CESifo WORKING PAPERS

10664
2023

September 2023

Contexts of Convenience: Generalizing from Published Evaluations of School Finance Policies

Danielle V. Handel, Eric A. Hanushek

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 10664*

Contexts of Convenience: Generalizing from Published Evaluations of School Finance Policies

Abstract

Recent attention to the causal identification of spending impacts provides improved estimates of spending outcomes in a variety of circumstances, but the estimates are substantially different across studies. Half of the variation in estimated funding impact on test scores and over three-quarters of the variation of impacts on school attainment reflect differences in the true parameters across study contexts. Unfortunately, inability to describe the circumstances underlying effective school spending impedes any attempts to generalize from the extant results to new policy situations. The evidence indicates that how funds are used is crucial to the outcomes but such factors as targeting of funds or court interventions fail to explain the existing pattern of results.

JEL-Codes: I210, H400.

Keywords: school finance, evaluation.

Danielle V. Handel
Stanford University / CA / USA
dvhandel@stanford.edu

*Eric A. Hanushek**
Stanford University / CA / USA
hanushek@stanford.edu

*corresponding author

September 5, 2023
We thank Doug Besharov and Jeff Smith for useful comments on an earlier draft.

1 Introduction

The "credibility revolution" in economics has led to a range of studies that revisit prior analyses into the causal impact of added educational funding but with more believable identification of the key impact parameters. This is an obviously important policy issue but one that cannot be reasonably addressed with randomized control trials (RCTs). As a substitute for RCTs, they employ "quasi-experimental methods" that are designed to mimic important aspects of RCTs. These approaches provide more internally valid parameter estimates but may simultaneously face questions about generalizability.

Generalizability of evidence about the impact of educational resources on student outcomes is especially important in matters of U.S. school finance. School finance decisions are primarily made by the separate state legislatures, although decisions are frequently affected directly by the courts.[1] Court decisions are generally limited to issues of the level of funding and its distribution across school districts in the state, even though the fundamental underlying issues center on student outcomes. Legislatures are also generally concerned about student outcomes, and they tend to specify details about the use of any additional funds that they provide in order to improve outcomes. But evidence on resource impacts introduced into either the state legislature or the state courts seldom involves research specific to the circumstances of the state but necessarily requires generalizing from existing quasi-experimental analyses related to other circumstances.

Disentangling the causal impact of funding on student outcomes from other factors is clearly very difficult, because the schools and their educational environments are themselves very complicated. The modern approach to evaluation analysis offers a way of cutting through these complications. In a variety of cases, it has been possible to find circumstances where the observed funding is plausibly exogenous and where it is possible to obtain unbiased estimates of the impact of funding. The downside of this line of research, however, is that the circumstances generating the exogenous funding variation are not developed according to an overarching study design but instead are embedded in the very different educational contexts that happen to support causal analysis.

Past discussions of research on how added funding influences student outcomes has been quite contentious, in large part because the past research did not provide compelling evidence that added funds would produce systematic improvements in student achievement. The more recent research has, as indicated below, more frequently (but far from uniformly) found that added funding causes improvements in student outcomes, but this is clearly insufficient for policy purposes since few people believe that added funding is actually likely to harm students. The sought after evidence is a clear indication of the magnitude of the causal impact (internal validity in analytical terms) in the context where new policy is contemplated (generalizability in analytical terms).

The growing body of work that produces well-identified estimates of the impact of funding on student performance provides the data for this investigation of the consistency and generalizability of the results. Comparing the magnitude of different estimates of funding impact is possible but not easy.

[1] The extent of judicial involvement in U.S. school finance is described in Hanushek and Joyce-Wirtz (2023).

Because educational inputs and educational outcomes are measured differently across studies and because the studies employ different estimation methodologies, a preliminary step is harmonizing the estimates from the individual studies.

If all of the impact estimates from the very different contexts of the underlying analyses were the same, one would not worry much about applying them to a new situation. This homogeneity would strongly suggest that the specific context of each study was not very important, implying that it might be reasonable to generalize the expected impact to a new circumstance. However, when put on a common scale, the large variation in estimated impact of funding across existing studies is apparent.[2] Sampling variation contributes to the range of estimates, but the majority of the variation appears to come from differences in the underlying true impacts of funding. While some of this variation may be independent of true parameter differences – reflecting quality problems in the underlying studies, design choices of the authors, or issues related to estimation approach – a more likely source is simply the influence of the educational contexts where incentives and constraints dictate that funds will have differential impacts on outcomes.

The important context differences revolve around the scope and restrictions surrounding specific spending under consideration and the quality of decision making and the incentives to decision makers. The federal nature of U.S. education gives primary authority over schools to the states, and this has led to a wide variety of institutional environments for schools where individual states attempt to establish a school system that effectively educates its youth. These policy and regulatory environments also interact with demographic and educational differences in the state populations.

We refer to the underlying educational decisions resulting from these distinct contexts as "how" funds are used. And this "how" appears to be a decisive force in determining the efficacy of any funding on schools. Unfortunately, little attention has been given to describing the key features of the context for the existing studies, and the limited replication of relevant quasi-experimental observations within common institutional frameworks makes direct analysis of the heterogeneity of impact estimates difficult.

Our initial attempts to understand structural features of the differences in impact of funding regrettably do not provide much overall guidance. Within our limited sample of existing studies, it does not appear that method of estimation in these quasi-experimental studies can explain the heterogeneity. Neither can whether or not spending is targeted, whether spending results from court orders, or whether spending is observed across broad experiences including state-level estimates. Looking within individual studies, there is an indication that spending has a larger impact on low SES children, but this does not solve the basic problem of how to use funding to get strong achievement results.

This reliance on contexts of convenience for the estimation of impact parameters leaves uncertainty about the effectiveness of spending in different contexts. Little progress has been made in

[2] Details of the underlying studies and the methodology for harmonizing the results are found in Handel and Hanushek (2023).

describing fundamental factors that influence the efficacy of spending in different contexts, thus severely limiting any attempt to generalize to other circumstances.

The next section describes the general evaluation problem surrounding school finance. Section 3 describes the search procedures to find relevant well-identified studies along with the methods for harmonizing the different studies. Section 4 presents the raw results of studies for test score outcomes and school attainment, while Section 5 introduces some of the interpretive issues. Sections 6 and 7 pursue meta-analytic approaches to understanding the overall outcomes and investigating some possible fundamental driving forces. Section 8 concludes.

2 Understanding Evidence on School Funding

The attention given to research on the impact of funding on student outcomes reflects the intensity of interest in improving school performance. The underlying issue that has been addressed for the last 50 years can be characterized by a single question: Under which circumstances will added funding reliably lead to improved student performance?

The research issue can be seen in a stylized linear model. Student outcomes (*O*) can be written as a function of funding (*F*) and other factors (ε) as in Eq. 1. The focus of estimation is the impact of funding on achievement (γ_c); this is indexed by the funding context of the funding (*c*) in order to facilitate subsequent comparisons across estimates.

$$O = \gamma_c F + \varepsilon \tag{1}$$

The key aspect to this depiction is that γ_c is not fixed across all possible funding circumstances but that instead:

$$\gamma_c = f(\mathbf{X_c}) + \nu_c \tag{2}$$

where γ_c is a function of an array of contextual factors, $\mathbf{X_c}$, that involve the objectives, regulations, constraints, and decision making dynamics surrounding the infusion of funds. Thus, a given magnitude of funding *F* may have quite different impacts on outcomes depending on the context.

The historical approach, beginning with the Coleman Report (Coleman et al. (1966)), was generally to estimate γ_c in a production function framework using standard regression techniques and including context with a variety of control variables in eq. 1 (see Hanushek (1979, 2003)). But, any correlation of *F* and ε arising from an inability to control for all of the factors affecting student outcomes implies that the estimates of γ_c will be biased.

Modern empirical approaches focus on quasi-experimental methods that rely on contexts where the variation in funding can reasonably be believed to be exogenous from other influences on student

outcomes. These approaches can provide more assurance that the estimated funding impact parameter is unbiased by being less ambitious in describing the range of contextual factors affecting outcomes.

Understanding the context of educational funding is especially challenging. School funding takes place within a range of institutional environments involving state differences in the laws and regulations surrounding schools, various possible restrictions on the policy choices of schools and districts, and differing quality of decision makers faced with different incentive structures. Some of the quasi-experiments involve fully prescribed use of funds while others involve considerable flexibility. Unfortunately, the full set of contexts is virtually never described in the analyses, in part because the analytical framework is designed to sidestep the need for specifying or understanding the full context of the funding choices.

Two aspects of these quasi-experimental methods offer an important perspective on the body of estimates of funding impacts. First, each study produces one estimate of the funding impact parameter and the sampling errors in this estimation may be large by the nature of the methods. Because the estimation generally relies on convenience samples that involve special circumstances that exhibit exogenous variations in spending, there is seldom replication of the specific funding situation being studied. Refining the impact estimates then calls for meta-analysis that combines related funding situations, but that aggregation then involves combining different impact factors, i.e., different γ_c. Second, the value of the impact estimates for policy purposes generally comes from application of the results to different circumstances than the samples from which they were generated.

Faced with alternative estimates of funding impacts, it would be useful to combine different estimates to improve the precision for any use in projecting outcomes from a specific funding program. Combining estimates obtained in the same context (*c*) is obviously desirable, but that requires being able to describe the context sufficiently to conclude that any two studies are drawn from the same context. Of course, even if contexts differed to some extent, the application of standard meta-analysis techniques could yield better estimates (in a mean square error sense) as long as the contexts were not too different.[3] Unfortunately, we have little guidance about how to describe and compare different contexts, making it difficult to know what combination of results can be used to improve our understanding of impacts as opposed to adding new distortions. We simply know little about the relevant moderators, a fact that has clear implications for the direction of future research.

3 Inclusion Criteria for Relevant Studies

This analysis considers well-identified studies of the impact of funding on student achievement or student attainment of different levels of schooling. We focus on studies of operating budgets and finance programs as distinct from studies of specific inputs to the production process such as capital expenditures for school construction or renovations and class size reduction. Studies of such specific inputs obviously involve school spending, but we are interested in the impact of more general funding

[3] This formulation is similar to that in Pritchett and Sandefur (2015) where the focus is combining RCT evidence garnered from different locations.

changes, potentially addressing the impacts of providing extra resources without further specification of how they are used.

This paper builds upon our prior review of quasi-experimental studies available through February 2022 (Handel and Hanushek (2023)). We searched for research pursuing quasi-experimental approaches to measuring the impacts of school spending on student achievement and attainment in the United States. We focused on studies adhering to modern quality standards for causal research. A central element of this research is an explicit description and justification of the counterfactual, or what would occur without the specific program under consideration.

We begin with a discussion of the search procedures used to find the relevant set of studies. We then turn to a description and compilation of impact results.

3.1 Study Selection

We followed a structured search of relevant sources and then systematically eliminated papers not meeting our pre-established selection criteria. The search began with journal articles published and available between 1999 and February 2022 using search engines covering the economics and education literatures: EconLit and the Education Resources Information Center (ERIC). We included the search term "education" along with a set of keywords: school spending; expenditure; resources; inputs; school finance; school finance reform; budget; funding; revenues; money matters. We then repeated the search for relevant working paper series: National Bureau of Economic Research (NBER); World Bank Policy Research; the Institute for the Study of Labor (IZA); the Center for Economic and Policy Research (CEPR); and the CESIfo Research Network.[4] We reviewed the abstracts of the English language articles and selected those papers whose abstracts met three criteria: 1) discussion of a quantitative causal analysis; 2) relevance to school spending, and 3) mention of effects on student outcomes, including test scores and various measures of attainment such as dropout rates, years of education, graduation rates, etc. From this set of studies, we selected those papers whose estimation strategies included sufficient treatment of possible omitted variable or endogeneity bias. These papers included those employing a randomized controlled trial (RCT), difference-in-differences (DD) regression, fixed effects (FE) estimators, regression discontinuity (RD) design, instrumental variables (IV), and variations on these methods.

We then identified additional papers either cited in the reference list of first round papers or citing first round papers (as identified using Google Scholar's "cited by" feature). The first round procedures were repeated with this second set of studies. Finally, we further narrowed the pool of studies by ensuring the inclusion of information relevant for producing comparable impact parameters.[5]

[4] While we review the international working paper series, we focus on studies of U.S. schools (which may appear in these series). Handel and Hanushek (2023) does include relevant studies for international schools.

[5] Those not providing either the base levels of per-pupil spending or the necessary inputs to calculate these levels were excluded. Studies that only provide effects of various policies on gaps in achievement or attainment (e.g., between white and black students or between low SES and high SES students), as opposed to levels, are likewise excluded. For more details on both selection of studies and choice of parameter estimates, see Handel and Hanushek (2023) and its data appendix at https://data-nber-org.stanford.idm.oclc.org/data-

We have found unpublished studies in major working paper series but may have missed articles not appearing in these restricted working paper series.

3.2 Creating comparable parameters

Because of differences in definitions and measurement of the fundamental inputs and outputs across studies, we harmonize the measurement so that the estimated impact parameters are as comparable as possible. This harmonization is not trivial but is crucial to obtaining reliable comparisons of different estimates of the impact of spending.[6] Following that, we return to the other issues of comparison.

We compute the effect of a 10 percent increase in real (inflation-adjusted) per-pupil school spending on standardized outcomes for the general population of students. This requires rescaling and transforming the estimates in order to make comparisons that allow conclusions across various contexts. Because of the sharp rise in spending per pupil over the past half century, it would be inappropriate to compare simple inflation-adjusted spending levels because the actual date of application, which varies widely across studies, would then be important. We select estimates taken four years after a policy change or from the beginning of the study period. If this is not available, we take the longest period up to four years.

We scale the estimates by the student level standard deviations of the outcomes.[7] This normalization is mostly straightforward for achievement levels because test score estimates are often provided in standardized terms. When effects on raw score are provided, they are simply divided by the standard deviation of test scores in the sample that is typically provided by the author.

It is generally impossible to put pass rates on a scale that is comparable to the estimated impact parameters based on standard deviations of test scores. When the outcome is a fraction of students above a proficient score threshold (i.e., a pass rate), we do not attempt to compare the magnitudes of changes to other test score estimates. Such proficiency rates depend on the cut scores chosen by a standard-setting process. Changes in cut scores placed at different points in the achievement distribution can vastly and unpredictably affect the interpretation of impacts (Holland (2002), Ho (2008)).[8]

In studies where effects are reported separately for different test score subjects, grade levels, or demographic populations, we use the reported standard deviation for that given subgroup if available. If

appendix/w30769/Appendix_tables.pdf or http://hanushek.stanford.edu/publications/us-school-finance-resources-and-outcomes.

[6] Here we provide an overview of the process; details can be found in Handel and Hanushek (2023) and its on-line appendices.

[7] This normalization of course implies that one standard deviation means the same across two different tests.

[8] It would be possible to translate the change in pass rates into a change in the SD of passing, using the formula for the standard deviation of a binomial variable, $\sqrt{p(1-p)}$, where p is the sample probability of passing. This calculation clearly varies with the underlying cut point for the passing score and is not the same as the standard deviation of student test performance. In other words, the same passing rate can come from distributions with wildly different standard deviations. It is thus inappropriate for standardizing effect sizes, leading us to drop consideration of the pass rate studies.

the student-level standard deviation is only available for the full sample, we use this general metric. To convert estimates into student-level standardized units if not already presented as such, we divide the raw effect by the standard deviation.

Some of the original studies focus on school attainment, school completion rates, or the like, which are obviously measures of time inputs into the educational process. They remain crude surrogates for student performance even if they are also frequently used as outcomes when there are no measures of achievement or learning. (The recent pandemic underscores the problems with these attainment measures, because school closures plus altered learning patterns make a year of schooling during the pandemic very different from a year of schooling outside of the pandemic period).[9] But this is a more general problem because the quality of schooling varies over time and across space.

We treat studies of school attainment as distinct from those of achievement, and we place less weight on them when thinking about school policies. We standardize attainment studies by transforming them into percentage change measures. We multiply the impact estimate in dropout studies by -1 to be able to compare them to other measures of attainment, in which positive estimates imply desired impacts. Nonetheless, the attainment studies are very heterogeneous. For example, much of the research and policy discussions generally treat concerns about high school dropouts as qualitatively different from college attendance – making any comparisons and aggregation of these impact parameters problematic.[10]

Throughout we report the estimates from the most general specification with regards to sample composition. If authors only provide separate estimates across grade levels, income levels, race, etc., we compute average estimates using a precision-weighted mean to combine estimates across grade levels. To combine estimates across populations with different demographic characteristics, we weight estimates with the relative share of their respective subgroups in the overall population.

3.3 Included studies

The list of studies by outcome measure is found in Appendix A. Details on each of the studies employed here can be found in Handel and Hanushek (2023) .

Applying these search and selection procedures, we have found 16 well-identified studies of funding and student test score performance. Ten of these came from the published literature. They are almost evenly split by estimation approach: instrumental variables (6), difference-in-differences (5), and regression discontinuity (5).

The 18 attainment studies are spread across decision points. Eight consider high school graduation, six consider college attendance, and four consider high school dropout rates. While four

[9] See, for example, Hanushek and Woessmann (2020), Halloran, Jack, Okun, and Oster (2021), Kuhfeld, Soland, and Lewis (2022),

[10] For example, Oreopoulos (2007) points to myopic behavior and lack of information in dropout decisions. While some informational issues about college entry are addressed in Page and Scott-Clayton (2016) and Dynarski, Nurshatayeva, Page, and Scott-Clayton (2023), the majority of discussion concerns financial aid and other barriers to entry (e.g., Dynarski, Page, and Scott-Clayton (2023)).

employ regression discontinuity methods, the remainder are evenly split between instrumental variables and difference-in-differences.

4 Overall Summary of Findings

The estimated impacts of spending on student achievement vary substantially across the 16 studies of U.S. outcomes. Not surprisingly, 14 show positive effects. Nine of them are statistically significant at conventional levels (Table 1). The overall median effect size for a 10 percent spending increase is 0.07 standard deviations. Estimates closer to the median tend to be more precise, although it is worth noting the wide range of estimates observed. The estimated impact on test scores, measured in standard deviations resulting from a 10 percent increase in spending ranges from -0.244 (not statistically significant at the 0.05 level) to 0.543 (statistically significant). Figure 1 visually summarizes the distribution of estimated effect sizes. We display the standardized effect size along with its 95 percent confidence interval using a forest plot for test scores (in standard deviations).

The 18 U.S. estimates of the effects of school spending on time spent in school attendance come from across different dimensions of attainment including such measures as high school completion, school dropouts, and college enrollment. While these measures are coarse indicators of student learning and skills, they have been widely used in labor economics and frequently appear in policy discussions. We present the results of these studies by quantifying the impact as a percentage change in the specific outcome for every 10 percent increase in spending. Figure 2 visually demonstrates the distribution of results. In Figure 2, it is evident that all 18 estimated effects for attainment, measured through graduation rates, dropout rates, and college enrollment rates, show positive impacts. Out of these, 14 reach statistically significant levels. The median impact suggests that a 10 percent increase in school spending leads to a 5.7 percent improvement in high school graduation, college enrollment, or other measures of attainment.

It is important to interpret this median impact cautiously because dropout rates and college enrollment rates may not respond to school spending in the same way and may be influenced by different factors. While most estimates align closely with the median, there are notable outliers. On the lower end, a 10 percent spending increase yields a modest 1.8 percent improvement in attainment, whereas on the higher end, there is an astonishing 85 percent improvement in dropout rates.

5 Interpretation of Overall Results

Ultimately we are interested in whether the results of the included studies provide us information about what might result from the infusion of resources into schools in a new setting outside of the ones analyzed in the currently available studies. Before getting into that issue, however, it is important to understand just what data we have. We first discuss the interpretation of estimates from the individual studies. We then discuss the relevant sample of results.

5.1 Factors affecting the study estimates

Possible publication bias introduces the first important caveat for our compilation of existing studies. Past observations and analyses indicate that the sign, size, and statistical significance of key parameters can influence publication of scientific research. This issue has been analyzed in a range of disciplines and has been found quite broadly to be a serious issue (see, for example, Nissen, Magidson, Gross, and Bergstrom (2016)). The problem has been linked both to the choices made by researchers and the choices made by journal editors. As an example, an early direct study of clinical trials using RCTs found that negative results systematically led to a lower probability of the findings being written up and submitted. As another example, Franco, Malhotra, and Simonovits (2014) analyze a cohort of NSF-sponsored projects in the social sciences and find that "Strong results are 40 percentage points more likely to be published than are null results and 60 percentage points more likely to be written up." Such selection of research into the record (and into the body of evidence compiled here) has direct implications for the generalizability of the various estimates but is obviously exceedingly difficult to assess.

One particular form of the publication-induced incentives, so called p-hacking, is more amenable to investigation. Head , Lanfear, Kahn, and Jennions (2015) summarize the issue as "A focus on novel, confirmatory, and statistically significant results leads to substantial bias in the scientific literature. One type of bias, known as "p-hacking," occurs when researchers collect or select data or statistical analyses until nonsignificant results become significant." Recent analysis, albeit not without controversy, attempts to quantify the extent of p-hacking in economics. The magnitude of it is under question, but its existence seems indisputable (Brodeur, Cook, and Heyes (2020, 2022), Kranz and Pütz (2022), Brodeur, Carrell, Figlio, and Lusher (2023)).[11] Thus, the subsequent reporting of statistical significance (and impact parameters themselves) may be directly influenced by the research bias induced by the publication process.

While seldom considered, the normal design and data preparation decisions of researchers introduces variation in estimated impact factors that go beyond the reported sampling errors (Huntington-Klein et al. (2021)). Well-intentioned researcher choices can introduce substantial variation in reported results and can alter judgments about statistical significance of individual estimates of impact parameters (Gelman and Loken (2014), Silberzahn and al. (2018), Dillon, Miller, and Smith (2023)).

While a variety of tests and corrections for publication bias have been proposed, we focus just on the author(s)'s findings from both published and unpublished studies where possible.[12]

Perhaps more importantly, no consideration is given to potential threats to the identification of causal impacts in the underlying studies. The application of the quasi-experimental methods does not

[11] See also Ioannidis, Stanley, and Doucouliagos (2017) for a somewhat different but related perspective on the influence of power of underlying estimates.

[12] See, for example, Andrews and Kasy (2019). The recent movement to pre-registration along with pre-analysis plans may ameliorate some of these problems (Brodeur, Cook, Hartley, and Heyes (2022)).

inherently guarantee unbiased estimates of the impact of resources on student outcomes. The quasi-experimental methods place clear restrictions on the underlying context behind their application if they are to deliver internally valid estimates.[13]

In a standard potential outcome evaluation model, we want to infer the difference in achievement of students when their school receives or does not receive a given infusion of resources. The concern is that any observed achievement differences are also influenced by other factors that are also correlated with the added resources. The possible problems are especially intense in the case of school funding because funding is often part of larger policy concerns that are difficult to separate from the resources per se. For example, in 1996, the State of California provided $650 per student in funding for classrooms in grades K-3 with 20 students or less (Stecher and Bohrnstedt (1999), Jepsen and Rivkin (2002, 2009)). One obviously would not want to correlate the added funding to a district with student achievement and interpret the differences in outcomes as what happens when the state increases general school funding.

The specific studies reviewed here are essentially efforts to find situations where it is possible to rule out non-resource factors in assessing the resource-achievement relationship. The "study-quality" issue is how successful the authors have been in selecting circumstances and employing the resultant data in ways that allow identification of the desired parameters. Josh McGee (2023a, 2023b), in evaluating a number of the studies that we have identified in our review, provides a number of examples that underscore the difficulty of identifying the causal impact of specific funding programs: some funding changes are directly correlated legislative programs; it is difficult to assume that either legislative or judicially-inspired funding changes are random; and, depending on the temporal nature of funding observations, it is possible that related demographic changes enter the picture. These issues of the validity of key underlying identification concerns interact with potential questions about the application of the estimation methodology, including but not restricted to the publication issues previously noted.[14]

The methodology for causal impact analysis can further enter into the resulting impact estimates. While the difference-in-differences approach is designed to identify the average treatment effect, the instrumental variables (IV) and regression discontinuity design (RDD) identify local average treatment effects, reinforcing concerns about the potential importance of study context when comparing estimates from different studies.

For this analysis, we again take the studies at face value. But, it is reasonable to presume that the distribution of results is affected by both these publication issues and empirical difficulties. For

[13] Note also that our understanding of the underlying estimation approaches has also changed over time. For example, recent advances in difference-in-differences methodology indicate that some early applications of this general approach could lead to substantial biases even if the studies were judged as being well-identified at the time of their execution (Roth, Sant'Anna, Bilinski, and Poe (2023)).

[14] Relatedly, there are also studies that attempt to replicate some prior work because of questions about the results, but we do not include separate estimates from these (e.g., the critique of Jackson, Wigger, and Xiong (2021) by Goldstein and McGee (2020)).

interpretative purposes, we simply note that these complications most likely lead to an upward bias of the estimated impacts of added resources.

5.2 The sample of estimated impacts

We have an array of estimated impact parameters that exhibit considerable variation, albeit with varying amounts of within-sample error. The set was not the outcome of a strategic sampling design but instead capitalized on specific circumstances that provided an opportunity to investigate the impact of funding differences.

Each of the studies has its genesis in special circumstances that permit identification of an impact parameter for added funding. As a result, they come from very different contexts. The studies include, for example, the use of funds for compensatory education (Title 1), reactions to legislative changes in district funding, court actions within and across states, and recessionary downturns in state budgets.

Some changes in funding come with highly prescribed conditions on spending. Title 1 spending is restricted to low-income students, while personnel decisions associated with budget reductions are generally highly prescribed by union contracts. Various judicial decisions about funding cover not only specific conditions to be remedied but also programmatic details of spending.

The issues go beyond this consideration of requirements placed on the additional funds because overall regulation, funding, and monitoring of school districts is the purview of the individual states. The states have enacted very different policies that will interact with the effectiveness of any added funding.[15] Specifically, the state institutional structure will determine the incentives facing district-level decision makers and is likely to influence the uses and impact of any additional funding. As a simple example, following prior approaches to identifying the funding impact parameter by state reactions to judicial decisions, Buerger, Lee, and Singleton (2021) show that the estimated funding impacts vary meaningfully with the existing state accountability policies. Brunner, Hyman, and Ju (2020) focus on differential achievement impacts resulting from the underlying degree of teacher unionization. In the different setting of comparing program impacts across the context of different developing countries, Pritchett and Sandefur (2015) conclude that "Social programs, in contrast [to understanding estimates of physical parameters], are embedded in contexts which encompass a long list of unknown factors which interact in often unknown ways."

As already seen, there appears to be considerable heterogeneity in the underlying impact parameters across both sets of estimated funding impact parameters. It is useful to consider the source of this variation. If it is just sampling error in the individual studies that drives these differences, we can improve on the estimates by aggregating across the different studies – a task the has been well-studied in a range of approaches including various versions of meta-analysis. On the other hand, if the heterogeneity reflects some of the driving forces discussed above, it is less clear how to treat the array of estimates. It makes sense to combine estimates from studies within contexts representing similar

[15] The potential importance of state policies as a moderator for the impact of funding is seen in prior estimates of educational production functions (Hanushek (2003)).

treatments such that the context has a small impact on the outcomes. But, if the context makes an important difference on the impacts of the treatment, such aggregation of results across disparate situations would provide biased forecasts of the impact of any funding changes.

6 Does Money Matter?

In order to formalize the summary analysis of these school spending studies, we begin by conducting a meta-analysis across the full pool of the separate impact estimates. The summary in section 4 indicated that 14 of 16 estimates of test score impact were positive, but only 60 percent give confidence that the true value is not zero. The appeal of meta-analysis is that by combining the estimates it may be possible to reduce some of the uncertainty.

Standard meta-analysis provides a methodology for aggregating the study results along with the ability to investigate the variations in estimated impacts that we observed. We summarize the estimates of the impact of spending on achievement and attainment separately. We apply the customary approach of using inverse variance weighting of the underlying estimates, which gives more precise studies more weight.[16] We use a random effects estimator, interpreting the variance used for each study weight as being composed of both within-study variance and between-study variance and implicitly allowing the true underlying funding impact to differ across studies. As suggested by the previous discussion, this allowance is particularly important given differences in contexts and the assumption that study quality differences are inconsequential. We also apply a Hartung-Knapp modification to the standard errors to incorporate the uncertainty in the estimation of the between-study variance.[17]

The summary effect for test scores implies that a 10% increase in funding leads to a .0647 standard deviation increase in test scores, with the 95% confidence interval spanning from .0394 SD to .0900 SD (Table 2). The summary effect is very similar to the median effect size discussed in Section 3, but the 95% confidence interval spans a much smaller range than the full distribution of estimates, reflecting the fact that studies with effect sizes farther from the median are typically less precise. For attainment, the summary value implies that a 10% increase in spending leads to a 5.5% increase in educational attainment, with the 95% confidence interval spanning from 2.25% to 8.75%. This meta-analytic summary effect is also similar to the median effect of spending on attainment.

This summary comes from pooling all of the estimated funding impact parameters that are found in the separate quasi-experiments. As such, it is not easy to specify the distribution from which they are drawn. The combined studies consider spending under a wide range of conditions, but it is not possible to extract what impact might be seen under any specific set of circumstances. From the test score studies, one can conclude that it is more likely than not that spending more yields a result that is different from zero.

Stopping with this summary leads to the naïve conclusion that “money matters” in the sense that the studies provide evidence that added funds are likely to have a positive impact on student

[16] This procedure makes more sense if the underlying impact parameters are the same but less if they differ.
[17] The Hartung-Knapp modification is used with meta-analyses for small numbers of estimates to correct for potential bias in the estimated between-study variance using standard procedures.

outcomes, i.e., that added funds are unlikely to harm students.[18] For public policy decisions, however, the issue is not whether there is an expected positive impact of added funds but whether these aggregate results provide a reliable indication of the magnitude of impact expected from introducing more funds. Simply knowing that the recipients of a governmental program are not likely to be harmed by itself is of course not sufficient justification for a governmental program. The magnitude and persistence of any impact along with the efficiency of program spending are vital policy-relevant metrics that are ignored when we reduce the discussion to that uninformative question.

When we estimate the source of the observed between-study variation in estimates, we provide new evidence about the inconsistency of impacts across study contexts. We find that heterogeneity in the true impact parameters accounts for one-half or more of the observed variance. The last column of Table 2 presents I^2, a standard measure of between-study heterogeneity for meta-analyses.[19] These findings suggest that for test scores, 50.5% of the variability in effect sizes reflects real differences in effect sizes across studies. For attainment, 77.6% of the variability between studies reflects real differences in effect sizes. In words, the context (or other more fundamental estimation concerns) drive much of the differences in estimated impacts across studies, making projection of any impact of funding highly dependent on the context – which is not well-specified at this point.

The "does money matter" debate in school finance reduces investigations of the impact of funding on student outcomes to an uninformative binary response.

7 What underlies the impact heterogeneity?

If we can identify the major sources of heterogeneity in the underlying impact of funding, it would be possible to provide more refined estimates of when and where funding has its largest impact. This would then facilitate assessment of alternative possible funding policies that would lead to enhanced student outcomes.

The range of factors that can be investigated is of course limited by the relatively small number of studies of funding impact that are available. Perhaps more importantly, the investigation is also limited by the generally cursory consideration of contexts from each of the studies. The estimation methods are designed to separate the pure funding impacts from other factors that might influence student outcomes, making such context considerations ancillary to the primary study objectives. However, a few studies explicitly incorporate elements of context and find that crude aggregate context factors are very important. Buerger, Lee, and Singleton (2021) show that the impact of funding varies importantly according to the existence of a student accountability program; Brunner, Hyman, and Ju

[18] The language reflects popular nonscientific discussions that address the rhetorical question of 'does money matter?' as opposed to the inherent policy questions about school funding (see, for example, Barnum (2023)). This rhetorical language also enters into the advocacy discussions surrounding school finance litigation (see, for example, Rebell (2019)).

[19] Note, however, that the I^2 calculation does not take into account the influence of normal researcher decisions on both the parameter estimates and the estimates of the underlying variation in these estimates as discussed in Section 5.1 above.

(2020) demonstrate that the impact of increased funding is strongly influenced by the extent of unionization of teachers.

We consider two major classes of heterogeneity: methods-induced and structural. Analytically, our subgroup analyses divide the studies along lines that potentially capture the key drivers of heterogeneity of impacts. This produces the meta-analytic analogue to an analysis of variance in a primary study, where we compare mean effect sizes across various subgroups of estimates or studies. As in our summary meta-analyses, we apply an inverse variance weighting with a random effects estimator. We also allow for the true between-study variance (used in constructing study weights) to differ across groups.

7.1 Funding and test scores

Heterogeneity may stem from factors related to study design or program and context characteristics. The alternative approaches to estimation might yield different estimates of impact parameters independent of any concerns about validity of their underlying assumptions and the quality of the analysis. In particular, the RDD and IV methods provide estimates of the local average treatment effect (LATE), and these estimates may not provide any direct estimates of the average treatment effects for the relevant populations. To identify the role that study design may play in influencing effect size, we first conduct a subgroup analysis using the alternative causal inference methodologies as the relevant subgrouping. We present these findings in Figure 3. Though the effect sizes differ slightly across studies using regression discontinuity, instrumental variables, and difference-in-differences, these differences are not statistically significant. We conclude that heterogeneity in effect sizes across studies in our sample is unlikely to stem primarily from differences in methodology or study design.[20]

The contexts for the studies as described differ in complex ways, and they do not fall into well-defined conceptual categories. We are challenged by small sample size with 16 estimates of the effects of school spending on test scores and 18 estimates of the effects of school spending on educational attainment. For this exploratory analysis, we choose three broad and readily defined context-related factors to explore.

First, we investigate whether funding impacts are affected by spending that can be categorized as "targeted" versus "non-targeted," where targeted funds originate from programs or policies with spending aimed at specific subgroups (low income, low scoring, etc.). As an illustrative example, compare NYC's Hold Harmless provision as leveraged by Gigliotti and Sorensen (2018) with Title I spending as examined by Cascio, Gordon, and Reber (2013). For the former, the authors examine the effects of excess funding that stem from a budget quirk – one which holds constant the total amount of funding even if enrollment falls. In this case, the additional funding is not necessarily targeted for any

[20] There are of course other factors not captured by this coarse approach, including sample size, more detailed methodological choices (RD bandwidth size, e.g.), and other unobservable research choices that are beyond our capacity at this point to analyze. But these attributes are more likely to influence the standard errors and precision of the estimates as opposed the parameter estimates themselves.

specific purpose or group of students. Title I spending provides a clear opposite case: spending here is very clearly targeted at low-income students (at least in de jure terms).

The context estimates imply that targeted spending (8 of the 16 estimates) was slightly less effective (see 4), though the difference in means is not statistically significant. Thus, we cannot attribute heterogeneous results to differences in effectiveness between targeted and non-targeted spending.

Second, we investigate whether it matters if a study's source of spending variation originates from a school finance court case.[21] Six of the studies of impacts on test scores rely on variation in funding engendered by court decisions in school finance litigation.[22] Such spending may involve specific court directives or may affect districts in the state differentially, yielding dissimilar impacts on student performance compared with other funding changes. Again, however, as evident in Figure 4, we cannot attribute the substantial amount of between-study heterogeneity to whether additional spending was tied to school finance litigation.

Finally, we explore whether a portion of between-study heterogeneity may be explained by the breadth of the sample of students, schools, and school districts. We compare estimated effects of spending on achievement from studies that look within a single state (where many regulations and incentives are constant) to those that combine or compare students across multiple states (where many more contextual attributes vary). The estimates from studies that span several states are more precisely estimated, likely a consequence of larger sample sizes. Still, the summary of estimated impact effects is strikingly similar for both groups.

None of our exploratory explanations of context differences - targeting of spending, origination in school finance litigation, or breadth of sample – explains the substantial between-study heterogeneity that we observe. This, we believe, reflects the incredibly complex network of factors that contribute to school spending effectiveness. With the small number of estimates and a dearth of studies replicating previous estimates in similar contexts, we are unable to identify policy-relevant drivers of differential effectiveness. In other words, how funds are spent is very important, but we do not have a good description of contexts that yield particularly effective (or ineffective) uses of funds.

7.2 Funding and School Attainment

We also conduct subgroup analyses for studies estimating the impact of school spending on education attainment, for which between-study heterogeneity represents over 75% of the differences across standardized effects. We first highlight the challenge of combining these estimates into a single parameter by exploring differential effect size by measure of attainment – graduation rates, dropout rates, or college-going rates. As evident in Figure 5, these measures respond differently to spending,

[21] See Hanushek and Joyce-Wirtz (2023) for the most recent discussion of the causes and consequences of school finance litigation.

[22] Categorizing the source of funding is itself subject to interpretation because, for example, Lafortune, Rothstein, and Schanzenbach (2018b) combine both court-induced and purely legislative increases. As noted, the reliance on court variation in funding assumes that court decisions and subsequent legislative action are exogenous to the funding responses and the effectiveness of such funding. If not, any subsequent bias in the estimates of impact parameters could contribute to the heterogeneity of results.

with dropout rates proving most sensitive to spending increases (although the estimates are the least precise) and graduation rates proving least sensitive (where the estimates are the most precise). Because of small sample sizes, it is difficult to make definitive conclusions about the magnitudes presented, but the difference in means is statistically significant, cautioning against treating all of these rates as interchangeable measures of educational attainment.

If we omit estimated effects of spending on dropout rates when calculating our mean effect on attainment, we get an estimate that implies that a 10% increase in spending leads to a 4.12% increase in attainment with the 95% confidence interval ranging from 1.71% to 6.53%. This point estimate is lower than the attainment summary effect including dropout rates, which implies a 5.5% increase in attainment with a 95% confidence interval ranging from 2.25% to 8.75%.

We also replicate the impact of targeted and court-induced and non-court-induced spending on attainment (Figure 5). We again find that the difference in subgroup means is not statistically or economically significant, meaning that we cannot attribute the substantial between-study heterogeneity to either of these dimensions of differing policy contexts.

7.3 Focus on SES

We have not identified any measures of study design or policy context that can explain the substantial between-study heterogeneity identified in our meta-analysis. We turn to another potential dimension of heterogeneity. The underlying studies exploring the impacts of school spending on student outcomes frequently performed a variety of their own tests for heterogeneous effects. Some studies did not report effect sizes for different groups, while others broke down effects by socioeconomic status of the school district or student, student gender, baseline achievement, baseline spending levels, and other margins. The most common margin studied is socioeconomic status, so we will focus on the studies that provided some breakdown of their estimated effects across this margin. From prior estimates of education production models, it is clear that both average socioeconomic status in a district and individual student status may have impacts on the effects of spending by serving as a proxy for family or community resources.

As seen in Table 3, 11 studies provide estimates of impacts of spending on achievement for low SES students or students living in low SES districts, albeit by the metrics chosen by the authors: percentage of students eligible for free lunch (FLE), childhood poverty rates, and mean income.[23] The

[23] In cases where the authors provide separate estimates for each of these subgroups, scaling the estimated effects using the standardization procedures outlined in Section 2.2 is straightforward. For those that present heterogeneity estimates by presenting the coefficient on the treatment interacted with a measure of SES, it is possible to obtain the effect size for each subgroup, but it is not possible to uncover the standard error of these estimates. For example, Abott, Kogan, Lavertu, and Peskowitz (2020) simply report the coefficient on referenda passage (variation in spending) and the coefficient on passage interacted with an indicator for high poverty districts. While this is useful for demonstrating the direction and magnitude of the difference in the estimated effect of additional spending between high- and low-poverty districts, it is difficult to use these results to construct summary metrics for the effect sizes for each group using evidence from many studies. Thus, we just include those estimates presented separately in our meta-analytic analysis.

studies that investigate the differential effects of spending across student-level SES use both childhood poverty and free lunch eligibility as measures of SES.

First, as in Table 3, we examine the distribution of estimated effects across low- and high-SES groups. For both test scores and attainment, the median effect size is higher for low SES groups than high SES groups. A 10% increase in spending leads to a 0.069 SD increase in test scores in low SES groups as compared to 0.046 in high SES groups. Similarly, these median values imply that a 10% increase in spending leads to a 12.3% increase in educational attainment in low SES groups and a 4.4% increase in attainment for high SES groups. Still, the sample sizes are quite small with few estimates in each grouping, so it is difficult to interpret these differences as differences in the true parameters.

These differences do motivate a meta-analytic assessment of the heterogeneous effects. We complete a subgroup analysis, similar to those outlined earlier in this section. Here, we again leverage a random effects model in which we allow true between-study variance to vary across subgroups. The summary parameters are very similar to the medians discussed above, which means that it is unlikely that any differences in effect sizes are driven solely by imprecisely estimated outliers. For test scores, we cannot identify a statistically significant difference in means. For attainment, however, the difference in means is statistically significant. When interpreting these findings, it is important to recall the earlier discussion that the different measures of attainment have varying levels of sensitivity to spending in this sample of studies. Similarly, it is also possible that dropout rates, which are the most sensitive to spending on average, are more likely to change in lower SES regions due to higher potential for improvement and higher base rates. In a more general sense, we might believe that the quantity of education (attainment) may be more sensitive in lower SES regions, but we cannot identify the mechanisms given the available information.

We present these results with caution given the small sample size and the risk of contamination due to publication bias. Although publication bias is a concern in any meta-analytic setting, its impact may be larger in this context. The study of differential effects by SES is an extremely policy-relevant exercise that one might have expected more universally. Eleven out of the 23 school spending studies presented some breakdown by a metric related to SES, meaning that 12 studies did not.

Conclusions

Recent research into school finance has greatly improved estimates of the impact of funding on student outcomes. By paying closer attention to the identification of the causal impact of funding on outcomes (i.e., to “internal validity”), the research has reduced many of the concerns with the historic research into this issue.

Research into the impact of funding on student outcomes has immediate policy relevance. Decision makers in both legislatures and the courts are most interested in student outcomes but rely upon funding of schools to achieve their desired results. Their task would be easier if they could focus just on the amount to be spent and then rely on local districts to produce high levels of performance.

The recent research employs quasi-experimental methods to estimate the impact of added funding on student achievement. This research centers on circumstances that involve exogenous variations in funding that can then be related to student outcomes – yielding unbiased estimates of the impact of funding. We have located 16 well-identified studies relating funding to student achievement and 18 relating funding to educational attainment.[24] Each of these has been harmonized to give an estimate of the change in test scores (attainment) that is related to a 10 percent increase in funding.

The estimated impact of funding is substantially different across the studies. Part of the difference is the result of sampling error, but half of the variation in estimated test score impacts and over three-quarters of the variation in estimated attainment impacts reflects variation in the true underlying parameters. This reflects essential differences in the context for each of the observed relationships where context reflects both the institutional/regulatory environment of the observed funding and the quality of decision making behind the use of the funds. We label the outcomes of decision making in each context as simply "how" money is spent.

The estimated impacts are almost all positive in sign, and 60 percent of the estimated test score parameters are statistically significant by conventional standards – implying that added funding most likely will not harm students. But not harming students is not a sufficient criterion to justify governmental spending. When we turn to consideration of the magnitude of funding impacts, we see that the efficacy of spending increases as uncovered by the existing studies depends crucially on how funds are spent.

While we can have more confidence than in earlier studies that the existing analyses provide internally valid estimates of the impact of funds within their specific context, we are left with serious questions about how to generalize from the study contexts to other circumstances relevant to educational policy. The extant studies have relied on circumstances where the observed funding is plausibly exogenous, not on whether the new estimates help in generalizing to other situations. The currently available evidence comes from wildly different funding situations – highly constrained spending formula to much less constrained, funding across very different states to funding within selected states, reductions in spending versus increases in spending, and more. It is natural to think about looking beyond the entire collection of estimates and – depending on the potential policy application – to look at a more homogeneous subset of circumstances. But the existing studies give few hints about what if any studies are relevant to a new potential application.

We pursued exploratory investigations where we attempted to understand more fundamental forces that dictated particularly effective use of funds. We dismissed the idea that the differences in estimated impacts depended on the particular estimation approach. But we also concluded that the targeting of funds, the role of court interventions, and the impacts across state or district programs could not explain the variations in impact parameters.

[24] We have assumed that issues involving publication bias, p-hacking, researcher analytical choices, and other issues of study quality do not affect our sample of impact estimates. While this is a strong assumption, we are not able to assess these possible challenges. These factors imply that are sample may overestimate potential funding impacts.

When we went deeper into differential impacts by student SES, we found evidence that low income students were more sensitive to added funding. But, the context was still important in the results.

The convenience samples employed in these analyses are clearly important in the ability to obtain well-identified estimates of the impact of funding. Yet the context of educational funding interventions is also important and makes it difficult to provide scientific estimates of the impact of funding in a different context.

The description of the relevant elements of how money is effectively used to improve student outcomes is unfortunately not likely to improve rapidly with research. Incentives to replicate existing quasi-experimental studies are not strong for either researchers or journal editors who value uniqueness. And, relying on contexts of convenience for identification and using estimation approaches that are designed to circumvent any context issues does not provide strong incentives for the development of scientific investigations of when and how funds are best used. This is also a research issue that goes beyond just queries into the impact of school funding but arises in a wide range of other policy areas.

References

Abott, Carolyn, Vladimir Kogan, Stéphane Lavertu, and Zachary Peskowitz. 2020. "School district operational spending and student outcomes: Evidence from tax elections in seven states." *Journal of Public Economics* 183.

Andrews, Isaiah, and Maximilian Kasy. 2019. "Identification of and Correction for Publication Bias." *American Economic Review* 109, no. 8 (August): 2766-94.

Barnum, Matt. 2023. An economist spent decades saying money wouldn't help schools. Now his research suggests otherwise. *Chalkbeat*, May 16.

Baron, E. Jason. 2022. "School Spending and Student Outcomes: Evidence from Revenue Limit Elections in Wisconsin." *American Economic Journal: Economic Policy* 14, no. 1: 1-39.

Baron, E. Jason, Joshua M. Hyman, and Brittany N. Vasquez. 2022. "Public School Funding, School Quality, and Adult Crime." NBER Working Paper No. 29855. Cambridge, MA: National Bureau of Economic Research (March).

Brodeur, Abel, Scott E. Carrell, David N. Figlio, and Lester R. Lusher. 2023. "Unpacking P-Hacking and Publication Bias." NBER Working Paper Series No. 31548. Cambridge, MA: National Bureau of Economic Research (August).

Brodeur, Abel, Nikolai Cook, Jonathan S. Hartley, and Anthony Heyes. 2022. "Do Pre-Registration and Pre-analysis Plans Reduce p-Hacking and Publication Bias?" SIEPR Working Paper No. 22-19. Stanford University (August).

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020. "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics." *American Economic Review* 110, no. 11 (November): 3634-60.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2022. "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics: Reply." *American Economic Review* 112, no. 9: 3137-39.

Brunner, Eric, Joshua Hyman, and Andrew Ju. 2020. "School Finance Reforms, Teachers' Unions, and the Allocation of School Resources." *The Review of Economics and Statistics* 102, no. 3: 473-489.

Buerger, Christian, Seung Hyeong Lee, and John D. Singleton. 2021. "Test-Based Accountability and the Eectiveness of School Finance Reforms." *AEA Papers and Proceedings* 111(May): 455-459.
Candelaria, Christopher A., and Kenneth A. Shores. 2019. "Court-Ordered Finance Reforms in the Adequacy Era: Heterogeneous Causal Effects and Sensitivity." *Education Finance and Policy* 14, no. 1 (Winter): 31-60.
Carlson, Deven, and Stéphane Lavertu. 2018. "School Improvement Grants in Ohio: Effects on Student Achievement and School Administration." *Educational Evaluation and Policy Analysis* 40, no. 3: 287-315.
Cascio, Elizabeth U., Nora Gordon, and Sarah Reber. 2013. "Local Responses to Federal Grants: Evidence from the Introduction of Title I in the South." *American Economic Journal: Economic Policy* 5, no. 3: 126-159.
Clark, Melissa. 2003. "Education Reform, Redistribution, and Student Achievement: Evidence From the Kentucky Education Reform Act." (mimeo) Mathematica Policy Research (October).
Coleman, James S., Ernest Q. Campbell, Carol J. Hobson, James McPartland, Alexander M. Mood, Frederic D. Weinfeld, and Robert L. York. 1966. *Equality of educational opportunity*. Washington, D.C.: U.S. Government Printing Office.
Dillon, Eleanor, Lois Miller, and Jeffrey Smith. 2023. "Quantifying Non-Sampling Variation: College Quality and the Garden of Forking Paths." Unpublished manuscript (September).
Dynarski, Susan, Aizat Nurshatayeva, Lindsay C. Page, and Judith Scott-Clayton. 2023. "Addressing nonfinancial barriers to college access and success: Evidence and policy implications." In *Handbook of the Economics of Education*, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann: Elsevier.
Dynarski, Susan, Lindsay Page, and Judith Scott-Clayton. 2023. "Financial Aid for College Students." In *Handbook of the Economics of Education, Vol. 7*, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann: Elsevier.
Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication bias in the social sciences: Unlocking the file drawer." *Science* 345, no. 6203 (August 28): 1502-1505.
Gelman, Andrew, and Eric Loken. 2014. "The Statistical Crisis in Science." *American Scientist* 102, no. 6 (November-December): 460.
Gigliotti, Philip, and Lucy C. Sorensen. 2018. "Educational resources and student achievement: Evidence from the Save Harmless provision in New York State." *Economics of Education Review* 66: 167-182.
Goldstein, Jessica, and Josh B. McGee. 2020. "Did Spending Cuts During the Great Recession Really Cause Student Outcomes to Decline?" EdWorkingPaper No. 20-303. Brown University: Annenberg (October).
Guryan, Jonathan. 2001. "Does Money Matter? Regression-Discontinuity Estimates from Education Finance Reform in Massachusetts." NBER Working Paper WP 8269. Cambridge, MA: National Bureau of Economic Research (May).
Halloran, Clare, Rebecca Jack, James C. Okun, and Emily Oster. 2021. "Pandemic Schooling Mode and Student Test Scores: Evidence from US States." *National Bureau of Economic Research Working Paper Series* No. 29497.
Handel, Danielle V., and Eric A. Hanushek. 2023. "U.S. School Finance: Resources and Outcomes." In *Handbook of the Economics of Education. Volume 7*, edited by Eric A Hanushek, Stephen Machin, and Ludger Woessmann. Amsterdam: North Holland.
Hanushek, Eric A. 1979. "Conceptual and empirical issues in the estimation of educational production functions." *Journal of Human Resources* 14, no. 3 (Summer): 351-388.
Hanushek, Eric A. 2003. "The failure of input-based schooling policies." *Economic Journal* 113, no. 485 (February): F64-F98.

Hanushek, Eric A., and Matthew Joyce-Wirtz. 2023. "Incidence and Outcomes of School Finance Litigation: 1968-2021." *Public Finance Review*.

Hanushek, Eric A., and Ludger Woessmann. 2020. *The Economic Impacts of Learning Losses*. Paris: OECD (September).

Head , Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. 2015. "The Extent and Consequences of P-Hacking in Science." *PlosBiology* 13, no. 3 (March 13).

Ho, Andrew Dean. 2008. "The Problem With "Proficiency": Limitations of Statistics and Policy Under No Child Left Behind." *Educational Researcher* 37, no. 6 (August 1, 2008): 351-360.

Holland, Paul W. 2002. "Two Measures of Change in the Gaps Between the CDFs of Test-Score Distributions." *Journal of Educational and Behavioral Statistics* 27, no. 1: 3-17.

Huntington-Klein, Nick, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, Paul Grieco, Godwin Ekpe, Todd Pugatch, Martin Saavedra, and Yaniv Stopnitzky. 2021. "The influence of hidden researcher decisions in applied microeconomics." *Economic Inquiry* 59, no. 3: 944-960.

Hyman, Joshua. 2017. "Does Money Matter in the Long Run? Effects of School Spending on Educational Attainment." *American Economic Journal: Economic Policy* 9, no. 4 (November): 256-80.

Ioannidis, John P. A., T. D. Stanley, and Hristos Doucouliagos. 2017. "The Power of Bias in Economics Research." *The Economic Journal* 127, no. 605 (October): F236-F265.

Jackson, C. Kirabo, Rucker C. Johnson, and Claudia Persico. 2016. "The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms." *Quarterly Journal of Economics* 131, no. 1 (February): 157-218.

Jackson, C. Kirabo, Cora Wigger, and Heyu Xiong. 2021. "Do School Spending Cuts Matter? Evidence from the Great Recession." *American Economic Journal: Economic Policy* 13, no. 2 (May): 304-335.

Jepsen, Christopher, and Steven Rivkin. 2002. *Class Size Reduction, Teacher Quality, and Academic Achievement in California Public Elementary Schools*. San Francisco: Public Policy Institute of California.

Jepsen, Christopher, and Steven Rivkin. 2009. "Class Size Reduction and Student Achievement: The Potential Tradeoff between Teacher Quality and Class Size." *The Journal of Human Resources* 44, no. 1 (2009): 223-250.

Johnson, Rucker C. 2015. "Follow the Money: School Spending from Title I to Adult Earnings." *RSF: The Russell Sage Foundation Journal of the Social Sciences* 1, no. 3 (December): 50-76.

Kranz, Sebastian, and Peter Pütz. 2022. "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics: Comment." *American Economic Review* 112, no. 9 (September): 3124-36.

Kreisman, Daniel, and Matthew P. Steinberg. 2019. "The effect of increased funding on student achievement: Evidence from Texas's small district adjustment." *Journal of Public Economics* 176: 118-141.

Kuhfeld, Megan, James Soland, and Karyn Lewis. 2022. "Test Score Patterns Across Three COVID-19-impacted School Years." EdWorkingPapers 22-521. Brown University: Annenberg.

Lafortune, Julien, Jesse Rothstein, and Diane Whitmore Schanzenbach. 2018a. "School finance reform and the distribution of student achievement." *American Economic Journal: Applied Economics* 10, no. 2: 1-26.

Lafortune, Julien, Jesse Rothstein, and Diane Whitmore Schanzenbach. 2018b. "School Finance Reform and the Distribution of Student Achievement: Online Appendix." *American Economic Journal: Applied Economics* 10, no. 2: 1-26.

Lee, Kyung-Gon, and Solomon W. Polachek. 2018. "Do school budgets matter? The effect of budget referenda on student dropout rates." *Education Economics* 26, no. 2 (March): 129-144.

McGee, Josh B. 2023a. "Researchers should be cautious when generalizing findings." *Journal of Policy Analysis and Management*.

McGee, Josh B. 2023b. "Yes, money matters, but the details can make all the difference." *Journal of Policy Analysis and Management*.

Miller, Corbin L. 2018. "The Effect of Education Spending on Student Achievement: Evidence from Property Values and School Finance Rules." *Proceedings. Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association* 111: 1-121.

Nissen, Silas Boye, Tali Magidson, Kevin Gross, and Carl T. Bergstrom. 2016. Publication bias and the canonization of false facts. *eLife*, December, e21451.

Oreopoulos, Philip. 2007. "Do Dropouts Drop Out Too Soon? Wealth, Health and Happiness from Compulsory Schooling." *Journal of Public Economics* 91, no. 11-12: 2213-2229.

Page, Lindsay C., and Judith Scott-Clayton. 2016. "Improving college access in the United States: Barriers and policy responses." *Economics of Education Review* 51: 4-22.

Pritchett, Lant, and Justin Sandefur. 2015. "Learning from Experiments When Context Matters." *American Economic Review* 105, no. 5: 471-75.

Rauscher, Emily. 2020. "Does Money Matter More in the Country? Education Funding Reductions and Achievement in Kansas, 2010-2018." *AERA Open* 6, no. 4 (Oct).

Rebell, Michael A. 2019. *Courts and kids: Pursuing educational equity through the state courts: 2019 Supplement*. Chicago.

Roth, Jonathan, Pedro H. C. Sant'Anna, Alyssa Bilinski, and John Poe. 2023. "What's trending in difference-in-differences? A synthesis of the recent econometrics literature." *Journal of Econometrics* 235, no. 2 (August): 2218-2244.

Rothstein, Jesse, and Diane Whitmore Schanzenbach. 2022. "Does Money Still Matter? Attainment and Earnings Effects of Post-1990 School Finance Reforms." *Journal of Labor Economics* 40, no. S1: S141-S178.

Silberzahn, R., and et al. 2018. "Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results." *Advances in Methods and Practices in Psychological Science* 1, no. 3: 337-356.

Stecher, Brian M., and George W. Bohrnstedt. 1999. *Class size reduction in California: Early Evaluation Findings, 1996-98*. Palo Alto: American Institutes for Research.

Weinstein, Meryle G., Leanna Stiefel, Amy Ellen Schwartz, and Luis Chalico. 2009. "Does Title I Increase Spending and Improve Performance? Evidence from New York City." Working Paper #09-09. New York: Institute for Education and Social Policy (August).

Appendix A. Studies Included in This Analysis

Studies meeting the selection criteria and employed in the analysis are described in detail in Handel and Hanushek (2023). Note that some studies include analysis of more than one outcome measures. The included studies:

Test score outcomes

Abott, Kogan, Lavertu, and Peskowitz (2020), Baron (2022), Baron, Hyman, and Vasquez (2022), Brunner, Hyman, and Ju (2020), Buerger, Lee, and Singleton (2021), Carlson and Lavertu (2018), Clark (2003), Gigliotti and Sorensen (2018), Guryan (2001), Jackson, Wigger, and Xiong (2021), Kreisman and Steinberg (2019), Lafortune, Rothstein, and Schanzenbach (2018a), Miller (2018), Rauscher (2020), Weinstein, Stiefel, Schwartz, and Chalico (2009)

School attainment

Abott, Kogan, Lavertu, and Peskowitz (2020), Baron (2022), Baron, Hyman, and Vasquez (2022), Candelaria and Shores (2019), Cascio, Gordon, and Reber (2013), Hyman (2017), Jackson, Johnson, and Persico (2016), Jackson, Wigger, and Xiong (2021), Johnson (2015), Kreisman and Steinberg (2019), Lee and Polachek (2018), Miller (2018), Rothstein and Schanzenbach (2022)

Table 1: Distribution of standardized school spending estimates

Outcome	Median	Min	Max	N	N pos.	N Significant
Test scores	0.070	-0.244	0.543	16	14	9
Pass rates	0.056	0.054	0.059	2	2	2
Attainment	0.057	0.011	0.850	18	18	14

Notes: The estimates presented here have been scaled by the authors as detailed in Section 5 and Appendix Table 2 (Handel and Hanushek (2023)) for the sake of reporting estimates in comparable terms. For test score estimates, results represent the effect of a 10% increase in spending on the change in test scores (in individual standard deviation units). For pass rates and all attainment outcomes, results represent the percent change in the outcome variable for a 10% increase in spending. For example, an estimate of 0.05 for graduation indicates that a 10% increase in spending led to a 5% increase in graduation rates. Estimates are significant if $p < 0.05$.

Table 2: Overall School Spending Meta-analysis

Outcome	N	MD	95% CI	p-value	I_2
Test scores	16	0.0647	[0.0394; 0.0900]	< 0.0001	50.5%
Attainment	18	0.0550	[0.0225; 0.0875]	0.0024	77.6%

Notes: This table presents the meta-analytic summary of effect sizes for studies covering the effect of school spending on test scores and education attainment in the U.S. The summary effect is computed using a random effects model with inverse variance weighting.

Table 3. Sample distribution for author-identified SES impacts

SES level	Median	Min	Max	N	N Significant
Panel A: Test scores (N=9)					
Low SES	0.069	0.005	0.354	5	3
High SES	0.046	0.021	0.054	4	1
Panel B: Attainment (N=10)					
Low SES	0.123	0.007	0.372	6	4
High SES	0.044	0.029	0.094	4	1

Notes: The estimates presented here have been scaled by the authors as detailed in Section 5 and Appendix Table 2 (Handel and Hanushek, forthcoming) for the sake of reporting estimates in comparable terms. For test score estimates, results represent the effect of a 10% increase in spending on the change in test scores (in individual standard deviation units). For pass rates and all attainment outcomes, results represent the percent change in the outcome variable for a 10% increase in spending. For example, an estimate of 0.05 for graduation indicates that a 10% increase in spending led to a 5% increase in graduation rates. Estimates are significant if $p < 0.05$. Samples marked as ”low SES” or ”high SES” are categorized as such by study authors using various measures of poverty at the student- and district-level as discussed in Section 7.3.

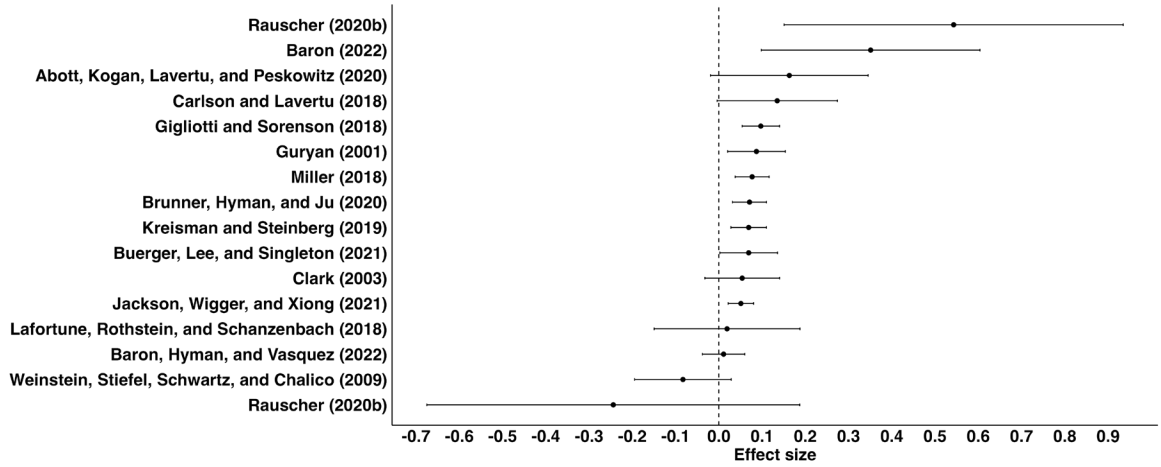

Figure 1: Effects of school spending on test scores, US

Notes: The estimates presented here have been scaled by the authors as detailed in Section 5 and Appendix Table 2 (Handel and Hanushek, forthcoming) for the sake of reporting estimates in comparable terms. Point estimates represent the effect of a 10% increase in spending on the change in test scores (in individual standard deviation units). Bars represent the 95% confidence interval.

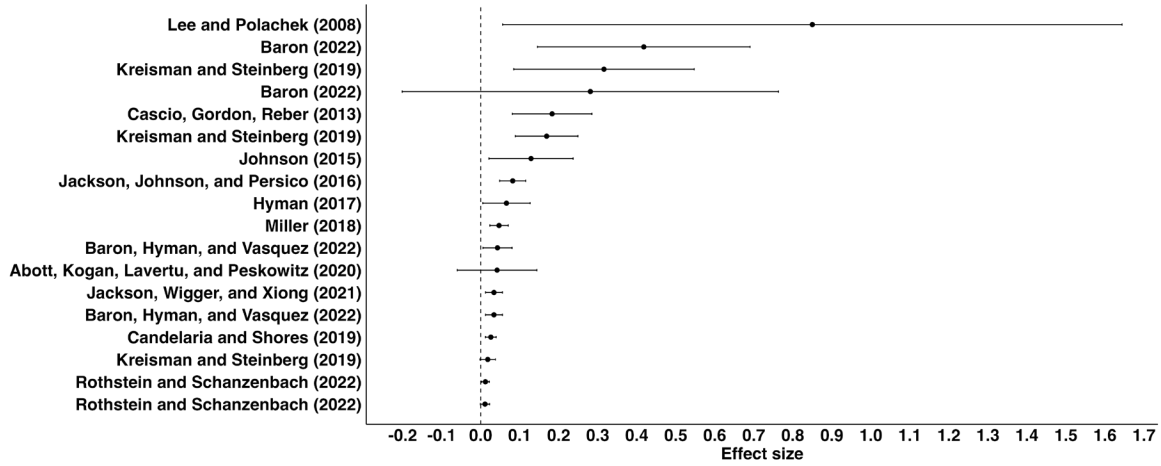

Figure 2: Effects of school spending on attainment

Notes: The estimates presented here have been scaled by the authors as detailed in Section 5 and Appendix Table 2 (Handel and Hanushek (2023)) for the sake of reporting estimates in comparable terms. The point estimates represent the percent change in the outcome variable for a 10% increase in spending. For example, an estimate of .05 for graduation indicates that a 10% increase in spending led to a 5% increase in graduation rates. Bars represent the 95% confidence interval.

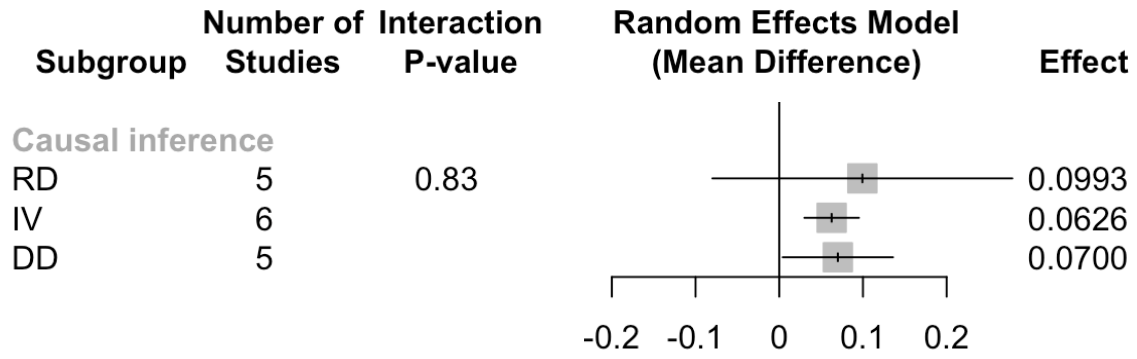

Figure 3: Differences in school spending and achievement impacts by estimation methods

Notes: This table presents the meta-analytic summary of effect sizes for studies covering the effect of school spending on test scores and compares summary effect sizes across policy types and motivations and whether studies covered sample within one state or across states in the U.S. The summary effects are computed using a random effects model with inverse variance weighting.

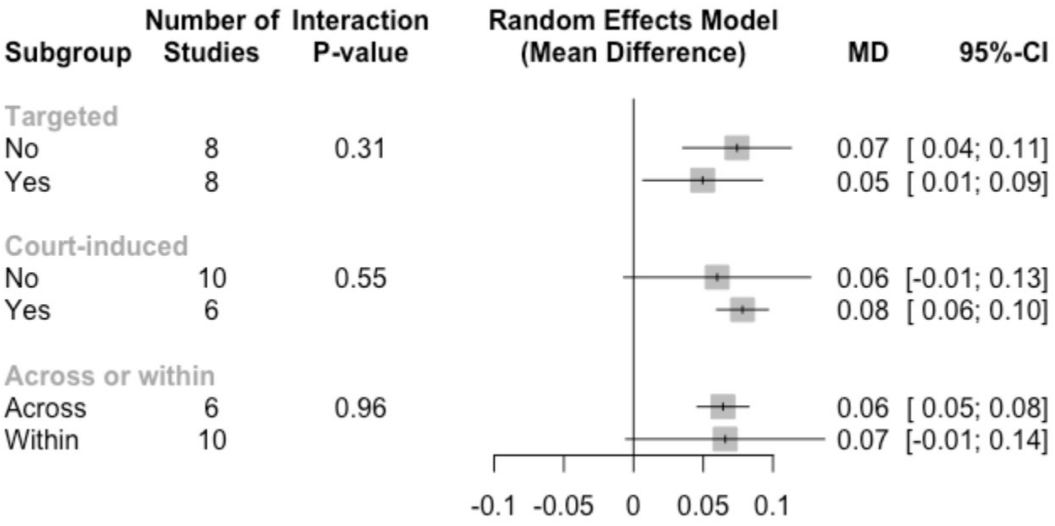

Figure 4: Differences in school spending and achievement impacts by context

Notes: This table presents the meta-analytic summary of effect sizes for studies covering the effect of school spending on test scores and compares summary effect sizes across policy types and motivations and whether studies covered sample within one state or across states in the U.S. The summary effects are computed using a random effects model with inverse variance weighting.

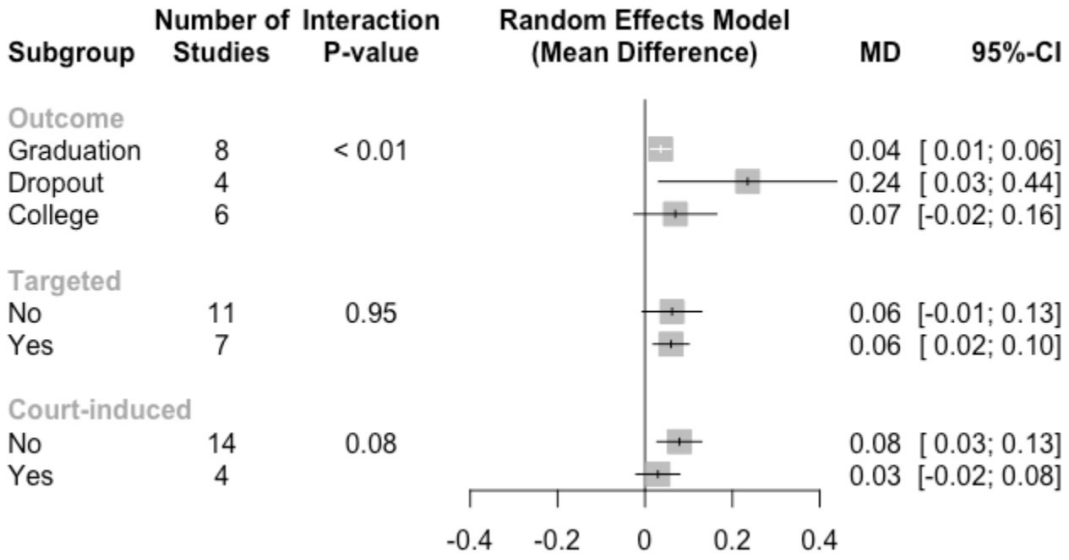

Figure 5: Differences in school spending and impacts by attainment level and context

Notes: This table presents the meta-analytic summary of effect sizes for studies covering the effect of school spending on educational attainment and compares summary effect sizes across measures of attainment and policy types and motivations. The summary effects are computed using a random effects model with inverse variance weighting.

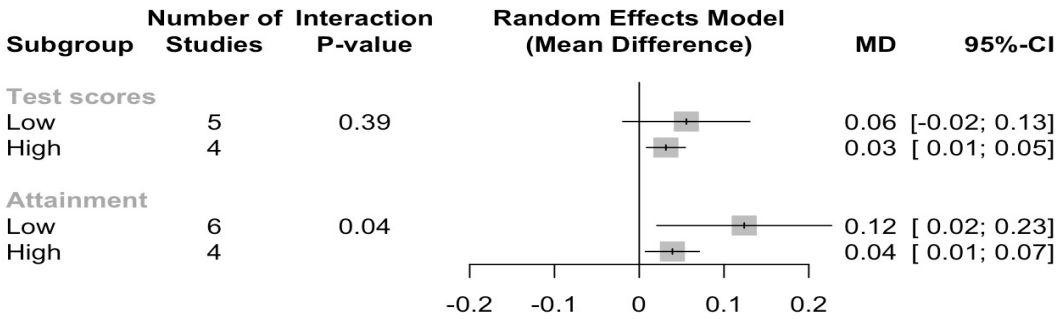

Figure 6: Differences in school spending and achievement impacts by SES group

Notes: This figure presents the meta-analytic summary of effect sizes for studies covering the effect of school spending on test scores and educational attainment and compares summary effect sizes across measures of low or high SES at the district- or student-level. The summary effects are computed using a random effects model with inverse variance weighting.