THE DISTRIBUTION OF SCHOOL SPENDING IMPACTS

C. Kirabo Jackson
Claire Mackevicius

WORKING PAPER 28517

NBER WORKING PAPER SERIES

THE DISTRIBUTION OF SCHOOL SPENDING IMPACTS

C. Kirabo Jackson
Claire Mackevicius

Working Paper 28517
http://www.nber.org/papers/w28517

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2021, Revised July 2021

This project was supported by the W.T. Grant Foundation. The statements made are solely the responsibility of the authors. We deeply appreciate feedback from Beth Tipton, James Pustejovsky, Jason Baron, and Sylvain Chabá-Ferret. We also thank cited authors who have verified that we accurately describe their work. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

The Distribution of School Spending Impacts
C. Kirabo Jackson and Claire Mackevicius
NBER Working Paper No. 28517
February 2021, Revised July 2021
JEL No. H0,I21,I26,J01,J58

ABSTRACT

We examine all known "credibly causal" studies to explore the distribution of the causal effects of public K-12 school spending on student outcomes in the United States. For each of the 31 included studies, we compute the same marginal spending effect parameter estimate. Precision-weighted method of moments estimates indicate that, on average, a $1000 increase in per-pupil public school spending (for four years) increases test scores by 0.0352 , high school graduation by 1.92 percentage points, and college-going by 2.65 percentage points. These pooled averages are significant at the 0.0001 level. The benefits to marginal capital spending increases take about five years to materialize, and are about half as large as (and less consistently positive than) those of non-capital-specific spending increases. The marginal spending impacts for all spending types are less pronounced for economically advantaged populations—though not statistically significantly so. Average impacts are similar across a wide range of baseline spending levels and geographic characteristics—providing little evidence of diminishing marginal returns at current spending levels.

To assuage concerns that pooled averages aggregate selection or confounding biases across studies, we use a meta-regression-based method that tests for, and removes, certain biases in the reported effects. This approach is straightforward and can remove biases in meta-analyses where the parameter of interest is a ratio, slope, or elasticity. We fail to reject that the meta-analytic averages are unbiased. Moreover, policies that generate larger increases in per-pupil spending tend to generate larger improvements in outcomes, in line with the pooled average.

To speak to generalizability, we estimate the variability across studies attributable to effect heterogeneity (as opposed to sampling variability). This heterogeneity explains between 76 and 88 percent of the variation across studies. Estimates of heterogeneity allow us to provide a range of likely policy impacts. Our estimates suggest that a policy that increases per-pupil spending for four years will improve test scores and/or educational attainment over 90 percent of the time. We find evidence of small possible publication bias among very imprecise studies, but show that any effects on our precision-weighted estimates are minimal.

C. Kirabo Jackson
Northwestern University
School of Education and Social Policy
Annenberg Hall, #204
2120 Campus Dr.
Evanston, IL 60208
and NBER
kirabo-jackson@northwestern.edu

Claire Mackevicius
Northwestern University
Evanston, IL 60208
cmackevicius@u.northwestern.edu

1 Introduction

For decades, social scientists have debated whether school spending affects student outcomes. This question is not just of academic importance, as public K–12 education is one of the largest single components of government spending (OECD, 2020). Additionally, current legal cases and policy decisions hinge on the extent to which, in what contexts, and how reliably increases in school spending causally impact students.[1] As such, understanding if, how much, and in what contexts increased school spending improves student outcomes is of considerable societal importance.

School spending impacts likely differ across studies due to differences in context, policy implementation, and treated populations. As a result, single estimates, *however well-identified*, may not meaningfully reflect the impacts of future policies in *other* contexts (DellaVigna and Linos (2021); Tipton et al. (2020); Vivalt (2020); Bandiera et al. (2021); Dehejia et al. (2021)). Without knowing the nature of heterogeneity across settings, there is no way to know how much the impacts of a particular study would generalize to a different setting (Tipton and Olsen (2018)). It is not the mere existence of heterogeneity that makes it difficult to generate policy predictions from existing studies, rather, the difficulty stems from a lack of understanding of that heterogeneity. Such an understanding can only be credibly obtained by examining impacts across several settings and contexts and among different populations.

Speaking to these issues, we perform a meta-analysis of all known "credibly causal" studies to quantify the averages and the spreads of the distributions of the causal effect of increased public K-12 school spending on test scores and educational attainment in the United States. This approach (a) generates pooled averages that are not driven by the particulars of any individual context or study, (b) provides greater statistical power than possible in any individual study to draw more precise conclusions, (c) facilitates more variability than available in individual studies to test new hypotheses, (d) allows one to measure and quantify treatment heterogeneity (i.e., the variability in impacts across studies not driven by sampling variability), which (e) facilitates the calculation of a plausible range of policy impacts that one may expect to observe in new settings. In sum, this approach allows us to provide several new insights.

Hanushek (2003) reviewed 163 studies published before 1995 that related school resources to student achievement. He documented more than ten times as many positive and significant studies than would be expected by random chance if spending had no impact, and almost four times as many positive and significant estimates than negative and significant – strong evidence of a positive association between school spending and student achievement in these older studies.[2] In a meta-analysis of these data, Hedges et al. (1994) concluded that "*it shows systematic positive relations*

[1] States with Supreme Courts cases challenging the funding of public schools in 2020 include Delaware, New York, Maryland, New Mexico, Illinois, and Tennessee.

[2] That is, Hanushek (2003) found that that 27 percent of these studies were positive and statistically significant and 7 percent negative and significant. By not considering the distribution of study impacts under the null hypothesis of no spending impact, Hanushek did not interpret this as evidence of a positive school spending effect. It is worth noting that the older studies may have overstated statistical significance irrespective of sign. However, a general overstatement of statistical significance would not explain the over-representation of positive estimates.

between resource inputs and school outcomes." However, these older studies were observational (based on partial correlations) and therefore unlikely to reflect causal impacts that are informative for policy (Hedges et al. (1994); Jackson (2020)). Understanding *causal* policy relationships requires an examination of studies that identify the *causal* impacts of school spending policies.

In the past decade there has been a notable increase in "credibly causal" papers using quasi-experimental variation (i.e., changes caused by specific identifiable policies) to identify the causal effect of school spending on student outcomes. However, there are sizable differences in reported impacts across studies, the reported effects are often noisy and not statistically significant, and nontrivial differences across studies (in reporting, measurement, policy context, etc.) make it difficult to directly compare one study's findings to another. Moreover, due to heterogeneity across studies, it is unclear what impact (or range of impacts) policymakers can expect from increasing school spending by a specific amount. We provide clarity on these points by using formal meta-analytic techniques on all "credibly causal" estimates that relate school spending changes to student outcomes. This analysis not only addresses the perennial question of "does money matter?" but it also quantifies, based on the best evidence available, (a) how much, *on average*, student outcomes would improve from a policy that increases school spending by $1000 per pupil sustained over four years, (b) how the marginal effects differ for non-capital and capital spending, (c) how the marginal effects differ for children from low- and non-low-income families, (d) whether marginal school spending impacts vary by baseline spending levels (i.e., if there are diminishing returns), (e) the extent to which estimates based on existing studies may generalize to other contexts, and (f) what range of policy impacts can be expected in a given context.

Conducting a rigorous meta-analysis involves several steps. First, we define the study inclusion criteria *ex ante*. To focus on credibly causal estimates, we require that the policy variation used in a study is a plausibly valid instrument (in the econometric sense) for school spending. We compile a list of all studies from the past 25 years that employ quasi-random or quasi-experimental variation in school spending and estimate impacts on student outcomes. Among these, we only include those studies that demonstrate that the variation used is plausibly exogenous (i.e., that the policy-induced changes in school spending are unrelated to other policies or other determinants of student outcomes) – this is analogous to the ignorability condition in an instrumental variables model.[3] We refer to this set of studies as "credibly causal." Because we are interested in the impacts of policies that change school spending, we focus on those "credibly causal" studies that demonstrate policy-induced variation in school spending. This second condition is analogous to the "instrument relevance" condition in an instrumental variables model.[4]

Meta-analysis is typically used to synthesize across randomized experiments where there is a well-defined treatment and estimate reporting is standard across studies. In contrast, school spend-

[3]This condition excludes all papers analyzed in well-known older literature reviews conducted in Hanushek (2003).

[4]Not all school spending *policies* actually change school spending (due to states or districts shifting other monies around in response to policy changes). For example, local areas may reduce property tax rates in response a state policy to provide additional money to some schools. In such a setting, the additional state funds are used for tax savings rather than being spent in the classroom. See Brunner et al. (2020) for an example of this kind of behaviour.

ing studies examine variation in spending based on a range of policies and they report effects on multiple outcomes in different ways. Direct comparison across studies requires that we define the treatment and the outcomes in the same way. To this end, we compute the same underlying empirical relationship for each study by capturing the estimated policy impact on K-12 per-pupil spending and the estimated impacts on outcomes. To make our estimates as comparable as possible, we (1) compute the average impacts for the full population (as opposed to particular sub-samples), (2) standardize all spending levels to 2018 CPI adjusted dollars, (3) convert all reported impacts into standardized effects, and (4), where possible, capture impacts four years after the spending change to keep student exposure to the spending increases consistent across studies. With these estimates, we compute the estimated effect of a $1000 per-pupil increase in school spending (over four years) on standardized educational outcomes. That is, for each paper and outcome, we construct an instrumental variables (IV) estimate of the marginal policy-induced impact on standardized outcomes of exposure to a $1000 per-pupil spending increase (CPI adjusted to 2018 dollars) over four years. Once summarized using the same relationship, studies that are reported in starkly different ways are remarkably similar – suggesting much less heterogeneity than one might expect at first blush.

Another innovation of our work is to propose a framework to compare the impacts of capital to non-capital spending. If school construction matters, a 40 million dollar construction project should affect student outcomes over the life of the building (about 50 years) and not just in the year the spending occurred. As such, a simple comparison of contemporaneous capital spending to contemporaneous outcomes would drastically understate the marginal impacts of capital spending on outcomes. To account for this, we amortize large one-time capital payments over the useful life of the capital asset. We then relate the change in outcomes to the present discounted "flow" value to obtain the marginal impacts of capital spending. This approach leads to annual spending increases comparable to those of non-capital spending increases.

Speaking first to the "*does money matter?*" question, we show that 90 percent of all included studies find a positive overall effect of increased school spending (irrespective of significance). If positive and negative impacts were equally likely (as is the case if school spending did not matter), the likelihood of observing this many positive estimates or more is less that one in 430 thousand. Next, we quantify the magnitude of the impact of increased school spending on student outcomes using formal meta-analysis. Some are skeptical of meta-analysis outside of randomized experiments because individual studies may vary considerably due to effect heterogeneity – making a naively pooled estimate difficult to interpret. However, rather than avoid direct comparison of studies because of heterogeneity, we seek to model and understand this heterogeneity to gain a deeper understanding of when, to what extent, and in what contexts, school spending affects student outcomes. To this aim, using the same relationship for each paper, we employ random effects meta-analysis that does not assume the existence of a single common effect, but rather explicitly estimates the extent of treatment effect heterogeneity across studies. This approach provides pooled average estimates that are robust to the inclusion of outlier and imprecise estimates, and produces a plausible range of predicted policy impacts informed by the variability both within and across

studies (see Meager (2019), Vivalt (2020), Bandiera et al. (2021), DellaVigna and Linos (2021), and Dehejia et al. (2021) for similar approaches).

The pooled meta-analytic estimate indicates that, *on average*, a $1000 per-pupil increase in school spending (sustained over four years) increases test scores by 0.0352σ - about 1.4 percentile points. We can reject that the pooled average is zero at the 0.0001 significance level. However, over 75 percent of the variability across studies reflects unobserved heterogeneity (due to different LATEs, policy contexts, etc.), so one may observe estimates well outside the confidence interval in new settings. Based on this information, a policy *in a different context* that increased per-pupil school spending by $1000 over a four-year period would have test score impacts between -0.007σ and 0.078σ ninety percent of the time, and would lead to positive test score impacts more than 91 percent of the time. Looking to educational attainment, our pooled meta-analytic estimate indicates that, *on average*, a $1000 per-pupil increase in school spending increases educational attainment by 0.0539σ (p-value$<$0.0001). This translates into a 1.92 percentage-point increase in high school graduation and a 2.65 percentage-point increase in the college-going rate.[5] In relative terms, this is a 2.3 percent increase in high school graduation and a 6.4 percent increase in college-going. Conservative estimates indicate that heterogeneity across the educational attainment studies may explain as much as 87% of the variability. Based on this estimate, a policy *in a different context* that increased per-pupil school spending by $1000 over a four-year period would be expected to have high-school graduation impacts between -0.2 and 4.1 percentage points and college-going impacts between -0.3 and 5.65 percentage points ninety percent of the time. Moreover, such a policy would lead to positive educational attainment impacts more than 92 percent of the time. We also examine the cumulative impacts on educational attainment and demonstrate that educational attainment impacts increase with years of exposure to a spending increase.

Another concern with meta-analysis of non-experimental studies is that a pooled average of individually biased studies may also be biased. We avoid this by focusing on credibly causal studies where individual biases should be small. Moreover, we show that by aggregating several studies, upward bias in some studies may cancel out downward bias in others – yielding an unbiased pooled average. However, the *possibility* of bias in our pooled average remains if the included studies all tend to suffer from a similar positive bias. To assuage this concern, we present a novel approach that relies on differences in the policy-induced changes in spending across studies to remove the influence of certain kinds of confounding bias in reported effects. We show that *even if all studies are biased upward* in ways that bias the meta-analytic average, our difference-based approach uncovers a bias-free average marginal impact as long as the size of the bias in a study is unrelated to the size of the policy effect on per-pupil spending. Using this test, we show that policies that generate larger increases in per-pupil spending also tend to generate larger improvements in outcomes, and we find no evidence of bias in our test score or educational attainment impacts. In addition, we show that

[5]These calculations multiply the standardized impact (0.0539σ) by the standard deviation of each outcome. The standard deviation of a binary variable is $\sqrt{p \times (1-p)}$, where p is the proportion of positive outcomes. We use a standard deviation of high-school graduation of 0.357 (rate = 0.85) and standard deviation of college-going of 0.492 (rate = 0.41) (Snyder et al. (2019).

the marginal spending impacts are similar for voluntarily adopted polices (where one may worry about selection and confounding) and those that are not – further evidence that problematic biases are unlikely among our set of included studies.

We benchmark our impacts against those of *other* well-known interventions with measurable effects on student outcomes (including class size reduction and "No Excuses" charter school attendance). Our pooled aggregate $1000 per-pupil spending estimates are on-par with these interventions. However, the benchmarked effects on educational attainment are consistently much larger than those on test scores – suggesting that test scores may not measure *all* of the benefits of increased school resources (Card and Krueger (1992); Krueger (1998)), and, more broadly, that test score impacts may only capture a portion of the overall benefits of educational inputs (Jackson (2018); Jackson et al. (2020)). We also examine observable predictors of differences in outcomes. Benefits to capital spending increases take a few years to materialize, and the average effects of increased capital spending on test scores are about half those of non-capital spending. And while we find that impacts of all spending types are quite stable along several observable dimensions (including urbanicity and geography), we *do* find that impacts are smaller, on average, for more economically advantaged populations.

While our results accurately describe the literature, the distribution of impacts reported may not reflect the true distribution of impacts if there is publication bias. Indeed, we find evidence of a relative paucity of imprecise negative point estimates – such that a *naively* estimated simple average (**which we do not employ**) might overstate the average marginal spending effects. To assess the extent to which potential publication bias impacts our precision-weighted pooled estimates, we implement several empirical approaches, including removing imprecise studies (which are more susceptible to biases) (Stanley et al. (2010)), employing the "trim and fill" method to impute potentially "missing studies" (Duval and Tweedie (2000)), adjusting for bias against non-significant effects (Andrews and Kasy (2019)), and implementing the precision-effect estimate with standard error (PEESE) approach (Stanley and Doucouliagos (2014)). In all cases, we find little evidence of significant impacts of possible publication bias. Additionally, we find no systematic differences between published and unpublished studies, or studies published in more or less selective outlets – further evidence of negligible impact of publication bias in our precision-weighted pooled averages.

Some have argued that while school spending may have mattered when baseline spending levels were low, the marginal impacts *may* be smaller at current spending levels (Jackson et al. (2016)). We test this with our data by examining whether the marginal spending impacts are larger in older studies (when national spending levels were lower) or in states with lower baseline spending levels (such as in the South). Precision-weighted models reveal that the marginal impacts are remarkably stable for a wide range of baseline per-pupil spending levels and across geographic contexts for both test scores and educational attainment. This pattern suggest that policy impacts at current spending levels are likely to be similar to those from the past (accounting for inflation).

This study moves beyond the question of whether money matters, and is the first to quantify the pooled average marginal impacts of an increase in per-pupil spending on student test scores and

educational attainment across studies. It is also the first study to measure and quantify the range of causal impacts supported by the existing literature. This allows us to measure the extent to which studies in this literature may estimate the same parameter, and then provide a plausible range of estimates one may observe in other contexts. We also show how one can compare the impacts of spending changes that affect students over different spans of time. Finally, we contribute to a small but growing literature (e.g., Hendren and Sprung-Keyser (2020)) showing how, by carefully computing the same parameter across studies, one can combine a variety of estimates outside of randomized controlled trials to provide new and important policy insights.

The remainder of this paper is as follows: Section 2 discuses how we identify and select the studies to create our dataset for analysis, Section 3 describes how we compute the same underlying parameter for each paper, Section 4 presents the formal meta-analytic methods, Section 5 presents our main results, Section 6 presents evidence of robustness to various assumptions and restrictions, Section 7 accounts for potential biases and shows negligible effect on our main results, Section 8 documents heterogeneity across population and study characteristics, and Section 9 concludes.

2 Data

We capture estimates from studies that employ quasi-experimental methods to examine the impacts of policy-induced changes (i.e., exogenous policy shocks) in K-12 per-pupil spending on student outcomes.[6] Our inclusion criteria requires that the variation in spending is driven by policy and (following best econometric practice) requires that the variation used in a study is plausibly a valid instrument for school spending. That is, the policy examined must lead to meaningful changes in per-pupil school spending (the treatment), and the variation used must be demonstrated to be plausibly exogenous to other determinants (i.e., non-school spending determinants) of student outcomes. Specifically, to be included, a study had to meet each of the following criteria:

1. The study relied on quasi-experimental or policy variation in school spending.[7] That is, the study used a quasi-experimental design (Regression Discontinuity, Event-Study, Instrumental Variables, or some combination of these) to isolate the impacts of *specific* school spending policy shocks (or features of a school spending policy) on student outcomes.

2. The study demonstrated that their analysis was based on policies (or policy-induced variation) that had a robust effect on school spending – enough to facilitate exploring the effect of school spending on student outcomes. That is, the study examined the effect of a particular policy that altered school spending or relied on an identifiable change in school spending caused by

[6] The two authors independently verified data captured from each study.

[7] Some well-known studies are excluded based on this criterion. For example, Husted and Kenny (2000) does not rely on an identifiable change in school spending due to a policy. As they state "Our preferred resource equalization measure... equals the change in resource inequality since 1972 relative to the predicted change (that is, the unexplained change in inequality). A fall in this variable reflects either the adoption of state policies that have reduced districts' ability to determine how much to spend in their district or an otherwise unmeasured drop in spending inequality" (298).

a specific policy. We included studies that showed a non-zero effect on spending at least at the 5 percent significance level.[8] However, we show that our results are robust to including only studies for which the effect on spending is significant at the 0.0000001 level (see Table A.1). We excluded studies of policies that did not demonstrate effects on school spending (as they are, by definition, uninformative of the effects of school *spending* on outcomes).

3. The study demonstrated that the variation in school spending examined was unrelated to other determinants of student outcomes, including other policies, student demographics, or underlying trend differences. *That is, the study maintains and provides evidence that the policy effect is only due to its effect on school spending.* This entails a formal test that the policy instrument is unrelated to other reasonable predictors of outcomes (such a demographics, socioeconomic measures, or other policies) and (if required) directly accounts for any potential imbalance.[9] That is, we include studies that provide evidence that (after including an appropriate set of controls) they make comparisons across entities with different levels of school spending but for which on average all else was equal.[10]

To locate studies that meet this inclusion criteria, we searched for all papers on the topic published or made public since 1995. We do not look before 1995 because, based on an initial search, no studies that meet the inclusion criteria existed before 1995.[11] Empirical practices in this literature were not focused on causal estimation until the early 2000s (see Angrist and Pischke (2010) for a discussion of the "*credibility revolution*" in empirical economics). Indeed, the earliest "credibly

[8]This corresponds to a first stage F-statistic of 3.85 for the policy instruments on per-pupil school spending. In a two-stage-least-squares (2SLS) framework, the typical threshold would be a first stage F-statistic of 10. We impose a weaker restriction. Still, some well-known studies are excluded based on this criterion. Specifically, van der Klaauw (2008) states that Title I "eligibility does not necessarily lead to a statistically significant increase in average per pupil expenditures" (750). Similarly, Matsudaira et al. (2012) do not find a robust association between the policy (Title I eligibility) and per-pupil spending. Some studies examine the effects of policies that influence school spending, but they do not report the effect of the policies on school spending in a way that allows us to construct a first-stage F-statistic. These include Downes et al. (1998), Figlio (1997), Hoxby (2001) and, more recently, Holden (2016). Given its prominence, we discuss Hoxby (2001) in more detail: Hoxby (2001) reports that *some* key policy parameters (such as the inverted tax price) do predict differences in school spending but that others do not (such as the income/sales tax rate in support of school spending, which has a t-statistic smaller than 1 in predicting per-pupil spending). In a 2SLS model, all the policy variables (including the weak predictors) are used and no first stage F-statistic is reported. As such, because a strong first stage is not demonstrated, the 2SLS model predicting spending effects on dropout rates does not satisfy our inclusion criteria. Having said this, two policy variables are *individually* significant at the 5 percent level in most first stage regressions (inverted tax price and the flat grant/median income). In reduced form models, both these variables individually indicate that increased school spending reduces dropout rates. As Hoxby concludes, "*while the estimated effects of equalization on student achievement are generally weak, it does appear that the drop-out rate falls in districts that are constrained to raise spending by the imposition of a per-pupil spending floor*" (p. 1229).

[9]For all models, this would include testing that the policy instrument is unrelated to observable predictors of the outcomes. In addition, for Diff-in-Diff studies they would also show no evidence of differential pre-trending or would include entity-specific time trends. For RD models this would also involve showing that there is smoothness in covariates through the cutoff point.

[10]Note that the seminal Hoxby (2001) paper is primarily focused on the effect of reform type on school spending. The additional analysis of the effect on student outcomes is not main focus of the paper, and explicit tests for bias were not conducted. As such, this important paper in the literature does not meet this component of our inclusion criteria for this particular analysis.

[11]Note that Hedges et al. (1994) find that "*most of the studies in Hanushek's data set are cross-sectional rather than longitudinal*" (12) – that is, relying on simple comparisons across locations or entities at a single point in time.

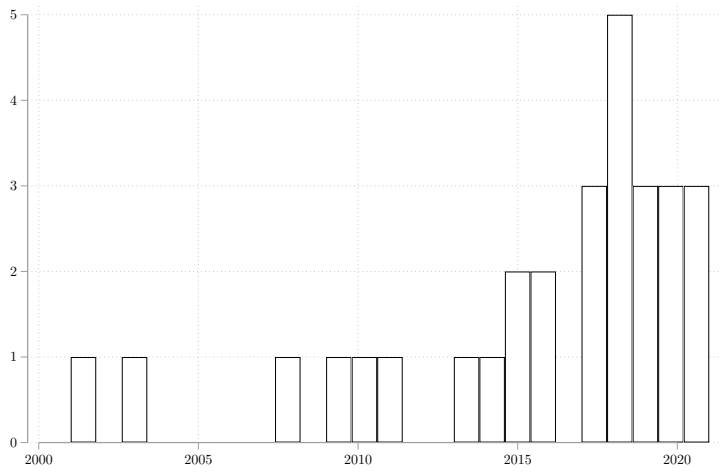


Figure 1: Count of Included Studies per Year

causal" study we located was published in 2001, the majority of studies meeting this criteria were published after 2010, and the lion's share were written or published after 2015 (Figure 1). We compiled this list by starting with all studies cited in the Jackson (2020) literature review. We then supplemented this with Google Scholar searches on relevant terms (including "school spending" and "causal"). We then consulted the cited papers and all papers that cite those on our list to identify other papers to possibly include. Finally, to avoid exclusion of unpublished papers or works in progress, we asked active researchers in the field to locate any additional papers beyond the list we compiled.[12] Using this approach, we identified 31 studies that met our conditions as of December 1, 2020. Where there are multiple versions of the same paper (e.g., a working paper and a published version) we use the most recent publicly-available version of each paper.[13]

2.1 Included Studies

Table 1 summarizes the 31 studies that satisfy the inclusion criteria.[14] We list the last names of the authors and the publication (or draft) year of each study (first column). We assign a unique Study ID to each of the 31 included studies (second column). Because we examine the impacts of school spending on different outcomes (test scores, educational attainment, and longer-run outcomes), we include multiple entries for studies that present impacts on multiple outcomes.[15] While we examine the sign of the impacts for all studies meeting the inclusion criteria, we only capture the estimated impacts on test scores and educational attainment for analyses that quantify the

[12] This was done using a broad appeal on Twitter to a large network of economists and education-policy scholars.

[13] When studies were updated (which happened with unpublished work) we updated our database to reflect the most up-to-date version of the paper's analysis.

[14] Given the use of the same data and identification strategy, to avoid double counting we categorize Jackson et al. (2016) and Johnson and Jackson (2019) as representing a single study.

[15] Note: Baron (2021) is the only study that reports distinct effects of both non-capital and capital spending, and in this table we report the average of the test score estimates across the two spending types. For our analyses that distinguish across spending types, we include both test score estimates separately.

relationship between spending and outcomes; there are too few studies of *other* outcomes to provide credible pooled estimates. As such, we provide a study-outcome Observation ID (third column) for test score and educational attainment outcomes only. Of 31 unique studies, 24 present estimates of test score impacts (either test scores or proficiency rates), 12 present estimates of impacts on educational attainment (high school dropout, high school graduation, or college enrollment), and 3 examine longer-run impacts (wages or income mobility).[16]

While some prominent studies have employed event-study designs to examine effects on school finance reforms (including Jackson et al. (2016), Lafortune and Schonholzer (2019), and others), the set of studies examined represent a diversity of estimation strategies and sources of variation (see Table 2). Of the 31 studies, 12 employ event-study type designs based on the implementation of a policy, 11 employ regression discontinuity designs to exploit sudden jumps in per-pupil spending at some known threshold, and 8 employ instrumental variables strategies to exploit changes in is per-pupil spending level embedded in school finance formulas that are unrelated to other policy decisions. These paper are also quite varied in terms of policies examined. There are 6 papers that examine school finance reforms nationally, 7 that examine particular state level school finance reforms, 3 that examine school spending referendum, 4 look at school improvement grants, 9 look at capital construction projects, and others identifying effects of Title I or impacts of economic shocks on spending. There is also good geographic coverage in terms of the populations treated in the included studies. The school finance reforms studies cover all districts in treated states (about 25 out of 51), several examine large urban school districts, and one paper (Kreisman and Steinberg, 2019) examines small school grants (focused on sparse, rural areas). Overall, the treated "super population" covered by these studies is largely representative of the United States as a whole. The fact that the pooled data cover a variety of setting makes this study well suited to measure the extent of heterogeneity one may expect across settings and to test for heterogeneous effects across settings (see Section 8).

Table 1 also reports the results of calculations to construct comparable estimates from each paper (detailed in Section 3). For each study-outcome combination, we report the sign of the relationship between school spending and the average student outcome for the study's full sample. Positive values indicate a positive association between school spending and *improved* outcomes.[17] For test score effects, we report the sign of the impact on the *average test scores across all subjects and grade levels* reported in the study. We also report, for each study-outcome Observation ID, the estimated marginal impact of a $1000 per-pupil spending increase (in 2018 dollars) sustained over four years on the standardized outcome (Effect) and its standard error (SE). The last columns report whether the average marginal effect is statistically significant at the 5 and 10 percent levels. Table 2 presents details on the estimation strategy and spending type examined in each study.

[16] One study (Card and Payne (2002)) examines test score inequality, which is not directly comparable to other studies and therefore not included in the formal meta analysis but is included as a "credibly causal" study on the topic in the coin test analysis.

[17] For example, Lee and Polachek (2018) and Cascio et al. (2013) examine impacts on dropout rates. The reported effects are reverse-coded so that a positive coefficient indicates improved outcomes (in these cases, reduced dropout).

Table 1: Summary of Studies

Study	Study ID	Obs ID	Outcome(s)	Sign	Effect	SE	5%	10%
Abott Kogan Lavertu Peskowitz (2020)	1	1	High school graduation	pos	0.0847	0.0876		
Abott Kogan Lavertu Peskowitz (2020)	1	2	Test scores	pos	0.1158	0.0667		*
Baron (2021)	2	3	College enrollment	pos	0.1869	0.0767	**	*
Baron (2021)	2	4	Test scores	neg	-0.0049	0.0877		
Biasi (2019)	3	.	Income mobility	pos	.	.		
Brunner Hyman Ju (2020)	4	5	Test scores	pos	0.0531	0.0173	**	*
Candelaria Shores (2019)	5	6	High school graduation	pos	0.0511	0.0133	**	*
Card Payne (2002)	6	.	Test score gaps	pos	.	.		
Carlson Lavertu (2018)	7	7	Test scores	pos	0.0902	0.0475		*
Cascio Gordon Reber (2013)	8	8	High school dropout	pos	0.5546	0.2056	**	*
Cellini Ferreira Rothstein (2010)	9	9	Test scores	pos	0.2120	0.0992	**	*
Clark (2003)	10	10	Test scores	pos	0.0148	0.0116		
Conlin Thompson (2017)	11	11	Test proficiency rates	pos	0.0084	0.0062		
Gigliotti Sorensen (2018)	12	12	Test scores	pos	0.0424	0.0098	**	*
Goncalves (2015)	13	13	Test proficiency rates	neg	-0.0019	0.0227		
Guryan (2001)	14	14	Test scores	pos	0.0281	0.0689		
Hong Zimmer (2016)	15	15	Test proficiency rates	pos	0.1159	0.0652		*
Hyman (2017)	16	16	College enrollment	pos	0.0552	0.0257	**	*
Jackson Johnson Persico (2015), Jackson Johnson (2019)	17	17	High school graduation	pos	0.0798	0.0163	**	*
Jackson Johnson Persico (2015), Jackson Johnson (2019)	17	.	Years of education, Wages, Poverty	pos	.	.		
Jackson Wigger Xiong (2021)	18	18	College enrollment	pos	0.0380	0.0133	**	*
Jackson Wigger Xiong (2021)	18	19	Test scores	pos	0.0499	0.0196	**	*
Johnson (2015)	19	20	High school graduation	pos	0.1438	0.0753		*
Johnson (2015)	19	.	Wages, Poverty	pos	.	.		
Kogan Lavertu Peskowitz (2017)	20	21	Test scores	pos	0.0190	0.0127		
Kreisman Steinberg (2019)	21	22	High school graduation	pos	0.0279	0.0146		*
Kreisman Steinberg (2019)	21	23	Test scores	pos	0.0779	0.0237	**	*
Lafortune Rothstein Schanzenbach (2018)	22	24	Test scores	pos	0.0164	0.0133		
Lafortune Schonholzer (2021)	23	25	Test scores	pos	0.2330	0.1032	**	*
Lee Polachek (2018)	24	26	High school dropout	pos	0.0640	0.0141	**	*
Martorell Stange McFarlin (2016)	25	27	Test scores	pos	0.0304	0.0270		
Miller (2018)	26	28	High school graduation	pos	0.0662	0.0169	**	*
Miller (2018)	26	29	Test scores	pos	0.0515	0.0137	**	*
Neilson Zimmerman (2014)	27	30	Test scores	pos	0.0248	0.0187		
Papke (2008)	28	31	Test proficiency rates	pos	0.0817	0.0121	**	*
Rauscher (2020)	29	32	Test scores	pos	0.0083	0.0049		*
Roy (2011)	30	33	Test scores	pos	0.3804	0.1563	**	*
Weinstein Stiefel Schwartz Chalico (2009)	31	34	High school graduation	pos	0.1595	0.1698		
Weinstein Stiefel Schwartz Chalico (2009)	31	35	Test scores	neg	-0.0541	0.0368		

Table 2: Summary of Estimation Strategy

Study	Est. Strategy	Spending Type
Abott Kogan Lavertu Peskowitz (2020)	Regression Discontinuity	operational
Baron (2021)	Regression Discontinuity	capital
Baron (2021)	Regression Discontinuity	operational
Biasi (2019)	Event Study	Any
Brunner Hyman Ju (2020)	Event Study DiD	Any
Candelaria Shores (2019)	Event-Study DiD	Any
Card Payne (2002)	Difference in Difference	Any
Carlson Lavertu (2018)	Regression Discontinuity	School Improvement Grant
Cascio Gordon Reber (2013)	Event Study	Title I
Cellini Ferreira Rothstein (2010)	Regression Discontinuity	capital
Clark (2003)	Event-Study DiD	Any
Conlin Thompson (2017)	Event Study	capital
Gigliotti Sorensen (2018)	Instrumental Variables	Any
Goncalves (2015)	Event Study	capital
Guryan (2001)	Instrumental Variables	Any
Hong Zimmer (2016)	Regression Discontinuity	capital
Hyman (2017)	Instrumental Variables	Any
Jackson Johnson Persico (2015), Jackson Johnson (2019)	Event-Study DiD	Any
Jackson Wigger Xiong (2021)	Instrumental Variables	Any
Johnson (2015)	Event-Study DiD	Title I
Kogan Lavertu Peskowitz (2017)	Regression Discontinuity	Any
Kreisman Steinberg (2019)	Instrumental Variables	Any
Lafortune Rothstein Schanzenbach (2018)	Event-Study DiD	Any
Lafortune Schonholzer (2021)	Event-Study DiD	capital
Lee Polachek (2018)	Regression Discontinuity	Any
Martorell Stange McFarlin (2016)	Regression Discontinuity	capital
Miller (2018)	Instrumental Variables	Any
Neilson Zimmerman (2014)	Event-Study DiD	capital
Papke (2008)	Instrumental Variables	Any
Rauscher (2020)	Regression Discontinuity	capital
Roy (2011)	Instrumental Variables	SFR
Weinstein Stiefel Schwartz Chalico (2009)	Regression Discontinuity	Title I

We assign a single, primary estimation strategy for each paper. Regression Discontinuity studies are those whose identification is dependent on a cutoff point for some running variable. Event Study studies are those whose identification strategy is driven by a policy or rollout over time. Instrumental Variables studies are those whose identification is driven by a change that occured conditional on a policy, but not RD.

3 Constructing The Same Parameter Estimate for All Papers

To assess a literature, one must compare studies to each other. However, unlike randomized control trials, studies on school spending policies are not reported in ways that facilitate direct comparison. For example, Lafortune et al. (2018) report the impacts (after ten years) of school finance reforms on 4^{th} and 8^{th} grade test-score gaps between high- and low income districts. In contrast, Hong and Zimmer (2016) report the impacts of passing a bond referenda on test proficiency rates 1 through 13 years after bond passage in 4^{th} and 7^{th} grades. While both studies report positive school spending impacts, the time frames are different, the time periods are different, the size of the spending increases are different, one reports relative changes while the other reports absolute changes, and

one reports impacts on standardized test scores while the other reports on proficiency rates. It is unclear which study implies larger marginal school spending impacts – or even how similar the study impacts are. Because studies report effects in different ways and on different scales, or define school spending differently, we extract information from each paper that allows us to standardize estimates for comparability across papers.[18]

3.1 The Common Parameter Estimate

For each study, we compute the effect of a $1000 per-pupil spending increase (in 2018 dollars), sustained for four years, on standardized outcomes for the full population affected by the policy. We compute separate estimates for test scores and educational attainment outcomes. We detail how we compute this empirical relationship (or parameter estimate) for each study. Because studies do not all report impacts in this form, this often requires several steps. We lay out these step and any additional required assumptions in the following subsections. We show that none of these assumptions change our final conclusions in any appreciable way (see Sections 6 and 7).

Step 1: Choice of outcomes

We report effects on student achievement (measured by test scores or proficiency rates) and educational attainment (measured by dropout rates, high school graduation, or college (postsecondary) enrollment). If multiple test score outcomes are reported (e.g., proficiency rates and raw scores) we use the impacts on raw scores. This allows for standardized test score effects that are more comparable across studies, and avoids comparing impacts across thresholds of differing difficulty (i.e., where some areas have higher proficiency standards than others).[19] For educational attainment outcomes, we capture impacts on high-school completion measures and college enrollment. For studies that report multiple of these measures, we take the highest level reported.[20]

Step 2: Computing Population Average Treatment Effects

For much of our analysis, we use one estimate per outcome per study. When studies report estimates for multiple specifications, we capture estimates from the authors' preferred specification. When there is a reported overall estimate across all populations (e.g., high-income and low-income), all subjects (e.g., Math and English), and all grade levels (e.g., 8th grade and 4th grade), we take the overall estimate as reported in the study. When studies report effects by subject, grade level, or population, we combine across estimates to generate an overall estimate and standard error for

[18] We detail the information we capture from each paper in Table A.2.

[19] In one case, Kogan et al. (2017), multiple raw score effects were reported. We took the estimates for the preferred outcome indicated by the authors.

[20] For example, if effects are reported for college enrollment and high school graduation, we take the college enrollment effects. If effects are reported for high school graduation and high school dropout, we take the high-school graduation effects. This particular decision rule of taking graduation over dropout outcomes is further justified because: (a) dropout rates are notoriously difficult to measure (Tyler and Lofstrom (2009)) and therefore a less reliable measure of educational attainment, and (b) different entities often measure dropout rates is very different ways.

analysis.[21] When we combine test score effects across subjects for the same grade, we assume these stem from the same population and use the simple average as our overall effect.[22] [23] We combine test score effects across grade levels using a precision-weighted average.[24] When we combine test score or educational attainment effects across populations (i.e., high- and low-income), we use the population-weighted average (i.e., put greater weight on the larger population) as our overall study effect.[25] This ensures that our overall estimate is as representative as feasible of what the effect would be for the entire population, and facilitates comparison across studies. In Section 6, We show that all of our results are remarkably similar to alternative ways to combine estimates.

Step 3: Standardize the Effect on the Outcome

Studies report effects on test scores with different scales, and may report impacts on different outcomes (e.g., district proficiency rates or high school graduation). To facilitate comparison across studies, we convert each estimated effect into student-level standardized units if not already reported in these units.[26]

[21] Note that we estimate our main models across a range of assumed correlations, displayed visually in Figure 5 and presented in Section A.3. These have little effect on our main results.

[22] We follow Borenstein (2009) Chapter 24 to compute the standard error of the average effect, and assume a correlation of 0.5 when combining subjects for the same grade.

[23] In the single paper (Baron (2021)) that presents impacts for two separate types of spending (non-capital and capital) on one outcome (test scores), we use the simple average of the impacts of both spending types as our single overall effect for the coin test analysis; we include both (non-capital and capital) distinct estimates of effects on test score outcomes for our meta-analysis. To compute the standard error of the overall test score effect for Baron (2021) we assume a correlation of zero.

[24] Precision weighting is a way to aggregate multiple estimates into a single estimate with the greatest statistical precision. Instead of a simple average, this approach more heavily weights more precise estimates (i.e., placing more weight on the estimates that are the most reliable). We follow Borenstein (2009) Chapter 23 to compute the standard error of the precision-weighted average as the reciprocal of the sum of the weights (inverse variances). This calculation of the standard error assumes a correlation of zero between the estimates.

[25] We follow Borenstein (2009) Chapter 24 to compute the standard error of the average effect, and assume a correlation of zero when combining outcomes for different populations. We use the relative sample sizes reported in the study to weight. For example, in Lafortune et al. (2018) we combine the estimates for the top and bottom income quintiles (using the relative sample sizes) and assume a correlation of zero between these estimates. We make an exception in one case: Cascio et al. (2013) report dropout rate estimates for Black and White students. For this study we population-weight by an estimated share White = 0.9 and share Black = 0.1 rather than the 0.68/0.32 shares reported for the study sample.

[26] When effects are not reported in student-level standardized units, we divide the reported raw effect, $\Delta\mathring{y}$, by the student-level standard deviation of the outcome to capture the estimated effect on the outcome in student-level standard deviation units (i.e. $\sigma_{\mathring{y}}$). To perform this standardization, we gather information from each paper on the standard deviation of the outcome of interest. This standard deviation is generally reported in summary statistics. In two cases (Rauscher (2020) and Kogan et al. (2017)), the standard deviation is reported at the school or district level. In these two exceptional cases, we convert the school- or district-level standard deviation into a student-level standard deviation by dividing the school or district-level standardized estimate impacts by the square root of the school or district size. Our results are robust to excluding these two studies (see Table A.5). For binary outcomes such as proficiency rates, graduation rates, or college-going rates, we use the fact that the standard deviation of a binary variable is $\sqrt{p \times (1-p)}$. In the three studies that report on graduation rates for relatively old samples (Jackson et al. (2016), Johnson (2015) and Weinstein et al. (2009)), we standardize estimated effects using graduation rates that prevailed at that time (77%) from national aggregate statistics, rather than using the baseline reported for the study sample. This choice makes studies more comparable by using the same standardization across studies of the same outcome and time period.

Step 4: Equalize the Years of Exposure

Because education is a cumulative process, one would expect larger effects for students exposed to school spending increases for a longer period of time. Indeed, we show evidence of this empirically in Section 8.3. To account for this, we standardize all effects to reflect (where possible) the effect of being exposed to a spending increase for four years. Several studies report the dynamic effects of a school-spending policy (i.e., the effect over time). For test scores, when the dynamic effects are reported, we take the outcome measured four years after the policy change.[27] Some papers do not report dynamic effects, and only report a single change in outcome after a policy-induced change in spending. In such cases, we take the average reported effect.[28] Because high school lasts four years, many papers report the effect on educational attainment of four years of exposure, but not all do.[29] [30] We adjust the captured effects to reflect four years of exposure by dividing the overall effect by the number of years of exposure and then multiplying by four. We test the assumption that the educational effects increase linearly with years of exposure, and find that this holds empirically in Section 5.3. Formally, the standardized four-year effect of policy j is Δy_j.

Step 5: Equalize the Size of the Spending Change

Each included study isolates the effect of the policy on spending (and that of the policy on outcomes) from other potential confounding factors and policies. We seek to determine the change in outcomes associated with a particular change in per-pupil spending. To ensure comparability of dollar amounts across time, we adjust reported dollars in each study into 2018 equivalent dollars using the Consumer Price Index (CPI).[31] Because we measure the impacts of exposure to four years of a spending change, we relate this four-year outcome effect to the change in spending during these same four years. For each study j we collect the *average* change in per-pupil spending (in 2018 CPI adjusted dollars) over the four years preceding the observed outcome, $\Delta\$_j$.[32] When the effect of spending on outcomes is directly reported in a study, we record this estimate directly. See Section 3.2 for a detailed description of accounting for capital spending.

[27] Note that some papers may refer to this as a year-three effect when they define the initial policy year as year zero, while others may refer to this at the year four effect if the initial policy year is year 1.

[28] In many cases, the average exposure is less than four years so that (if at all) we may *understate* the magnitude of any school spending effects for these studies.

[29] Papers that report effect for years of exposure other than 4 are: Abott et al. (2020), Jackson et al. (2016)/Johnson and Jackson (2019), and Kreisman and Steinberg (2019).

[30] We capture the effect of referendum passage on college enrollment 10 years post-election in the case of Baron (2021) to ensure comparability with other studies which report on the same outcome.

[31] We adjust based on the article's reported $ year, and the last year of data if no $ year reported.

[32] For a policy that leads to a permanent shift in spending, the *total* four-year change in spending is 4 times the permanent shift and the *average* is the permanent shift. However, because spending can vary across years following policy enactment, the duration of exposure and duration of the policy may not be the same. In these cases, we use the average increase in spending during the four years preceding the outcome. For example, a policy may have increased per-pupil spending by $100 in the first year, and increased linearly up to a $400 increase in the 4^{th} year. In this case, we would use the *average* increase in spending during the four years, which is $250. If a study does not report spending change in the four years preceding the observed outcome, we capture the change in spending and the contemporaneous measured outcome. This decision likely *understates* the true spending effect because these models may not account for the benefit of spending in previous years.

Step 6: The Standardized 4-Year $1000 Spending Effect

For each study, we obtain an estimate of the change in the standardized outcome per $1000 policy-induced change in school spending (averaged over four years and in 2018 dollars). Our standardized effect on outcome y from study j is $\mu_{yj} = (\Delta y_j)/(\Delta \$_j)$. For 5 out of 31 study-outcomes, we compute this ratio manually after standardizing the impact of the policy on both student outcomes and per-pupil spending. For the 26 out of 31 study-outcomes that report marginal spending effects directly, we take the reported marginal effect and adjust it (where needed) for exposure, CPI, and student-level standardization. Importantly, this parameter estimate is comparable across studies.[33] μ_{yj} can be interpreted an Instrumental Variables (IV) estimate of the marginal impacts of school spending on outcomes using the exogenous policy-induced variation in school spending as the instrument.[34]

To illustrate the importance of computing the same parameter from each paper, consider the following two papers: Lafortune et al. (2018) report that the "*implied impact is between 0.12 and 0.24 standard deviations per $1,000 per pupil in annual spending*" while Clark (2003) reports that "*the increased spending [...] had no discernable effect on students' test scores*", reporting small, positive, statistically insignificant impacts. At first blush, these two studies suggest very different school spending impacts. However, when compared based on the same empirical relationship, the papers are similar. Specifically, precision aside, μ_{yj} for Clark (2003) is 0.0148σ. By comparison, the large positive impact in Lafortune et al. (2018) is based the change in the **test-score gap** between high- and low-income groups (a *relative* achievement effect) over ten years. Their estimates of absolute overall test score impacts over 4 years yields a μ_{yj} of 0.0164σ.[35] Despite the two studies coming to **very** different conclusions and reporting their results is very different ways, when compared based on a common parameter, they are, in fact, remarkably similar.

3.2 Making Capital Spending Comparable to Non-Capital

A key contribution of this work is to provide a framework, informed by theory, to allow for a direct comparison of the marginal impacts of capital spending to those of non-capital spending.

[33] We also capture the associated standard error of the estimate. When studies report the effects on spending and then on outcomes, our standardized effect μ is a ratio of two estimates: the estimated change in the outcome divided by the estimated change in spending. In these cases, where studies report the effect of a *policy* and not of a specific *dollar change*, we account for this in computing the standard error. We follow Kendall et al. (1994) and use a Taylor expansion approximation for the variance of a ratio. If β and δ are both estimates, if $Corr(\beta, \delta) = 0$, the standard deviation of $\frac{\beta}{\delta}$ is approximately $\sqrt{\frac{\mu_\beta^2}{\mu_\delta^2}[\frac{\sigma_\beta^2}{\mu_\beta^2} + \frac{\sigma_\delta^2}{\sigma_\beta^2}]}$. In Appendix Tables A.7 and A.8 we run our main specifications across the range $Corr(\beta, \delta) = [-1, 1]$ and our overall results are largely identical.

[34] For the 16 study-outcomes that report population average IV estimates, we simply re-scale the reported effects (and standard errors) to equalize exposure, and CPI-adjust policy spending changes. For 15 study-outcomes, our overall effect combines estimates across subjects (e.g., math and reading) and/or populations (e.g., grade-levels, high and low-income, or Black and White students). In all but 1 of these cases we compute the average of the sub-population IV estimates – as opposed to computing the ratio of the average effects. We only compute the ratio of the average effects when we combine estimates across grades levels and subjects. In these cases, because there are no reported differences in spending changes by grade or subject, the ratio of the average effects and the average of the individual IV ratios are identical.

[35] In their study, using relative versus absolute achievement gains matters. Specifically, they report test-score *declines* for high-income areas which makes the relative gains larger but the absolute gains smaller.

Increases in non-capital spending go toward educational inputs that are used in the same year (such as teacher salaries or transportation fees). In contrast, because capital spending goes toward durable assets that are used for years after the initial financial outlay, it is inappropriate to relate outcomes in a given year to spending on capital *that same year*. To account for the difference in timing between when capital spending occurs and when the inputs purchased may affect outcomes, we use the annualized accounting value of the one-time increase in spending as the spending change associated with estimates of student outcomes.

To assess the value of $1000 in capital spending as comparable to the same in non-capital spending requires some reasonable assumptions. Specifically, a one-time (i.e., non-permanent) $1000 increase in spending to hire an additional teacher for a single year may be reflected in outcomes in that year. In contrast, such spending on a building should be reflected in improved outcomes for the life of the building. In a simplistic case, where the asset does not depreciate (i.e., there is no wear and tear and the asset is equally valuable over its life), one would distribute the total cost of the asset equally over the life of the asset. For example, if the life of a building is 50 years and the building costs $25,000,000, the one-time payment of $25,000,000 would be equally distributed across the 50-year life span and be equivalent to spending $25,000,000/50=$500,000 per year. Note that, with no depreciation, for a typical school of 600 students, this seemingly large one-time payment of $25M would be equivalent to $500,000/600= $833.33 per-pupil per year.

In a more realistic scenario with depreciation, during the first year of a building's life, it is more valuable than in its 50^{th} year, due to wear and tear and obsolescence. In our example, the building's value in its first year would be greater than $500,000 and in its last year less than $500,000. To account for this, we follow convention in accounting and apply the depreciated value of capital spending projects over the life of the asset. We assume annual depreciation of 7%, representing the asset losing 7% of its value each year. We depreciate expenses that went primarily to new building construction or sizable renovations over 50 years.[36] We depreciate expenditures of less durable assets (such as equipment or upgrading electrical wiring for technology) over 15 years, and for studies that report the proportion of capital spending that went to new building construction, we depreciate the capital amount proportionally between 50 and 15 years.[37] In Section A.3 we show that our main conclusions are robust to using lower and upper bounds of years depreciated, as well as to assuming no depreciation.

For each study of capital spending, we compute the change in student outcomes for each $1000 in average flow value of the capital spending in the years preceding the measured effect.[38] We

[36] In 2013-14, the average age of school buildings in since original construction was 44 years (NCES 2016). Studies report on building age, including: Lafortune and Schonholzer (2019) (44.5 years), Martorell et al. (2016) (36 years), and Neilson and Zimmerman (2014) (well over 50 years).

[37] For example, Martorell et al. (2016) report that most of the spending went to renovations, and Cellini et al. (2010) provide an example of specific capital projects funded by a bond referenda, including to "improve student safety conditions, upgrade electrical wiring for technology, install fire doors, replace outdated plumbing/sewer systems, repair leaky rundown roofs/bathrooms, decaying walls, drainage systems, repair, construct, acquire, equip classrooms, libraries, science labs, sites and facilities..." (220). We describe capital paper coding in Table A.3.

[38] Depreciating the asset puts more value on the early years when test scores are measured and less on the years for which outcomes are not measured (many studies do not evaluate what the effect is more than 6 years after the funds

illustrate this depreciation in Figure 2, which shows the 15-year depreciation of a $7,800 per-pupil ($4.7 million per school) expenditure (as in Martorell et al. (2016)) and the 50-year depreciation of a $70,000 per-pupil ($42 million per school) expenditure (as in Neilson and Zimmerman (2014)). This transforms the extraordinarily large one-time expenditure over the projected life of the asset, which falls in value over time. After computing the flow value of the capital outlay for each year after initial payment, we can relate observed student outcomes associated with the average depreciated value of the asset in the years preceding measured outcomes.[39]

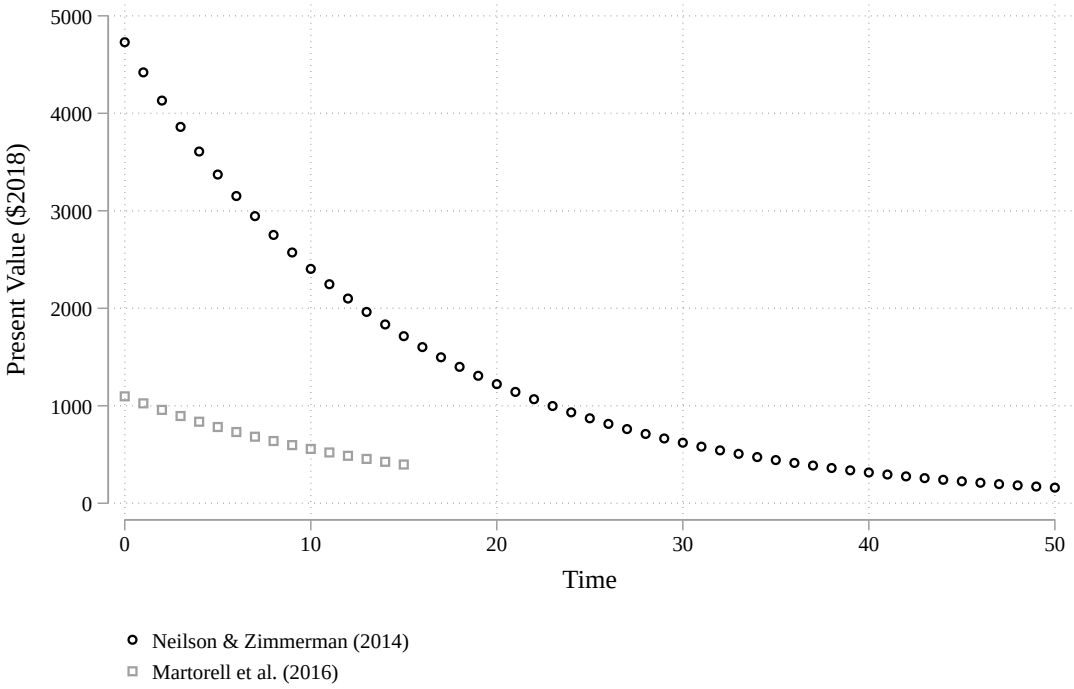


Figure 2: Exemplar Capital Expenditure Depreciation

Accounting for Construction Time

Because the typical capital project does not lead to contemporaneous changes in classroom experiences, it is reasonable to expect any possible student improvements to take several years to materialize after the capital outlay. Indeed, large capital projects that involve entirely new construction or major upgrades to a new wing of a building can take multiple years to complete. Moreover, capital projects often entail some temporary disruption to everyday operations during

are used). Because our parameter includes the spending change in the denominator, this reduces the reported school spending effect relative to not depreciating the asset. Accordingly, our approach may be considered conservative.

[39] Because we use the size of the overall capital spending amount to compute the policy effect on spending, (Δ\$) is not an estimate. As such, the standard error of the IV estimate is simply the standard error of the policy effect on the outcome divided by the actual spending change. The one exception is Rauscher (2020), who does not report an average bond amount but provides an estimated policy effect on capital spending during the six years following bond passage. In this case, we do adjust our IV estimate standard error to account for this estimated spending change.

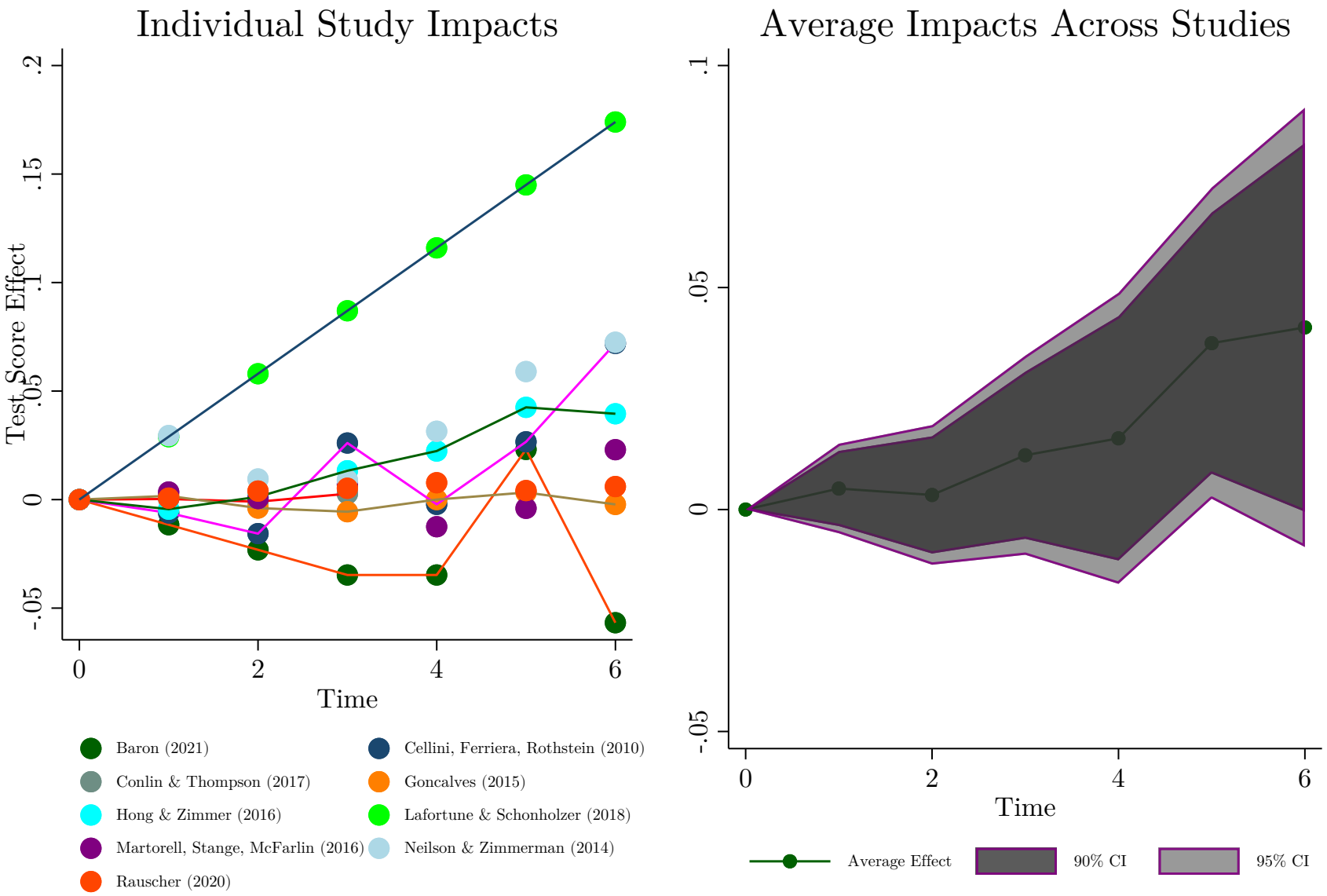
the renovation/construction period, which may be deleterious to student outcomes. For these reasons, we assign the first two years of a capital spending project to a "construction/adjustment period" and capture outcomes six years after the increase in capital spending.[40]

To assess whether this temporal decision is reasonable, Figure 3 presents the dynamic effects of the nine studies estimating changes in capital spending on student test score outcomes. The left panel plots the *raw* effects for each study, not the marginal per-$1000 effects, over time as relative to a baseline year zero (t = 0) in which there should be no effect of the policy (the year of the construction or the policy change).[41] Consistent with an initial disruption, in several cases there is an immediate dip in outcomes. Consistent with long-run benefits to capital spending, this initial dip is followed by a gradual increase in outcomes in most studies. By about 5 or 6 years after a capital spending increase, one observes improved outcomes in most cases. To more formally assess the evolution of outcomes over time, we present the average dynamic effect in the right panel of Figure 3.[42] We plot the average (across the nine studies) effects 1 through 6 years after the capital project or construction along with the 90 and 95 percent confidence intervals. This shows the same per-study pattern of no change (or possibly a slight dip) in the first two years and then improving outcomes after about 5 or 6 years. Indeed, one rejects that the effect of capital spending is zero at the 5-percent level by year five. This pattern validates our assigning the first two years of these studies to a "construction/disruption" period and using the six-year effect for capital spending increases as the most comparable to non-capital spending four-year effects. Overall, the pattern indicates that (a) capital spending *does* improve outcomes on average, and (b) these benefits take between 4 and 6 years to materialize. We present more formal statistical tests in Section 5 that quantify the extent to which capital spending may affect outcomes.

[40] Eight of nine papers report six-year estimates of the effect of capital spending changes on student outcomes. When the six-year effect is not reported, we use the latest year reported. Conlin and Thompson (2017) reports only 3 years after capital spending, so we capture their year-three effect as our estimated effect. As a conceptual matter, if capital spending does not improve student outcomes over both the year-four and the six-year effects, the impacts of spending would be zero. This distinction matters only if one rejects the null hypothesis of zero spending impacts.

[41] For Lafortune and Schonholzer (2019), Neilson and Zimmerman (2014), and Goncalves (2015), year one (t = 1) represents first year of occupancy at a new or renovated school. In the case of Conlin and Thompson (2017) year one is the first year of program eligibility. For all other studies, year one (t = 1) represents the first year after a capital bond was passed.

[42] Figures A.1 and A.2 aggregate effects over time by precision and random effects weightings.



Note: Lafortune & Schonholzer (2018) reports linear effects over time.

For both figures, time is year relative to construction or spending increase. Effect is the standardized effect, relative to year 0.

Figure 3: Capital Spending Effects Over Time

4 Meta-Analytic Methods

We quantitatively describe the distribution of school spending impacts on test scores and educational attainment. While the simple average and standard deviation of the study impacts (μ_{yj}) provides information about the distribution of impacts, this approach can be very misleading in two important ways that a formal meta-analysis can address.

First, while the simple average across studies is an unbiased estimate of the center of the distribution of impacts, an inverse-variance weighted average is the minimum variance unbiased estimate (Hedges (1983), Hartung et al. (2008)). Intuitively, when forming an average, one would place less weight on less reliable estimates (i.e., those which have larger standard errors due to underpowered methods or small samples) and more weight on those that are more precisely estimated.[43] As such, the inverse-variance weighted average (or precision-weighted average) is be a more reliable measure (i.e., less sensitive to imprecise outliers) of the center of the distribution of impacts.

Second, because of sampling variability, the spread of the raw estimates may drastically overstate the spread of the distribution of *true* impacts. To inform policy, one must know how much of the spread across studies can be attributed to sampling variability (i.e., the chance variability across studies that is due to the choice of sample) versus real contextual differences (due to different treated populations, different policy types, and different estimation strategies) across studies. Understanding the role of these contextual differences is critical to being able to predict what one might expect to observe in a new context, and a failure to account for cross-study heterogeneity could lead to overconfidence in the ability to extrapolate to other settings.

To address both these limitations, we perform random effects meta-analysis to generate overall pooled estimates of the average effect of spending on student outcomes and to estimate heterogeneity across studies. We detail this approach below.

4.1 The Formal Model of the Distribution of Study Impacts

Where μ_{yj} is the observed effect of a \$1000 spending increase (in 2018 dollars) over four years on outcome y = {test scores, educational attainment} in study j, each study-outcome can be represented as in (1).

$$\mu_{yj} = \theta + \delta_j + \epsilon_j \tag{1}$$

In Equation (1), θ is the pooled average effect across all studies (not necessarily the effect estimated by any individual study). There are two reasons that a study estimate would deviate from this average. The first is sampling variability (or within-study error), represented by δ_j. The second is treatment effect heterogeneity (or the between-study error), represented by ϵ_j. Where $\sigma^2_{\mu,yj}$ is the within-study variance for study j, and τ^2 is the variance of the study-specific deviations from the pooled mean, the study impacts are distributed around a grand mean θ with variance $\sigma^2_{\mu,yj} + \tau^2$.

[43] This logic assumes that the precision of an estimate is unrelated to the the study estimate. In Section 7 we show that this is reasonable in our setting.

Where $\sigma^2_{\mu,yj}$ is treated as known and approximated by the squared standard error, $se^2_{\mu,yj}$, one can estimate τ^2 empirically by method of moments. Specifically, the estimated heterogeneity parameter $\hat{\tau}^2$ is identified based on the difference between the observed variability across studies and that which would be expected due to sampling variably alone.[44] Intuitively, if the confidence intervals for the individual studies tend to overlap, it would suggest that τ is small, while non-overlapping intervals would suggest heterogeneity. Accounting for both sources of variability, the optimal inverse-variance weighted average across all J studies is $\hat{\theta}_{pw} = \frac{\sum \mu_{yj} w_{yj}}{\sum w_{yj}}$, where each study receives weight w_{yj} as in (2).

$$w_{yj} = \frac{1}{(\sigma^2_{\mu,yj} + \tau^2)} \tag{2}$$

To form the empirical analog of (2), and therefore $\hat{\theta}_{pw}$, one can use the square of the standard error ($se^2_{\mu,yj}$) as an estimate of $\sigma^2_{\mu,yj}$, and estimate τ^2 by method of moments. The variables μ_{yj} and $se^2_{\mu,yj}$ come from the individual studies, while the parameters τ^2, $\hat{\theta}_{pw}$, and the standard error of the weighted average ($se_{\hat{\theta}_{pw}}$) can be estimated. We estimate this random effects model using weighted least squares, with inverse-variance weights. We estimate standard errors using robust variance estimation (RVE), a meta-analytic analog to heteroskedasticity and cluster-robust standard errors (Hedges et al. (2010)). Following best practice, we use small-sample corrections, including degrees-of-freedom adjustments, that result in confidence intervals with good coverage even with fewer than ten studies (Tipton (2015)).[45] Another parameter from this estimation is the relative amount of between-study heterogeneity. This is referred to as I^2, and is the ratio of the variance of the between-study heterogeneity and the overall variance (reported in regression tables as % Cross-Study Var.).[46]

4.2 Confidence Intervals and Prediction Intervals

To answer "does money matter?" one can test the hypothesis that the average pooled effect is zero. For this, one would use the standard error of the estimate to form a t-test.[47] Similarly, one can use the standard error of the mean to compute a confidence interval for the pooled average.

$$CI = \hat{\theta}_{pw} \pm t^* \times se_{\hat{\theta}_{pw}} \tag{3}$$

[44] Formally, where $\tau = 0$, the precision-weighted grand mean is M=$\frac{\sum \mu_{yj}(1/se^2_{yj})}{\sum(1/se^2_{yj})}$. The sum of the standardized square deviations from M across studies is $Q = \sum_1^J (\frac{\mu_{yj} - M}{se_{yj}})^2$. Importantly, Q follows a χ^2 distribution with an expected value of degrees of freedom (df), which is the number of studies j minus 1. As such, $Q - df$ measures the extent to which the observed dispersion is greater than can be explained by sampling variability alone. This forms the basis for an estimate of τ^2. That is, the method of moments estimate of τ^2 is $(Q - df)/C$, where C is a factor based on the study weights used to compute Q. Dividing by C reverses this process so that the τ^2 units are the same as those used in the studies (see Borenstein et al. (2017) for a full derivation).

[45] We implement these estimators using the "robumeta" package in Stata (Hedberg et al. (2017)).

[46] Where $\tilde{\sigma}^2_\mu$ is a precision-weighted average of the individual within-study variances, $I^2 = \hat{\tau}^2/(\tilde{\sigma}^2_\mu + \hat{\tau}^2)$.

[47] All tests we present use the t-distribution with the appropriate degrees of freedom adjustment.

This confidence interval pertains to the pooled average across all studies and *does not* account for treatment heterogeneity (i.e., that different studies provide estimates of different true causal impacts). As such, it does not provide a sense of what to expect in *future* studies. For this, one would form a prediction interval that includes both sources of error. The prediction interval is given by Equation (4).

$$PI = \hat{\theta}_{pw} \pm t^* \times \sqrt{se^2_{\hat{\theta}_{pw}} + \hat{\tau}^2} \tag{4}$$

The prediction interval is necessarily wider than the confidence interval because it also accounts for heterogeneity across studies. It represents the range of values that one can expect to observe in a randomly sampled new study. This is an important policy parameter. That is, while one can be near certain that, *on average*, policies that increase school spending improve student outcomes, policymakers may wish to know how likely they are to have a positive effect of a future policy in their particular context. While the confidence interval speaks to the former, the prediction interval speaks to the latter. We discuss both as we interpret our results.

Conservative Prediction Intervals

Recent studies have shown that prediction intervals may be too narrow with fewer than 20 studies (Nagashima et al. (2019); Kontopantelis et al. (2013)). There are only 12 studies of educational attainment outcomes, so the estimate of τ^2 for this set of studies may understate the true underlying heterogeneity in the population.[48] To account for this, we also compute a conservative predication interval based on an "overestimate" of τ^2. Specifically, we take bootstrap samples from our data to get 400 estimates of τ^2. We then take the 99^{th} percentile of this bootstrap distribution ($\hat{\tau^2_{99}}$) as our upper bound estimate of τ^2, and use this to form a conservative prediction interval as in (5).

$$PI = \hat{\theta}_{pw} \pm t^* \times \sqrt{se^2_{\hat{\theta}_{pw}} + \hat{\tau^2_{99}}} \tag{5}$$

Our alternate estimate of the heterogeneity parameter ($\hat{\tau^2_{99}}$) is near the maximum amount of underlying heterogeneity that is consistent with the studies in the sample. While prediction intervals based on this may overestimate variability for test score effects (where there are more than 20 studies), it may be reasonable for educational attainment effects (for which there are only 12 studies).

4.3 Instrument Validity and Interpretation of Our Estimate

Because each standardized effect is an IV estimate, our overall meta-analytic average is an average of IV estimates. As such, we consider the standard IV assumptions in the context of our meta-analytic pooled average. For each paper's estimate to be valid, it must be that the instrument is relevant (i.e., each policy lead to a real change in per-pupil spending) and it must be excludable (i.e., each policy only influences the outcomes through its effects on per-pupil spending). As we

[48] Intuitively, if one happens to have 12 studies with similar estimates, the model may suggest that there is no heterogeneity in other studies.

discuss in Section 7, in our setting, for the pooled average to be unbiased and consistent *does not* require that each paper have a valid instrument. For our pooled average (of IV estimates) to be a consistent estimate of the true average marginal spending effect requires the two following conditions:

Relevance (*on average*): On average, the policies examined lead to meaningful changes in per-pupil spending. By focusing only on papers that demonstrate a meaningful policy effect on spending (our inclusion restrictions), the first condition is satisfied for each study. It follows that this condition is also satisfied in our meta-analytic average.

Excludable (*on average*): Our pooled average requires that *on average* the relationship between the policies and outcomes only operates through the policy effect on school spending. This condition *does not* require that each study satisfy the exclusion restriction, but that it is satisfied on average. Intuitively, even if all the individual studies are biased due to violations of the exclusion restriction within each study, *if the biases are largely random*, some studies will be biased upward while other will be biased downward. In this case, the average of the biases will be zero in expectation – yielding an unbiased pooled average. Even though our inclusion criteria reduces the likelihood that any single study is severely biased, our method does not require that all studies are actually unbiased, just that any bias in the included studies is essentially random. We present a novel test of this condition in Section 7.1 and shows that it holds empirically.

5 Results

5.1 Does School Spending Matter? The Coin Test

Before quantifying the extent to which increased spending affects outcomes, we perform a simple count-based test of whether the causal evidence supports the notion that increased school spending improves student outcomes. It is well-known that the standard vote-count approach (i.e., counting the share of *statistically significant* effects above some pre-specified threshold) as used in Hanushek (2003) is inconsistent (Hedges and Olkin (1980)). We present an alternative counting approach that does not suffer these consistency problems, based on simple counts of positive and negative estimates irrespective of statistical significance.[49] The value of this test is that it is intuitive, and can be used when little is known about a study other than the sign of the study impacts.

Specifically, under the null hypothesis that the true effect is zero (i.e., $\mu_{yj} = 0 \ \forall j \in J$), each study's reported effect is simply the error term (ϵ_{yj}). By the central limit theorem, ϵ_{yj} is approximately normal (and hence symmetric) so long at asymptotics apply. It follows that, if studies are independent and there were no association between school spending and student outcomes, half the studies would be positive while half would be negative. As such, the probability of observing X positive estimates out of N studies follows a binomial distribution with probability $p = 0.5$. By comparing the number of positive studies out of all studies to the binomial distribution

[49] See Section 12.2.1.3 of Higgins et al. (2021) for a discussion of this basic idea.

(with $p = 0.5$), we can quantify the likelihood of the data under the null hypothesis of no spending impacts. Given the similarity to a series of fair coin tosses, we refer to this as the "*coin*" test.

To implement this test, we classify each study as reporting a positive or negative effect of school spending on outcomes. To be conservative (i.e., stacking the deck *against* finding a positive association), studies that examine impacts on multiple outcomes are classified as negative if the average impacts for any outcome is negative (even if the impacts on all other outcomes is positive).[50] Note that with weak instruments, the distribution of the error terms may be non-normal Stock et al. (2002), so this test may be inappropriate. To address this potential concern, we also present the coin test using only studies that have strong first stages (first stage F-statistic > 10) where this is less of a concern. While this mechanically reduces the level of confidence (owing to less data), the main conclusions are unchanged.

Across all 31 studies, 28 report positive impacts of school spending on student outcomes. If there were no relationship between spending and outcomes, the likelihood of observing 28 or more positive effects out of 31 is 1 in 430,185. This is the same likelihood of flipping a fair coin (i.e., a coin that has a 50/50 chance of head or tails) 31 times and and getting 28 (or more) heads. While one could quibble with this test on the grounds that each study is not a purely independent draw (given that some studies examine overlapping policy changes), this is compelling evidence that, on average, policies that increase school spending improve student outcomes. Looking only to studies with first stage F-statistics greater than 10, so that the distributional assumption of the test are satisfied (in expectation), 24 out of 26 papers (92%) are positive, which would happen with with a 1 out of 190,650 chance, or probability of 0.000005245 under the null hypothesis that there is no effect of spending on student outcomes.

We present the same test separated by outcome in Table 3. For each specific outcome, the number of available credible studies is more limited, which leads to a lower level of confidence about the relationship. Despite this, for all outcomes, most papers find positive impacts of school spending. Of 24 studies that report effects on test scores, 21 find that increased school spending increases scores. If there were no effect, observing this high a number of positive studies (or more) would occur with probability 1 in 7,216 – extremely unlikely. Of the 12 studies that estimate effects of school spending on educational attainment, all 12 find that increased school spending leads to increased educational attainment. If there were no effect, this high number of positive studies would occur with probability 1 in 4,096 – compelling evidence that specific policies that increase school spending improve students' educational attainment. The final outcome studied is adult earnings. All three independent studies that link changes in school spending to adult earnings find positive impacts. With only 3 studies, there is the possibility that this occurred by chance. Even so, if there were no effect, this high a number of positive studies would occur by chance with probability 1 in 8. Looking only to studies with strong first stages, the likelihood of observing the observed distribution of test score results under the null of no impacts is 1 in 2,745, and for

[50] Using this conservative definition, we classify Weinstein et al. (2009) as negative even though they find positive effects on educational attainment.

educational attainment it is 1 in 256 and for wages it is the same (1 in 8). In sum, the pattern of results is statistically incompatible with the notion that "money does not matter" and provides overwhelming evidence that policies that increase school spending improve student outcomes *on average*.

Table 3: Coin Test by Outcome Examined

Panel A: All Studies					
Outcome	Papers	Positive	Positive & Significant	% Positive	1 in X Chance
All Studies	31	28	14	0.90	430185
Test Score	24	21	9	0.88	7216
Educational Attainment	12	12	8	1	4096
Wages (income mobility)	3	3	2	1	8
Panel B: Fstat > 10					
Outcome	Papers	Positive	Positive & Significant	% Positive	1 in X Chance
All Studies	26	24	14	0.92	190650
Test Score	19	17	9	0.89	2745
Educational Attainment	7	7	6	1	128
Wages (income mobility)	3	3	2	1	8

5.2 How Much Does School Spending Matter?

To assess the extent to which school spending matters, we examine the distribution of effects for each outcome type. We first present a forest plot to visualize all the estimates, and describe the distribution of raw estimates with simple averages and medians. We then provide a more rigorous analysis of the center and spread of the distributions of causal impacts, along with a discussion of the importance of treatment effect heterogeneity, based on random-effects meta-analysis.

We present forest plots for the test score and educational attainment impacts separately in Figure 4. For each included study that examines impact on test scores or educational attainment, we plot the estimate of a $1000 increase in per-pupil school spending (2018 CPI adjusted) sustained over four years. We also plot the 95% confidence interval associated with each study estimate. Studies are presented in descending order by estimated impact. To show the meta-analytic results visually, the 90% confidence interval for the pooled average is in dark blue, and the 90% prediction interval (using τ) for a new study in a different context is in pink, and the conservative 90% prediction interval (using τ_{99}) is in light blue (see Section 4.2 for details).

5.3 Test Score Impacts

The forest plot in the top panel of Figure 4 indicates that the most precisely estimates studies are those in the middle of the distribution, and the most imprecise estimates tend to be at the extremes of the distribution of effects. This suggests that the most reliable estimates are near the median of the distribution. The 25^{th} percentile of school spending effects on test scores is 0.0156 and the 75^{th} percentile is 0.103. This range of estimates underscores (a) that school spending effects are largely positive, and (b) it is important to look at the literature as a whole to gauge magnitudes

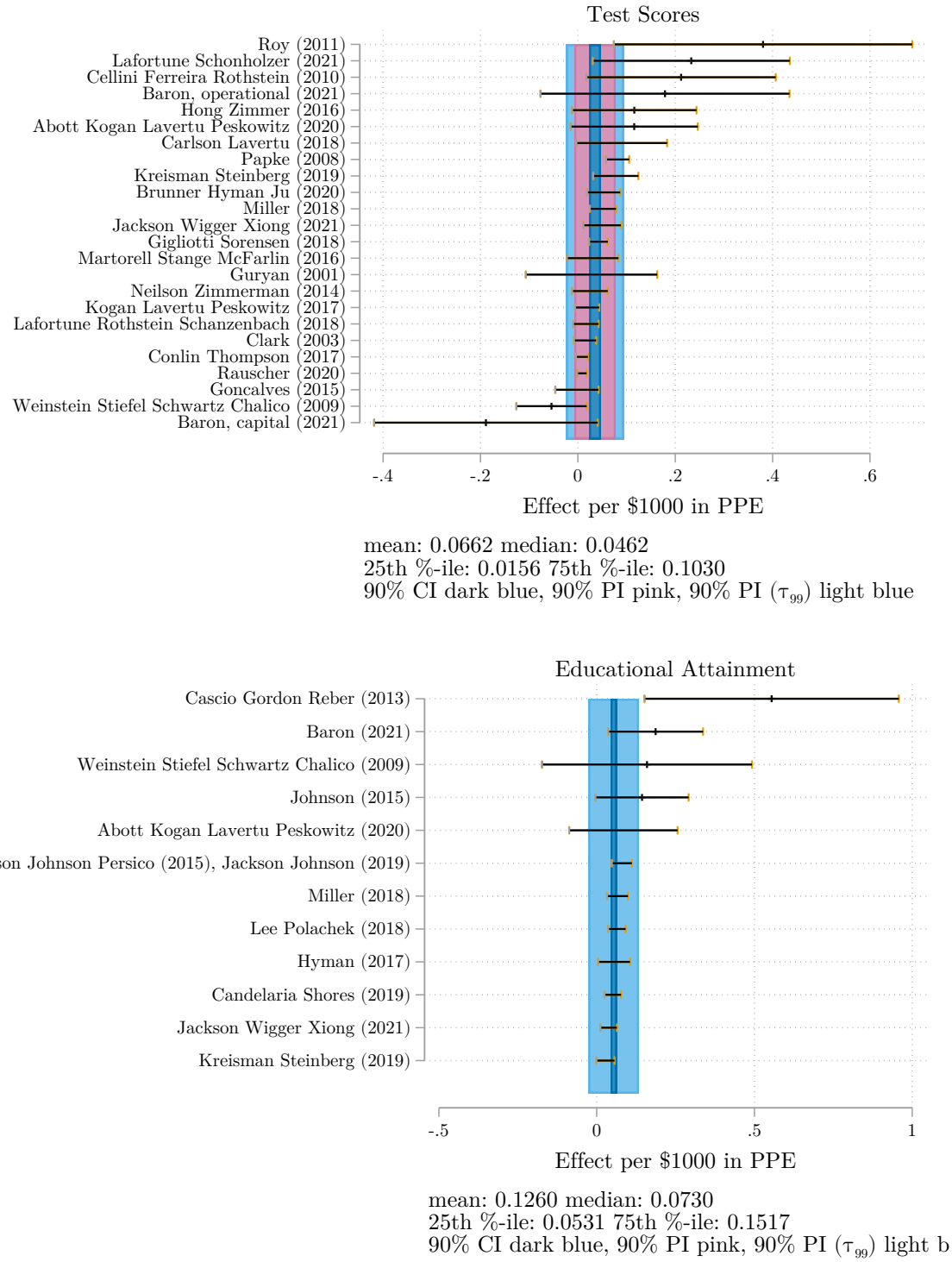


Figure 4: Overall Estimates

of impacts. The simple average is 0.0662σ, while the median is 0.0462σ. The fact that the median is notably smaller than the mean suggests that some studies with large effects are inflating the

average. Given that the largest estimates are also the least precise, a precision-weighted average may be more appropriate than a simple average.

We present meta-regression results in Table 4. We report the pooled average impacts for all test score studies in column (1), for non-capital spending on test scores in column (2), capital spending on test scores in column (3), and the effects of non-capital spending on educational attainment in column (4). For each model, we report the pooled effect in addition to the standard error of the pooled effect. Importantly, we also report τ, an estimate of the between study variability – which is critical to helping estimate what one could expect in *other* settings.

Table 4: Meta-Analysis Estimates

	(1) Overall Test Scores	(2) Non-Capital Test Score	(3) Capital Test Score	(4) Overall Educational Attainment
Panel A: Full sample				
Average Effect	0.0352***	0.0431***	0.0150***	0.0539***
	(0.00723)	(0.00878)	(0.00536)	(0.00577)
N	24	15	9	12
τ	0.0247	0.0246	0.0139	0
% Cross-Study Var.	0.761	0.759	0.501	0
90% PI	[-0.00725,0.07767]	[0.000,0.086]	[-0.010,0.040]	[0.044,0.063]
Prob. Pos	0.914	0.951	0.843	1
τ_{99}	0.0355	0.0399		0.0362
% Cross-Study Var. (τ_{99})	0.868	0.844		0.873
90% PI (τ_{99})	[-0.02462,0.09505]	[-0.02424,0.11042]		[-0.007,0.115]
Prob. Pos (τ_{99})	0.834	0.854		0.929
Panel B: Fstat > 10				
Average Effect	0.0386***	0.0442***	0.0252**	0.0553***
	(0.00737)	(0.00836)	(0.0122)	(0.00724)
N	19	11	8	7
τ	0.0257	0.0222	0.0275	0.0161
% Cross-Study Var.	0.771	0.716	0.755	0.308
Panel C: Cluster like studies				
Average Effect	0.0337***	0.0386***	0.0235**	0.0536***
	(0.00693)	(0.00805)	(0.0107)	(0.00573)
N	24	15	9	12
τ	0.0227	0.0207	0.0245	0
% Cross-Study Var.	0.729	0.595	0.831	0
90% PI	[-0.006,0.073]	[0.002,0.075]	[-0.021,0.068]	[0.044,0.063]
Prob. Pos	0.922	0.959	0.810	1

Standard errors in parentheses

Full sample coefficient on capital= -.0247 (p-val = .0567)

* $p < .1$, ** $p < .05$, *** $p < .01$

Looking at our preferred test score estimate (column 1), the pooled effect across all studies implies that a $1000 increase in per-pupil spending (in 2018 dollars and sustained over four years)

would increase average test scores by roughly 3.52 percent of a standard deviation. The 95 percent confidence interval for this pooled average lies well above zero and is between 0.021σ and 0.049σ. Note the narrower 90 percent confidence interval for this pooled average is depicted in dark blue in Figure 4.

While the model indicates that the pooled average of the impacts is greater than zero (at the one-tenth of one percent significance level), this does not mean that one should expect positive spending impacts more than 99 percent of the time. The models estimates that 76.1 percent of the variability in impacts reflects heterogeneity across studies (i.e., not all contexts will have the same treatment effect), which suggests there is uncertainty about what one would observe in other settings. More specifically, the standard deviation of heterogeneity across studies (i.e., τ) is 0.0247. This implies that any two studies may have true causal impacts that differ by about 0.0247σ simply due to treatment heterogeneity. Intuitively, the model estimates this level of heterogeneity because relatively precise studies like Papke (2008) and Rauscher (2020) both have positive effects but *do not* have overlapping margins of error. The model takes this as evidence that these studies likely come from contexts with different effects (both of which are positive), and infers a positive average effect with nontrivial heterogeneity across contexts. An implication of this estimate is that even though the pooled effect is 0.0352, in other contexts one would expect estimates between -0.00725σ and 0.0777σ about 90 percent of the time. The 90 percent prediction interval for what one would expect in a new study is depicted in pink in Figure 4. This prediction interval contains the point estimates of 14 of the 24 studies, and (with the exception of Papke (2008)) those that lie outside this range are very imprecise. Another policy-relevant summary of the predicted impacts is that a policy that increases school spending by $1000 over a four-year period would increase test scores 91 percent of the time - more than 9 times out of 10.[51]

Capital Versus Non-Capital Spending Impacts on Test Scores

In a recent review, Jackson (2020) points out that while the impacts of operational spending are consistently positive, the impacts for capital spending are less clear. Additionally, Baron (2021) finds positive test score impacts for operational spending increases but no such pattern for capital spending.[52] However, it is possible that many capital studies may not *individually* have the statistical power to detect reasonable effects. As such, it may be useful to formally examine whether marginal capital and non-capital spending impacts differ across several studies.

Looking at the pooled estimates for non-capital and capital spending (columns 2 and 3), the average effect is larger for non-capital spending (0.0431σ) than for capital (0.0150σ). However, the average impacts for both spending types are individually significant at the 1 percent level. A formal test of the difference in effects involves estimating a meta-regression with a capital indicator

[51] In Panel C of Table 4, we present our main meta-analytic models using the conservative approach of clustering estimates from studies based on the same policies as if they came from the *same* study. This adjustment increases the precision of our estimates, tightening the prediction interval of the likelihood of positive effects from 91.4 to 92.2.

[52] It is noteworthy that, as shown in Figure 3, Baron's results stands in contrast to most other studies of the effects of capital spending on outcomes.

variable – representing the difference between the average for capital spending and others, which yields a p-value slightly greater than 0.05 (see Table 4 note).[53] This suggests that both capital and non-capital spending matters for student outcomes, and that after a few years the economic value of spending is of a similar order of magnitude across the two types. While the point estimate for capital spending is about half the size of that for non-capital-specific spending, the similarity also suggests that our modelling decisions to generate comparable marginal impacts of per-pupil capital to non-capital spending were reasonable.

To put these capital estimates in perspective, we consider two typical kinds of projects. A new elementary school construction would typically cost about \$27.5M and house 624 students (Abramson (2015)). This is a one-time expense of about \$44,000 per pupil. Assuming a 50 year life of the asset, and distributing the value of this capital spending over the life of the asset (while accounting for depreciation at 7% per year), this would be associated with an average per-pupil flow value in the first four years of about \$2693. Using the estimates from column 3, one would expect test scores to increase by about $2.69 \times 0.015 = 0.04\sigma$ six years after the capital outlay. Given the depreciation of the building, *an extrapolation beyond the variation in the data*, suggests that the marginal effect might fall to about half this amount after 15 years. By way of comparison, a modest set of upgrades (i.e., a \$1,000,000 renovation project) may cost 1000000/600=\$1667 per pupil. Assuming a 15 year life of the asset, this would be associated with an average per-pupil flow value in the first four years of about \$150. This would increase test scores by about $0.15 \times 0.015 = 0.00225\sigma$ six years after the capital outlay. This is smaller than what most individual studies can detect. Moreover, with a study that has a standard error greater than 0.02σ (about the median standard error of the sample), true effect sizes of this magnitude could yield negative points estimates over a third of the time *by random chance alone*. This calculation highlights that, given the economic life of capital assets, even though the expected annual marginal benefits are relatively small (often smaller than most individual studies have power to detect), lifetime benefits may be similar to those for non-capital spending. These facts reinforce the importance of (a) calculating the flow value of large one-time capital outlays, and (b) the increased statistical precision afforded by formal meta-analysis that facilitates more conclusive statements than those possible from any individual study.

5.4 Educational Attainment Impacts

The forest plot of all the estimated impacts on educational attainment outcomes is the bottom panel of Figure 4. The 25^{th} percentile of school spending effects on educational attainment is 0.0531σ and the 75^{th} percentile is 0.1517σ. This range of positive estimates underscores the importance of looking at the literature as a whole to gauge magnitudes. The more precisely estimated studies lie close to the median of the distribution and some of the larger estimated impacts are imprecise. The simple average of the educational attainment effects is 0.1260σ, while the median is 0.073σ. As with test scores, the nontrivial difference between the mean and the median reflects the fact that

[53] See Section A.5 for separate forest plots by spending types.

the mean is more heavily influenced by several large, imprecise, positive estimates. This suggests that the precision-weighted average is more appropriate than a simple average and would likely be more similar to the median.

The pooled meta-analytic average school spending effect for educational attainment (column 4 of Table 4) is 0.0539σ. This is similar to the median across all studies. To aid interpretation, we convert the pooled impacts to high school completion and college going rates. For high-school graduation (with a standard deviation of 0.357 in 2018) the estimates suggests that, on average, increasing school spending by $1000 (sustained for 4 years) would increase high-school graduation rates by $0.357\times0.0539=1.92$ percentage points. For college-going (with a standard deviation of about 0.492 in 2018) this suggests that on average, increasing school spending by $1000 (sustained for 4 years) would increase postsecondary attendance rates by $0.49\times0.0539=2.65$ percentage points.

Estimates for educational attainment suggest similar levels of contextual heterogeneity to test scores.[54] Over 85 percent of the variability in causal impacts across studies can be explained by heterogeneity. A policy that increases school spending by $1000 per pupil sustained for four years *in some other context* would lead to educational attainment impacts between -0.007σ and 0.115σ about 90 percent of the time. This implies high school completion impacts between $-0.007\times.357=-0.2$ percentage points and $0.115\times.357=4.1$ percentage points about 90 percent of the time and, college completion impacts between $-0.007\times0.49=-0.3$ percentage points and $0.115\times0.49=5.6$ percentage points 90 percent of the time. Put differently, a policy that increased school spending by $1000 over a four year period would conservatively be expected to increase educational attainment over 92 percent of the time.

Benchmarking the Impacts on Test Scores and Educational Attainment

To put theses estimates into perspective, it is helpful to compare the magnitude of the school spending impacts to those of other interventions. We show this for three separate interventions.

Class Size: Using Project STAR, Chetty et al. (2011) find that reducing class size by roughly seven students increases test scores by 0.12σ (4.76 percentile points) and college-going (by age 20) by 1.8 percentage points. Also, Dynarski et al. (2013) find that reducing class size by seven students increases college-going (by age 30) by 2.7 percentage points. Using this as a benchmark, our test score impacts of 0.0352σ are equivalent to reducing class size by $7\times0.0352/0.12=2.05$ students, while our college-going impacts of 2.65 percentage points are equivalent to reducing class size by between $7\times2.65/1.8=10.3$ students and $7\times2.65/2.7=6.87$ students.

Teacher Quality: Chetty et al. (2014) find that increasing teacher quality by one standard deviation increases test scores by 0.12σ and college going by 0.82 percentage points. Using this as a benchmark, our test score impacts of 0.0352σ would be equivalent to increasing teacher quality by $0.0352/0.12=0.293$ standard deviations, while our college-going impacts of 2.65 percentage points would be equivalent to increasing teacher quality by $2.65/0.82=3.23$ standard deviations.

[54] For our educational attainment estimates, as described in Section 4.2, we use the 99^{th} percentile of the distribution of 400 boootstrap samples to estimate τ^2.

High-achieving Charter Schools: High-achieving charter schools increase test scores by over 0.3σ (Angrist et al. (2016)) and increase college going by as much as 10 percentage points (Booker et al. (2011); Davis and Heller (2019)). As such, our $1000 school spending impacts on test scores are equivalent to about 11.7 percent of the impacts of attending a high-achievement charter school, while our college-going impacts are equivalent to over 25 percent. One may worry that these comparisons are skewed by the large test score impacts in Angrist et al. (2016) or by pulling estimates from different sets of schools. To assuage this concern, we take estimates from Dobbie and Fryer (2020) who find that "No Excuses" charter schools in Texas increase test scores by 0.093σ and college going by 2.5 percentage points. Our $1000 school spending impacts on test scores are equivalent to about a third of the impacts of attending a Texas "No Excuses" Charter, while our college-going impacts are about the same.

For all three benchmarking interventions, our school spending effects are economically meaningful. However, a consistent pattern is that irrespective of the benchmark, the spending impacts on educational attainment are at least twice as large as those on test scores. Importantly, these differences in magnitude between test score and educational attainment impacts are not driven by a comparison across studies, because *this same pattern holds within those studies that examine impacts on both outcomes*. Among the 6 studies that report on both test scores and educational attainment, 5 indicate larger educational attainment impacts than on test scores (Jackson et al. (2021), Baron (2021), Miller (2018), Weinstein et al. (2009), and Kreisman and Steinberg (2019)), while only one does not (Abott et al. (2020)). This suggests that school spending impacts as measured by test scores may not capture the full benefits of school spending policy (Card and Krueger (1992); Jackson et al. (2016)). It is also consistent with the view that educational output is only partially measured by test scores, and that a focus on test score impacts may lead one to understate the benefits of school quality on student outcomes (Beuermann et al. (2020); Jackson (2018); Jackson et al. (2020)).

6 Robustness to Modelling Assumptions and Sample Restrictions

To construct the same parameter for each study, we make several modelling assumptions. It is, therefore, important to assess the sensitivity of our results to these choices, given that alternative choices could have been made. In this section, we show our main estimates under different modelling assumptions and sample restrictions - demonstrating the robustness of our estimates to these assumptions and restrictions.

1. **Strong First Stage:** It is well understood that when the first stage relationship between the treatment and the instrument (in this case the policy) is weak, the resulting estimates may be biased and have unreliable standard errors (Bound et al. (1995) and Conley et al. (2012)). We are relatively liberal in our inclusion of studies, using any study with a first stage F-statistic of 3.85. Because we use precision weighting, and studies with weak first stages are likely to have larger standard errors, our method of moment estimator should be relatively

robust to this problem. However, to assuage concerns, Table 4, Panel B presents our main specifications for those studies with a first stage F-statistic greater than 10, as constructed by the strength of the policy impact on the change in school spending. The results are very similar to our main results. Moreover, the dark orange bars in Figure 5 present results for those studies with a first stage F-stat>20 and reveal very similar results.

2. **Clustering similar policies:** While we are careful to include a single overall population estimate for each study, one may worry that some studies are based on the same underlying policies and should not be treated as independent.[55] We present our main meta-analyses, using a conservative approach to assigning dependence between estimates of the same policies (across different studies) by clustering those estimates as if they stemmed from the *same* study.[56] We assign dependence for studies of an Ohio capital subsidy program (Conlin and Thompson (2017), Goncalves (2015)), Michigan's Proposal A (Hyman (2017), Papke (2008), Roy (2011)), recent School Finance Reforms (Lafortune et al. (2018), Brunner et al. (2020)), and the introduction of Title I (Cascio et al. (2013), Johnson (2015)).[57] These estimates are in Panel C of Table 4. The estimated effects are very similar to our main estimates.

3. **Coding Proficiency Rates:** To make all test score estimates comparable, we converted reported effects into standardized effects. This is common practice for tests that are given on different scales, but less common for test score outcomes such as proficiency rates. For these outcomes, we reported standardized proficiency rate changes by dividing the effect by the student-level standard deviation of the proficiency rate $\sqrt{p(1-p)}$, where p is the proficiency rate. Improvements in student outcomes above or below the proficiency threshold may lead to very small changes in the proficiency rate, even if they reflect large changes in standardized raw scores. Or conversely, concentrated changes right around the proficiency threshold may appear much larger as proficiency rate increases than they reflect changes in standardized raw scores. As such, one may worry that our modelling of outcomes for these studies may skew our results. To assess this, we estimate test score models that remove the 3 studies that report effects on proficiency rates. We plot this effect and the confidence interval in the blue bar on the left panel of Figure 5. Dropping these studies has no appreciable effect on our results – indicating that this modelling choice does not affect our conclusions in a meaningful way.

4. **Combining Effects**: For our main analysis we seek to have one single effect per study-outcome. As such, in many cases we combine impacts across subjects, grade levels, and populations making different assumptions about the correlation between effects. To ensure

[55] It is worth noting that while several paper examine the effect of school finance reforms, these do not all overlap. For example, Jackson et. al. study reforms between 1972 and 1990, while Lafortune et. al. examine reform starting in 1991. In such cases, we treat these estimates as independent.

[56] We use the "study()" option in the "robumeta" Stata command.

[57] Note that Lafortune et al. (2018) and Candelaria and Shores (2019) both examine recent school finance reforms, but look at different outcomes are are therefore not included in the same regression models.

that these assumptions do not drive our conclusions, we re-estimate combined studies under very different assumptions and show that they all yield very similar results. We summarize these alternative approaches below.

Our main analysis assumes 0 correlation between independent effects (across grades or populations), but these correlations could reasonably lie between 0 and 0.5. Our main analysis assumes 0.5 correlation between dependent effects (math/reading), but the correlations between dependent effects could reasonably range from 0.25 to 0.75. To show the practical importance of these assumptions on our estimates, we estimate our main models assuming all four combinations of the upper and lower bound assumed correlations. We plot the resulting estimates in grey, blue, green, and pink bars in Figure 5.[58] The stability of our results indicates that our main estimates and conclusions are largely insensitive to reasonable assumptions about the correlations between effect across subjects, grades, and populations.

5. **Capital Depreciation:** To directly compare the effects of operational and capital spending, we depreciate capital expenditures following commonly accepted accounting approaches. To assess the robustness to different assumptions about length of time capital projects depreciated over, we re-run our main specifications with lower and upper bounds on years across which capital investments are depreciated. At a lower bound, we depreciate buildings at 30 and non-buildings at 10 years. At an upper bound, we depreciate buildings at 50 and non-buildings at 30 years. Additionally, one may also worry that the 7 percent depreciation rate is too high and that the value of the asset should be more evenly distributed over time. To gauge the importance of this choice, we estimate models that assume the the value is uniform over the life of the asset (or that there is no depreciation). We report the estimated effects in Appendix Table A.13. Irrespective of the assumptions made, our estimates of the marginal effect of capital spending are largely similar (ranging between 0.0121σ and 0.0198σ) and cannot be distinguished from each other nor from our preferred approach using formal statistical tests.

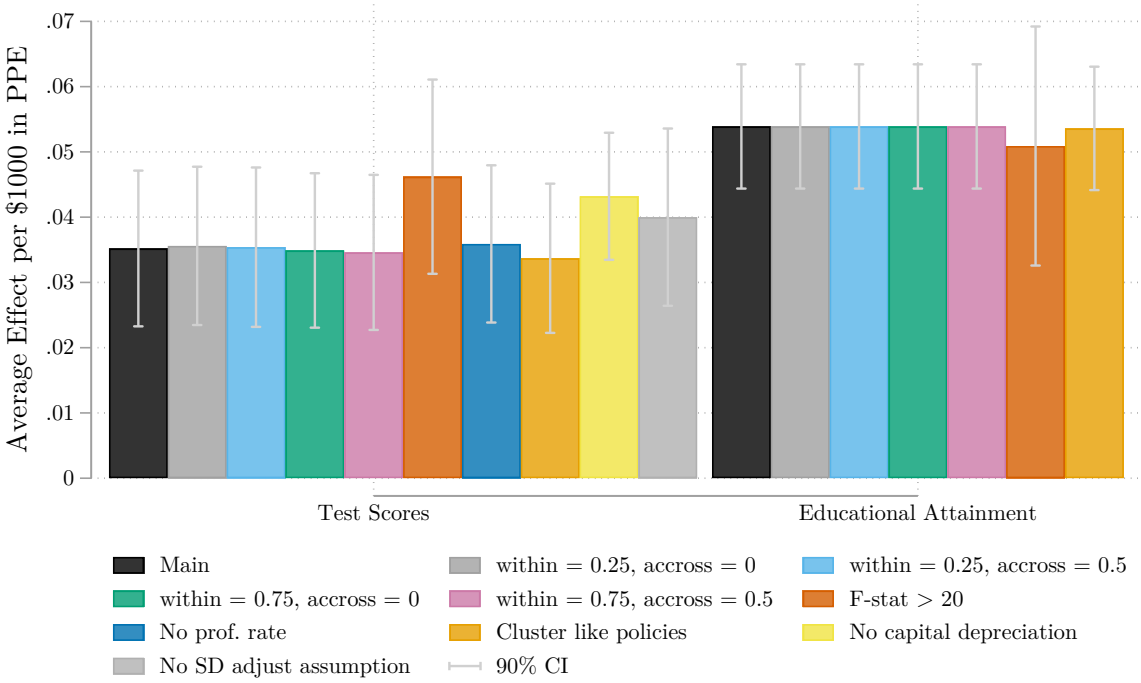
6. **First and Second Stage Standard Errors Correlations:** While many studies report marginal spending effects that we can take directly, for 5 study-outcomes, we must form the IV effects manually using the policy effects on spending and on outcomes.[59] When computing the standard error of this IV estimate, we assume zero correlation between the spending effect and the outcome effect. To provide bounds on the importance of this assumption, we estimate models that assume correlations of -1 and 1 (See Tables A.7 and A.8). The effects are largely unchanged under either assumed upper and lower bound correlations – underscoring the robustness of our meta-analytic average to this assumption.

7. **Student Level Standard Deviations:** For two studies (Kogan et al. (2017) and Rauscher

[58] See full results in Tables A.9, A.10, A.11, and A.12

[59] These include Brunner et al. (2020), Johnson (2015), Kogan et al. (2017), Lafortune and Schonholzer (2019), and Rauscher (2020).

(2020)), we convert school- or district-level standard deviations to standardize the effect size at the student standard deviation level. Because this conversion relies on some assumptions, to assuage any concerns that this drives our results, we drop these two studies and re-estimate our model, resulting in very similar effects to including them (see Figure 5).



Each bar represents a precision-weighted average estimate for each outcome type, comparing our main specification to different modelling assumptions.

Figure 5: Modelling Assumptions

7 Assessing Bias in Individual Studies and Publication Bias

In a meta-analysis, one reports on the average of the reported study effects. However, this reported average may not reflect the true average if (a) the individual studies are biased by confounding or specification errors, and/or (b) the set of studies is somehow systematically selected. We address the possibility of **both** sources of bias and fail to reject that our meta-analytic averages are unbiased.

7.1 Testing for Bias in Individual Studies

A common criticism of meta-analysis is that the end result is only credible if the studies included are themselves credible. For this reason, we are careful to only include studies that employ methods that may yield credibly causal effects. However, one may reasonably worry that even these individual studies may still suffer from bias – potentially biasing our meta-analytic average. In this section,

we formalize a discussion of these biases and discuss when they may bias our meta-analytic average. We also present empirical tests to assess the existence and extent of such possible biases. Finally, we also propose a new meta-analytic approach that is robust to the existence of bias in individual studies under certain reasonable conditions.

A Framework For Assessing Confounding Bias

In our setting, there is a concern that the change in outcomes observed reflect not just the effect of school spending *per se*, but also other factors. This would occur if there were violation of the exclusion restriction as laid out in Section 4.3. In this section we lay out a framework within which to think about such violations, clearly define when such violations may lead to a biased meta-analytic average, and motivate an alternative estimation approach that can uncover average marginal spending effects even when biases may influence the meta-analytic average. For ease of exposition, we abstract away from treatment heterogeneity.

Consider a single outcome y. The change in the standardized outcome due to policy j is Δy_j, which is a function of the change in spending caused by the policy $\Delta\$_j$, plus some noise υ_j, plus possible bias b_j. Where the average mean effect is μ, the observed policy effect on outcome y is:

$$\Delta y_j = \underbrace{(\mu \times \Delta\$_j)}_{Effect of Spending} + \underbrace{\upsilon_j}_{Noise} + \underbrace{b_j}_{Bias} \tag{6}$$

To compute a comparable statistic for each policy/paper, we use each study's marginal effect:

$$\hat{\mu_j} \equiv \frac{\Delta y_j}{\Delta\$_j} = \mu + \frac{\upsilon_j}{\Delta\$_j} + \frac{b_j}{\Delta\$_j} \tag{7}$$

This is the true average marginal effect, plus the error to treatment ratio, plus the bias to treatment ratio. Where w_{yj} is the weight for study j for outcome y, our meta-analytic average ($\hat{\theta}_{pw}$) is a weighted average of each study's reported standardized effect as below:

$$\hat{\theta}_{pw} = \frac{\sum(\frac{\Delta y_j + upsilon_j}{\Delta\$_j})w_{yj}}{\sum w_{yj}} \equiv \underbrace{\mu}_{\text{True Average}} + \underbrace{\frac{\sum(\frac{\upsilon_j}{\Delta\$_j})w_{yj}}{\sum w_{yj}}}_{\text{Average of Noise Ratio}} + \underbrace{\frac{\sum(\frac{b_j}{\Delta\$_j})w_{yj}}{\sum w_{yj}}}_{\text{Average of Bias Ratio}} \tag{8}$$

The observed average is comprised of three pieces; the true effect, the average of the random noise ratios (noise divided by the change in spending) across all the papers, and the average of the bias ratio terms (bias divided by the change in spending) across all the papers. Equation 8 makes clear that the meta-analytic average is an unbiased estimate of the true average (i.e., $E[\hat{\theta}_{pw}] = \mu$) so long as (1) the average of the noise terms is equal to zero in expectation, and (2) the average of the bias terms is equal to zero in expectation. The first condition implies that while some studies' impacts may be overstated due to measurement error or sampling variability, others will be understated for the same reasons so that on average the errors cancel each other out.

So long as there are enough studies in the pooled sample and the random errors are unrelated to the policy-induced spending change, this condition will be satisfied. The second condition is less straightforward. It would trivially be satisfied if the individual studies are themselves unbiased. However, *even with bias in individual studies*, the second condition would hold if some studies' impacts are biased upward while others are biased downward so that the average bias is zero *and* the bias is unrelated to the policy-induced spending change so that the *average* bias ratio is approximately zero. We will present empirical evidence that this holds in our setting.

Differences by Strength of The First Stage

It is known that biases due to violations of the exclusion restriction tend to be more severe when the first stage relationship is weak (Bound et al. (1995) and Conley et al. (2012)). In our context, one can see this clearly because the individual bias-ratio for study j represented by $b_j/\Delta\$_j$ in equation (7) is smaller for policies that generate larger changes in spending. It follows that if biases exist in the included studies, the marginal spending effects should be systematically different (a) among studies that have strong first stages, and (b) among studies based on policies that generate larger versus smaller spending changes.

We examine this in two ways. First, we show the main effects based on studies that have first stage F-statistics over 20 (Table A.1). The results are very similar to models that have first stage F-statistics above 3.85 and above 10 (Table 4) – suggesting minimal bias in the individual studies. As a second test, we examine if the marginal policy impact varies by the magnitude of the spending change induced by the policy. If there were biases (which one expects to be larger in studies with weaker first stages), then the average marginal effects would be larger for small spending changes than for larger spending changes. We test this by regressing the marginal effect of the study against the magnitude of the spending change (see associated scatterplot in Figure A.8). Such a model yields a slope of 0.000012 (p-value of 0.259) for test scores and slope of -0.00002 (p-value = 0.2390) for educational attainment outcomes – indicating no relationship between the marginal effect and the size of the spending change. While these tests are not dispositive on their own, they suggest that the individual studies included (which were specifically chosen *because* they are credibly causal) are by-and-large not appreciably biased on average.

An Approach to Testing For and Removing Bias

The test above suggests that the meta-analytic average likely does not suffer from considerable bias. However, taking the possibility of bias seriously, we present a novel approach to estimating a meta-analytic average that is robust to the existence of the bias laid out in Equation (8) *even if the average of the bias is non-zero*. The meta-analytic average in Equation (7) is an estimate of the average marginal effect across all papers. However, the equation predicting the policy effect on outcomes laid out in Equation (6), reveals that one could also estimate the average marginal effect of *differences* in spending increases by estimating the relationship between the policy-induced changes in outcomes and the policy-induced changes in spending. Equation (6) indicates that a

regression of the change in outcomes for a given policy against the change in spending may yield an estimate of the true average under certain conditions. Abstracting from precision-weighting, the simple linear regression of (6) would yield:

$$\theta_{diff} = \mu + \frac{cov(v_j, \Delta\$_j)}{var(\Delta\$_j)} + \frac{cov(b_j, \Delta\$_j)}{var(\Delta\$_j)} \tag{9}$$

This difference-based approach is a consistent estimate of the true pooled average so long as the random errors are unrelated to the change in spending change induced by a policy and the bias in each study is unrelated to the spending change induced by a policy. Importantly, the difference-based approach does not require that the individual studies be unbiased (which is needed to believe any individual study), nor does it require that the biases in the individual studies average out to zero (which is needed to believe the meta-analytic average), but relies on a weaker identifying assumption that the bias in individual studies in unrelated to the spending changes induced by the policy under study.

We argue that this weaker identification condition is plausible, and we test it empirically.

- **Test 1:** Some difference-in-difference based studies may be biased due to a violation of the common trends assumption (Rambachan and Roth (2020)), studies using regression discontinuity designs may have bias due to extrapolation away from the cutoff point, and credible instrumental variables-based studies may have modest violations of the exclusion restriction (Conley et al. (2012)). Because some of our included studies may be underpowered, such violations may not have been detected. This motivates a simple test. If underpowered studies are less able to detect bias, then in the presence of bias, well-powered studies will be less susceptible to bias. We can assess the importance of this bias by seeing how robust our estimates are to the exclusion of underpowered studies. That is, we estimate models only among studies that would have detected (based on the standard error) our main meta-analytic averages at the 5% level. Using this approach, we obtain a test score effect of 0.0312 and an educational attainment effect of 0.0529 – both similar to our main estimates (see Table A.14).

- **Test 2:** There is no reason to expect that bias of this sort would be correlated with the policy effect on spending. However, the most plausible cause for concern regarding correlated bias is for policies that involve voluntary adoption. One may expect that places that voluntarily implement policies that lead to larger spending increase also are more likely to do other things that improve student outcomes. Such dynamics would generate bias correlated with the spending increase and would inflate the marginal estimate. While we cannot entirely rule out this form of bias, because we *can* distinguish studies that rely on variation induced by the voluntary adoption of policies, we are able to test for its potential presence. Specifically, we compare the average marginal effect for studies that rely on a new policy implementation (e.g., budget-increasing referenda) versus those that rely on variation conditional on policies being in place (e.g., differential impacts of the recession or fluctuating student enrolment). We find that studies based on a voluntary policy adoption are similar to other studies and

(See Table A.6), suggesting little bias of this form.

Because the meta-analytic average may be biased by b while the difference-based estimate is not, the extent to which the difference-based estimates differ from the meta-analytic averages may be indicative of systematic bias in all studies. This motivates a formal test of bias, whether the meta-analytic average differs from the difference-based estimate (i.e., that $\hat{\theta}_{diff} = \hat{\theta}_{pw}$). **Note that while this is a useful test, it comes with an important caveat.** The estimators may differ even when there is no bias *if any treatment heterogeneity is correlated with the size of the spending change.*[60] Because bias is not the only reason the meta-analytic average and the difference-based estimates may differ, one should take equality of effects as compelling evidence of no bias, but should not take differences in these estimates as an indication of bias.

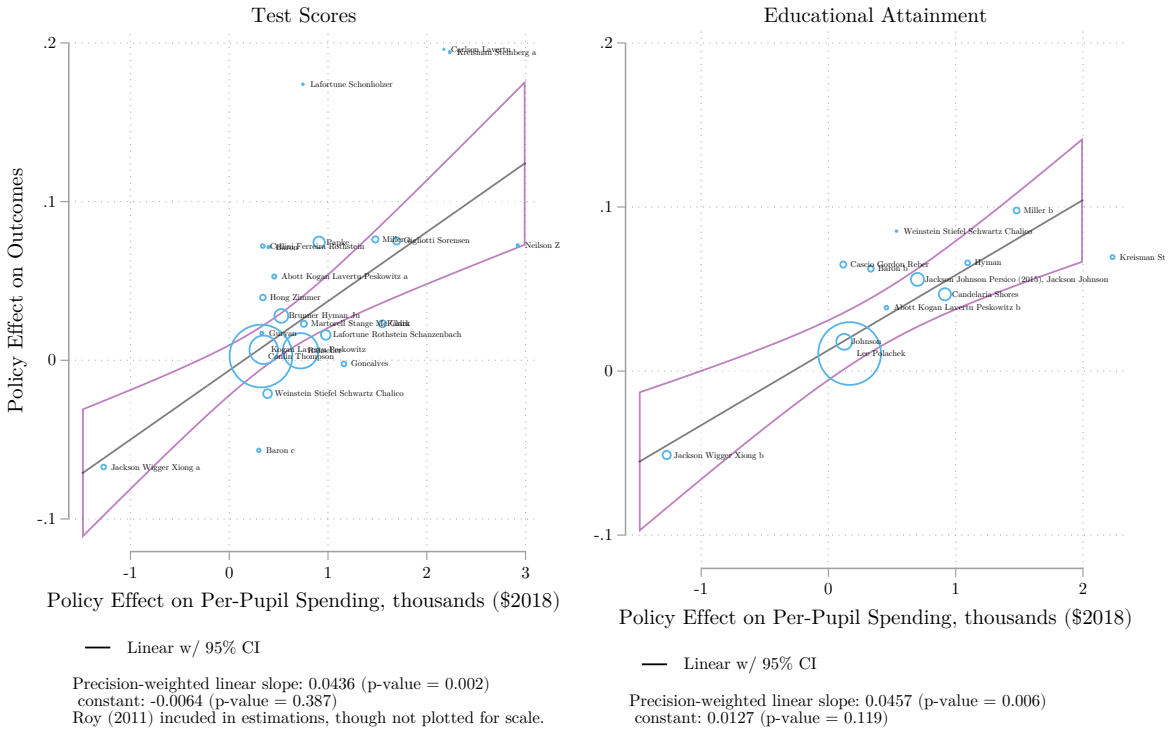


Figure 6: Policy Impacts against Increase in Spending

To assess this in our setting, in Figure 6 we plot the raw, standardized overall effect of each policy on student outcomes against the change in per pupil expenditures ($2018) caused by the

[60] To give a concrete example, image that there were only two studies, of Policy A and Policy B. Policy A increases per-pupil spending by $1000 and test scores by 0.05σ (leading to $\mu_A = 0.05$), while Policy B increases per-pupil spending by $2000 and test scores by 0.04σ (leading to $\mu_B = 0.02$). Both policies have a within-study positive relationship between school spending and test scores (so that $\hat{\theta}_{pw} > 0$). However, the policy with the larger spending increase (Policy B) had a smaller improvement in test sores, so that the difference-based relationship is negative (i.e, $\hat{\theta}_{diff} < 0$). While this may seem counter-intuitive, if there is some correlation between the size of the policy and other contextual factors that determine policy efficacy, this could occur.

the same policy.[61] Each study is represented by a circle, and larger circles indicate more precise outcome estimates. We also plot the fitted values from a precision-weighted regression relating the two, along with the 95 percent confidence interval. There is a clear positive relationship between the size of the spending increase caused by a policy and the increase in outcomes associated with that policy. Using random effects meta-regression, the slope is 0.0436σ/\$1000 for test scores and 0.0457σ/\$1000 for educational attainment – both significant at the 0.01 level. Remarkably, for both outcomes, one fails to reject that the averages of the *within-study* relationships are the same as the *across-study* relationships at the 5 percent significance level.[62] This suggests that, for both outcomes, the documented positive causal relationships between school spending and outcomes are robust. For both test scores and educational attainment, those policies that lead to larger spending increases also lead to larger outcome improvements, on average, and the magnitudes of the differences across policies are similar to those documented within studies. To ensure that our finding is robust, we conduct the same tests (1) excluding studies for which we were forced to make assumptions about the size of the policy effect of spending, and (2) excluding (Jackson et al., 2021) which is an influential point in the model (given its large negative spending effect). As we show in Table A.15, our finding is robust across these alternate specifications.

A Direct Test of the Exclusion Restriction

The difference-based model allows for a direct and intuitive test of the exclusion restriction on average. Specifically, the exclusion restriction is that the only mechanism through which the policies examined affect outcomes is through school spending. If this condition holds, the regression line relating the effect of the policy on outcomes to the effect of the policy on spending should go through the origin. That is, the regression model should predict that a policy that has no effect on school spending should have no effect on outcomes. One can see this mathematically by the fact that the constant term in (6) reflects the average of the noise plus the average of the bias. Given that the average of the noise is zero in expectation, this largely reflects the average of the bias. This is a simple test that the constant in the regression is zero. For test scores, the constant is -0.0064 with a p-value of 0.387, while for educational attainment is is 0.0127 with a p-value of 0.119. The signs of the constants are different for the two outcomes, suggesting no systematic bias. Taken together, the data suggest that the exclusion restriction is satisfied for both outcomes.

[61] There are 6 studies which report policy effects on student outcomes translated into \$1000-increases, already having made assumptions about the linear relationship between effect size and per-pupil spending change. For these studies, if possible, we capture the reported average policy effect on per-pupil spending, and adjust the reported policy effect on outcomes assuming linearity in the dollars-effect relationship (Gigliotti and Sorensen (2018), Guryan (2001), Jackson et al. (2021), Kreisman and Steinberg (2019)). For the two papers that study Michigan's Proposal A (Hyman (2017) and Roy (2011)), there is no one clear policy effect on per-pupil spending, and we rely on effects-per-\$1000 as reported. In Figure 6 we plot and report regression results for all studies—adjusted for the four we can adjust. Our results do not change appreciably when we exclude those studies which do not report one average policy effect on per-pupil spending.

[62] We perform two-sample unpaired t-tests for the hypothesis of equality of the pooled meta-analytic average effect and the slope relating the policy-induced spending changes to the policy related impacts on outcomes.

7.2 Publication Biases

Our analysis may be biased if certain kinds of studies – especially those which find no effect of a policy or intervention – are systematically not published. There are two kinds of publication biases that one may worry about in our context. First, journals may be less likely to publish studies that are not statistically significant. If so, assuming that there is an overall positive effect, those studies with larger positive impacts (and therefore larger t-statistics) will be more likely to be published – such that the average among published studies may overstate effects. Second, if researchers and journals are more likely to publish results consistent with "preferred" results, precisely estimated impacts of all signs will be published (because they are credible), but imprecise studies (where the results are are more ambiguous) of the non-preferred sign will be disproportionately not published. This would lead to a meta-analytic average biased toward the preferred result. We conduct several tests to assess the extent to which these are a concern in our setting. We visually represent estimates from these approaches in Figure 8, describe the tests in more detail in Section A.7 (Table A.21), and summarize them here:

1. One simple approach to assessing publication bias is to compare the average estimates of published and unpublished studies (Lipsey (2009)). In a metaregression, we observe no difference in estimated effects by publication status. If certain kinds of studies were more likely to be published, then these two groups would differ – but this is not the case in our sample.

2. We compare the average impacts of studies published in the most elite journals (where selection biases may be most severe (Brodeur et al. (2016)) to other journals, and find no evidence of differences by journal prestige.

3. To assess whether there is a bias toward statistically significant impacts among the included studies, we show that there is no excess density (i.e., overrepresentation) of studies right at the significance threshold (i.e., t-statistic of 1.96). A histogram of all studies shows slightly *less* density above the significance threshold (Figure A.5) and regression evidence uncovers no indication of a discontinuous shift in density at that threshold (Table A.23).

4. To explicitly adjust for bias against insignificant impacts, we implement the selection adjustment described in Andrews and Kasy (2019). They show how bias from selective publication can be corrected if the probability of publication, as a function of a study's results, is known. They propose estimating the publication probabilities (based on the t-statistics) for studies, and using these to produce bias-corrected estimators and confidence sets. Using estimated publication probabilities (allowing for differences between significant and not-significant studies), this approach re-weights the distribution of studies to account for differences in publication probability (up-weighting studies that are least likely to be observed). When allowing for a discontinuous change in publication probability at t-statistics of 1.96 and -1.96, their approach yields similar estimates to our preferred model (Figures A.6 and A.7).[63]

[63] We also follow Mathur and VanderWeele (2020) and adjust our estimates assuming extreme levels of selection

5. To examine whether there is evidence of bias against imprecise studies with a negative sign, we test for asymmetry in a plot of study impacts against their precision. In a stylized world, with no publication bias, a scatter plot of study impacts and precision of each study should be symmetric around the grand mean (Borenstein (2009)). However, with publication bias, the scatter plot around the grand mean will be asymmetric – suggesting that there are some "missing" studies. In this stylized world with publication bias, while all of the most precise studies will be published, there may be an over-representation of published imprecise estimates in the "desired" direction and no (or few) published imprecise estimates in the "undesirable" direction. Figure 7 plots the study effects depicted by the black circles. Both outcomes *do* indicate some asymmetry among very imprecise studies (i.e., there are some very imprecise positive estimates but few imprecise negative estimates) – suggestive of possible publication bias. While it is important to note that our approach uses precision weights, which yields results relatively robust to missing imprecise estimates, we account for possible publication bias in three ways. We show the result from each approach in Figure 8 and Table A.21. In sum, using all three approaches to potential publication bias yields estimates within the 90% confidence interval of our main specification, and for each one rejects that the pooled average is zero.

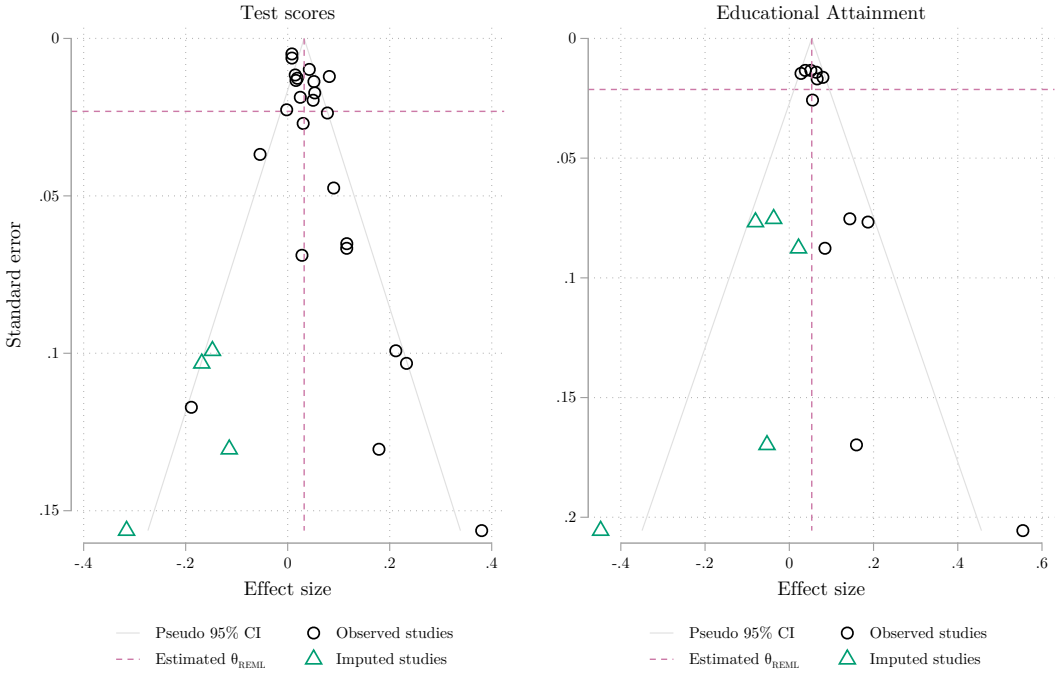


Figure 7: Funnel Plots

to report "worst case" scenario lower-bound estimates. Under selection of this form, test score impacts fall by less than 15 percent, educational attainment impacts fall by only 22 percent, and both remain positive and significant at the 5 percent level.

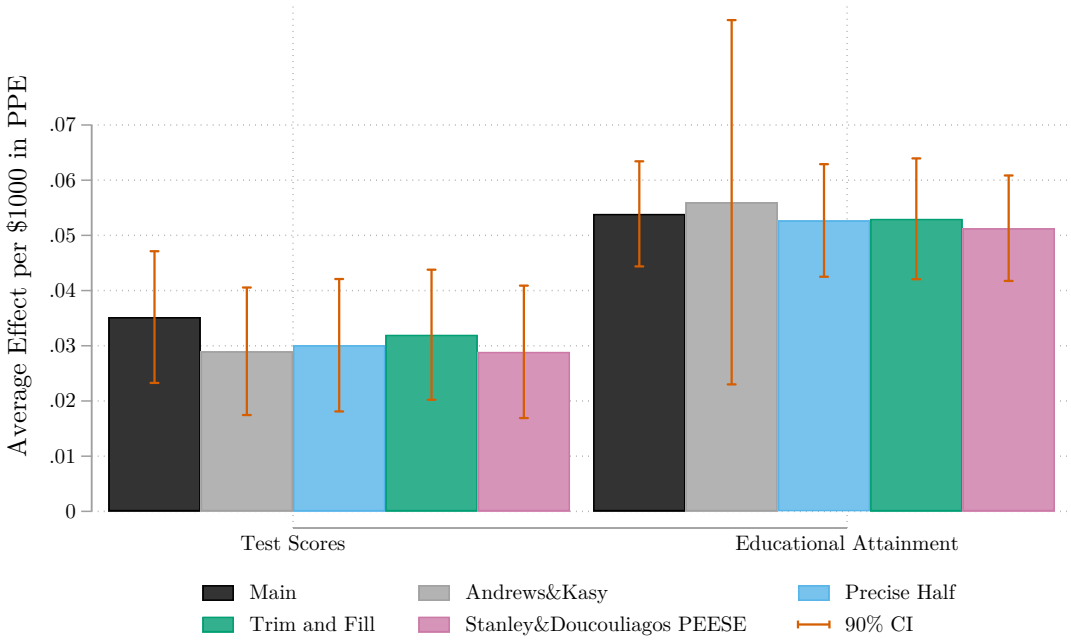
(5.1) First we implement the "trim and fill" method (Duval and Tweedie (2000)). Specifically, to create symmetry in the scatterplot, the "trim and fill" approach imputes additional "missing" studies. The imputed studies are depicted by the green triangles in Figure 7. The method imputes four "missing" studies of test score outcomes and five for educational attainment – all but one of which are negative and rather imprecise. For both outcomes, the re-estimated pooled effects are very similar to our original estimates including all observed estimates.

(5.2) Second, we estimate our main model using a conservative approach of dropping much of the data (Stanley et al. (2010)). Specifically, we drop the least precise half of studies shown above the pink dashed lines in Figure 7. Note that the asymmetry detected lies well below the level of the included studies – suggesting that there would be little bias among the most precise half of studies. Indeed, above this cut-off, for both outcomes the estimates are tightly clustered around the pooled average, and formal tests for asymmetry fail to reject the null of a symmetric distribution. Using the precise half, for both outcomes the results are very similar to our preferred estimates – indicating minimal bias.

(5.3) Finally, we follow both Stanley and Doucouliagos (2014) and Ioannidis et al. (2017) and implement the precision-effect estimate with standard error (PEESE) approach. This approach estimates the relationship between the precision of the estimates and the estimates reported in each study. Under the assumption that the most precise estimates will yield the true relationship, one can empirically model the relationship between the precision of the estimates and the reported estimates and then infer what the most precise estimate would be. In practice, this involves regressing the reported effect on the square of its precision and taking the constant term as the bias-adjusted estimate.[64] Using this approach, for both outcomes, our results are very similar to our preferred estimates – indicating minimal bias.

While no single test can entirely rule out publication bias, taken as a whole the empirical evidence is consistent with minimal bias. That is, across several empirical tests and adjustments for potential publication bias, we find little evidence that our estimates are appreciably impacted by publication bias. Indeed, in all models that adjust for possible publication bias, the point estimates lie within the confidence interval for our main estimate. Given the consistent pattern of results (i.e., 90 percent of study impacts are positive), the fact that publication bias is unlikely to explain our positive overall association is not entirely surprising. The robustness of our effect is also driven by the fact that we employ precision-weighed estimates that down-weight those studies most susceptible to bias. We note that there is no prefect test for publication biases and we cannot entirely rule out the possibility of selection biases in ways these tests are unable to detect.

[64] This approach has been found to preform well in simulations.



Each bar represents a precision-weighted average estimate for each outcome type, comparing our main specification to different approaches to account for potential publication bias.

Figure 8: Four Approaches to Publication Bias

8 Testing For Additional Patterns

8.1 Assessing Heterogeneous Effects by Income Level

An important policy question is the extent to which school spending impacts vary for students from more or less economically advantaged backgrounds. While some studies document larger policy impacts for low-income students (or schools and districts that enroll large shares of low-income students), because many policies may lead to larger spending increases for low-income students (such as many school finance reforms), the policy effect may reflect a combination of differences in spending changes experienced across income groups and differences in the marginal response to spending changes across income groups.

We disentangle these two channels by exploiting the fact that some studies provide separate estimates of policy impacts by income status, and some policies (such as Title I) are targeted to schools that enroll large shares of low-income students. Because some studies report impacts by the income status of the student, while others report impacts by the income status of the school or district, low-income estimates are not perfectly comparable across studies. As such, while the students informing the low-income estimates will disproportionately be from low-income families, the share of low-income students can vary across studies. This introduces a kind of measurement error that may bias us away from detecting significant impacts. Another source of error stems from the fact that the definition of low-income status differs across studies – some define low-

income as being in the bottom quintile of the income distribution, while others define low-income based on free-lunch eligibility.[65] Changing income distributions across time further complicates comparisons. Caveats aside, the question is sufficiently important that the hypothesis is worth testing, albeit imperfectly.

To avoid confounding differences in spending changes with differences in responsiveness to spending changes, we compute marginal spending impacts for low-income and non-low-income groups separately for those studies which report effects by income status.[66] We first perform a simple coin test analysis for the 15 study-outcome combinations that provide impacts by income status (see Table A.16).[67] Of these 15 studies, 11 have larger marginal impacts for the low-income groups. The likelihood of observing this many or more studies with this pattern by random chance (under a null hypothesis of no difference) is just under 6 percent, or 1 in 17 – suggesting that marginal impacts are larger for low-income groups than non-low-income groups. When looking at the outcomes separately, 6 out of 8 test score estimates are larger for low-income groups, and 5 out of 7 educational attainment estimates are larger for low-income groups – suggesting that the pattern may be stronger for test scores than for educational attainment.

We use meta-regression to quantify the magnitude of these differences. For studies that report impacts by income level, we compute separate estimates of μ_{yj} by income. To connote this, we add the subscript inc such that $\mu_{yj,inc}$ is the effect of an increase in per pupil spending of \$1000 (sustained over four years) for study j on outcome y for population $inc \in \{average, high, low\}$. We then estimate a random effects meta-regression, as described in Equation (10), where each study-outcome is weighted by the inverse of its precision: [68]

$$\mu_{yj,inc} = \theta + (LowIncome_{j,inc} \times \beta_1) + (NonLowIncome_{j,inc} \times \beta_2) + \delta_{j,inc} + \epsilon_{j,inc} \qquad (10)$$

The variable $LowIncome_{j,inc}$ is equal to 1 for observations pertaining to a low-income population, which we define in two ways (specified below), $NonLowIncome_{j,inc}$ is equal to 1 for observations pertaining to a higher-income population. β_1 and β_2 indicate the *difference* between the effect for the average student and those from low-income populations and non-low-income populations, respectively. Because those studies that report impacts by income may incidentally differ from those that do not (particularly if the number of studies is small) to avoid confounding differences across studies with differences in responsiveness by income, we control for an indicator δ for whether the study reports estimates by income level. We report results in Table A.17, which shows consistently lower estimated effects for economically advantaged populations compared to the average overall population, and a consistent pattern of larger impacts for less economically advantaged populations

[65] We detail how studies define low-income in Table A.4.

[66] For two papers, Baron (2021) and Goncalves (2015), we capture low-income but not non-low-income estimates, so this analysis compares their low-income to overall estimates.

[67] Two additional studies examine impacts on achievement or attainment gaps by income (Biasi (2019) and Card and Payne (2002)). These studies do show that school spending policies reduced gaps in student outcomes by income status, but they do not allow one to disentangle spending differences from response differences.

[68] Because this model includes multiple estimates per study (if the study reports effects for different income levels), we adjust for dependent effects within studies following Hedges et al. (2010).

than for economically advantaged populations.[69]

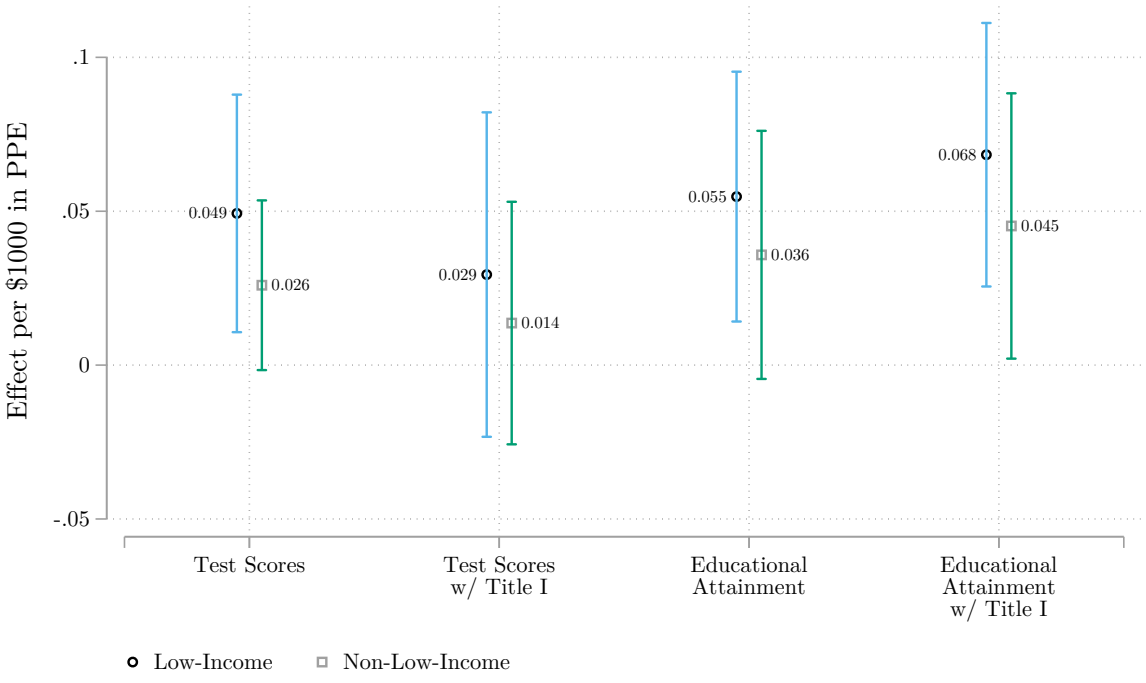
We summarize the results in Figure 9 by plotting the estimated marginal effects for low-income ($\theta + \beta_1$) and not-low-income ($\theta + \beta_2$) groups, along with 95% percent confidence intervals, from the regression models. We include two different categorizations of low-income. Our first categorization includes only those studies with distinct estimates for low-income populations (models 1 and 3), and our second also includes studies whose overall estimates are of effects of Title I, a program explicitly aimed at providing funding to schools with low-income students (models 2 and 4). In model 1, the marginal test score effect for low-income students is 0.049, and that for non-low-income students is just over half the size at 0.026. While the difference between these two estimates is economically significant, the formal test that these estimates are different yields a p-value of 0.113. Model 2, which expands the definition of low-income to also include overall Title I estimates, shows a similar pattern but lower estimates in general.[70] In this model (using the Title I-inclusive definition of low-income), the low-income estimate is more than twice as large as the for the non-low-income group, but the formal test that these estimates are different yields a large p-value of 0.3 – suggestive, but not conclusive evidence of differences by income status.

Our results for differential impacts by income status for educational attainment are directionally similar. Using our more restrictive definition of low-income (model 3), we find no statistically significant difference between effects of spending for low-income populations and for non-low-income populations. However, the marginal effect for low income groups is 0.055 which is more than 50 percent larger than that for the non-low-income group (which is 0.036). With an expanded definition of low-income to also include overall Title I studies (last set of estimates), the results are similar. While the results do indicate that the marginal effects are smaller for the higher-income groups, a formal test of whether the marginal impacts on educational attainment differ across income groups fails to reject the null that there is no difference.

Taken as a whole, these results show lower marginal effects for economically advantaged populations compared to the average overall population – patterns consistent with marginal school spending impacts varying by socioeconomic status. Importantly, our results suggest that larger policy impacts of school spending is not only due to lower income populations receiving larger increases in spending (which does often happen), but also likely reflects more responsiveness to the same increases in spending by less economically advantaged student populations compared to more economically advantaged populations. In Tables A.18 and A.19, we show that these results are robust to clustering like studies and restricting low-income to only include those studies for which estimated impacts on spending are clearly reported separately by income, respectively.

[69] In Table A.18, we present our main models with a conservative approach to account for possible correlations across studies which identify changes form the same policies by clustering estimates of the same policy as if they stemmed from the same study. Our results become more pronounced and precise.

[70] As a robustness check, in Appendix Table A.19 we estimate models that exclude two studies for which the estimated impacts on spending were not clearly reported separately by income – potentially biasing our estimates. These studies are: Brunner et al. (2020)), and Goncalves (2015). The results are very similar.



Low-Income - Non-LI = 0 (p-val):
Test Scores (0.113), Test Scores w/ Title I (0.305), Educational Attainment (0.493), Educational Attainment w/ Title I (0.396)
See Table A.17 for detailed regression results.

Figure 9: Low-Income versus Non-Low-Income Estimates

8.2 Are There Systematic Differences by Geography?

Given the wide range in average per-pupil spending across regions of the United States, and differences in institutional contexts between urban and rural communities, we consider whether impacts of spending vary by specific geographic characteristics. For studies which allow for categorization, in Table A.20, we document that there do not appear to be systematic differences across geographic dimensions. First, we test for whether there are differences between multi-state studies and studies based on smaller levels of geography. In a meta-regression, with a "multi-state" indicator, the point estimate is a statistically insiginificant 0.0118. We also test for differences across different regions, including regional indicators. In this model, none of the regional indicators are significant,and the point estimates are small in magnitude. Finally, we explore whether there are differences by urbanicity by including indicators for rural and urban. Such models suggest that marginal effects do not vary systematically by urbanicity. An implication of these results is that the variability in marginal effects are not explained by geography and are a result of other factors.

8.3 Do Longer-Run Impacts Increase with Exposure?

Given that learning is a cumulative process, one would expect that the benefits to increased spending would increase with exposure. Indeed, to make studies comparable to each other, we assumed that

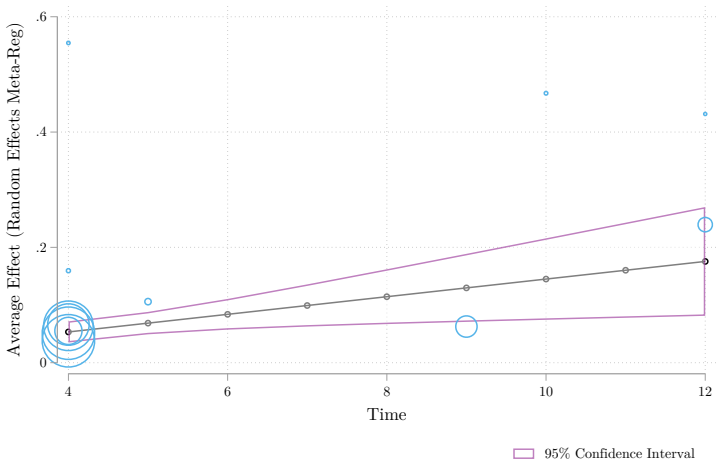


Figure 10: Educational Attainment by Years of Exposure

the impacts are linear in the years of exposure and adjusted all estimates to reflect four-year impacts. Because some studies of educational attainment outcomes show the effects of four year of exposure to a spending increase, while others present effects of 9 years and 12 years, we can test if our assumption is reasonable.

First, we plot the estimates (not adjusted for exposure) on educational attainment in Figure 10. We represent more precise studies with larger circles. There are several relatively precise estimates pertaining to four years of exposure centered around 0.15. There are two observations pertaining to 9 and 10 years of exposure that are both above the center of the four-year impacts, and two studies (one very imprecise large estimate) that relate 12 years of exposure to increased spending with even larger overall impacts. The pattern indicates larger overall impacts for estimates that relate to more years of exposure (per $1000 per-pupil spending increase).

To formally test this notion, we run a meta-regression on the years-unadjusted effects (denoted $\ddot{y}_j$), and include the years of exposure underlying each estimate as a covariate. If impacts are increasing with years of exposure, as suggested visually, then studies that report the impacts of more years of exposure should report larger educational attainment impacts. In addition, we can directly test if the average four-year effect (the shortest exposure reported) is similar to four times the average impact of an additional year of exposure. This is a direct statistical test of the notion that the educational attainment impacts increase linearly with years of exposure. In a regression this is achieved by estimating equation (11) by random effects meta-regression:

$$\ddot{y}_j = \alpha + \beta \times (Exposure_j - 4) + \epsilon \tag{11}$$

In this model $(Exposure_j - 4)$ is the years of exposure to the spending change minus 4 so that α is the average estimated 4-year impact (identified off those studies with four years of exposure). The parameter β is the increase in impact associated with each additional year of exposure. The formal test for whether there in greater educational attainment with more years of exposure to increased

spending is whether $\beta = 0$. This test yields a p-value of 0.03 – suggesting that the effects increase with years of exposure. As described above, a formal test for linearity is whether $\alpha - (4 \times \beta) = 0$. This test yields a p-value of 0.79 – suggesting that the impacts may increase linearly with years of exposure. In sum, the data indicate that the educational attainment impacts increase with years of exposure and that the increase is approximately linear in years of exposure. This is both (a) a substantively important result to inform policy, and (b) validates our modelling assumptions.

8.4 Examining Evidence of Diminishing Returns

Under optimizing behaviour, schools would spend their first thousand dollars on inputs that produce the most output, the next thousand dollars on the second most productive input, and so on. If so, school spending would exhibit diminishing marginal returns. Informed by this notion, some scholars hypothesize that the level of school spending in United States is sufficiently high that the marginal impact of spending is approaching zero. To shed light on this, we examine if the marginal impacts of school spending depend on the baseline spending level in the study context. Per-pupil school spending levels have more than doubled in the past thirty years (Hill and Zhou (2006)), and at any given point in time some states spend much more per pupil than others. In principle, studies based on recent policies in high spending states such as New York (e.g., Gigliotti and Sorensen (2018) and Lee and Polachek (2018)) would have smaller marginal impacts on average than studies of old policies (such as the roll-out of Title I in 1965 examined in Cascio et al. (2013)) or in lower-spending states such as Texas (e.g., Martorell et al. (2016)). To assess this, in Figure 11 we plot the marginal spending impact against the baseline spending for all papers. Each circle represents a single study-outcome, and larger circles connote more precise estimates. We also include the precision-weighted linear relationship along with the 95% confidence interval.

The scatter-plot of marginal test score impacts (left) shows little evidence that marginal impacts are smaller at higher baseline spending levels. While there are some large positive marginal impacts at lower spending levels (e.g., Hong and Zimmer (2016) and Roy (2011)), these studies are all very imprecise relative to those with smaller estimated impacts at similar baseline spending levels (e.g., Clark (2003) or Brunner et al. (2020)). A precision-weighted linear regression of the scatter-plot yields a slightly *positive* slope with a p-value above 0.1. The scatter-plot for educational attainment (right panel) follows a similar pattern. There is evidence of larger estimates at very low levels of spending, but these estimates are also imprecise. A precision-weighted linear regression of the scatter-plot yields a slightly negative slope with a p-value above 0.1. For both outcomes, the marginal impacts are remarkably similar across a wide range of per-pupil spending levels. After accounting for the precision of the estimates, there is no evidence of diminishing returns between \$8,000 and \$20,000 per-pupil. Given a national average of \$14,439 per-pupil (NCES (2020)), these patterns suggests that educational spending in the United States is not yet "*on the flat.*"

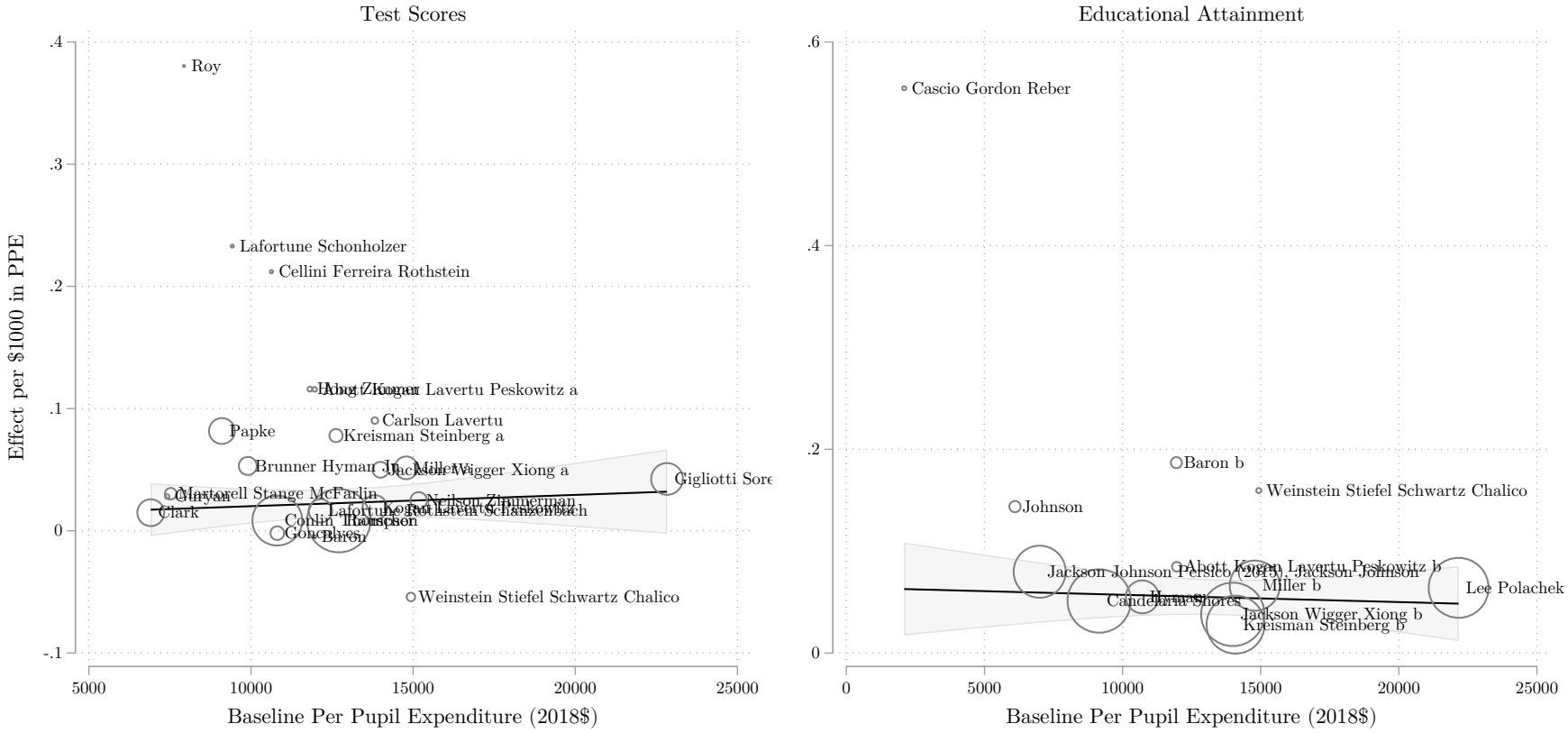


Figure 11: Marginal Impacts by Baseline Spending Level

Note that because education is a very labor intensive field, as wages rise in many sectors, wages for educators will also rise with minimal ability to reduce workers (Baumol and De Ferranti (2012)). This could explain rising education costs that would not represent movement along the productivity schedule (i.e., going from the most to the least productive input) – potentially explaining the constant marginal impacts, on average, across a wide range of spending levels. Another explanation is that, because public educators do not have a profit motive, school spending is not allocated to the **most** productive inputs on the margin, but rather based on rules of thumb or heuristics so that additional monies go toward bundles of inputs that are generally similarly productive.

9 Discussion and Conclusions

We collect and classify all known credible causal studies of the impact of public school spending on student outcomes in the United States. Of these 31 studies, 28 find positive impacts of policies that increased school spending on student outcomes. That is, the most credible evidence to date is extraordinarily consistent with the notion that money does matter. To shed light on magnitudes, we estimate the centers and spreads of the distributions of causal school spending impacts on test scores and educational attainment. *On average*, a $1000 increase in school spending (sustained over four years) increases test scores by 0.0352σ, high-school graduation by 1.9 percentage points, and college-going by 2.65 percentage points. In relative terms, this is a 2.3 percent increase in high school graduation and a 6.5 percent increase in college-going. These within-study relationships hold across studies such that policies that generate larger per-pupil spending increases also tend to generate larger increases in outcomes – bolstering a causal interpretation of these results. We find little indication that these effects are skewed by confounding biases or publication biases.

We find that school spending impacts on educational attainment are larger than on test scores – when benchmarked against the impacts of other interventions – suggesting that using test scores to estimate school spending impacts, while informative, may understate the long-term benefits of school spending. Another key result of this analysis is that marginal school spending effects are very similar across a wide range of baseline spending levels – suggesting little evidence of diminishing returns to school spending at current levels. We present an approach that allows for an economically meaningful direct comparison of the causal effects of large one-time capital spending increases to those of annual (mainly operational) spending increases. We find that capital spending increases take about 5-6 years to materialize into improved outcomes, at which point the marginal effects are about half as large as other forms of school spending. We find little evidence of larger impacts for low-income populations as compared to the overall average population, though we *do* find lower effects for more economically advantaged populations. Overall, we find that the marginal school spending effects are remarkable stable across populations, geography, and time – as such, much of the variability across studies remains unexplained.

Accounting for underlying variability due to context and differences in policy implementation indicates that not all policies will have similar impacts in the future. We find evidence of con-

siderable treatment heterogeneity (i.e., variability unexplained by sampling variability). Using our estimates of the underlying heterogeneity, we "*predict*" that a policy that increases per-pupil spending $1000 for at least four years will lead to positive test-score impacts over 91 percent of the time, and positive educational attainment impacts more than 92 percent of the time. Because we document relatively consistent estimates across a variety of observable dimensions *on average*, further research uncovering *why* impacts are larger in some contexts than others (such as Brunner et al. (2020) and Johnson and Jackson (2019)) may be fruitful.

References

Abott, C., Kogan, V., Lavertu, S., and Peskowitz, Z. (2020). School district operational spending and student outcomes: Evidence from tax elections in seven states. *Journal of Public Economics*, 183:104142.

Abramson, P. (2015). 20th Annual School Construction Report: National Statistics, Building Trends & Detailed Analysis. *School Planning & Management*.

Andrews, I. and Kasy, M. (2019). Identification of and Correction for Publication Bias. *American Economic Review*, 109(8):2766–2794.

Angrist, J. D., Cohodes, S. R., Dynarski, S. M., Pathak, P. A., and Walters, C. R. (2016). Stand and Deliver: Effects of Boston's Charter High Schools on College Preparation, Entry, and Choice. *Journal of Labor Economics*, 34(2):275–318.

Angrist, J. D. and Pischke, J.-S. (2010). The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics. *Journal of Economic Perspectives*, 24(2):3–30.

Bandiera, O., Fischer, G., Prat, A., and Ytsma, E. (2021). Do women respond less to performance pay? Building evidence from multiple experiments. *American Economic Review: Insights*.

Baron, J. (2021). School Spending and Student Outcomes: Evidence from Revenue Limit Elections in Wisconsin. *American Economic Journal: Economic Policy*.

Baumol, W. J. and De Ferranti, D. M. (2012). *The cost disease: why computers get cheaper and health care doesn't*. Yale University Press, New Haven London. OCLC: 796276883.

Beuermann, D., Jackson, C. K., Navarro-Sola, L., and Pardo, F. (2020). What is a Good School, and Can Parents Tell? Evidence on the Multidimensionality of School Output. *NBER Working Paper*.

Biasi, B. (2019). School Finance Equalization Increases Intergenerational Mobility: Evidence from a Simulated-Instruments Approach. Technical Report w25600, National Bureau of Economic Research, Cambridge, MA.

Booker, K., Sass, T. R., Gill, B., and Zimmer, R. (2011). The Effects of Charter High Schools on Educational Attainment. *Journal of Labor Economics*, 29(2):377–415.

Borenstein, M., editor (2009). *Introduction to meta-analysis*. John Wiley & Sons, Chichester, U.K. OCLC: ocn263294996.

Borenstein, M., Higgins, J. P. T., Hedges, L. V., and Rothstein, H. R. (2017). Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*, 8(1):5–18.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with Instrumental Variables Estimation when the Correlation between the Instruments and the Endogenous Explanatory Variable is Weak. *Journal of the American Statistical Association*, 90(430):443–450.

Brodeur, A., Cook, N., and Heyes, A. (2020). Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics. *American Economic Review*, 110(11):3634–3660.

Brodeur, A., Lé, M., Sangnier, M., and Zylberberg, Y. (2016). Star Wars: The Empirics Strike Back. *American Economic Journal: Applied Economics*, 8(1):1–32.

Brunner, E., Hyman, J., and Ju, A. (2020). School Finance Reforms, Teachers' Unions, and the Allocation of School Resources. *The Review of Economics and Statistics*, 102(3):473–489.

Candelaria, C. A. and Shores, K. A. (2019). Court-Ordered Finance Reforms in the Adequacy Era: Heterogeneous Causal Effects and Sensitivity. *Education Finance and Policy*, 14(1):31–60.

Card, D. and Krueger, A. B. (1992). Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States. *Journal of Political Economy*, 100(1):1–40.

Card, D. and Payne, A. A. (2002). School finance reform, the distribution of school spending, and the distribution of student test scores. *Journal of Public Economics*, 83(1):49=82.

Cascio, E. U., Gordon, N., and Reber, S. (2013). Local Responses to Federal Grants: Evidence from the Introduction of Title I in the South. *American Economic Journal: Economic Policy*, 5(3):126–159.

Cellini, S. R., Ferreira, F., and Rothstein, J. (2010). The Value of School Facility Investments: Evidence from a Dynamic Regression Discontinuity Design. *Quarterly Journal of Economics*, 125(1):215–261.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., and Yagan, D. (2011). How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star. *The Quarterly Journal of Economics*, 126(4):1593–1660.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review*, 104(9):2593–2632.

Christensen, G. and Miguel, E. (2018). Transparency, Reproducibility, and the Credibility of Economics Research. *Journal of Economic Literature*, 56(3):920–980.

Clark, M. A. (2003). Education Reform, Redistribution, and Student Achievement: Evidence from the Kentucky Education Reform Act. *PhD diss. Princeton University.*

Conley, T. G., Hansen, C. B., and Rossi, P. E. (2012). Plausibly Exogenous. *Review of Economics and Statistics*, 94(1):260–272.

Conlin, M. and Thompson, P. N. (2017). Impacts of new school facility construction: An analysis of a state-financed capital subsidy program in Ohio. *Economics of Education Review*, 59:13–28.

Davis, M. and Heller, B. (2019). No Excuses Charter Schools and College Enrollment: New Evidence from a High School Network in Chicago. *Education Finance and Policy*, 14(3):414–440.

Dehejia, R., Pop-Eleches, C., and Samii, C. (2021). From Local to Global: External Validity in a Fertility Natural Experiment. *Journal of Business & Economic Statistics*, 39(1):217–243.

DellaVigna, S. and Linos, E. (2021). RCTs to Scale: Comprehensive Evidence from Two Nudge Units. *NBER WP*.

Dobbie, W. and Fryer, R. (2020). Charter Schools and Labor Market Outcomes. *Journal of Labor Economics*, 38(4).

Downes, T. A., Dye, R. F., and McGuire, T. J. (1998). Do Limits Matter? Evidence on the Effects of Tax Limitations on Student Performance. *Journal of Urban Economics*, 43(3):401–417.

Duval, S. and Tweedie, R. (2000). Trim and Fill: A Simple Funnel-Plot-Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis. *Biometrics*, 56(2):455–463.

Dynarski, S., Hyman, J., and Schanzenbach, D. W. (2013). Experimental Evidence on the Effect of Childhood Investments on Postsecondary Attainment and Degree Completion. *Journal of Policy Analysis and Management*, 32(4):692–717.

Figlio, D. N. (1997). Did the "tax revolt" reduce school performance? *Journal of Public Economics*, 65(3):245–269.

Franco, A., Malhotra, N., and Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203):1502–1505.

Frisvold, D. E. (2015). Nutrition and cognitive achievement: An evaluation of the School Breakfast Program. *Journal of Public Economics*, 124:91–104.

Gigliotti, P. and Sorensen, L. C. (2018). Educational resources and student achievement: Evidence from the Save Harmless provision in New York State. *Economics of Education Review*, 66:167–182.

Goncalves, F. (2015). The Effects of School Construction on Student and District Outcomes: Evidence from a State-Funded Program in Ohio. *SSRN Electronic Journal*.

Guryan, J. (2001). Does Money Matter? Regression-Discontinuity Estimates from Education Finance Reform in Massachusetts. *NBER Working Paper*.

Hanushek, E. A. (2003). The Failure of Input-based Schooling Policies. *The Economic Journal*, 113(485):F64–F98.

Hartung, J., Knapp, G., and Sinha, B. K. (2008). *Statistical meta-analysis with applications*. Wiley series in probability and statistics. Wiley, Hoboken, N.J. OCLC: ocn212627347.

Hedberg, E., Pustejovsky, J., and Tipton, E. (2017). robumeta: A macro for Stata.

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin*, 93(2):388–395.

Hedges, L. V., Laine, R. D., and Greenwald, R. (1994). An Exchange: Part I: Does Money Matter? A Meta-Analysis of Studies of the Effects of Differential School Inputs on Student Outcomes. *Educational Researcher*, 23(3):5–14.

Hedges, L. V. and Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin*, 88(2):359–369.

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

Hendren, N. and Sprung-Keyser, B. (2020). A Unified Welfare Analysis of Government Policies*. *The Quarterly Journal of Economics*, 135(3):1209–1318.

Higgins, J., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M., and Welch, V., editors (2021). *Cochrane Handbook for Systematic Reviews of Interventions version 6.2 (updated February 2021)*. Cochrane.

Hill, J. and Zhou, L. (2006). Documentation for the NCES Common Core of Data School District Finance Survey (F-33), School Year 1991–92 (Fiscal Year 1992) (NCES 2007-322).

Holden, K. L. (2016). Buy the Book? Evidence on the Effect of Textbook Funding on School-Level Achievement. *American Economic Journal: Applied Economics*, 8(4):100–127.

Hong, K. and Zimmer, R. (2016). Does Investing in School Capital Infrastructure Improve Student Achievement. *Economics of Education Review*, page 44.

Hoxby, C. M. (2001). All School Finance Equalizations are Not Created Equal. *The Quarterly Journal of Economics*, 116(4):1189–1231.

Husted, T. and Kenny, L. (2000). Evidence on the Impact of State Government on Primary and Secondary Education and the Equity-Efficiency Trade-Off. *The Journal of Law and Economics*, 43(1):285–308.

Hyman, J. (2017). Does Money Matter in the Long Run? Effects of School Spending on Educational Attainment. *American Economic Journal: Economic Policy*, 9(4):256–280.

Ioannidis, J. P. A., Stanley, T. D., and Doucouliagos, H. (2017). The Power of Bias in Economics Research. *The Economic Journal*, 127(605):F236–F265.

Jackson, C. K. (2018). What Do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes. *Journal of Political Economy*, 126(5):2072–2107.

Jackson, C. K. (2020). Does school spending matter? The new literature on an old question. In Tach, L., Dunifon, R., and Miller, D. L., editors, *Confronting inequality: How policies and practices shape children's opportunities.*, pages 165–186. American Psychological Association, Washington.

Jackson, C. K., Johnson, R. C., and Persico, C. (2016). The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms. *The Quarterly Journal of Economics*, 131(1):157–218.

Jackson, C. K., Porter, S. C., Easton, J. Q., Blanchard, A., and Kiguel, S. (2020). School Effects on Socioemotional Development, School-Based Arrests, and Educational Attainment. *American Economic Review: Insights*, 2(4):491–508.

Jackson, C. K., Rockoff, J. E., and Staiger, D. O. (2014). Teacher Effects and Teacher-Related Policies. *Annual Review of Economics*, 6(1):801–825.

Jackson, C. K., Wigger, C., and Xiong, H. (2021). Do School Spending Cuts Matter? Evidence from the Great Recession. *American Economic Journal: Economic Policy*, 13(2).

Johnson, R. C. (2015). Follow the Money: School Spending from Title I to Adult Earnings. *RSF: The Russell Sage Foundation Journal of the Social Sciences*, 1(3):50.

Johnson, R. C. and Jackson, C. K. (2019). Reducing Inequality through Dynamic Complementarity: Evidence from Head Start and Public School Spending. *American Economic Journal: Economic Policy*, 11(4):310–349.

Kendall, M. G., Stuart, A., and Ord, J. K. (1994). *The advanced theory of statistics in 3 volumes. 1 1.* Griffin, London. OCLC: 1071028235.

Kogan, V., Lavertu, S., and Peskowitz, Z. (2017). Direct Democracy and Administrative Disruption. *Journal of Public Administration Research and Theory*, 27(3):381–399.

Kontopantelis, E., Springate, D. A., and Reeves, D. (2013). A Re-Analysis of the Cochrane Library Data: The Dangers of Unobserved Heterogeneity in Meta-Analyses. *PLoS ONE*, 8(7):e69930.

Kraft, M. A. (2020). Interpreting Effect Sizes of Education Interventions. *Educational Researcher*, 49(4):241–253.

Kreisman, D. and Steinberg, M. P. (2019). The effect of increased funding on student achievement: Evidence from Texas's small district adjustment. *Journal of Public Economics*, 176:118–141.

Krueger, A. B. (1998). Reassessing the view that American schools are broken. *Economic Policy Review*, 4(1).

Krueger, A. B. (1999). Experimental Estimates of Education Production Functions. *The Quarterly Journal of Economics*, 114(2):497–532.

Lafortune, J., Rothstein, J., and Schanzenbach, D. W. (2018). School Finance Reform and the Distribution of Student Achievement. *American Economic Journal: Applied Economics*, 10(2):1–26.

Lafortune, J. and Schonholzer, D. (2019). Measuring the Efficacy and Efficiency of School Facility Expenditures. *Working Paper*, page 85.

Lee, K.-G. and Polachek, S. W. (2018). Do school budgets matter? The effect of budget referenda on student dropout rates. *Education Economics*, 26(2):129–144.

Lipsey, M. W. (2009). Identifying interesting variables and analysis opportunities. In Cooper, H., Hedges, L. V., and Valentine, J. C., editors, *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation.

Martorell, P., Stange, K., and McFarlin, I. (2016). Investing in Schools: Capital Spending, Facility Conditions, and Student Achievement (Revised and Edited). *Journal of Public Economics*.

Mathur, M. B. and VanderWeele, T. J. (2020). Sensitivity analysis for publication bias in meta-analyses. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(5):1091–1119.

Matsudaira, J. D., Hosek, A., and Walsh, E. (2012). An integrated assessment of the effects of Title I on school behavior, resources, and student achievement. *Economics of Education Review*, 31(3):1–14.

Meager, R. (2019). Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments. *American Economic Journal: Applied Economics*, 11(1):57–91.

Miller, C. L. (2018). The Effect of Education Spending on Student Achievement: Evidence from Property Values and School Finance Rules. *Working Paper*.

Nagashima, K., Noma, H., and Furukawa, T. A. (2019). Prediction intervals for random-effects meta-analysis: A confidence distribution approach. *Statistical Methods in Medical Research*, 28(6):1689–1702.

NCES, N. C. f. E. S. (2001). Common Core of Data, "Public Elementary/Secondary School Universe Survey," 1999-2000.

NCES, N. C. f. E. S. (2020). The Condition of Education 2020 (NCES 2020-144).

NCES, N. C. f. E. S., 21st Century School Fund, Council, U. G. B., and on School Facilities, N. C. (2016). State of Our Schools: America's K–12 Facilities.

Neilson, C. A. and Zimmerman, S. D. (2014). The effect of school construction on test scores, school enrollment, and home prices. *Journal of Public Economics*, 120:18–31.

OECD (2020). *Education at a Glance 2020: OECD Indicators*. Education at a Glance. OECD.

of Education Data Reporting Office, C. D. (2020). Fingertip Facts on Education in California - CalEdFacts. Technical report.

Papke, L. E. (2008). The Effects of Changes in Michigan's School Finance System. *Public Finance Review*, 36(4):456–474.

Rambachan, A. and Roth, J. (2020). Design-Based Uncertainty for Quasi-Experiments. *arXiv preprint arXiv:2008.00602*.

Rauscher, E. (2020). Delayed Benefits: Effects of California School District Bond Elections on Achievement by Socioeconomic Status. *Sociology of Education*, 93(2):110–131.

Roy, J. (2011). Impact of School Finance Reform on Resource Equalization and Academic Performance: Evidence from Michigan. *Education Finance and Policy*, 6(2):137–167.

Snyder, T., de Brey, C., and Dillow, S. (2019). Digest of Education Statistics 2018. *Institute of Education Sciences*.

Stanley, T. D. and Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias: T. D. STANLEY AND H. DOUCOULIAGOS. *Research Synthesis Methods*, 5(1):60–78.

Stanley, T. D., Doucouliagos, H., and Ioannidis, J. P. A. (2017). Finding the power to reduce publication bias: Finding the power to reduce publication bias. *Statistics in Medicine*.

Stanley, T. D., Jarrell, S. B., and Doucouliagos, H. (2010). Could It Be Better to Discard 90% of the Data? A Statistical Paradox. *The American Statistician*, 64(1):70–77.

Stock, J. H., Wright, J. H., and Yogo, M. (2002). A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments. *Journal of Business & Economic Statistics*, 20(4):518–529.

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, 20(3):375–393.

Tipton, E., Bryan, C., and Yeager, D. (2020). To change the world, behavioral intervention research will need to get serious about heterogeneity. *Manuscript in Preparation*.

Tipton, E. and Olsen, R. B. (2018). A Review of Statistical Methods for Generalizing From Evaluations of Educational Interventions. *Educational Researcher*, 47(8):516–524.

Tyler, J. and Lofstrom, M. (2009). Finishing High School: Alternative Pathways and Dropout Recovery. *The Future of Children*, 19(1):77–103.

van der Klaauw, W. (2008). Breaking the link between poverty and low student achievement: An evaluation of Title I. *Journal of Econometrics*, 142(2):731–756.

Vivalt, E. (2020). How Much Can We Generalize From Impact Evaluations? *Journal of the European Economic Association*, 18(6):3045–3089.

Weinstein, M. G., Stiefel, L., Schwartz, A. E., and Chalico, L. (2009). Doest Title I Increase Spending and Improve Performance? Evidence from New York City. *IESP working paper series*.

A Appendix

A.1 Strength of First Stage

Table A.1: Meta-Analysis, F-stat > 20

	(1) Overall Test Scores	(2) Non-Capital Test Score	(3) Capital Test Score	(4) Overall Ed. Attainment
Average Effect	0.0462***	0.0567***	0.0264**	0.0509***
	(0.00902)	(0.00791)	(0.0132)	(0.0111)
N	13	7	6	4
τ	0.0296	0.0153	0.0277	0
% Cross-Study Var.	0.836	0.579	0.796	0

Standard errors in parentheses

* $p < .1$, ** $p < .05$, *** $p < .01$

A.2 Estimates Captured per Paper

Table A.2: Summary of per-study steps

study	outcome	effect per $1000	$ Δ: source	outcome Δ: source
Abott Kogan Lavertu Peskowitz (2020)	High school graduation	0.0850	$417 (2012$): Table 8 Expend. P.P. Operations, $\leq 5yrs$, Bandwidth $+/-10$	0.0174: Table 8 Grad. Rate, $\leq 5yrs$, Bandwidth $+/-10$, standardized (Table 2 Grad. Rate (4yr), Passed)
Abott Kogan Lavertu Peskowitz (2020)	Test scores	0.1160	$417 (2012$): Table 8 Expend. P.P. Operations, $\leq 5yrs$, Bandwidth $+/-10$	0.066: Table 8 Math/ELA (SDs), $\leq 5yrs$, Bandwidth $+/-10$
Baron (2021)	College enrollment	0.1870	$289.743 (2010$): Figure 1 (b) Total Operational Expenditures, averaged across 1-10yrs Relative to the Election (exact estimates provided by author)	0.195: Figure 2 Panel (d) Log(Postsecondary Enrollment) Year 10 relative to election (exact estimates provided by author), multiplied by baseline rate (.39, Table 2), standardized
Baron (2021)	Test scores	-0.1890	$4400 (2010$): "the median per-pupil bond campaign approved in Wisconsin is only approximately $4,400 per pupil" (24), depreciated over 15 years and averaged over first 6 years	-0.0567: Figure 6 panel (c) Average 10th Grade Math Score, cubic Year 6 relative to election (exact estimates provided by author), divided by student-level SDs (43.2, footnote 28)

Baron (2021)	Test scores	0.1790	$346 (2010$): Figure 1 (b) Total Operational Expenditures, averaged across 1-4yrs Relative to the Election (exact estimates provided by author)	3.084: Figure 2 Panel (c) Average 10th Grade Math Score Year 4 relative to election (exact estimates provided by author), divided by student-level SDs (43.2, footnote 28)
Brunner Hyman Ju (2020)	Test scores	0.0530	$498 (2015$): Table 2 Current Expenditures, State Aid, Expanded controls Yes	0.007: Table 7 All Districts Years postreform, multiply by 4 (years)
Candelaria Shores (2019)	High school graduation	0.0510	$795.02 (2010$): .1xbaseline (Table 2 Weighted Mean Total revenues)	0.197: Table 5, Full log(Rev/Pupil), standardized (Table 2, Graduation rates)
Carlson Lavertu (2018)	Test scores	0.0900	$2048.79 (2014$): Table 8 Dynamic RD model SIG eligibilty, average Year 1-4	0.221, 0.171: Table 5 Dynamic model SIG eligibility Year 4 of SIG, average Reading and Math
Cascio Gordon Reber (2013)	High school dropout	0.5550	$100 (2009$):"each additional $100 increase in annual current expenditure per pupil..." (pg. 152)	-3.46, 0.66: Table 7 Δ White and Black high school dropout (reverse sign), population weighted (0.9/0.1) and translated to SD units based on baseline (pg 147, population-weighted)

Cellini Ferreira Rothstein (2010)	Test scores	0.2120	\$6300 (2010\$): "the average bond proposal in close elections is about \$6,300 per pupil" (249), depreciated over 15 years and averaged over first 6	0.103, 0.160: Table VII, Academic achievement 6 yrs later Reading and Math, standardized ("the year-six point estimates correspond to effects of roughly 0.067 student-level standard deviations for reading and 0.077 for mathematics" (252)
Clark (2003)	Test scores	0.0150	\$1094.28 (2001\$): Table 3 Current expenditures per pupil Post-reform (1=yes)	0.023: Table 6 Composite, Kentucky x post model (3)
Conlin Thompson (2017)	Test proficiency rates	0.0080	\$4000 (2013\$): "Capital expenditure and capital stock variables in Panels A and B are listed in \$1000s" (Table 3 note) x4 (years), depreciated 15 years averaged over first 3	0.081, 0.07: Table 3 Capital Exp PP_t model (2) Percent Proficient in Math and Reading, relative to time t-3, standardized (Table 1 Percent Proficient in Math and Reading)
Gigliotti Sorensen (2018)	Test scores	0.0420	\$1000 (2016\$): "models... measure the effect of a \$1000 spending increase" (175)	0.0468, 0.042: Table 4 PPE Math and Reading

Goncalves (2015)	Test proficiency rates	-0.0020	$23740.4 (2010$): Table 1 Construction Cost Per Pupil Total, depreciated over 36.875 (weighted between 15 and 50 based on "60-65% of projects are new facilities" (6), averaged across first 6 years	1.266, -1.442: Table 4 6+ yr. Completion Exposure Math and Reading, standardized (baseline Avg. Proficiency Table 4)
Guryan (2001)	Test scores	0.0280	$1000 (1991$): "median estimate...implies that a one standard deviation increase in per-pupil spending ($1,000)..." (21)	0.039, 0.032, -0.034, -0.026: Table V and Table VI Math and Reading, subject-combined and standardized (assumed student-level SD of 100), then precision-weighted across grades
Hong Zimmer (2016)	Test proficiency rates	0.1160	$8123 (2000$): Table 1 Avg. bond amount per pupil, depreciated over 26.9 years (weighted between 15 and 50 based on Table 4 Passed a measure New building) averaged over 6 years	2.13, 1.44: Table 5 4th7th proficiency Relative year 6, standardized based on Table 3 proficiency baseline

Hyman (2017)	College enrollment	0.0550	$1000 (2012$): "interpretation... is that $1,000 of additional spending during each of grades four through seven..." (269)	0.03: Table 4 model (4) Enroll in postsecondary schooling, standardized (baseline Table 1 All districts and cohorts Enrolls in postsecondary school)
Jackson Johnson Persico (2015), Jackson Johnson (2019)	High school graduation	0.0800	$480 (2000$): Table I All Per pupil spending (avg., ages 5-17) ($4,800) x0.1	0.07053: Table III Prob(High School Graduate) model (7), standardized based on avg. national baseline graduation rate of 0.77
Jackson Wigger Xiong (2021)	College enrollment	0.0380	$1000 (2015$): "preferred model, a $1000 reduction in per-pupil spending..." (14)	0.0201: Table A19 model (8) 4-Year Avg Per-Pupil Spending (thousands), standardized based on Table 1 College Enrollment Rate baseline
Jackson Wigger Xiong (2021)	Test scores	0.0500	$1000 (2015$): "preferred model, a $1000 reduction in per-pupil spending..." (14)	0.0529: Table A19 model (4) 4-Year Avg Per-Pupil Spending (thousands)

Johnson (2015)	High school graduation	0.1440	\$85 (2000\$): "results indicate that a \$100 increase in per-pupil Title I funding..." (66) times 0.85 passed through in real dollars seen by students (Figure 9)	0.0225: Table 2 first column County Title I per-pupil spending (00s), average ages five to seventeen, standardized based on avg. national baseline graduation rate of 0.77
Kogan Lavertu Peskowitz (2017)	Test scores	0.0190	-\$303.096 (2010\$): Table 3 Total average Election year-3 years after, times 12000 ("District spending per pupil is just under \$12,000 annually" (384))	-0.14: Table 7 3 years after, to student-level SD units based on footnote 34
Kreisman Steinberg (2019)	High school graduation	0.0280	\$1000 (2011\$): specification, abstract	0.021: Table 8 Graduation, standardized based on Table 1 Graduation rate baseline
Kreisman Steinberg (2019)	Test scores	0.0780	\$1000 (2011\$): specification, abstract	0.097, 0.077: Table 5 Reading and Math
Lafortune Rothstein Schanzenbach (2018)	Test scores	0.0160	\$907 (2013\$): Table 4 Mean Total exenditures	0.004: Table 8 Post event x years elapsed times 4 (years)

Lafortune Schonholzer (2021)	Test scores	0.2330	$15000 (2013$): "projects we study...$15,000 per pupil" (footnote 6), depreciated over 50 years average across first 6 years	0.031xyear - 0.016, 0.027xyear - 0.004: Table 3 2SLS New School + Newschool Trend, Math and English Language Arts, 6 years
Lee Polachek (2018)	High school dropout	0.0640	$169.40 (2018$): Table 2 (percent change) times baseline spend by authors' calculations ($16939.79)	-0.1837: Table 3 9th-12th Grade Cubic, standardized based on baseline dropout rate Table 1 Mean Dropout Rate 9-12th Grade
Martorell Stange McFarlin (2016)	Test scores	0.0300	$7800 (2010$): "average per-pupil size of capital campaigns in Texas, the state we study in this paper, is about $7800" (14), depreciated over 15 years averaged over first 6 years	0.016, 0.03: Table 5 Standardized Test Scores 6 years after bond passage Reading and Math
Miller (2018)	High school graduation	0.0660	$1371.9 (2013$): specification, 0.1 times baseline spend $13,719.24 (pg. 30)	0.384: Table 4 10th Grade Cohort 1-4 years, standardized based on Table 1 Graduation Rate 4-year lag

Miller (2018)	Test scores	0.0520	$1371.9 (2013$): specification, 0.1 times baseline spend $13,719.24 (30)	0.775, 0.879, 0.929, 0.477: Table 5 4th Grade Math and Reading and 8th Grade Math and Reading, subject-combined then precision-weighted across grades
Neilson Zimmerman (2014)	Test scores	0.0250	$70000 (2005$): "about $70,000 in the New Haven SCP" (25), depreciated over 50 years averaged over first 6 years	0.153, 0.031: Table 6 > 5 Reading and Math, FE
Papke (2008)	Test proficiency rates	0.0820	$684.75 (2004$): 0.1 times baseline spend $6847.5 (Table 3 Average Expenditure per Pupil 1992-2004)	36.77: Table 7 Fixed Effects-Instrumental Variables log(average eral per pupil expend), standardized based on baseline Table 5 average 50th percentile first three years

Rauscher (2020)	Test scores	0.0080	$9600 (2014$): average capital outlays years 1-6 post election (Table 5), depreciated over 15 years averaged across first 6 years	47.77, 12.36: Table 4 models (3) and(6) 6 Years after election Low-SES achievement and High-SES achievement, to student-level standard deviation units extrapolating from "These estimates amount to 0.40 to 0.57 standard deviations..." (119), distributed across estimated students per school (NCES)
Roy (2011)	Test scores	0.3800	$1000 (2010$): specification, "estimates imply...for every $1,000" (159)	0.057, 0.061: Table 8 Instrumental variables regressions Lagged spending 1998-2001 Reading and Math, standardized based on baseline SE (Footnote 35)
Weinstein Stiefel Schwartz Chalico (2009)	High school graduation	0.1600	$391.7 (2003$): Table 6 Direct Expenditure Title I model (2)	3.59: Table 8 Graduation Rate Title I model (2), standardized based on avg. national baseline graduation rate of 0.77

Weinstein Stiefel Schwartz Chalico (2009)	Test scores	-0.0540	$284.3 (2003$): Table 5: Direct Expenditure Title I model (2)	-0.011, -.031: Table 7 Title I Math and Reading

This describes the steps per *overall* study-outcome (and by spending type, relevant for Baron (2020)).

Table A.3: Summary of capital depreciation decision

Study	Depreciate over (years)	Life of project description
Baron (2021)	15	"the median per-pupil bond campaign approved in Wisconsin is only approximately $4,400 per pupil, and bond funds are frequently used to repair, maintain, and modernize existing structures, rather than to build new schools" (24)
Cellini Ferreira Rothstein (2010)	15	"Anecdotally, bonds are frequently used to build new permanent classrooms that replace temporary buildings (e.g., Sebastian (2006)), although repair, maintenance, and modernization are common uses as well' (220) // Table 1 average amount per pupil is of smaller magnitude than full-building construction
Conlin Thompson (2017)	15	this paper doesn't specify, and they translate effects into per-$1000 but the OH program was for both new construction and renovations
Goncalves (2015)	36.875	"I corresponded with an OSFC employee who reported that about 60-65
Hong Zimmer (2016)	26.9	for the three years of data they have more detailed spending, percent new building is about 34

Lafortune Schonholzer (2021)	50	"We restrict attention only to large new school construction project" // "Nearly $11 billion was spent over this period, about 86 percent of which went to new school openings, while the rest went to additions, renovations, and equipment delivery at existing schools" (7)
Martorell Stange McFarlin (2016)	15	"typical capital campaigns deliver only modest facility improvements for the average student" (14) // "evidence is stronger for the claim that capital campaigns increase exposure to renovated schools" (20)
Neilson Zimmerman (2014)	50	"Of 42 school buildings, 12 had been rebuild completely by 2010, and 18 had been significantly renovated. . . school renovations were generally substantial, incurring costs similar to those of new construction" (20)
Rauscher (2020)	15	looks at CA bonds, which "can be used only for construction, rehabilitation, equipping school facilities, or acquisition/lease of real property for school facilities" (113)

Table A.4: Studies with LI and non-LI estimates

Study	Outcome	non-LI $	LI $	non-LI effect	LI effect	LI definition
Abott Kogan Lavertu Peskowitz (2020)	Test scores	279.99	609.19	0.2572	0.0460	"compare spending and educational outcomes between districts that are above or below our sample median in terms of poverty rates among 5–17-year-olds (according to the American Community Survey)" (9)
Abott Kogan Lavertu Peskowitz (2020)	High school graduation	279.99	609.19	0.1396	0.0295	"compare spending and educational outcomes between districts that are above or below our sample median in terms of poverty rates among 5–17-year-olds (according to the American Community Survey)" (9)

Baron (2021)	College enrollment	.	428.72	.	0.2566	"I classify a school district as having an initially-high share of economically disadvantaged students if its share falls above the median of the Wisconsin 2000-01 school district distribution (the earliest year this variable is made publicly available)." (18)
Baron (2021)	Test scores	275.52	328.42	-0.4197	-0.1697	"I classify a school district as having an initially-high share of economically disadvantaged students if its share falls above the median of the Wisconsin 2000-01 school district distribution (the earliest year this variable is made publicly available)." (18)

Baron (2021)	Test scores	.	532.74	.	0.1760	"I classify a school district as having an initially-high share of economically disadvantaged students if its share falls above the median of the Wisconsin 2000-01 school district distribution (the earliest year this variable is made publicly available)." (18)
Brunner Hyman Ju (2020)	Test scores	527.60	527.60	0.0303	0.0682	"We separate the effects of SFRs by within-state 1980 income terciles because reforms were designed to differentially impact state aid for low- and high-income districts, with the goal of equalizing school funding" (478)
Candelaria Shores (2019)	High school graduation	915.52	915.52	0.0188	0.1313	"state-specific poverty quartiles, defined using free lunch eligibility status" (39)

Goncalves (2015)	Test proficiency rates	.	1160.92	.	0.0031	Poorest 25% (Table 3)
Hyman (2017)	College enrollment	1093.70	1093.70	0.0791	0.0055	"districts with below-median 1995 district-level fraction receiving free lunch" (276)
Jackson Johnson Persico (2015), Jackson Johnson (2019)	High school graduation	710.59	686.24	0.0275	0.1140	"... a child is defined as low income if parental family income falls below two times the poverty line for any year during childhood" (165)
Johnson (2015)	High school graduation	123.95	123.95	0.0556	0.3406	
Kreisman Steinberg (2019)	Test scores	1116.33	1116.33	0.0264	0.0618	tercile of poverty (economically disadvantaged) (Table 6)
Kreisman Steinberg (2019)	High school graduation	1116.33	1116.33	-0.0053	0.0571	tercile of poverty (economically disadvantaged) (Table 6)

Lafortune Rothstein Schanzenbach (2018)	Test scores	672.62	1484.28	-0.0059	0.0189	"bottom or top quintile, respectively, of the state district-level income distribution" (Table 5)
Rauscher (2020)	Test scores	766.29	766.29	0.0047	0.0182	"The CDE defines low-SES students as those who are eligible for free or reduced-price lunch *or* whose parents both have less than a high school diploma...I refer to the distinction as SES throughout the article" (114)

This represents all studies included in our meta-analyses which report separate effects for LI and non-LI populations (Except Baron (2021) operational and Goncalves (2015), which report for LI but not non-LI). The studies not included in our analyses, but relevant for identifying whether effects of spending are generally larger for LI populations include: Biasi (2019) on income mobility, Card & Payne (2002) on test score gaps, JJP (2015) on wages and poverty, Johnson (2015) on wages and poverty. These papers all find either a decrease in outcome gaps between LI and non-LI groups, or specifically more pronounced effects for LI individuals exposed to increased spending. This assumes the *same* dollar change for LI and non-LI districts in Hyman (2017). Without additional information about within- and across-district demographic heterogeneity, we are unable to capture (potentially) different spending changes for LI and non-LI students despite evidence in the paper which suggests money was distributed disproportionately to non-LI schools within districts. Analogous to our inclusion criteria for studies, we include only low-income estimates from Baron (2021) and not non-low-income estimates because (estimates provided by author) indicated no detectable spending change associated with operational referendum change for that population.

A.3 Sensitivity and Robustness Analyses

Without assumed SD adjustment

Table A.5: Meta-Regressions w/o Papers Assuming SD adjustment

	(1) Overall Test Scores
Average Effect	0.0400***
	(0.00823)
N	22

Standard errors in parentheses

Omits Rauscher (2020) and Kogan et al. (2017)

* $p < .1$, ** $p < .05$, *** $p < .01$

By policy categories

Table A.6: Meta-Regressions by Policy Categories

	(1) Overall Test Scores	(2) Overall Educational Attainment
Average Effect	0.0428***	0.0487***
	(0.0122)	(0.0119)
Voluntary Policy	-0.0108	0.00952
	(0.0148)	(0.0129)
N	24	12

Standard errors in parentheses

Voluntary Policy includes: Equalization, Referenda,

School Finance Reform, New Construction, and School Improvement Grants.

* $p < .1$, ** $p < .05$, *** $p < .01$

Constructing IV when SE underreported

Table A.7: Meta-Analysis Estimates, Corr = −1

	(1) Overall Test Scores	(2) Non-Capital Test Score	(3) Capital Test Score	(4) Overall Test Scores	(5) Overall Educational Attainment
Average Effect	0.0356***	0.0437***	0.0157***	0.0434***	0.0539***
	(0.00742)	(0.00899)	(0.00576)	(0.00883)	(0.00577)
Capital				-0.0251**	
				(0.0115)	
N	24	15	9	24	12
τ	0.0258	0.0249	0.0151	0.0208	0
% Cross-Study Var.	0.740	0.649	0.593	0.649	0
90% PI	[-0.009,0.080]	[0.000,0.087]	[-0.011,0.042]		[0.044,0.063]
Prob. Pos	0.908	0.950	0.834	0.966	1

Standard errors in parentheses

* $p < .1$, ** $p < .05$, *** $p < .01$

Table A.8: Meta-Analysis Estimates, Corr = 1

	(1) Overall Test Scores	(2) Non-Capital Test Score	(3) Capital Test Score	(4) Overall Test Scores	(5) Overall Educational Attainment
Average Effect	0.0346***	0.0422***	0.0140***	0.0416***	0.0592***
	(0.00695)	(0.00843)	(0.00482)	(0.00828)	(0.00710)
Capital				-0.0241**	
				(0.0107)	
N	24	15	9	24	12
τ	0.0224	0.0228	0.0124	0.0195	0.0151
% Cross-Study Var.	0.838	0.779	0.711	0.796	0.309
90% PI	[-0.004,0.073]	[0.002,0.082]	[-0.008,0.036]		[0.032,0.087]
Prob. Pos	0.930	0.959	0.855	0.969	1.000

Standard errors in parentheses

* $p < .1$, ** $p < .05$, *** $p < .01$

Correlation bounds

Our preferred analysis assumes 0.5 correlation between dependent effects (math/reading) and 0 correlation between independent effects (across grades or populations). We re-run our main specifications with updated assumed correlations between effects within studies to generate one overall effect per study. We re-run our main specifications with assumed correlations for dependent effects from 0.25 to 0.75 and for independent effects from 0 to 0.5.

Table A.9: Meta-Analysis (w/in pop. low (0.25) // across pop. low (0))

	(1)	(2)	(3)	(4)	(5)
	Overall Test Scores	Non-Capital Test Score	Capital Test Score	Overall Test Scores	Overall Educational Attainment
Average Effect	0.0356***	0.0432***	0.0159***	0.0427***	0.0539***
	(0.00735)	(0.00898)	(0.00584)	(0.00872)	(0.00577)
Capital				-0.0239**	
				(0.0115)	
N	24	15	9	24	12
τ	0.0251	0.0251	0.0147	0.0206	0
% Cross-Study Var.	0.780	0.697	0.656	0.705	0

Standard errors in parentheses

* $p < .1$, ** $p < .05$, *** $p < .01$

Table A.10: Meta-Analysis (w/in pop. low (0.25) // across pop. high (0.5))

	(1)	(2)	(3)	(4)	(5)
	Overall Test Scores	Non-Capital Test Score	Capital Test Score	Overall Test Scores	Overall Educational Attainment
Average Effect	0.0354***	0.0433***	0.0155***	0.0428***	0.0539***
	(0.00740)	(0.00914)	(0.00565)	(0.00889)	(0.00577)
Capital				-0.0244**	
				(0.0115)	
N	24	15	9	24	12
τ	0.0253	0.0256	0.0148	0.0212	0
% Cross-Study Var.	0.764	0.696	0.626	0.693	0

Standard errors in parentheses

* $p < .1$, ** $p < .05$, *** $p < .01$

Table A.11: Meta-Analysis (w/in pop. high (0.75) // across pop. low (0))

	(1) Overall Test Scores	(2) Non-Capital Test Score	(3) Capital Test Score	(4) Overall Test Scores	(5) Overall Educational Attainment
Average Effect	0.0349***	0.0430***	0.0142***	0.0424***	0.0539***
	(0.00717)	(0.00866)	(0.00497)	(0.00855)	(0.00577)
Capital				-0.0253**	
				(0.0109)	
N	24	15	9	24	12
τ	0.0244	0.0242	0.0132	0.0198	0
% Cross-Study Var.	0.745	0.654	0.574	0.658	0

Standard errors in parentheses

* $p < .1$, ** $p < .05$, *** $p < .01$

Table A.12: Meta-Analysis (w/in pop. high (0.75) // across pop. high (0.5))

	(1) Overall Test Scores	(2) Non-Capital Test Score	(3) Capital Test Score	(4) Overall Test Scores	(5) Overall Educational Attainment
Average Effect	0.0346***	0.0430***	0.0136***	0.0424***	0.0539***
	(0.00720)	(0.00880)	(0.00461)	(0.00870)	(0.00577)
Capital				-0.0259**	
				(0.0108)	
N	24	15	9	24	12
τ	0.0244	0.0245	0.0129	0.0201	0
% Cross-Study Var.	0.723	0.653	0.519	0.639	0

Standard errors in parentheses

* $p < .1$, ** $p < .05$, *** $p < .01$

Depreciation

Our preferred analysis assumed buildings are depreciated 50 years and non-buildings are depreciated 15 years. We re-run our main specifications with lower and upper bounds on years across which capital investments are depreciated. At a lower bound, we depreciate buildings at 30 and non-buildings at 10 years. At an upper bound, we depreciate buildings at 50 and non-buildings at 30 years.

Table A.13: Depreciation Sensitivity Meta-Analysis

	(1) Baseline	(2) Low Bound (years dep.)	(3) High Bound (years dep.)	(4) No Depreciation
Average Effect	0.0150***	0.0121***	0.0185***	0.0198**
	(0.00536)	(0.00450)	(0.00617)	(0.00830)
N	9	9	9	9
τ	0.0139	0.0113	0.0170	0.0193
% Cross-Study Var.	0.613	0.622	0.587	0.662

Standard errors in parentheses

* $p < .1$, ** $p < .05$, *** $p < .01$

Restricted to Models w/ Power to Detect Main Effects

Table A.14: Meta-Analysis Estimates w/ Power to Detect Main Effects

	(1) Overall Test Scores	(2) Overall Educational Attainment
Average Effect	0.0312***	0.0529***
	(0.00859)	(0.00572)
N	9	7

Standard errors in parentheses

* $p < .1$, ** $p < .05$, *** $p < .01$

A.4 Capital Spending Effects Over Time

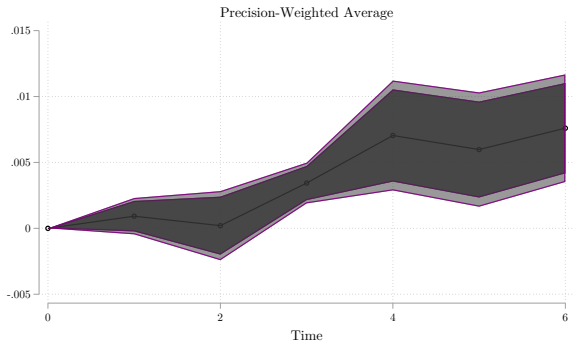


Figure A.1: Precision-Weighted

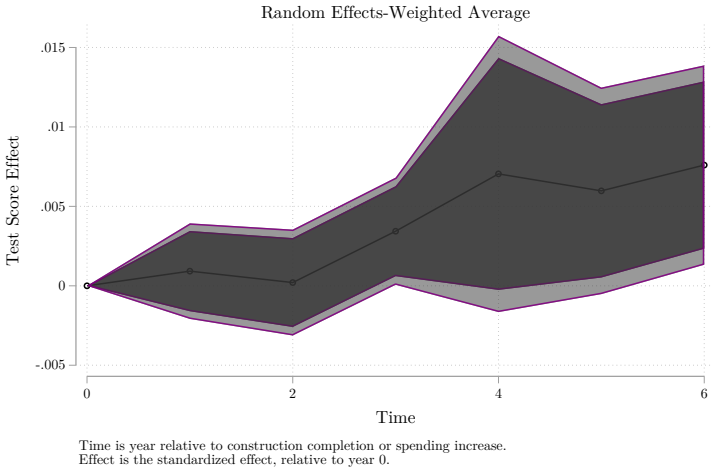


Figure A.2: Random Effects Precision-Weighted)

A.5 Forest Plots by Spending Type

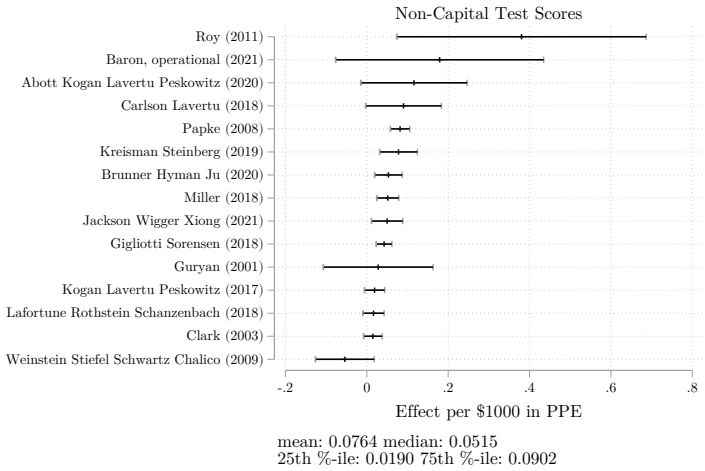


Figure A.3: Non-Capital Test Score

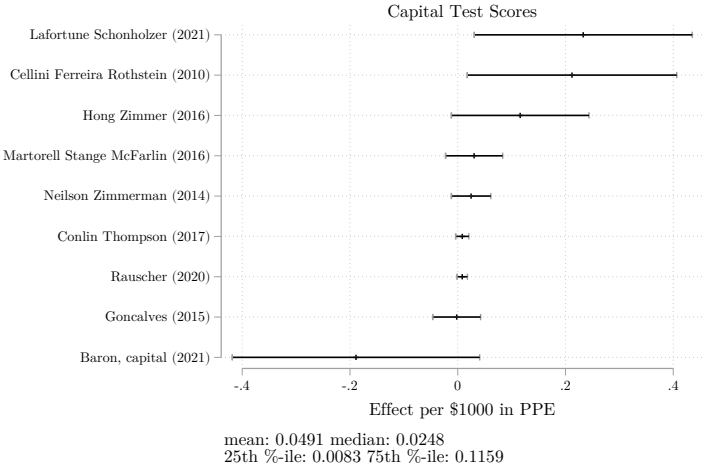


Figure A.4: Capital Test Score

A.6 Testing for Additional Patterns

Linearity in Spending

Table A.15: Linearity in spend

	(1) Test Scores All	(2) Test Scores w/o assumed	(3) Test Scores w/o JWX	(4) Ed Attain All	(5) Ed Attain w/o assumed	(6) Ed Attain w/o JWX
Policy on Exp. ($1000s)	0.0436***	0.0323	0.0422***	0.0457***	0.0514***	0.0401***
	(0.00820)	(0.0136)	(0.0119)	(0.00652)	(0.0141)	(0.0136)
Constant	-0.00636	-0.0000539	-0.00496	0.0127	0.0129	0.0171
	(0.00686)	(0.0103)	(0.00998)	(0.00584)	(0.0105)	(0.0112)
N	24	19	23	12	9	11
Pr(slope = pooled avg.)	0.451	0.890	0.587	0.367	0.519	0.213

Standard errors in parentheses

* $p < .1$, $p < .05$, *** $p < .01$

Additional Tests by Income Level

We present by-outcome coin test comparisons between low-income and non-low-income estimates in Table A.16.

Table A.16: Coin Test for Studies w/ LI and non-LI Estimates

Outcome	Papers	LI> non-LI	% LI> non-LI	1 in X Chance
All Studies	15	11	0.73	17
Test Score	8	6	0.75	7
Educational Attainment	7	5	0.71	4

LI > non-LI represents the count (or percent) of studies whose effect per $1000 for non-LI populations is larger than the effect for LI populations.

We present our main models estimating heterogeneous effects by income level in Table A.17, which are shown in Figure 9.

We present models which restrict low-income estimates to only include those studies for which estimated impacts on spending are clearly reported separately by income in columns 2 and 4 of Table A.19.

Table A.17: Meta-Regressions w/ LI

	(1) Test Scores	(2) Test Scores	(3) Educational Attainment	(4) Educational Attainment
Average Effect	0.0385***	0.0389***	0.0554***	0.0553***
	(0.00894)	(0.00875)	(0.0101)	(0.0101)
Low-Income	0.0108		-0.000601	
	(0.0175)		(0.0181)	
Low-Income (w/ Title I)		-0.00949		0.0131
		(0.0241)		(0.0192)
Non-Low-Income	-0.0125	-0.0252	-0.0195	-0.0100
	(0.0109)	(0.0168)	(0.0179)	(0.0194)
Has Estimates by Income (Indicator)	-0.0139	-0.00163	-0.00399	-0.0134
	(0.0160)	(0.0211)	(0.0145)	(0.0155)
N	38	38	25	25
τ	0.0248	0.0250	0.00174	0
% Cross-Study Var.	0.765	0.768	0.00349	0
Low-Income = Non-LI = 0 (p-val)	0.170	0.154	0.537	0.692

Standard errors in parentheses

All Low-Income Estimates are comparisons with Non-Low-Income except in the case of Goncalves (2015) and Baron (2021) operational.

Low-Income w/ Title I is an indicator that additionally captures all Title I studies, even those which do not present distinct by-income effects.

* $p < .1$, ** $p < .05$, *** $p < .01$

Table A.18: Meta-Regressions w/ LI, Cluster Same Policies

	(1) Test Scores	(2) Test Scores	(3) Educational Attainment	(4) Educational Attainment
Average Effect	0.0363***	0.0374***	0.0546***	0.0545***
	(0.00778)	(0.00725)	(0.00999)	(0.00998)
Low-Income	0.00693		0.00824	
	(0.0216)		(0.0145)	
Low-Income (w/ Title I)		-0.0212		0.0132
		(0.0297)		(0.0145)
Non-Low-Income	-0.0165	-0.0340*	-0.0245*	-0.0212
	(0.0121)	(0.0201)	(0.0139)	(0.0146)
Has Estimates by Income (Ind.)	-0.00775	0.00859	-0.00428	-0.00762
	(0.0185)	(0.0251)	(0.0153)	(0.0143)
N	38	38	25	25
τ	0.0234	0.0238	0	0
% Cross-Study Var.	0.744	0.751	0	0
Low-Income = Non-LI (p-val)	0.163	0.453	0.186	0.172

Standard errors in parentheses

All Low-Income Estimates are comparisons with Non-Low-Income except in the case of Goncalves (2015) and Baron (2021).

Low-Income w/ Title I is an indicator that additionally captures all Title I studies, even those which do not present distinct by-income effects.

* $p < .1$, ** $p < .05$, *** $p < .01$

Table A.19: Meta-Regressions w/ LI

	(1) Test Scores	(2) Test Scores	(3) Test Scores	(4) Test Scores	(5) Educational Attainment	(6) Educational Attainment
Average Effect	0.0386***	0.0389***	0.0389***	0.0397***	0.0553***	0.0553***
	(0.00897)	(0.00904)	(0.00880)	(0.00870)	(0.0101)	(0.0101)
Low-Income	0.00970	0.00591			-0.000939	
	(0.0177)	(0.0231)			(0.0151)	
Low-Income (w/ Title I)			-0.00745	-0.0180		0.00515
			(0.0218)	(0.0276)		(0.0150)
Non-Low-Income	-0.0173	-0.0257	-0.0254	-0.0371*	-0.0251	-0.0233
	(0.0150)	(0.0193)	(0.0165)	(0.0221)	(0.0256)	(0.0256)
Has Estimates by Income (Ind.)	-0.00431	-0.00102	0.00341	0.00962	-0.00361	-0.00535
	(0.0152)	(0.0188)	(0.0174)	(0.0220)	(0.0106)	(0.0108)
N	38	33	38	33	25	25
τ	0.0255	0.0268	0.0254	0.0265	0	0
% Cross-Study Var.	0.670	0.697	0.669	0.693	0	0
Low-Income = Non-LI (p-val)	0.0779	0.0984	0.284	0.335	0.411	0.329

Standard errors in parentheses

All Low-Income Estimates are comparisons with Non-Low-Income except in the case of Goncalves (2015) and Baron (2021).

Low-Income w/ Title I is an indicator that additionally captures all Title I studies, even those which do not present distinct by-income effects.

* $p < .1$, ** $p < .05$, *** $p < .01$

Geographic Dimensions

Table A.20: Meta-Analysis Estimates by Geographic Characteristics

	(1) Test Scores by Multistate	(2) Test Scores by Region	(3) Test Scores by Urbanicity	(4) Educational Attainment by Multistate	(5) Educational Attainment by Region
Average Effect	0.0321***	0.0452***	0.0377***	0.0585***	0.0523***
	(0.00913)	(0.00949)	(0.00799)	(0.00756)	(0.00748)
Capital					
Multistate	0.0118			-0.00660	
	(0.0132)			(0.0106)	
South		-0.00660			-0.00962
		(0.0217)			(0.0874)
North		-0.00575			0.00570
		(0.0203)			(0.0205)
Northeast		-0.0284			0.0128
		(0.0225)			(0.0128)
West		-0.00137			
		(0.0765)			
Urban			-0.0260		
			(0.0325)		
Rural			0.00221		
			(0.0319)		
N	24	24	24	12	12

Standard errors in parentheses

* $p < .1$, ** $p < .05$, *** $p < .01$

A.7 Details of Publication Bias Tests

Table A.21 presents our preferred estimates (columns 1 and 6) along with estimates using several approaches to potential publication bias.

Table A.21: Meta-Regressions w/ Approaches to Potential Biases

	Test Scores					Educational Attainment				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Avg. Effect	0.0352***	0.029***	0.0301***	0.032 ***	0.0289***	0.0539***	0.056***	0.0527***	0.053***	0.0513***
	(0.00723)	(0.007)	(0.00727)	(0.00714)	(0.00727)	(0.00577)	(0.020)	(0.00618)	(0.00663)	(0.00579)
N	24	24	12	28	24	12	12	6	17	12

Standard errors in parentheses

Test Score: (1) Robumeta (2) Andrews & Kasy (3) SE $< .023$ (4) Meta Trim&Fill (5) PEESE

Educational Attainment: (6) Robumeta (7) Andrews & Kasy (8) SE $< .021$(9) Meta Trim&Fill (10) PEESE

* $p < .1$, ** $p < .05$, *** $p < .01$

1. Studies that find null results may be less likely to be published than studies that find significant effects (Franco et al. (2014), Christensen and Miguel (2018)). If one is able to observe studies that are not published, a simple test for publication bias compares estimates from studies that are published to those that are not published. In line with this, we compare average estimates of published and unpublished studies and find no difference in impacts.[71] In Table A.22, the coefficients on the indicator for "Unpublished" show no evidence that there is any difference in average effects reported in published versus unpublished papers for both test scores and educational attainment outcomes.

2. Related to the first test, if there are biases against publication of certain kinds of studies, one might expect these biases to be most pronounced at the most selective journals (Brodeur et al. (2016)). Informed by this notion, we compare the average impacts of studies published in the most elite journals to studies published in other journals, and similarly find no differences across journal prestige (in columns 2 and 4 of Table A.22, the formal tests of equality across publication type and publication status yield p-vals of 0.871 and 0.622 for test scores and educational attainment, respectively). That is, we do not find evidence that publication status or type have any bearing on the estimates reported in studies of effects of school spending.

3. Publication bias is thought to be most prevalent among imprecise studies (Andrews and Kasy (2019)), and when there are biases against publication of insignificant studies, one might observe an over-representation of studies right at the significance threshold (in social sciences this would be the 5 percent level pertaining to a t-statistic of 1.96) and an under-representation of studies right below the significance threshold (Brodeur et al. (2020)). To test for this in our data, we test for a discontinuity in the cumulative density of t-statistics at 1.96. We show that there is no over-representation of studies right at the significance threshold (t-statistic = 1.96) in Figure A.5. In Table A.23, we show that there is no significant jump in density, by outcome type or combining across both test score and education attainment outcomes, at the significance threshold (t-stat > 1.96).

4. Even though we find limited evidence of selection of significant impacts, we implement a model that accounts for any such selection (should it exist). To this aim, we show results for the Andrews and Kasy (2019) selection adjustment using their web application in Figures A.6 and A.7. They propose estimating the publication probabilities (based on the t-statistics) for studies, and using these probabilities to produce bias-corrected estimators and confidence sets. More specifically, using the relative publication probabilities, this approach re-weights the distribution of studies to account for differences in publication probability (up-weighting studies that are least likely to be observed). For both test scores and educational attainment, their model fails to reject the null of no selection at the 1.96 t-statistic threshold. Reassuringly,

[71] Of course, we cannot observe the unobservable – or those papers which are fully not shared in any form, published or not.

their adjustment approach yields similar estimates to our preferred model (columns 2 and 7 of Table A.21).

5. We test whether there is bias against imprecise, negative estimates. In a stylized world, with no publication bias, a scatter plot of study impacts against the precision of each study should be roughly symmetric around the grand mean (Borenstein (2009)). However, with publication bias, the scatter plot around the grand mean will be asymmetric – suggesting that there are some "missing" studies. In this stylized world with publication bias, while all or most precise studies will be published, there may be an over-representation of published imprecise estimates in the "desired" direction and no (or few) published imprecise estimates in the "undesirable" direction. We account for this kind of publication bias in two ways: First, we impute "missing" (imprecise, negative) studies and re-estimate our models. Second, we separately drop the least precise estimates (the least-precise half) and re-estimate our models. Neither appreciably impacts our estimates.

   We visualize the Duval and Tweedie (2000) "trim and fill" approach in Figure 7, where black circles indicate the individual study impacts. The distribution of effects are largely symmetrical around the mean for very precise studies (at the top of the figures), but the distribution may be asymmetric for studies with standard errors greater than about 0.1 and 0.06 for test scores and education attainment, respectively (the bottom of the plots). That is, while there is little visual evidence of publication bias among precisely estimated studies, there is some suggestive evidence that imprecise positive studies with large impacts may be more likely to be published (or written) than imprecise studies with negative or small impacts. To be clear, because (a) our inclusion criteria requires that the policy has meaningful impacts on school spending and (b) one would expect there to be some effect heterogeneity across states and policies, some asymmetry is likely even absent publication bias. Even so, to be conservative one can assume that any asymmetry is due to publication bias, and assess the impacts of this asymmetry on the estimated pooled average. We follow this approach.

   In the left panel of Figure 7, to create symmetry, the "trim and fill" approach imputes four "missing" studies of test score outcomes (green triangles) – both of which are negative and very imprecise. These imputed studies are outside of the more precise range employed for our first test of bias – validating that approach. The re-estimated pooled effect that includes these two additional imputed studies is 0.032 (Table A.21 column 4) – very similar to our original estimate including all observed estimates. Following this same approach for educational attainment, "trim and fill" imputes five additional negative and relatively imprecise estimates. The re-estimated pooled effect that includes the three additional imputed studies is 0.053 (Table A.21 column 9) – also similar to our original estimate including all observed estimates. The fact that estimates do not change very much with the imputed data also reflects the fact that the evidence of asymmetry is only among very imprecise estimates, which receive lower weight in our precision-weighted pooled average. This suggests that the impacts of any

potential publication bias on our estimates are small (at most creating a bias of 5 percent).

When we estimate our main model on all studies using a drastic approach of dropping the majority of the data (Stanley et al. (2010)), specifically those test score studies with an estimated standard error of 0.023 or less (Table A.21 column 3) and educational attainment studies with estimated standard errors of 0.021 or less (Table A.21 column 8), our results are similar to our main models. We indicate these precision levels in the higher horizontal lines in the funnel plot in Figure 7. Above this cut-off, estimates are very tightly clustered around the pooled average.[72] In this most precise sample (where there is no evidence of asymmetry), the coefficient estimate for test scores is 0.0301 Table A.21 column 3). This is very similar to our preferred estimate – indicating minimal bias. Following this same approach for educational attainment, when we restrict our sample to studies with standard errors below 0.021, the Egger's tests indicates no asymmetry, and the regression estimate is 0.0527 (Table A.21 column 8).[73]

Finally, we follow both Stanley and Doucouliagos (2014) and Ioannidis et al. (2017) and implement the precision-effect estimate with standard error (PEESE) approach. This approach estimates the relationship between the precision of the estimates and the estimates reported in each study. Under the assumption that the most precise estimates will yield the true relationship, one can empirically model the relationship between the precision of the estimates and the reported estimates and then infer what the most precise estimate would be. In practice this involves regressing the reported effect on the square of its precision and taking the constant term as the bias-adjusted estimate. This approach has been found to preform well in simulations. This approach yields a meta-regression estimate which takes into account the influence of publication bias – based on estimate precision. In columns (5) and (9) of Table A.21 we report meta-regression results. For test scores, the PEESE method estimates a precision-weighted pooled average of 0.0289 and for educational attainment of 0.0513.

In sum, across multiple approaches to testing and accounting for potential publication bias, our main results hold, suggesting that if this bias exists it is minimal – our conclusions do not change.

[72] The p-values on both the intercept and slope associated with the Egger's test for this sample are both above 0.1.

[73] The Egger's test is the simply the p-value associated with the y-intercept being different from zero in a regression on the study effects against its precision. When the funnel is asymmetric, this p-value will be small.

Table A.22: Meta-Regressions w/ Publication Type

	(1) Test Score	(2) Test Score	(3) Educational Attainment	(4) Educational Attainment
Unpublished	-0.00923	-0.00346	0.00570	0.00967
	(0.0183)	(0.0213)	(0.0159)	(0.0196)
Top Field Journal		0.00223		
		(0.0165)		
Field Journal		0.0149		0.00918
		(0.0204)		(0.0143)
Average Effect	0.0380***	0.0323**	0.0530***	0.0484***
	(0.00834)	(0.0128)	(0.00659)	(0.0128)
N	24	24	12	12
τ	0.0257	0.0298	0.00461	0.00872
% Cross-Study Var.	0.775	0.823	0.0389	0.126
Top Field = Field = Unpublished = 0 (p-val)		0.855		0.806
Unpublished = 0 (p-val)	0.615	0.871	0.719	0.622

Standard errors in parentheses

Reference category High Impact ommitted.

High Impact: American Economic Journal, Quarterly Journal of Economics, Review of Economics and Statistics, Sociology of Education.

Top Field: Journal of Econometrics, Journal of Public Economics.

Field: Economics of Education Review, Education Economics, Education Finance and Policy, Educational Evaluation and Policy Analysis, Public Finance Review, Russell Sage Foundation Journal of the Social Sciences, Journal of Public Administration Research and Theory, Journal of Urban Economics

* $p < .1$, ** $p < .05$, *** $p < .01$

Table A.23: Regressions to test for jump at 5% significance, Outcome: Cumulative T-stat density

	(1) Test Scores $1 < tstat < 3$	(2) Educational Attainment $1 < tstat < 3$	(3) All Outcomes $1 < tstat < 3$
Sig, 5%-level (ind)	-0.0442***	0.0937	-0.0163
	(0.0131)	(0.0550)	(0.0276)
N	15	6	21

Standard errors in parentheses

All models include controls for the t-stat and the square of the t-stat.

In column 3 pooled models (with both outcome types) we include an indicator for the outcome and interact t-stat and t-stat squared with the outcome.

* $p < .1$, ** $p < .05$, *** $p < .01$

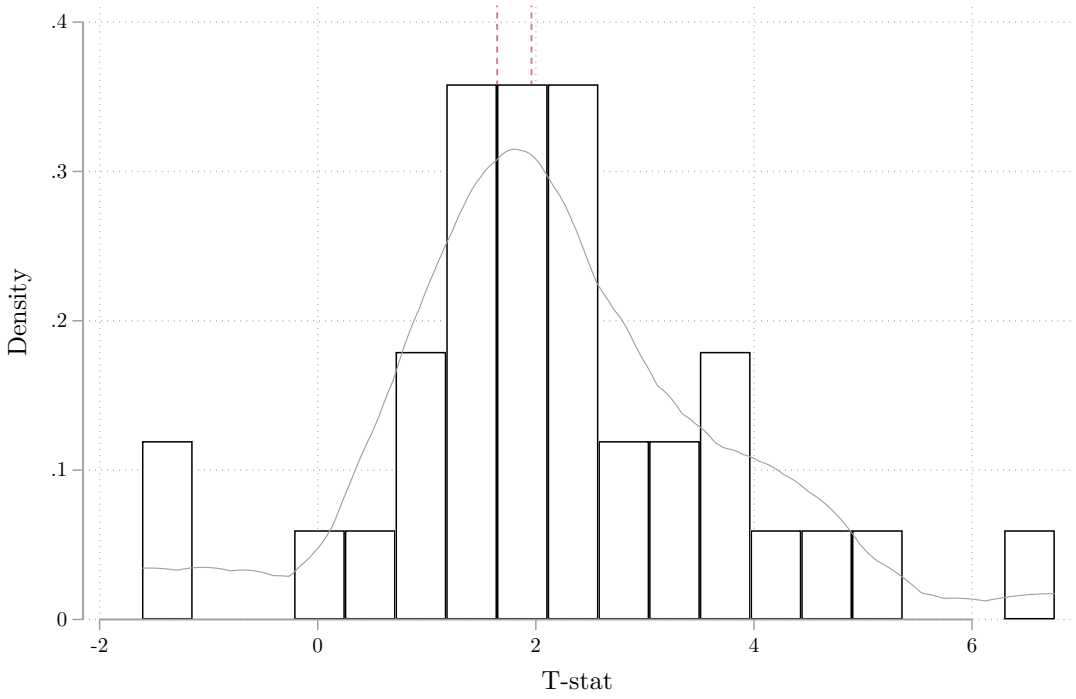


Figure A.5: Histogram of *all* effects

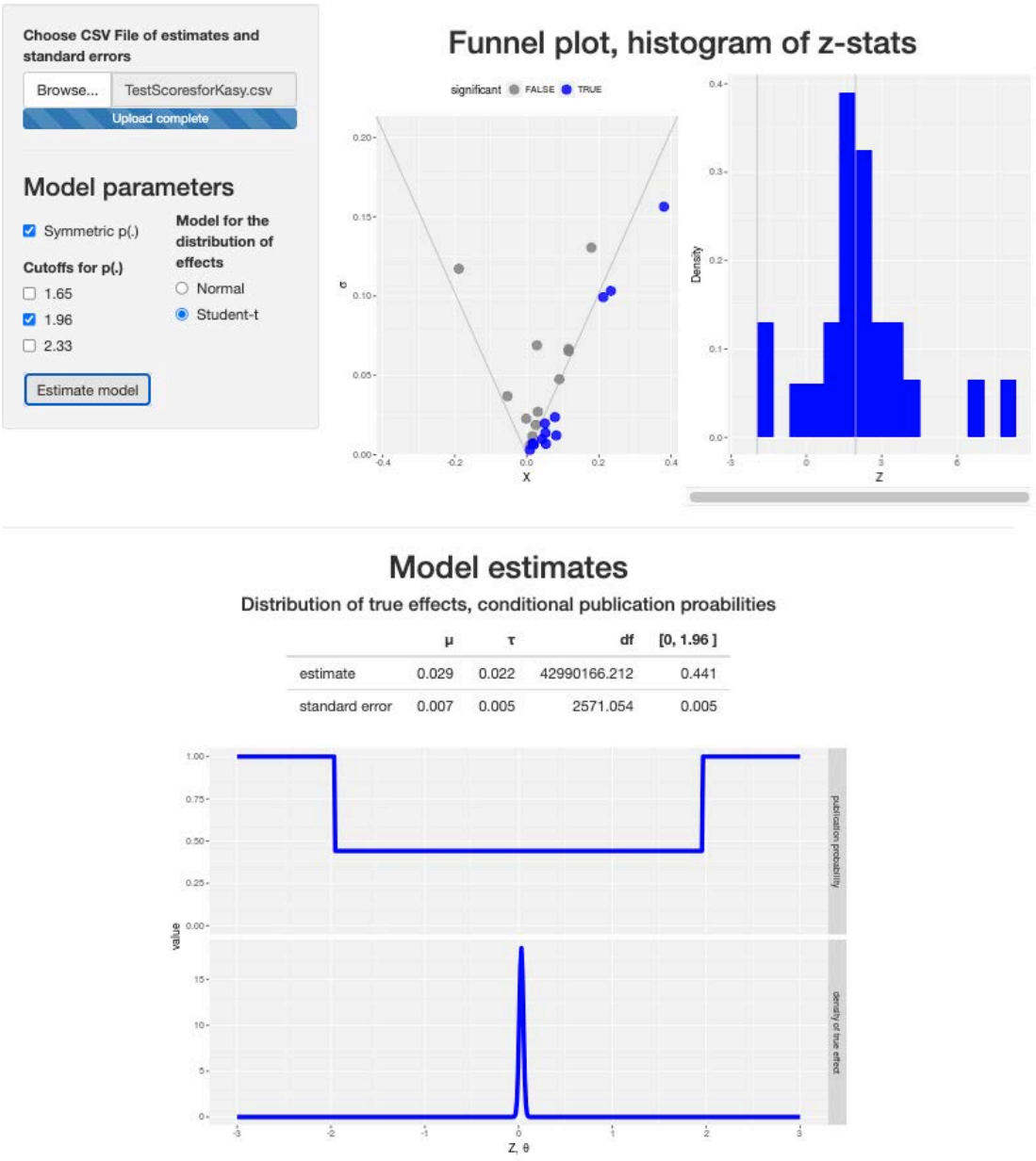


Figure A.6: Test Scores

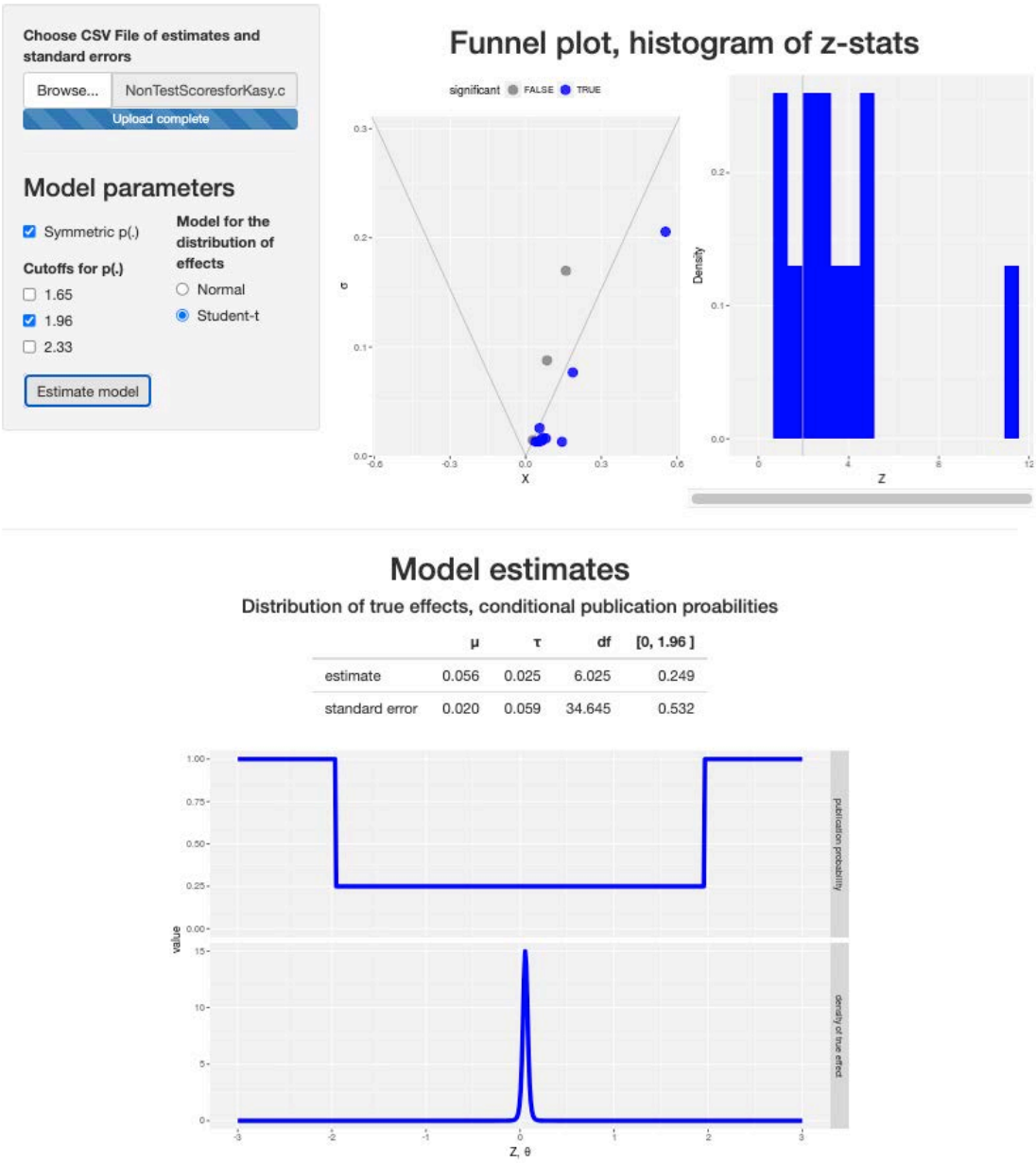


Figure A.7: Non-Test Scores

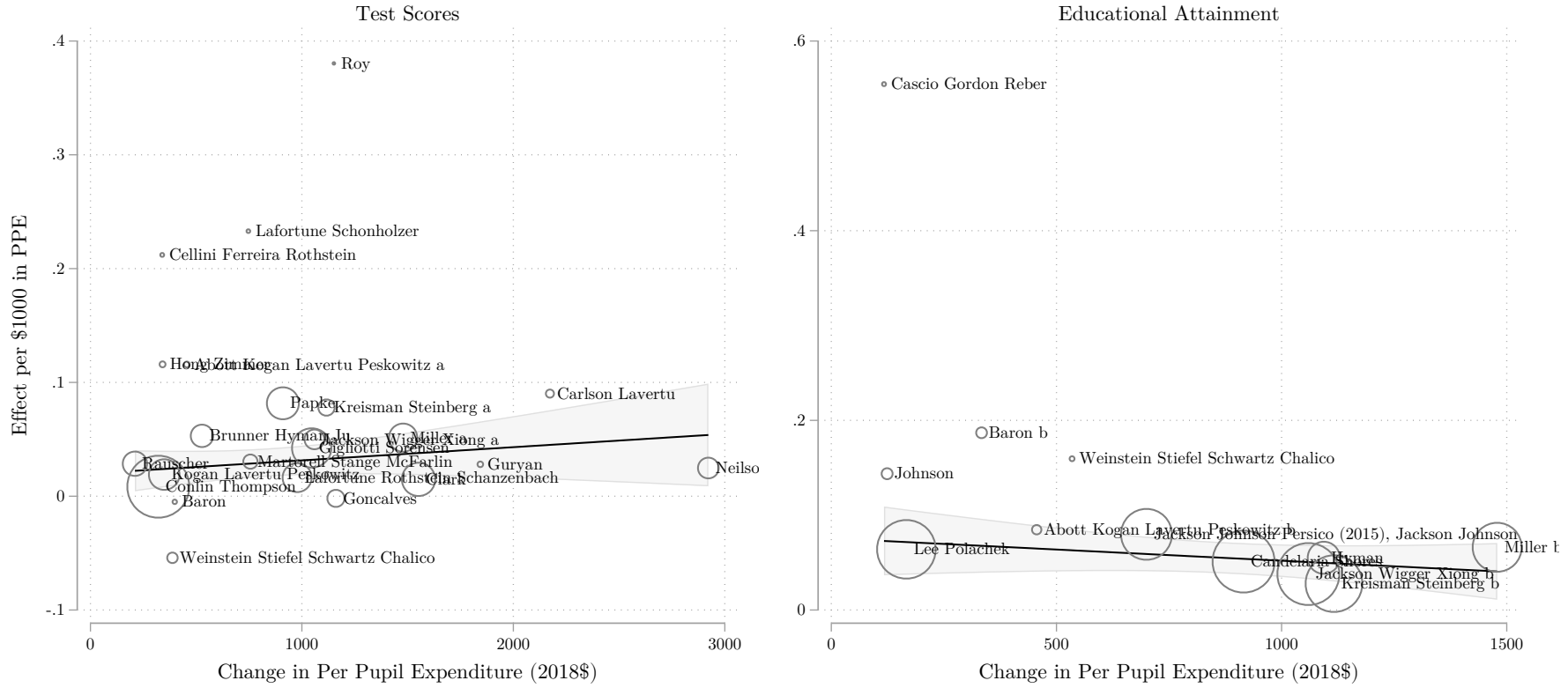


Figure A.8: Marginal Impacts by Change in Spending Level

A.8 Estimation Strategy

Table A.24: Meta-Regressions w/ Estimation Strategy

	(1) Overall Test Scores	(2) Overall Educational Attainment
RD	0.00126	0.00977
	(0.0133)	(0.0186)
IV	0.0363***	-0.00795
	(0.0131)	(0.0182)
Capital	-0.00701	
	(0.0124)	
Average Effect	0.0240**	0.0562***
	(0.00985)	(0.0156)
N	24	12
τ	0.0193	0.00615
% Cross-Study Var.	0.659	0.0671
RD = IV = 0 (p-val)	0.0194	0.447

Standard errors in parentheses

Event Study (strategy) omitted

* $p < .1$, ** $p < .05$, *** $p < .01$