2013a). Explaining Charter School Effectiveness. *American Economic Journal: Applied Economics*. Registry ID: 36.

Angrist, J. D., Dynarski, S. M., Kane, T. J., Pathak, P. A., and Walters, C. R. (2010). Inputs and impacts in charter schools: KIPP lynn. *American Economic Review: Papers & Proceedings*, 100(2):239–243. Registry ID: added.

Angrist, J. D., Pathak, P. A., and Walters, C. R. (2013b). Explaining charter school effectiveness. *American Economic Journal: Applied Economics*, 5(4):1–27. Registry ID: added; key angrist2012 maps here.

Bacher-Hicks, A., Billings, S. B., and Deming, D. J. (2019). The school to prison pipeline: Long-run impacts of school suspensions on adult crime. *NBER Working Paper*, (26257).

Backes, Cowan, Goldhaber, and Theobald (2024). Heterogeneous Impacts of Teacher Value-Added on Postsecondary Outcomes. *NBER Working Paper*. Registry ID: 79.

Bailey, Sun, and Timpe (2021). Prep School for Poor Kids: The Long-Run Impacts of Head Start on Human Capital and Economic Self-Sufficiency. *American Economic Review*. Registry ID: 17.

Benner, Boyle, and Sadler (2016). Parental Involvement and Adolescents' Educational Success: The Roles of Prior Achievement and Socioeconomic Status. *Journal of Research on Adolescence*. Registry ID: 101.

Bhatt, Cook, Guryan, and Ludwig (2024). Scaling High-Dosage Tutoring: Evidence from a Hybrid Model. *NBER Working Paper*. Registry ID: 52.

Blazar (2018). Validating Teacher Effects on Students' Attitudes and Behaviors: Evidence from Random Assignment of Teachers to Students. *Education Finance and Policy*. Registry ID: 78.

Borman and Dowling (2010). Schools and Inequality: A Multilevel Analysis of Coleman's Equality of Educational Opportunity Data. *Teachers College Record*. Registry ID: 63.

Campbell, Conti, Heckman, Moon, Pinto, Pungello, and Pan (2014). Early Childhood Investments Substantially Boost Adult Health. *Science*. Registry ID: 13.

Card and Krueger (1992). Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States. *Journal of Political Economy*. Registry ID: 32.

Cascio and Schanzenbach (2013). The Impacts of Expanding Access to High-Quality Preschool Education. *Brookings Papers on Economic Activity*. Registry ID: 18.

Castles, Rastle, and Nation (2018). Ending the Reading Wars: Reading Acquisition From Novice to Expert. *Psychological Science in the Public Interest*. Registry ID: 47.

Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan (2011a). How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR. *Quarterly Journal of Economics*. Registry ID: 16.

Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan (2011b). How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR. *Quarterly Journal of Economics*. Registry ID: 26.

Chetty, Friedman, and Rockoff (2014a). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review*. Registry ID: 1.

Chetty, Friedman, and Rockoff (2014b). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*. Registry ID: 2.

Clark, A. E., Nong, H., Zhu, H., and Ward, R. (2020). Compensating for academic loss: Online learning and student performance during the covid-19 pandemic. *China Economic Review*, 68:101619.

Cohen, Kulik, and Kulik (1982). Educational Outcomes of Tutoring: A Meta-Analysis of Findings. *American Educational Research Journal*. Registry ID: 93.

Cohodes, Setren, and Walters (2021). Can Successful Schools Replicate? Scaling Up Boston's Charter Schools. *American Economic Journal: Economic Policy*. Registry ID: 39.

Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld, and York (1966). Equality of Educational Opportunity (Coleman Report). *U.S. Department of Health, Education, and Welfare*. Registry ID: 62.

Cook, Dodge, Farkas, Fryer, Guryan, Ludwig, Mayer, Pollack, and Steinberg (2015). Not Too Late: Improving Academic Outcomes for Disadvantaged Youth. *IPR Working Paper*. Registry ID: 54.

Credé, M., Tynan, M. C., and Harms, P. D. (2017). Much Ado About Grit: A Meta-Analytic Synthesis of the Grit Literature. *Journal of Personality and Social Psychology*. Registry ID: 59.

CREDO at Stanford University (2015). Urban Charter School Study: Report on Charter School Performance in 41 Urban Regions. *Stanford University*. Registry ID: 43.

Darling-Hammond (2010). The Flat World and Education: How America's Commitment to Equity Will Determine Our Future. *Teachers College Press*. Registry ID: 106.

Darling-Hammond, L. and Cook-Harvey, C. M. (2018). Educating the whole child: Improving school climate to support student success. *Learning Policy Institute*.

Deming (2009). Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start. *American Economic Journal: Applied Economics*. Registry ID: 14.

Dobbie and Fryer (2011). Are High-Quality Schools Enough to Increase Achievement Among the Poor? Evidence from the Harlem Children's Zone. *American Economic Journal: Applied Economics*. Registry ID: 37.

Duckworth, Peterson, Matthews, and Kelly (2007). Grit: Perseverance and Passion for Long-Term Goals. *Journal of Personality and Social Psychology*. Registry ID: 99.

Duckworth and Quinn (2009). Development and Validation of the Short Grit Scale (Grit-S). *Journal of Personality Assessment*. Registry ID: 58.

Duckworth, A. (2016). Don't grade schools on grit. *The New York Times*, Opinion. Available at: https://www.nytimes.com/2016/03/27/opinion/sunday/dont-grade-schools-on-grit.html.

Durlak, Weissberg, Dymnicki, Taylor, and Schellinger (2011). The Impact of Enhancing Students' Social and Emotional Learning: A Meta-Analysis of School-Based Universal Interventions. *Child Development*. Registry ID: 55.

Ehri, Nunes, Stahl, and Willows (2001). Systematic Phonics Instruction Helps Students Learn to Read: Evidence from the National Reading Panel's Meta-Analysis. *Review of Educational Research*. Registry ID: 48.

Fredriksson, Öckert, and Oosterbeek (2013). Long-Term Effects of Class Size. *Quarterly Journal of Economics*. Registry ID: 24.

Fryer (2014). Injecting Charter School Best Practices into Traditional Public Schools: Evidence from Houston's Apollo 20 Program. *Quarterly Journal of Economics*. Registry ID: 87.

García, Heckman, and Ronda (2022). The Lasting Effects of Early Childhood Education on Promoting the Skills and Social Mobility of Disadvantaged African Americans. *NBER Working Paper*. Registry ID: 12.

Goldhaber and Brewer (2000). Does Teacher Certification Matter? High School Teacher Certification Status and Student Achievement. *Educational Evaluation and Policy Analysis*. Registry ID: 9.

Goldhaber, D. and Chaplin, D. (2015). Assessing the "rothstein falsification test": Does it really show teacher value-added models are biased? *Journal of Research on Educational Effectiveness*, 8(1):8–34. Challenges Rothstein's (2010) falsification test; finds bias in VAM estimates is small (approx. 3%) and that Rothstein's test has its own limitations.

Goldhaber, D., Kane, T. J., McEachin, A., Morton, E., Patterson, T., and Staiger, D. O. (2022). The consequences of remote and hybrid instruction during the pandemic. NBER Working Paper 30010, National Bureau of Economic Research. Registry ID: added.

Gray-Lobe, G., Pathak, P. A., and Walters, C. R. (2023). The long-term effects of universal preschool in boston. *Quarterly Journal of Economics*, 138(1):363–411.

Greenberg, S. A. (2009). How citation distortions create unfounded authority: Analysis of a citation network. *BMJ*, 339:b2680. Registry ID: added.

Greenwald, Hedges, and Laine (1996). The Effect of School Resources on Student Achievement. *Review of Educational Research*. Registry ID: 86.

Guryan (2001). Does Money Matter? Regression-Discontinuity Estimates from Education Finance Reform in Massachusetts. *NBER Working Paper*. Registry ID: 33.

Guryan, Ludwig, Bhatt, Cook, Davis, Dodge, Farkas, Mayer, Pollack, and Steinberg (2023). Not Too Late: Improving Academic Outcomes Among Adolescents. *American Economic Review*. Registry ID: 51.

Hanford (2018). Hard Words: Why Aren't Kids Being Taught to Read? *APM Reports*. Registry ID: 46.

Hansford, Buckingham, and Meeks (2025). Structured Literacy vs. Balanced Literacy: A Systematic Review and Meta-Analysis. *Working Paper*. Registry ID: 90.

Hanushek (1997). Assessing the Effects of School Resources on Student Performance: An Update. *Educational Evaluation and Policy Analysis*. Registry ID: 31.

Hanushek (2003). The Failure of Input-Based Schooling Policies. *Economic Journal*. Registry ID: 85.

Hanushek (2011). The Economic Value of Higher Teacher Quality. *Economics of Education Review*. Registry ID: 5.

Hanushek and Woessmann (2015). The Knowledge Capital of Nations: Education and the Economics of Growth. *MIT Press*. Registry ID: 105. Full text not directly accessed; cited from publisher description and secondary sources.

Hanushek, E. A. (1986). The Economics of Schooling: Production and Efficiency in Public Schools. *Journal of Economic Literature*, 24(3):1141–1177. Registry ID: 125.

Hanushek, E. A. and Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *The American Economic Review*, 100(2):267–271.

Harris and Jones (2017). Leading Schools as Learning Organizations. *School Leadership Management*. Registry ID: 70.

Heckman, Moon, Pinto, Savelyev, and Yavitz (2010). The Rate of Return to the HighScope Perry Preschool Program. *Journal of Public Economics*. Registry ID: 11.

Heckman, Stixrud, and Urzua (2006). The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. *Journal of Labor Economics*. Registry ID: 60.

Howell, W. G., Wolf, P. J., Campbell, D. E., and Peterson, P. E. (2002). School vouchers and academic performance: Results from three randomized field trials. *Journal of Policy Analysis and Management*, 21(2):191–217. Registry ID: added.

Hoxby (2000). The Effects of Class Size on Student Achievement: New Evidence from Population Variation. *Quarterly Journal of Economics*. Registry ID: 25.

Hoxby and Murarka (2009). Charter Schools in New York City: Who Enrolls and How They Affect Their Students' Achievement. *NBER Working Paper*. Registry ID: 42.

Hyman (2017). Does Money Matter in the Long Run? Effects of School Spending on Educational Attainment. *American Economic Journal: Economic Policy*. Registry ID: 35.

Jackson (2018a). What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes. *Journal of Political Economy*. Registry ID: 4.

Jackson (2018b). What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes. *Journal of Political Economy*. Registry ID: 61.

Jackson, Johnson, and Persico (2016). The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms. *Quarterly Journal of Economics*. Registry ID: 29.

Jepsen and Rivkin (2009). Class Size Reduction and Student Achievement: The Potential Tradeoff between Teacher Quality and Class Size. *Journal of Human Resources*. Registry ID: 27.

Jones, S. M. and Kahn, J. (2017). The evidence base for how we learn: Supporting students' social, emotional, and academic development. *National Commission on Social, Emotional, and Academic Development*.

Kane and Staiger (2008). Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. *NBER Working Paper*. Registry ID: 6.

Kim (2006). Effects of a Voluntary Summer Reading Intervention on Reading Achievement: Results from a Randomized Field Trial. *Educational Evaluation and Policy Analysis*. Registry ID: 103.

Kjeldsen, Kärnä, Niemi, Olofsson, and Witting (2014). Gains from Training in Phonological Awareness in Kindergarten Predict Reading Development in the First 9 School Years. *Reading and Writing*. Registry ID: 91.

Koedel and Betts (2009). Value Added to What? How a Ceiling in the Testing Instrument Influences Value-Added Estimation. *Education Finance and Policy*. Registry ID: 76.

Kraft and Falken (2021). A Blueprint for Scaling Tutoring Across Public Schools. *AERA Open*. Registry ID: 53.

Kraft and Lovison (2025). The Effects of Tutoring Group Size on Student Learning. *Educational Evaluation and Policy Analysis*. Registry ID: 96.

Krueger (1999). Experimental Estimates of Education Production Functions. *Quarterly Journal of Economics*. Registry ID: 21.

Lafortune, Rothstein, and Schanzenbach (2018). School Finance Reform and the Distribution of Student Achievement. *American Economic Journal: Applied Economics*. Registry ID: 30.

Levin (1997). Accelerated Schools for Disadvantaged Students. *Educational Leadership*. Registry ID: 72.

Lipsey, Farran, and Durkin (2018). Effects of the Tennessee Voluntary Prekindergarten Program on Children's Achievement and Behavior through Third Grade. *Early Childhood Research Quarterly*. Registry ID: 19.

May, Sirinides, Gray, and Goldsworthy (2016). Reading Recovery: An Evaluation of the i3 Scale-Up. *Consortium for Policy Research in Education*. Registry ID: 92.

May, Sirinides, Gray, and Goldsworthy (2023). Reading Recovery: An Evaluation of the i3 Scale-Up. *Consortium for Policy Research in Education*. Registry ID: 45.

Morgan and Jung (2016). Still No Effect of Resources, Even in the New Gilded Age? *Sociology of Education*. Registry ID: 66.

Morrissey and Vinopal (2018). Neighborhood Poverty and Children's Academic Skills and Behavior in Early Elementary School. *Journal of Marriage and Family*. Registry ID: 67.

Mourshed, Chijioke, and Barber (2010). How the World's Most Improved School Systems Keep Getting Better. *McKinsey Company*. Registry ID: 107.

Neilson and Zimmerman (2014). The Effect of School Construction on Test Scores, School Enrollment, and Home Prices. *Journal of Public Economics*. Registry ID: 34.

Nickow, Oreopoulos, and Quan (2020). The Impressive Effects of Tutoring on PreK-12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence. *NBER Working Paper*. Registry ID: 50.

Nickow, Oreopoulos, and Quan (2024). The Promise of Tutoring for PreK-12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence. *American Educational Research Journal*. Registry ID: 95.

Nye, Konstantopoulos, and Hedges (2004). How Large Are Teacher Effects? *Educational Evaluation and Policy Analysis*. Registry ID: 84.

Osher, D., Cantor, P., Berg, J., Steyer, L., and Rose, T. (2020). Drivers of human development: How relationships and context shape learning and development. *Applied Developmental Science*, 24(1):6–36.

Pages, Lukes, Bailey, and Duncan (2020). Revised Estimates of the Impacts of Head Start: A Target Population Approach. *Educational Evaluation and Policy Analysis*. Registry ID: 20.

Panel, N. R. (2000). Teaching Children to Read: An Evidence-Based Assessment of the

Scientific Research Literature on Reading and Its Implications for Reading Instruction. *NICHD*. Registry ID: 44.

Papay and Kraft (2015). Productivity Returns to Experience in the Teacher Labor Market: Methodological Challenges and New Evidence on Long-Run Effects. *Journal of Public Economics*. Registry ID: 10.

Porter, Fusarelli, and Fusarelli (2022). Implementing the Every Student Succeeds Act: Accountability, Devolution, Democratic Agency. *Educational Policy*. Registry ID: 98.

Puma, Bell, Cook, Heid, and Shapiro (2010). Head Start Impact Study: Final Report. *U.S. Department of Health and Human Services*. Registry ID: 15.

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. *Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances*, pages 91–116. Registry ID: added.

Rivkin, Hanushek, and Kain (2005). Teachers, Schools, and Academic Achievement. *Econometrica*. Registry ID: 7.

Rockoff (2004). The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *American Economic Review (Papers Economic Review (Papers Proceedings) Proceedings)*. Registry ID: 8.

Rothstein (2009). Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables. *Education Finance and Policy*. Registry ID: 73.

Rothstein (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *Quarterly Journal of Economics*. Registry ID: 3.

Rothstein (2016). Revisiting the Impacts of Teachers. *Working Paper*. Registry ID: 74.

Rouse (1998). Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program. *Quarterly Journal of Economics*. Registry ID: 41.

Sahlberg (2020). Finnish Lessons 3.0: What Can the World Learn from Educational Change in Finland? *Teachers College Press*. Registry ID: 69. Full text not directly accessed; cited from publisher description and secondary sources.

Scherer, R. and Siddiq, F. (2019). The relation between students' socioeconomic status and ict literacy: Findings from a meta-analysis. *Computers Education*, 138:13–32.

Shanahan and Lonigan (2010). The National Early Literacy Panel: A Summary of the Process and the Report. *Educational Researcher*. Registry ID: 49.

Sims, S., Anders, J., Inglis, M., and Lortie-Forgues, H. (2023). Quantifying "promising trials bias" in randomized controlled trials in education. *Journal of Research on Educational Effectiveness*, 16(4):663–680. Registry ID: added.

Sims, S. and Fletcher-Wood, H. (2020). Identifying the characteristics of effective teacher professional development: A critical review. *School Effectiveness and School Improvement*, 32(1):47–63.

Sisk, Burgoyne, Sun, Butler, and Macnamara (2018). To What Extent and Under Which Circumstances Are Growth Mind-Sets Important to Academic Achievement? Two Meta-Analyses. *Psychological Science*. Registry ID: 57.

Tan, C. Y., Lyu, M., and Peng, B. (2020). Academic benefits from parental involvement are stratified by parental socioeconomic status: A meta-analysis. *Parenting: Science and Practice*, 20(4):241–287. Published online 2019; print 2020. Meta-analysis of 371 studies (N $>$ 2.7 million). Parental involvement benefits academic outcomes but effects are moderated by SES: higher-SES parents show stronger returns. Relevant to family background and parental involvement sections.

Walters (2018). The Demand for Effective Charter Schools. *Journal of Political Economy*. Registry ID: 88.

Wang, M.-T., Degol, H., and Amemiya, J. (2020). Classroom climate and children's academic and psychological wellbeing: A systematic review and meta-analysis. *Developmental Review*, 57:100912.

Wei, Darling-Hammond, Andree, Richardson, and Orphanos (2012). Professional Learning in the Learning Profession: A Status Report on Teacher Development in the United States and Abroad. *National Staff Development Council*. Registry ID: 77.

Wilson, Darling-Hammond, and Berry (2001). A Case of Successful Teaching Policy: Connecticut's Long-Term Efforts to Improve Teaching and Learning. *Center for the Study of Teaching and Policy*. Registry ID: 104.

Wodtke, White, and Zhou (2026). How Much Do Schools Matter? Machine Learning Evidence on School Effects and the Neighborhood-Poverty Achievement Gap. *Sociological Science*. Registry ID: 65.

Wolf, Magnuson, and Kimbro (2017). Family and Neighborhood Socioeconomic Status and Cognitive Development in Early Childhood. *Developmental Psychology*. Registry ID: 102.

Wößmann, L. and West, M. (2006). Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS. *European Economic Review*, 50(3):695–736. Registry ID: added.

Wößmann and West (2006). Class-Size Effects in School Systems Around the World: Evidence from Between-Grade Variation in TIMSS. *European Economic Review*. Registry ID: 28.

Yeager, Hanselman, Walton, Murray, Crosnoe, Muller, Tipton, Schneider, Hulleman, Hinojosa, Paunesku, Romero, Flint, Roberts, Trott, Iachan, Buontempo, Yang, Carvalho, Hahn, Gopalan, Mhatre, Ferguson, Duckworth, and Dweck (2019). A National Experiment Reveals Where a Growth Mindset Improves Achievement. *Nature*. Registry ID: 56.

Yeager, D. S., Carroll, J. M., Buontempo, J., Cimpian, A., Woody, S., Crosnoe, R., Muller, C., Murray, J., Mhatre, P., Kersting, N., et al. (2021). Teacher mindsets help explain where a growth-mindset intervention does and doesn't work. *Psychological Science*, 133(1):18–32.

Yeager, D. S. and Dweck, C. S. (2020). What can be learned from growth mindset controversies? *American Psychologist*, 75(9):1269–1284.