American Economic Journal: Applied Economics 2024, 16(1): 412–446
https://doi.org/10.1257/app.20220279

What Impacts Can We Expect from School Spending Policy? Evidence from Evaluations in the United States[†]

By C. Kirabo Jackson and Claire L. Mackevicius*

We conduct meta-analysis on a comprehensive set of studies of the impacts of US K-12 public school spending on student outcomes–estimating average marginal impacts and heterogeneity across contexts. On average, a policy increasing spending by $1,000 per pupil for four years improves test scores by 0.0316σ and college-going by 2.8 pp. Moving beyond averages, we use estimates of heterogeneity and observable policy differences to produce informative probability distributions of policy effects. Effects are smaller for economically advantaged populations, marginal effects of capital spending are similar to noncapital, and effects are similar across baseline spending levels and geography. Confounding and publication biases are minimal.(*JEL* H75, I21, I22, I26, I28)

Social scientists have long debated the effect of school spending on student outcomes. This debate is important as public K-12 education is one of the largest single components of government spending (OECD 2020), and current legal cases (see Lecker 2020) and policy decisions hinge on the extent to which, in what contexts, and how reliably increases in school spending causally impact students. Fortunately, 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 effect of school spending on student outcomes. While this new literature shows that increased school spending tends to improve student outcomes (Jackson 2020), there is little clarity on **how much** money matters, on average, how the marginal effects may vary **across contexts**, and what **range of policy impacts** one may expect to see in the future across different populations.

Summarizing estimates from these recent studies in ways that are actionable for policy is challenging in part because impacts are likely heterogeneous across studies due to differences in context, policy implementation, and treated populations, so that no single estimate, *or average of estimates*, directly informs likely impacts of future

*Jackson: School of Education and Social Policy, Northwestern University (email: kirabo-jackson@northwestern.edu); Mackevicius: School of Education and Social Policy, Northwestern University (email: cmackevicius@u.northwestern.edu). Christopher Walters was coeditor for this article. This project was supported by the W.T. Grant Foundation (190178). The statements made are solely the responsibility of the authors. We deeply appreciate feedback from Beth Tipton, James Pustejovsky, Jason Baron, and Sylvain Chabé-Ferret. We also thank cited authors who have verified that we accurately describe their work.

[†]Go to https://doi.org/10.1257/app.20220279 to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

policies in *other* contexts (Tipton and Olsen 2018; Tipton, Bryan, and Yeager 2020; Vivalt 2020; Bandiera et al. 2021; Dehejia, Pop-Eleches, and Cyrus Samii 2021; Meager 2019). Providing evidence on heterogeneity across contexts is important because policymakers often dismiss academic research because they are "*not sure it applies*" to them (Clarke 2019; Nakajima 2021). Relatedly, if school spending exhibits diminishing marginal returns, studies of older policies (when spending levels were lower) may report very different impacts than policies being considered today. Moreover, individual studies are often underpowered, leading to large estimation errors and a wide range of statistically insignificant estimates—limiting the ability to make precise policy statements. Understanding true heterogeneity and the role of estimation errors requires examining impacts across settings, contexts, and populations—something impossible in a single study of a single policy.

Speaking to these challenges, we perform a meta-analysis of a comprehensive set of design-based studies (i.e., those that might credibly identify causal impacts) covering a broad range of contexts, time periods, and populations to quantify treatment heterogeneity—allowing us to describe the distribution of true effects of increased US public K-12 school spending on test scores and educational attainment. We shed light on what range of policy impacts can be expected in a new context and how marginal effects may differ by spending type, population, geography, and baseline spending levels.

Our analyses involved several steps. First, we only included studies that (a) employed quasi-random or quasi-experimental variation in school spending to estimate impacts on student outcomes, (b) demonstrated the spending variation was plausibly exogenous,[1] and (c) demonstrated meaningful policy-induced variation in school spending. Next, to facilitate direct comparison, for each study we constructed an estimate of the marginal policy-induced impact on standardized outcomes of exposure to a $1,000 per pupil spending increase (in 2018 dollars) over four years. We made impacts of capital comparable to noncapital spending by amortizing large one-time capital payments over the useful life of the capital asset and relating changes in outcomes to the present discounted "flow" value. Finally, using this comparable estimate for each paper, we employed random effects meta-analysis to generate a precision-weighted, pooled average and quantify (*by computing variability across study estimates unexplained by sampling variability*) an estimate of true treatment heterogeneity. Under testable assumptions about the distribution of this heterogeneity, we use these estimates to produce a plausible range of predicted policy impacts across settings.

The pooled meta-analytic average estimate indicates that, *on average*, a $1,000 per pupil increase in school spending sustained over four years increases test scores by 0.0316σ (p-value < 0.001). Once harmonized, the individual study estimates are much more similar to each other than a simple comparison of reported effects would suggest. While almost all the point estimates are positive, there is meaningful heterogeneity across studies. The standard deviation of underlying true heterogeneity in test-score effects is roughly 0.021σ—about one-fifth the size of the raw

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

spread of estimates and two-thirds of the mean. This implies that, *on average*, two randomly chosen policies would have true marginal effects that differ by 0.021σ. Flexible deconvolution estimates and formal statistical tests indicate that the distribution of true effects is approximately normal. As such, we assume normality and compute ranges of estimates for that same policy in a different average context (i.e., a setting with average levels of urbanicity and income levels and a mix of capital and operational spending). Increasing per pupil spending by $1,000 sustained for four years in an *average* setting would have positive test score impacts more than 90 percent of the time, effects larger than 0.05σ about 20 percent of the time, and effects larger than 0.07σ about 3 percent of the time. Our pooled meta-analytic estimate indicates that, *on average*, a $1,000 per pupil increase in school spending increases educational attainment by 0.0573σ (p-value < 0.0001). This translates into a 2.0 percentage point increase in high school graduation and a 2.8 percentage point increase in college-going. The individual study estimates on educational attainment are remarkably consistent with each other such that the pooled average lies within the 95 percent confidence interval for almost all estimates. The estimate of heterogeneity (0.027σ) is about one-quarter the spread of raw estimates and just under half the mean effect. As with test scores, formal testing indicates that the distribution of true effects is approximately normal. Using these estimates along with the tested assumption of normality, the aforementioned policy in a different average context would have positive high school graduation and college-going impacts over 97 percent of the time, increased high school completion and college-going by about 2.5 and 3.5 percentage points 30 percent of the time, and increased high school completion and college-going by about 3.2 and 4.5 percentage points just over 10 percent of the time, respectively.

School spending impacts on educational attainment are larger than on test scores when benchmarked against the impacts of other interventions—suggesting that test scores may understate the benefits of school spending, consistent with claims in Krueger (1998) and Card and Krueger (1992a) and recent work highlighting what test scores miss (Jackson 2018; Jackson et al. 2020; Beuermann et al. 2023). Average effects of increased capital spending on test scores are smaller than those of noncapital-specific spending, but not statistically significantly so. Importantly, most individual studies of capital spending are underpowered to detect the pooled effect—explaining why the literature on capital spending *appears* mixed. Average impacts are similar across several observable dimensions (including urbanicity and geography). Also, while some studies of older reforms (when baseline spending levels were lower) report large marginal effects, these are imprecisely estimated; we find little statistical evidence that school spending exhibits diminishing returns. This implies that current impacts are likely similar to our pooled average. We do, however, find that marginal impacts are significantly smaller for economically advantaged populations than low-income populations (particularly for educational attainment)—with very different predicted policy impacts. For example, increasing per pupil spending by $1,000 over four years would increase college-going by 2 pp among low-income groups over 85 percent of the time, compared to less than 30 percent of the time for higher-income groups. Also, large effects above 5 pp would occur about 20 percent of the time for low-income groups, but almost never for more-advantaged groups.

One may worry that a pooled average of individually biased studies may be biased. We seek to avoid this by including only design-based studies that provide evidence supporting a causal interpretation. However, we also present several empirical tests indicating that problematic biases are unlikely among these studies. We also implement several tests for, and find little evidence of, any impact of possible publication bias.

This study moves beyond whether money matters. It is the first study to quantify the 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 quantify the extent of true treatment heterogeneity, quantify the range of causal impacts supported by the existing literature, and provide a plausible range of predicted impacts expected in other contexts.

The remainder of this paper is as follows. Section I discusses how we identify and select the studies to create our dataset for analysis. Section II describes how we compute the same underlying parameter for each paper. Section III presents the formal meta-analytic methods. Section IV presents our main results. Section V presents evidence of robustness to various assumptions and restrictions. Section VI accounts for potential biases and shows negligible effect on our main results. Section VII documents heterogeneity across population and study characteristics, and Section VIII concludes.

I. Data

We capture estimates from studies that examine the effect of policy-induced changes in K-12 per pupil spending on student outcomes. We seek to shed light on the causal impact of school spending policy on student outcomes. Accordingly, our inclusion criteria require that the variation in spending is driven by policy and is plausibly a valid instrument for school spending. To be included, a study had to (a) use a quasi-experimental design (Regression Discontinuity, Event-Study, Instrumental Variables, or a combination) to isolate the impacts of *specific* school spending policy shocks (or features of a school spending policy) on student outcomes; (b) demonstrate that analyses were based on policy-induced variation that had a statistically significant effect (i.e., p-value < 0.05)[2] on school spending—enough to facilitate exploring the effect of school spending on student test scores or educational attainment; and (c) provide evidence that the policy effect was only due to its effect via school spending, which entails including rich controls for plausible sources of bias and/or conducting a formal test that the policy instrument (after including controls) is unrelated to other reasonable predictors of outcomes (e.g., demographics or other policies).[3]

Our requirement that the policy has a statistically significant effect on spending (i.e., first stage $F > 3.85$) is not entirely innocuous. Studies with a weak first stage (i.e., first stage $F < 10$) may not be normally distributed and may have inaccurately

[2] All results are robust to restricting to first-stage F-stat above 20 (see Table 4).

[3] For all models, this includes testing that the policy instrument is unrelated to observable predictors of the outcomes, pre-trends, or other forms of selection.

reported sampling variability (Bound, Jaeger, and Baker 1995). The motivation for being more permissive rather than less is to include as representative a set of studies as possible. Only including studies with very strong first stages could lead to an unrepresentative sample of included studies and introduce potential sample selection bias.[4] That is, there is an inherent tension; we should not exclude papers with weak instruments, but one cannot trust the reported inference in these papers. To assuage concerns that our more inclusive standard biases our results, we are careful to show that our results are largely the same among studies with first-stage *F*-statistics above 10 and even above 20.

We identified 31 studies that met our conditions as of December 1, 2020. To identify potential studies, we started with the set in Jackson (2020) and those returned in Google Scholar searches on relevant terms (including "school spending" and "causal"). Those that met our inclusion criteria became a set of seed papers, each of which we fed into connectedpapers.com to obtain a list of related papers (exemplar in online Appendix A), and we evaluated each of these based on our inclusion criteria. Connected Papers reviews the Semantic Scholar Paper Corpus of papers across disciplines, and for a given seed paper reports papers according to "similarity," operationalized based on co-citations and bibliographic coupling. We iterated our process until there were no more new connected papers that met our inclusion criteria. This approach located any paper that cited other papers in the design-based school spending literature, so this should represent a near-comprehensive set of studies on the topic. 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 economics), and while the earliest study we located meeting these criteria was published in 2001, most studies meeting these criteria were written or published after 2015 (online Appendix Figure A.2).

Included Studies.—Table 1 summarizes basic information about each included study for each outcome analyzed in that study.[5] We provide a Study ID (second column) for each study-outcome estimate. Of 32 unique study-outcomes, 25 present estimates of test score impacts (either test scores or proficiency rates) and 12 present estimates of impacts on educational attainment (high school dropout, high school graduation, or college enrollment).[6] The studies represent a range of estimation strategies and sources of variation. These papers are also varied in terms of policies examined; six papers examine school finance reforms nationally, seven examine particular state-level school finance reforms, three examine school spending referenda, four look at school improvement grants, nine look at capital construction projects, and others identify effects of Title I or impacts of economic shocks on spending. They present widespread geographic coverage in terms of the populations treated: school finance reform studies cover all districts in treated states (about 25 out of 51); several studies examine large urban school districts, and two papers

[4] Angrist and Kolesár (2021) warn of the potential biases introduced by screening on the first-stage *F*-statistic.
[5] See online Appendix A for notes on specific studies that we did not include.
[6] Note that one study (Baron 2022) conducts independent analyses of the effects of capital and operational spending and in analyses, we treat these as two studies.

TABLE 1—SUMMARY OF STUDIES

Study	Study ID	Outcome	Spending type	Estimation strategy	Raw estimate ($\hat{\theta}_j$)	SE of $\hat{\theta}_j$ (se_j)	Bayes estimate ($\tilde{\theta}_j$)
Abott et al. (2020)	1	High school graduation	Operational	RD	0.0847	0.0876	0.0596
Abott et al. (2020)	1	Test scores	Operational	RD	0.1158	0.0667	0.0439
Baron (2022)	2	College enrollment	Operational	RD	0.1869	0.0767	0.0681
Baron (2022)	2	Test scores	Operational	RD	0.1790	0.1305	0.0388
Baron (2022)	3	Test scores	Capital	RD	−0.1579	0.0979	0.0216
Brunner, Hyman, and Ju (2020)	4	Test scores	Any	ES DiD	0.0531	0.0173	0.0469
Candelaria and Shores (2019)	5	High school graduation	Any	ES DiD	0.0511	0.0133	0.0529
Carlson and Lavertu (2018)	6	Test scores	Any	RD	0.0902	0.0475	0.0461
Cascio, Gordon, and Reber (2013)	7	High school dropout	Any	ES	0.5546	0.2056	0.0638
Cellini, Ferreira, and Rothstein (2010)	8	Test scores	Capital	RD	0.1773	0.0829	0.0458
Chaudhary (2009)	9	Test scores	Any	IV	0.0179	0.0424	0.0294
Clark (2003)	10	Test scores	Any	ES DiD	0.0148	0.0116	0.0181
Conlin and Thompson (2017)	11	Test proficiency rates	Capital	ES	0.0063	0.0047	0.0072
Gigliotti and Sorensen (2018)	12	Test scores	Any	IV	0.0424	0.0098	0.0413
Goncalves (2015)	13	Test proficiency rates	Capital	ES	−0.0017	0.0197	0.0117
Guryan (2001)	14	Test scores	Any	IV	0.0281	0.0689	0.0329
Hong and Zimmer (2016)	15	Test proficiency rates	Capital	RD	0.0911	0.0512	0.0448
Hyman (2017)	16	College enrollment	Any	IV	0.0552	0.0257	0.0568
Jackson, Johnson, and Persico (2016)	17	High school graduation	Any	ES DiD	0.0798	0.0163	0.0723
Jackson, Wigger, and Xiong (2021)	18	College enrollment	Any	IV	0.0380	0.0133	0.0432
Jackson, Wigger, and Xiong (2021)	18	Test scores	Any	IV	0.0499	0.0196	0.0438
Johnson (2015)	19	High school graduation	Any	ES DiD	0.1438	0.0753	0.0650
Kogan, Lavertu, and Peskowitz (2017)	20	Test scores	Any	RD	0.0190	0.0127	0.0219
Kreisman and Steinberg (2019)	21	High school graduation	Any	IV	0.0279	0.0146	0.0369
Kreisman and Steinberg (2019)	21	Test scores	Any	IV	0.0779	0.0237	0.0572
Lafortune, Rothstein, and Schanzenbach (2018)	22	Test scores	Any	ES DiD	0.0164	0.0133	0.0201
Lafortune and Schonholzer (2022)	23	Test scores	Capital	ES DiD	0.0504	0.0223	0.0430
Lee and Polachek (2018)	24	High school dropout	Any	RD	0.0640	0.0141	0.0623
Martorell, Stange, and McFarlin (2016)	25	Test scores	Capital	RD	0.0254	0.0226	0.0290
Miller (2018)	26	High school graduation	Any	IV	0.0662	0.0169	0.0632
Miller (2018)	26	Test scores	Any	IV	0.0515	0.0137	0.0474
Neilson and Zimmerman (2014)	27	Test scores	Capital	ES DiD	0.0314	0.0236	0.0324
Papke (2008)	28	Test proficiency rates	Any	IV	0.0817	0.0121	0.0728
Rauscher (2020)	29	Test scores	Capital	RD	0.0070	0.0041	0.0076
Rauscher (2020)	30	Test scores	Operational	DiD	0.0161	0.0271	0.0254
Roy (2011)	31	Test scores	Any	IV	0.3804	0.1563	0.0424
Weinstein et al. (2009)	32	High school graduation	Any	RD	0.1595	0.1698	0.0597
Weinstein et al. (2009)	32	Test scores	Any	RD	−0.0541	0.0368	0.0054

Notes: RD = Regression Discontinuity; ES = Event Study; DiD = Difference in Differences; IV = Instrumental Variable. Clustering by study-outcome, as well as studies of same policies, including OH capital subsidy program (Conlin and Thompson (2017) and Goncalves (2015), MI Proposal A (Chaudhary 2009; Hyman 2017; Papke 2008; and Roy 2011), Same-years SFRs (Lafortune, Rothstein, and Schanzenbach (2018) and Brunner, Hyman, and Ju 2020), and Title I (Johnson 2015 and Cascio, Gordon, and Reber (2013). Jackson, Johnson, and Persico (2016) and Johnson and Jackson (2019) are combined and study older school finance reforms. Johnson and Jackson (2019) omitted for duplication reasons with Jackson, Johnson, and Persico (2016).

focus on rural schools. The studies also cover policies implemented from the 1960s through the 2010s. The treated "super population" covered is broadly representative of the United States; this set of studies is well-suited to measuring heterogeneity, and to testing for heterogeneous effects.

II. Constructing the Same Parameter Estimate for All Papers

To make estimates comparable to each other, for each study we compute the overall effect of a $1,000 per pupil spending increase (in 2018 dollars using CPI 2020), sustained for four years, on standardized outcomes $y \in$ (test scores, educational attainment), $\hat{\theta}_j$. This is an instrumental variables (IV) estimate of school spending on outcome y using policy j as an instrument.[7] We present analyses for each study's single overall estimated effect, and separately for subsamples by income or urbanicity (when reported).[8]

Our study inclusion criteria require that each policy instrument is plausibly valid (relevant and excludable), so the meta-analytic average should be unbiased. While the relevance condition holds, on average (mechanically based on our inclusion criteria), the excludability condition may not. However, we present evidence that this condition is satisfied, on average (online Appendix F.1). We compute separate estimates for test score and educational attainment outcomes (detailed calculations in online Appendix Table A.1).[9] Because studies do not all report impacts as IV estimates, this often requires several steps (i.e., standardizing outcomes, aggregating across populations, equalizing the spending change, equalizing the duration of policy exposure, computing marginal spending impacts, and computing standard errors). We detail all steps and assumptions in online Appendix C, and show robustness to modeling assumptions in online Appendix E.

Table 1 summarizes our data.[10] For each study-outcome, we report the estimated marginal impact ($\hat{\theta}_j$) and its standard error, where positive values indicate a positive association between school spending and *improved* outcomes.

A. *Making Capital Comparable to Noncapital*

Noncapital spending goes toward inputs used in the same year (e.g., teacher salaries), while capital spending goes toward durable assets used for several years (e.g., building construction). As such, it is inappropriate to relate outcomes in a given year to spending on capital *that same year*. To account for this, we amortize capital

[7] While we do as much as possible (with the available information) to make estimates comparable, small differences may remain due to (a) different tests measuring different dimensions of student learning, and (b) different school districts having different high-school graduation standards. While we are careful to show that our conclusions are robust to alternative modeling decisions, any such differences will manifest as unobserved heterogeneity across studies. In the ideal, one would have sufficient information on each test to perform a linking across settings (as discussed in Reardon, Kalogrides, and Ho 2021) and sufficient information about graduation standards in each school district to ensure comparability. However, this is infeasible given the information provided in each study.

[8] In cases where there is an overall effect and effects for subsamples, we use both estimates. Note, however, to avoid duplication, we never include estimates for subsamples that are a linear combination of other estimates.

[9] We combine dropout rates, high school graduation rates, and college-going by standardizing these outcomes.

[10] We used the most recent version of each paper, and updated if unpublished working papers were updated.

outlays over the life of the asset. Specifically, we depreciate new buildings at 4.7 percent and renovations at 16.5 percent per year so that 90 percent of the value of the overall value is depreciated within 50 and 15 years, respectively. We then relate the amortized flow value of the capital outlay to student outcomes (detailed in online Appendix C.1).

Because capital projects typically take two years to complete, we take the effect in the sixth post-completion year as comparable to our four-year noncapital spending treatment effect. To assess whether this temporal decision is reasonable, Figure 1 presents the *raw*[11] dynamic effects of the nine studies estimating changes in capital spending on student test score outcomes. These are plotted over time relative to a baseline year zero ($t = 0$, the year of construction or policy change) in which there should be no effect of the policy on outcomes.[12] Consistent with an initial disruption caused by construction, in several cases there is an initial dip in outcomes, which is followed by a gradual increase in outcomes several years after increased capital spending. To more formally assess the evolution of outcomes over time, We present the average dynamic effect in panel B of Figure 1, plotting one through six years after the baseline year along with 90 and 95 percent confidence intervals. This suggests no change in the first two years and then gradually improving outcomes (which become statistically significant). This pattern validates our assigning the first two years of these studies to a "construction/disruption" period and using the year-six effect for capital spending increases as the most comparable to noncapital spending year-four effects.[13]

III. Meta-Analytic Methods

We perform random effects meta-analysis to generate precise overall pooled estimates of the average effect of spending on student outcomes *and* to estimate heterogeneity across studies (i.e., heterogeneity not due to sampling variability). This general approach allows us to credibly predict policy impacts one might expect to observe in new contexts (see Cochrane Handbook (Higgins et al. 2019) and Hedges 1983 for early discussions of this) and is increasingly being used in economics (e.g., Meager 2019; Vivalt 2020; Bandiera et al. 2021; and Dehejia, Pop-Eleches, and Samii 2021).

A. *A Model of the Distribution of Estimates and True Effects*

We suppress the subscript y, and let each study-outcome j have one estimate, $\hat{\theta}_j$, the *estimated* harmonized and standardized effect (on test scores or educational attainment). Each study has a real effect θ_j, and due to sampling error, estimates are

[11] These are not the marginal per $1,000 effects.

[12] For Lafortune and Schonholzer (2022); Neilson and Zimmerman (2014); and Goncalves (2015); year one ($t = 1$) represents the 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.

[13] One study, Conlin and Thompson (2017), only reports estimates three years after baseline, so we use that estimate in place of a year-six estimate.

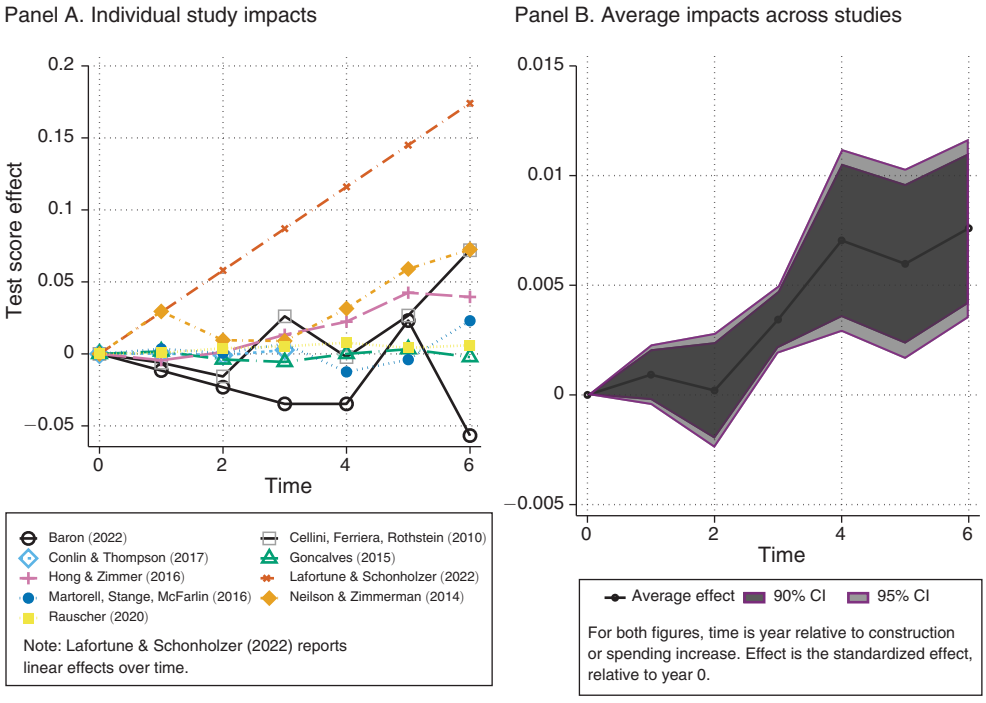

FIGURE 1. CAPITAL SPENDING EFFECTS OVER TIME

Notes: Panel A depicts the dynamic treatment effects on standardized test scores for the nine studies of policies that increased capital spending. All results are relative to the year before construction or the increase in spending (whichever is provided by the authors). Panel B plots the precision-weighted average of these policy effects across the studies in each treatment year (along with the 90 and 95 percent confidence intervals for each treatment year effect).

distributed as in (1). This distributional assumption follows from the central limit theorem.

$$\hat{\theta}_j \sim \mathcal{N}\left(\theta_j, \sigma_j^2\right) \tag{1}$$

The true effects of individual studies deviate from the grand mean (Θ) due to heterogeneity with variance τ^2 according to some distribution $g\left(\Theta, \tau\right)$. In Section IVA, we show that this distribution is approximately normal. Accordingly, we follow convention and assume normality, such that the true effects are distributed as in (2).

$$\theta_j \sim \mathcal{N}\left(\Theta, \tau^2\right) \tag{2}$$

From (1) and (2), study estimates deviate from the grand mean (Θ) due to within-study sampling error and between-study heterogeneity, and are distributed as in (3).

$$\hat{\theta}_j \sim \mathcal{N}\left(\Theta, \sigma_j^2 + \tau^2\right) \tag{3}$$

If the marginal effects are independent of the precision of the estimates, then the optimal inverse-variance, precision-weighted pooled average is $\hat{\Theta}_{pw} = \sum \hat{\theta}_j w_j / \sum w_j$, where each estimate receives weight w_j as in (4).

$$w_j = \frac{1}{\left(\sigma_j^2 + \tau^2\right)} \tag{4}$$

In Section IV, we show that this independence assumption is satisfied for test score effects, but that larger educational attainment estimates tend to be less precise. We discuss the implications of this dependence and show that, if anything, our estimates are conservative.

To form the empirical analog of $\hat{\Theta}_{pw}$, and (4), we use the square of the standard error (se_j^2) as an estimate of σ_j^2, and estimate τ^2 empirically. $\hat{\tau}^2$ is identified based on the difference between the observed variability across estimates and that which would be expected due to sampling variably alone.[14] Intuitively, overlapping confidence intervals for individual estimates would suggest that τ^2 is small, while nonoverlapping intervals would suggest heterogeneity. This model is implemented using weighted least squares, which provides estimates $\hat{\tau}^2$, $\hat{\Theta}_{pw}$, and the standard error of the pooled average ($se_{\hat{\Theta}_{pw}}$). To shed light on the potential variability of the hyperparameter $\hat{\tau}$, we report the standard error of $\hat{\tau}$ estimated by the standard deviation of 500 bootstrap replications.[15] Also, note that we obtain very similar results using full Bayesian methods reported in online Appendix Table A.17.

When studies report separate effects by income level, urbanicity, or spending type, we include multiple (nonindependent) estimates per study. Also, some *different* studies examine the same policy (Table 1 notes) and are therefore not independent. To account for both sources of correlated estimates, we cluster estimates from the same study *and related policies from different studies* using Robust Variance Estimation (as in Hedges, Tipton, and Johnson 2010). This model accounts for a possible small number of clusters by using small-sample adjustments (as in Tipton 2015).[16]

The confidence interval for the pooled average uses the standard error of the mean as in (5).

$$CI = \hat{\Theta}_{pw} \pm t^* \times se_{\hat{\Theta}_{pw}} \tag{5}$$

[14] See Borenstein et al. (2017) for the formal derivation of the method of moment estimate we employ. The estimate of τ reflects the variance of weighted estimates around the pooled mean minus a weighted average of the observed sampling variability. Importantly, this estimate does not rely on the normality of the distribution of θ_js around Θ.

[15] Kernel density plot for the bootstrap distribution of $\hat{\tau}$ are in online Appendix Figure A.4.

[16] We implement these estimators using the "robumeta" package in Stata (Hedberg, Pustejovsky, and Tipton 2017). This uses the DerSimonian and Laird (1986) method of moments approach to estimating τ^2. This approach yields estimates similar to Bayes models (online Appendix G).

The prediction interval, which represents the range of true effects one can expect to observe in a new setting, is given by (6). This is wider than the confidence interval because it also accounts for heterogeneity across studies.

$$PI = \hat{\Theta}_{pw} \pm t^* \times \sqrt{se^2_{\hat{\Theta}_{pw}} + \hat{\tau}^2} \tag{6}$$

B. *Better Estimates of Individual Studies*

A key benefit of the Bayesian approach to meta-analysis is that one can use information from other similar studies to learn about the real impact on any individual study (Efron and Morris 1973; Morris 1983). Intuitively, if other studies are informative about what estimates are more or less plausible, we can use this information to better predict the true effect of individual studies. Formally, under (1) and (2), the expected value of the true effect for study j is (7), where $B = \left(\sigma_j^2\right)/\left(\sigma_j^2 + \tau^2\right)$.

$$E\left[\theta_j \middle| \hat{\theta}_j, \sigma_j, \tau\right] = B \times \Theta + \left(1 - B\right) \times \hat{\theta}_j \tag{7}$$

Replacing, σ_j, τ, and Θ with their estimates, (7) yields the Best Linear Unbiased Prediction (BLUP) of the true effect for study j given the data (i.e., the best estimate of the *true effect* in a similar setting). The BLUPs ($\tilde{\theta}_j$)—also empirical Bayes—are weighted averages of the individual estimates and the pooled average, where more precise estimates receive greater weight (see Table 1).

IV. Results

In Figure 2 we plot each study-estimate $\hat{\theta}_j$, the 95 percent confidence interval for each estimate, and the 95 percent prediction interval for the grand mean Θ.

Test Score Impacts.—The forest plot at the top of Figure 2 shows that most test score estimates are positive, the most precise estimates are in the middle of the distribution, and the noisiest estimates are at the extremes—indicating that the spread of *raw* estimates vastly overstates true heterogeneity across studies. A naïve equal-weighted average of the effects (ignoring the fact that noisier estimates are less reliable) is 0.0451σ—indicating that increasing per pupil spending by \$1,000 over four years, in an average setting, would improve test scores by 4.5 percent of a standard deviation (See Table 2). However, the fact that the largest positive estimates are the least precise suggests that our precision-weighted estimate (which is more reliably estimated) will be conservative relative to this raw average.

We present meta-regression results in Table 3. For each model, we report the pooled average (and its standard error) and the estimated between-study variability ($\hat{\tau}$). The pooled test score estimate (column 1) implies that, on average, a \$1,000 increase in per pupil spending (in 2018 dollars and sustained over four years) would increase average test scores by 0.0316σ (p-value < 0.0001). The 95 percent confidence interval for this pooled average is between 0.021σ and 0.043σ, which overlaps with the 95 percent confidence interval for *almost all* individual studies—indicating considerable consistency across studies. As expected, the precision-weighted

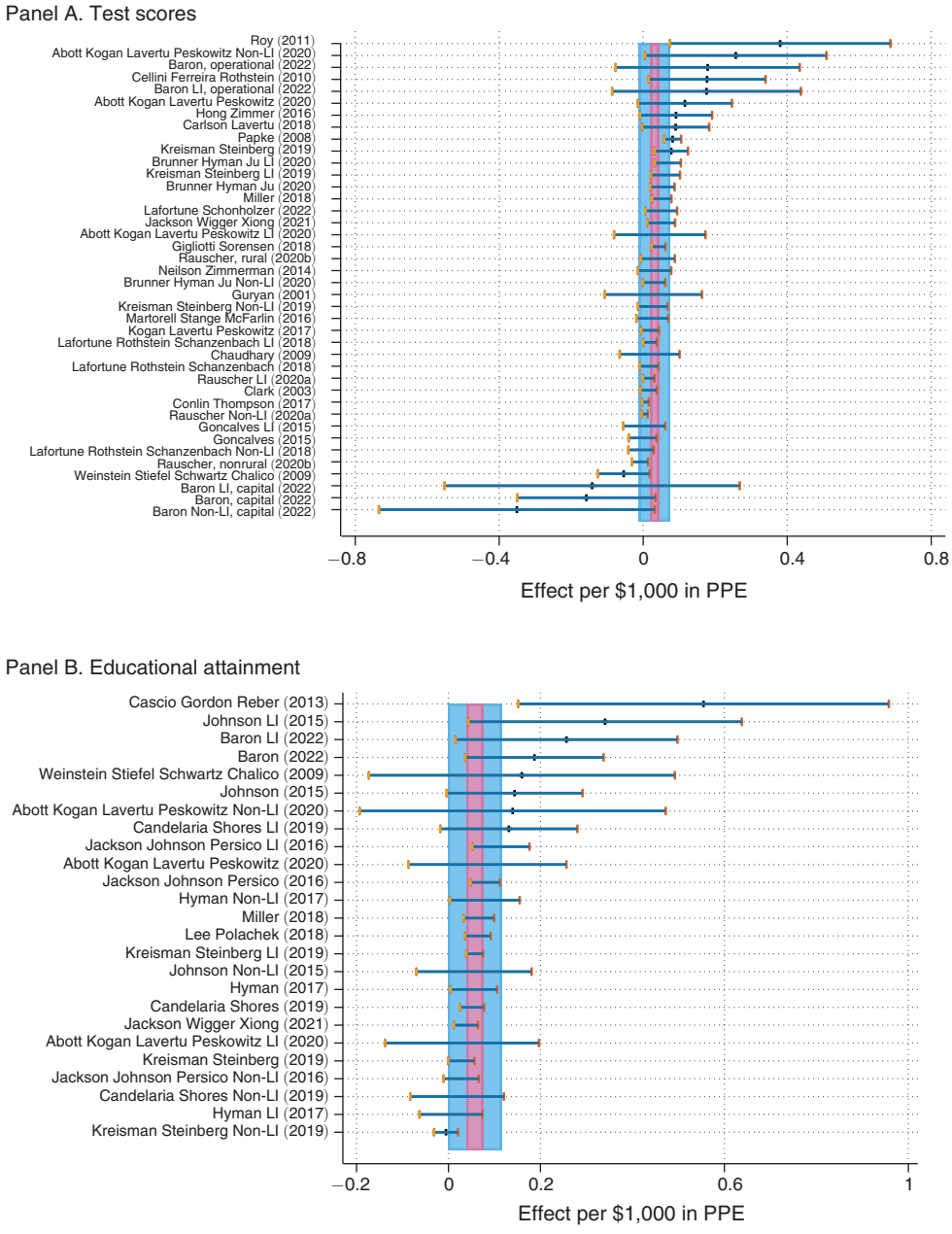

FIGURE 2. FOREST PLOT: MULTIPLE ESTIMATES PER PAPER

Notes: The top panel shows estimates from papers that examine effects on test scores, and the bottom shows estimates from papers that examine spending effects on educational attainment. Each estimate represents the marginal effect of a $1,000 per pupil spending increase sustained over four years on standardized outcomes. The error bars represent the 95 percent confidence interval for each estimate. We show the 95 percent confidence interval for the pooled overall effect in pink and the 95 percent prediction interval in blue.

Table 2—Equal Weighting Comparison

	Test scores				Educational attainment			
	RE (multiple) (1)	Equal weight (multiple) (2)	RE (single) (3)	Equal weight (single) (4)	RE (multiple) (5)	Equal weight (multiple) (6)	RE (single) (7)	Equal weight (single) (8)
Average effect	0.0316	0.0451	0.0323	0.0478	0.0573	0.102	0.0568	0.106
	(0.00559)	(0.0178)	(0.00638)	(0.0159)	(0.00860)	(0.0243)	(0.00722)	(0.0285)
Observations	40	40	26	26	25	25	12	12
SD of $\hat{\theta}_j$'s	0.037	0.111	0.038	0.093	0.047	0.122	0.034	0.144
τ	0.0207	0.0630	0.0221	0.0606	0.0267	0.0608	0.0167	0.0739

Notes: Columns 1, 3, 5, and 7 are all precision-weighted estimates from a Random Effects (RE) meta-regression. In the Random-Effects models, the errors are adjusted for clustering within related papers. These are our preferred estimates. To provide a basis for comparison, columns 2, 4, 6, and 8 present naïve equal-weighted averages. For the equal-weighted models, we assign the average of the standard error to all papers and use the robumeta command. Note that the estimates for τ are almost identical to the square root of the raw variance of the estimates minus the square of the average standard error. Standard errors in parentheses. For random effects models, standard errors are adjusted for clustering of related papers.

average is somewhat smaller than (i.e., a more conservative estimate than) the equal-weighted pooled average.[17]

Despite a precisely estimated positive pooled estimate, some studies have nonoverlapping margins of error (e.g., Papke 2008 and Rauscher 2020a)—suggesting heterogeneity. The standard deviation of heterogeneity across studies ($\hat{\tau}$) is 0.0207σ. That is, two randomly chosen policies would have true impacts that differ, on average, by 0.0207σ due to this heterogeneity. An implication of our pooled estimate 0.0316σ, our estimated heterogeneity 0.0207σ, and the normality of true effects is that while the pooled effect is 0.0316σ, in other contexts one can expect estimates between -0.004σ and 0.067σ about 90 percent of the time (Figure 2). This prediction interval overlaps the 95 percent confidence intervals of *all* the studies, and estimates that lie far from this range are imprecise.

To shed further light on likely future policy effects in an average setting similar to that of the super population (and with a mix of capital and operational spending), we take our estimated effects as given, and quantify (based on the distribution of observed effects) the probability of observing an effect of a particular size (or larger) in a different context (Figure 3). More concretely, we report on the distribution of true effects implied by equation (2). Specifically, we take our estimate of τ as given (i.e., we assume that $\hat{\tau} = \tau$) and report the cumulative probabilities associated with $N\left(\hat{\Theta}_{pw}, se^2_{\hat{\Theta}_{pw}} + \tau^2\right)$ to account for noise in our estimate of Θ. We provide more detailed prediction ranges for specific policy types in Section VII. In an average setting, a policy that increases school spending by $1,000 over a four-year period would increase test scores over 93 percent of the time. An implication of this result is that we cannot rule out the possibility of small negative marginal effects

[17] Because the equal-weighted approach does not down-weight imprecise estimates (which tend to be large), the raw variance of the estimates is higher than the precision-weighted variance, leading to much larger estimates of τ.

Table 3—Meta-Regression Estimates

	Test scores				Educational attainment		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Overall	0.0316	0.0343	0.0359	0.0323	0.0573	0.0574	0.0568
	(0.00559)	(0.00681)	(0.00747)	(0.00638)	(0.00860)	(0.00786)	(0.00722)
Capital		−0.00769	−0.00533				
		(0.0119)	(0.0112)				
Low income			−0.00470			0.0217	
			(0.0107)			(0.0215)	
Non-low income			−0.0198			−0.0336	
			(0.0102)			(0.0193)	
Capital		0.027	0.031				
(SE)		(0.010)	(0.009)				
LI − non-LI			0.015			0.055	
(SE)			(0.008)			(0.028)	
Observations	40	40	40	26	25	25	12
Clusters	22	22	22	22	11	11	11
τ	0.0207	0.0211	0.0210	0.0221	0.0267	0.0219	0.0167
$\hat{\sigma}_\tau$	(0.004)	(0.004)	(0.005)	(0.005)	(0.009)	(0.007)	(0.006)
Average 80% PI	[0.004,0.059]	[0.003,0.065]	[0.006,0.066]	[0.003,0.062]	[0.022,0.093]	[0.018,0.097]	[0.034,0.080]
Average 90% PI	[−0.004,0.067]	[−0.006,0.074]	[−0.003,0.075]	[−0.006,0.070]	[0.011,0.104]	[0.007,0.108]	[0.027,0.087]
Average 95% PI	[−0.011,0.074]	[−0.013,0.082]	[−0.010,0.082]	[−0.013,0.077]	[0.002,0.112]	[−0.003,0.118]	[0.021,0.093]

Notes: Standard errors in parentheses are adjusted for clustering of related papers. Out of 26 total estimates of test score outcomes, 9 are estimates of effects of capital spending.. The reported $\hat{\sigma}_\tau$ is estimated by bootstrap. Columns 4 and 7 report for one estimate per study outcome.

in an average setting.[18] This same policy would improve test scores by at least 0.01σ 85 percent of the time, 0.0316σ half the time, 0.05σ about 19 percent of the time, and 0.08σ about 1 percent of the time. These prediction bounds are consistent with the fact that most estimates are positive, and the largest estimates are very imprecise. Because there are multiple studies with estimates greater than 0.1σ, one may naïvely expect that this is likely in other settings and make poor policy predictions. However, meta-analysis reveals that these estimates are likely imprecise estimates of smaller "true" effects, allowing for *much* more reliable policy predictions. Indeed, the largest raw estimate (Roy 2011) is 0.38, while its predicted *real* effect is only 0.038 (Table 1).

Educational Attainment Impacts.—Panel B of Figure 2 shows estimated impacts on educational attainment, *all* of which are positive, and the largest estimates are the least precise. Indeed, the largest raw estimate (Cascio, Gordon, and Reber 2013) is 0.55, though the BLUP of its true effect is 0.0638—underscoring the importance of accounting for the role of noise when interpreting these seemingly large estimates. As with test scores, the fact that the largest estimates are the least precise suggests that our precision-weighted estimate is conservative relative to the raw average. The pooled precision-weighted effect for educational attainment (Table 3, column 5)

[18] We confirm this with formal tests that all effects are weakly positive—comparing the lowest t-statistic in our data the distribution of the minimum of 26 standard normal random draws.

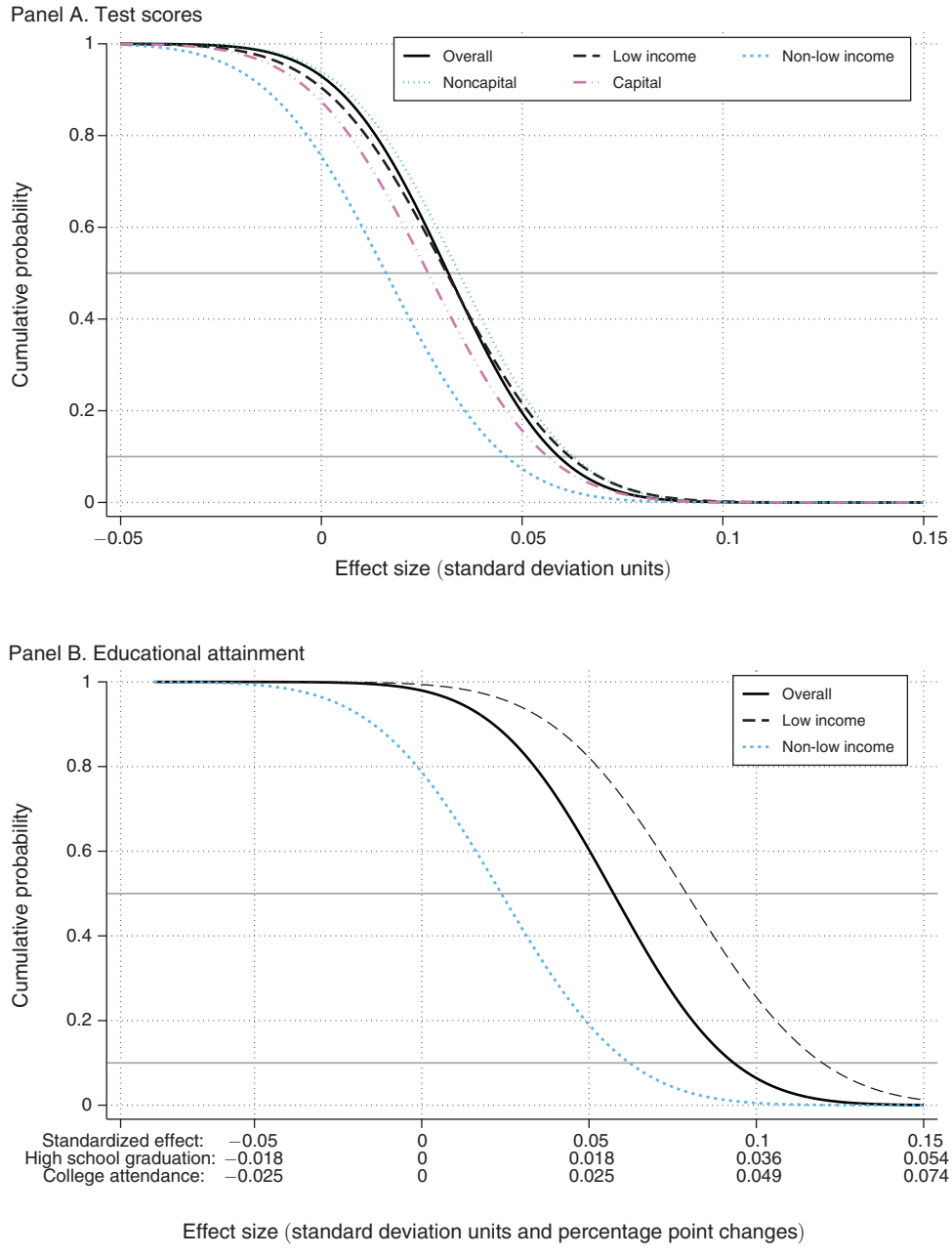

FIGURE 3. PROBABILITIES OF EFFECT SIZES

Notes: Each curve shows the probability that $\theta_j >$ Effect Size, based on estimated τ and $se_{\hat{\Theta}_{pw}}$ from Table 3. Each line reports the probability that a real effect is greater than a particular effect size based on the pooled average, its standard error, the estimated heterogeneity, and the assumption that true effects are normally distributed around the pooled average. Specifically, we (a) take our estimate of τ as given (so that $\hat{\tau} = \tau$), (b) include $se_{\hat{\Theta}_{pw}}$ to account for noise in our estimate of the grand mean Θ, and then (c) plot the cumulative probabilities for $N\left(\hat{\Theta}_{pw}, se^2_{\hat{\Theta}_{pw}} + \tau^2\right)$ for a range of marginal effect sizes. All predictions come from models in Table 3. For test scores, the overall predictions (for an average population and mix of capital and operational spending) correspond to model (1), low income and non-low income predictions come from model (2), and capital and noncapital predictions come from model (3). For educational attainment, the overall estimates correspond to model (4), and low income versus non-low income correspond to model (5).

is 0.0573σ (p-value < 0.001). As expected, this is considerably smaller than the simple equal-weighted average of 0.102σ (Table 2).[19]

To aid interpretation, we convert these standardized impacts into high school completion and college-going rates. For high school graduation (with a 2018 standard deviation of 0.357), on average, increasing school spending by \$1,000 (sustained for four years) would increase rates by $0.357 \times 0.0573 = 2.05$ pp. For college-going rates (with a 2018 standard deviation of 0.492), that same policy would increase postsecondary attendance rates by $0.49 \times 0.0573 = 2.81$ pp. Despite a wide range of raw estimates, 23 out of 25 estimates have 95 percent confidence intervals that include the pooled average—underscoring how ignoring estimation errors can lead to a *very* inflated view of heterogeneity.[20] Indeed, the range of BLUPs (Table 1) is considerably narrower than the range of study estimates. However, we do find evidence of heterogeneity ($\hat{\tau} = 0.0267\sigma$), yielding a prediction interval of the likely range of estimates in a different setting (Figure 2).

Taking our estimates as given and assuming normality of the true effects, a policy that increases school spending by \$1,000 per pupil for four years, in an average setting with some average mix of operational and capital spending, would lead to educational attainment impacts between 0.002σ and 0.112σ about 95 percent of the time. This implies high school completion impacts between $0.002 \times 0.357 = 0.07$ pp and $0.112 \times 0.357 = 3.99$ pp about 95 percent of the time, and college-going impacts between $0.002 \times 0.49 = 0.9$ pp and $0.112 \times 0.49 = 5.51$ pp 95 percent of the time. We report likely policy impacts in Figure 3. The magnitude of the average effect relative to heterogeneity is sufficiently large that a policy that increased school spending by \$1,000 over a four-year period would increase educational attainment over 98 percent of the time. In fact, formal tests that compare the minimum observed z-score to the distribution of the minimum of 25 random normal variables fail to reject that all the effects are weakly positive. That policy which increases school spending by \$1,000 per pupil for four years is expected to increase high school completion by at least 1 pp over 85 percent of the time, 2 pp half the time, and 3 pp over 15 percent of the time. That same policy will increase college-going rates by at least 1 pp over 90 percent of the time, 2.9 pp half the time, and 4.5 pp over 10 percent of the time.

Benchmarking.—To put these estimates into perspective, we compare the magnitudes of school-spending impacts to the effects of other educational interventions.

Project STAR: Reducing class size by roughly seven students increased test scores by 0.12σ, and college-going rates by between 1.8 pp (by age 20) and 2.7 pp (by age 30) (Chetty et al. 2011; Dynarski, Hyman, and Whitmore Schanzenbach 2013). As such, our pooled \$1,000 test score effects are equivalent to reducing class

[19] As with test scores, giving equal weight to imprecise estimates (which tend to be large and positive) inflates the pooled average and the raw variance—leading to large estimates of heterogeneity.

[20] Consistent with this, the estimated heterogeneity in the equal-weighted model is almost three times that of the preferred estimate.

size by 1.8 students, while our college-going impacts are equivalent to reducing class size by between 10 and 7.3 students.

Teacher Quality: Chetty, Friedman, and Rockoff (2014) find that increasing teacher quality by one standard deviation increases test scores by 0.12σ and college going by 0.82 pp. Our $1,000 test score impacts on test scores and college-going are equivalent to increasing teacher quality by 0.26 and 3.4 standard deviations, respectively.

For both benchmarking interventions, the spending impacts on educational attainment are at least twice as large as those on test scores. This same pattern holds within five of the six studies that examine impacts on both outcomes. 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, Johnson, and Persico 2016) or, more generally, the benefits of school quality on student outcomes (Beuermann et al. 2023; Jackson 2018; Jackson et al. 2020). However, we do note an alternative interpretation of this pattern is that both these benchmarking interventions happen to raise test scores and have relatively smaller effects on educational attainment.

A. *Testing Modelling Assumptions*

Assessing the Normality Assumption.—The distributional claims made in this paper rely on the assumption that the true effects are normally distributed around the grand mean (Θ). We implement several tests to justify this modeling decision (detailed in online Appendix D). First, following Wang and Lee (2020), we implement the Shapiro-Wilk normality tests on appropriately standardized effect sizes—a simple approach has been found to detect deviations from normality. For both outcomes, one fails to reject the null hypothesis of normality. Second, we implement meta-analytic models that allow for outliers by modeling the distribution of effects using a *t*-distribution (e.g., Lee and Thompson 2008), or a mixture of normal distributions (e.g., Beath 2014).[21] These models yield similar results to the normal model, and likelihood-ratio tests fail to reject the null of normally distributed effects.

We also implement a deconvolution kernel approach following Delaigle, Hall, and Meister (2008) to fit the distribution of true effects using a Fourier transform (as in Wang and Wang 2011). We set the tuning parameter (a bandwidth of 0.008) to closely match the variance estimate in the one per study sample. The deconvolved distributions do exhibit some mild positive skew, but these deviations cannot be distinguished from chance. The deconvolved distributions are well approximated by the normal distribution for both outcomes (see Figure 4), being single-peaked, roughly symmetrical, with relatively flat tails. Indeed, the cumulative densities are very similar using a normal distribution and the flexible approach (Figure 4). Finally, we implement Efron (2016) empirical Bayes (EB) deconvolution of the *z*-scores that allows for confidence bounds (following Kline, Rose, and Walters 2022). This

[21] This is implemented using the metaplus package in R.

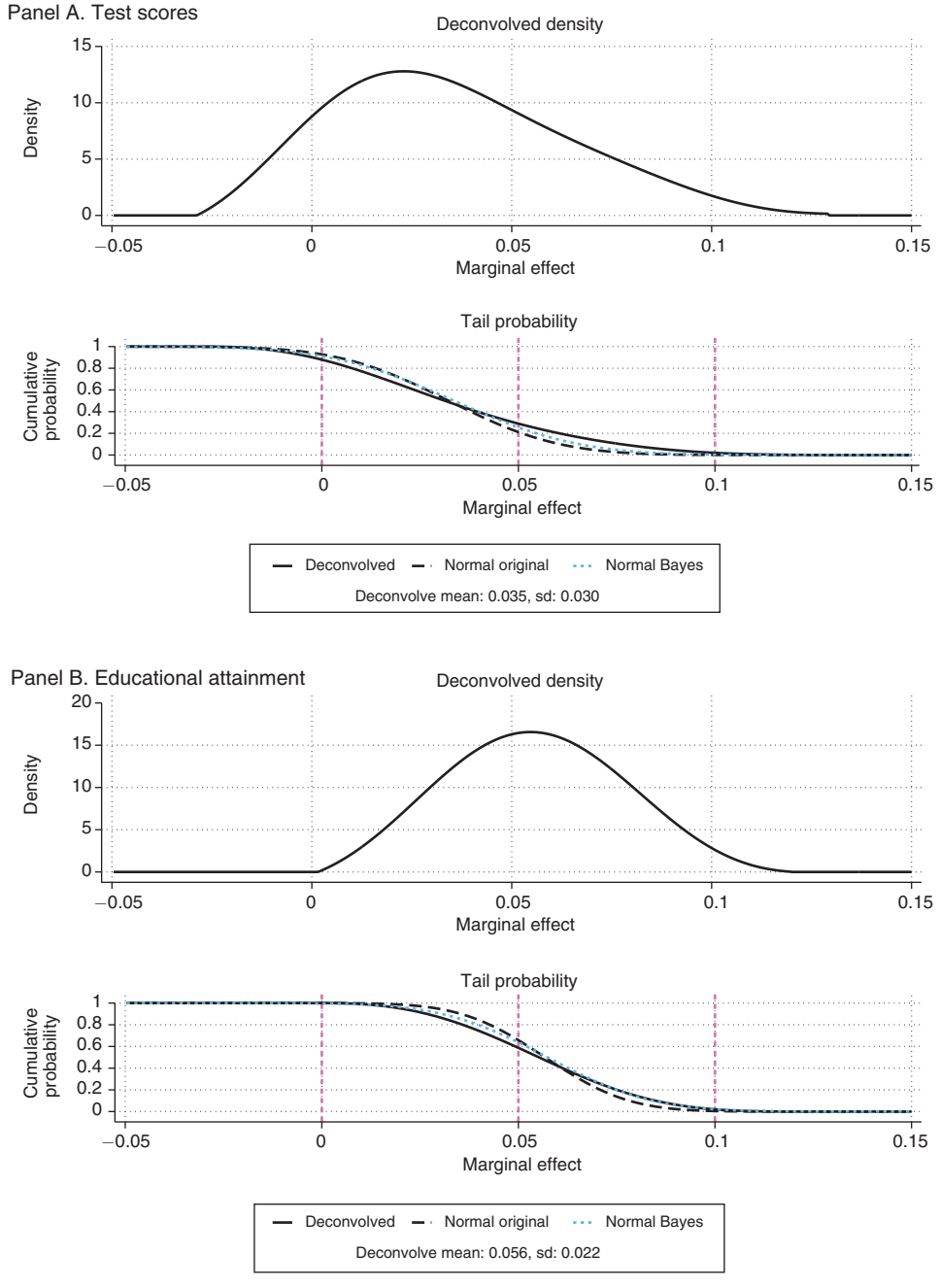

FIGURE 4. DECONVOLVED DENSITY AND CUMULATIVE DENSITY

Notes: The top of Panel A presents the deconvolved density distribution of true test score effects using a Fourier transform. The bottom of panel A is the cumulative probability based on this same deconvolved distribution. To provide a basis for comparison, we also plot the cumulative probability for a normal distribution with mean $\hat{\Theta}_{pw}$ and standard deviation $\hat{\tau}$ from model (6) in Table 3. Panel B presents analogous figures for education attainment, where the normal estimates come from model (8) in Table 3.

approach models the underlying distribution with an exponential family flexibly parameterized by an eighth-order spline. For both outcomes, the deconvoluted distribution of z-scores is approximately normal (Figure 4)—bolstering our modeling decision and the plausibility of the resulting predicted policy impacts.

Dependence between Estimates and Variance.—In Section III we assume that the marginal effects are independent of the sampling error. If there is some dependence such that more positive estimates tend to be noisier, then the pooled average will disproportionately down-weight larger positive estimates leading to a downward-biased pooled average, and vice versa. It is well understood that there could be some dependence if noisy estimates of a particular sign are systematically not published (discussed in Section VI). However, such dependence could also exist due to nonlinearity of effects (e.g., larger marginal effects may occur with smaller spending changes, which would tend to lead to larger estimation errors in an IV setting) or other forms of heterogeneity (e.g., policies in small locations may be more targeted and therefore yield larger marginal effects but have smaller datasets, resulting in larger standard errors). While we cannot know the source of any dependence, we can assess the extent of any dependence and determine the likely direction and magnitude of any bias.

To test for dependence between our estimates and their precision, we regress our estimates on the natural log of the estimated standard error. To facilitate interpretation of the constant, we subtract the sample median from the log standard error. We estimate this relationship in a few different samples (all estimates, those with first-stage F-statistics above 20, one estimate per study, and removing outlier estimates) and find no systematic relationship for test scores, but we do find a robust positive relationship for educational attainment (see online Appendix Figure A.18 and Table A.6). The independence assumption approximately holds for test score studies, but there *is* a robust positive relationship between the marginal effect and precision for educational attainment. As such, we focus our discussion on the implications for educational attainment results.

If the positive dependence we uncover is due to heterogeneity or nonlinearity (and not sample selection bias), then our pooled average would be biased down and would be a lower bound. To assess the extent to which this may affect our estimates, we use the estimated mean ($\hat{\Theta}$) implied by the deconvolution discussed in the previous subsection, which should recover the true pooled average even with violations of normality and with dependence between precision and the estimates.[22] The mean and standard deviation of the deconvoluted distribution are 0.056σ and 0.022σ, respectively. These estimates are very similar to those reported in Table 3 and are well within the estimated variability of the estimates. That is, insofar as our reported effects on educational attainment are biased down, the impact of this is minimal.[23] However, if this dependence reflects publication bias, then our reported

[22] The deconvolution approach yield estimated densities for particular values of θ. This can then be used to compute the mean and standard deviation of true effects.

[23] The deconvolution-based estimates for test scores are also similar to those reported in Table 3.

Table 4—Meta-Analysis, by Strength of First Stage

	F-stat $>$ 10		F-stat $>$ 20					
	Overall test scores	Overall ed. attainment	Equal weight test scores	Overall test scores	Noncapital test score	Capital test score	Equal weight ed. attainment	Overall ed. attainment
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Average effect	0.0325	0.0537	0.0342	0.0329	0.0480	0.0121	0.0561	0.0551
	(0.00592)	(0.00853)	(0.0263)	(0.00799)	(0.00333)	(0.00609)	(0.00833)	(0.00978)
Observations	30	13	18	18	9	9	8	8
SD of $\hat{\theta}_j$'s	0.113	0.143	0.142	0.037	0.022	0.025	0.034	0.030
τ	0.0217	0.0281	0.0809	0.0197	0.00752	0.00890	0.0181	0.0191

Notes: These models are estimated only among the set of papers with first-stage F-statistics above 10 (columns 1 and 2) and above 20 (columns 3–8). Columns 3 and 7 report naïve equal-weighted averages and the resulting estimates of heterogeneity (analogous to those presented in Table 2). Columns 4, 5, 6, and 8 present the precision-weighted, random-effects meta-regression results (for different outcomes and subsamples). Standard errors in parentheses are adjusted for clustering of related papers.

estimate would be biased upward. In Section VIB we assess this possibility in detail and find that any such bias (if it exists) is likely minimal.

V. Robustness to Modeling Decisions and Sample Restrictions

To harmonize estimates across studies, we make several assumptions. In online Appendix E.2 we show that these choices, while important, do not alter our conclusions. Our results (a) do not change if we only include studies with strong first stages (see columns 1 and 2 of Table 4), have a single estimate per study (see Table 3), or remove studies with binary test-score measures; and (b) are robust to upper and lower bounds for a variety of assumptions including the correlation between effects across subjects, grades or income groups; the time horizon and depreciation rate used to convert large one-time capital outlays to a flow value; the correlation between the errors for impacts on outcomes and impacts on spending; or how we convert outcomes to student-level standard deviations. Our results are also robust to only focusing on those studies that report IV estimates directly, for which we do not have to make any modeling decisions. As a final robustness check, in Table 4 (columns 3–8) we replicate our main approach only on the sample of studies with strong first stages (F-stat $>$ 20), and present funnel plots for this sample in online Appendix B. Reassuringly, for both outcomes, the random-effects pooled average and heterogeneity estimate are very similar to those using the full sample—assuaging concerns that our results are skewed by the inclusion of weakly identified studies.[24]

[24] Remarkably, among only well-identified studies, the equal-weighted and precision-weighted pooled averages are almost identical. The equal-weighted and precision-weighted pooled averages for test scores are 0.034σ and 0.033σ, respectively (Table 4). However, as one would expect, estimates of τ are much smaller in the precision-weighted model.

VI. Assessing Bias in Individual Studies and Publication Bias

A. *Confounding Bias*

One may worry about bias due to confounding or specification errors in individual studies. We summarize a simple framework within which to think through this point (see online Appendix F for greater detail) and provide empirical tests showing this is not a major concern in our setting.[25] For ease of exposition, we abstract away from treatment heterogeneity. The change in the standardized outcome y due to policy j is Δy_j. This is the total effect of the change in spending caused by the policy $(\Theta \times \Delta \$_j)$, plus random noise (υ_j), plus possible bias (b_j) as in (8):

$$\Delta y_j = \left(\Theta \times \Delta \$_j\right) + \upsilon_j + b_j \tag{8}$$

Dividing by the change in spending $(\Delta \$_j)$, we get each study's marginal effect:

$$\hat{\theta}_j \equiv \frac{\Delta y_j}{\Delta \$_j} = \underbrace{\Theta}_{\text{True Average}} + \underbrace{\frac{\upsilon_j}{\Delta \$_j}}_{\text{Noise Ratio}} + \underbrace{\frac{b_j}{\Delta \$_j}}_{\text{Bias Ratio}} \tag{9}$$

Our pooled estimate $\hat{\Theta}_{pw}$ is a precision-weighted average of (9), which is the true average (Θ) plus an average noise ratio plus an average bias ratio. If estimation errors are random such that $E\left[\upsilon_j \middle| \Delta \$_j\right] = 0$, then $\hat{\Theta}_{pw}$ is unbiased so long as the average bias term is equal to zero in expectation. This is satisfied if the *average* bias ratio is approximately zero (which is trivially satisfied if the individual studies are unbiased). We show that this condition likely holds in our setting.

Test 1: Only Look at Studies with Strong First Stage. Bound, Jaeger, and Baker (1995); Conley, Hansen, and Rossi(2012), and others have pointed out that confounding biases tend to be more severe when the first stage relationship is weak. Relatedly, Raudenbush, Reardon, and Nomi (2012) points out that precision-weighted averages of IVs will be (approximately) unbiased so long as individual studies do not have weak first stages. When we remove studies that have weak first stages (F-stat < 20), we find little difference in our results (Table 4). This suggests no significant bias in our full sample.

Test 2: Relate the Magnitudes of the Spending Change to the Marginal Effect. In equation (9) the bias-ratio for study j represented by $b_j/\Delta \$_j$ is smaller for policies that generate larger changes in spending. If there were biases, then the marginal effects would be larger for small spending changes than for larger ones. We test this by regressing the marginal effect $(\hat{\theta}_j)$ against the magnitude of the spending change $(\Delta \$_j)$ (see online Appendix Figure A.7). For both outcomes the

[25] Note that there are some similarities between this framework and that laid out by Raudenbush, Reardon, and Nomi (2012), but the issues addressed here are about bias, whereas Raudenbush Reardon, and Nomi (2012) is primarily concerned with disambiguating heterogeneity from differences in, compliance rates for binary treatments.

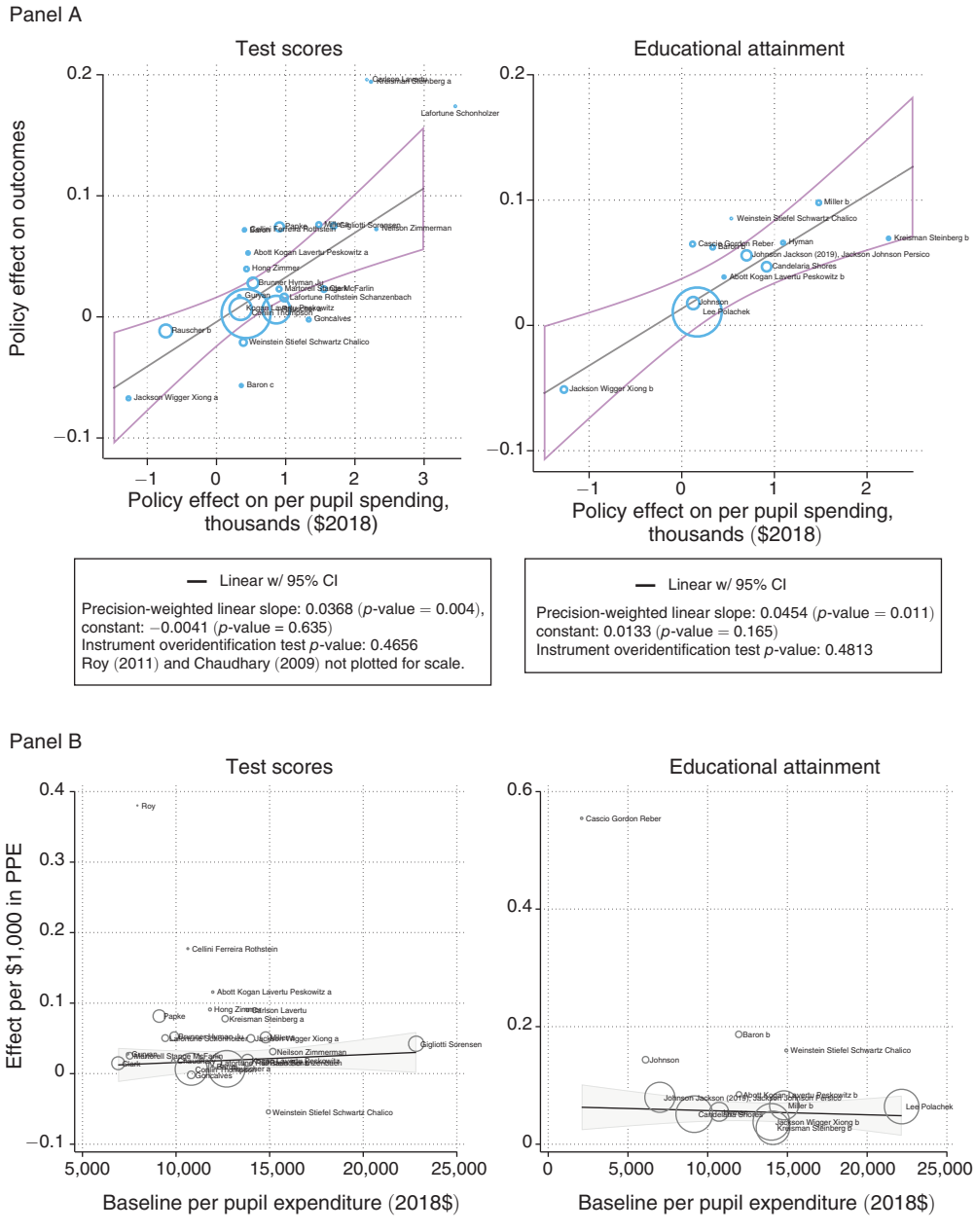

FIGURE 5. POLICY IMPACTS AGAINST INCREASE IN SPENDING AND MARGINAL IMPACTS AGAINST BASELINE SPENDING

Notes: Panel A is a scatterplot of each policy's effect on standardized outcomes (Δy_j) against its effect on spending ($\Delta \$_j$). Panel B is a scatterplot of each policy's marginal effect on standardized outcomes ($\hat{\theta}_j$) against its effect on spending ($\Delta \$_j$). For both plots, more precise estimates are depicted with larger circles and we plot the precision-weighted slope and its 95 percent confidence interval. For parsimony, each plots one estimate per study (see A.5 and A.6 in the online Appendix for figures with multiple estimates per study).

slopes are very close to zero and are not statistically significant, suggesting minimal bias.

Test 3: Only Look at Well-Powered Studies. While many studies tested for violations of the exclusion restriction, because some of our included studies may be underpowered, such violations may not have been detected. 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 assess this by excluding underpowered studies (i.e., those that do not have the power to detect the average effect). Using this approach, we obtain effects similar to our main estimates (online Appendix Table A.13).

Test 4: Examine Voluntary Policies versus Involuntary Changes. One may worry that places that voluntarily increase spending may also be more likely to do other things that improve student outcomes. Such dynamics would generate bias correlated with the spending increase and inflate the marginal estimate. We propose a test for this kind of bias. Specifically, we compare the average marginal effect for studies that rely on a new "voluntary" policy implementation (e.g., bond referendum resulting from a vote among those in the affected areas) versus those that rely on variation conditional on policies being in place (e.g., changes due to differential impacts of the recession or fluctuating student enrollment). Studies based on voluntary policy adoption are similar to other studies (online Appendix Table A.13), suggesting little bias of this form.

An Alternative Approach.—Equation (8) shows that one can also estimate Θ with a linear regression relating the overall policy effect (Δy_j) to the overall spending effect ($\Delta \$_j$). This bivariate regression approach does not require that the biases in the individual studies average out to zero (which is needed for an unbiased pooled average), but is a consistent estimate of Θ under the weaker assumption that the bias in individual studies is unrelated to the spending changes induced by the policy. For example, if all studies had the same positive bias (i.e., $b_j > 0$, $\forall j$), it would inflate the pooled average (since $E\left[b \mid \Delta_j\right] > 0$), but would not affect the bivariate regression slope (since $\text{corr}\left(b, \Delta \$_j\right) = 0$). As such, the extent to which the bivariate regression-based estimates differ from the pooled averages may be informative about systematic bias in all studies.[26]

To assess this in our setting, in Figure 5 we plot the raw, standardized overall effect of each policy on student outcomes against the change in per pupil expenditures (\$2018) caused by the same policy along with a precision-weighted regression relating the two and its 95 percent confidence interval. Each study is represented by a circle, and larger circles indicate more precise outcome estimates. For both

[26] 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*. To give a concrete example, imagine that there were only two studies: of Policy A and Policy B. Policy A increases per pupil spending by \$1,000 and test scores by 0.05σ (leading to $\hat{\theta}_A = 0.05$), while Policy B increases per pupil spending by \$2,000 and test scores by 0.04σ (leading to $\hat{\theta}_B = 0.02$). 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 scores, so the difference-based relationship is negative (i.e., $\hat{\Theta}_{diff} < 0$). While this may seem counterintuitive, if there is some correlation between the size of the policy and other contextual factors that determine policy efficacy, this could occur. 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.

outcomes, the relationship is positive and statistically significant, and one fails to reject that the bivariate regression estimate differs from the pooled average—suggestive of no bias.[27] To test whether all the data fall along the line, we use individual study indicators as instruments for $\Delta\$_j$ and report the p-value for the test of overidentifying restrictions. The model fails to reject that all the effects fall along the line for both outcomes–indicating that each policy's effect on spending generates achievement gains commensurate with the regression line. This implies that for any given policy change in per pupil spending, the linear relationship from the pooled model will provide a good predictor of the policy impact on outcomes.

A Suggestive Test of the Exclusion Restriction: The exclusion restriction for all included studies 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 (i.e., the bias terms in (8) should be equal to zero, on average). This motivates a simple test for whether the constant in the regression is zero. For test scores, the constant is -0.0026 with a p-value of 0.757, while for educational attainment it is 0.0133 with a p-value of 0.165. The signs of the constants are different for the two outcomes, suggesting no systematic bias (consistent with all the other tests).

B. *Publication Biases*

Our analysis may be biased if certain kinds of studies are systematically not published. We conduct several tests to assess whether publication biases are a concern. We visually represent estimates from these approaches in Figure 6, present regression results in online Appendix Table A.14, and summarize them here. While no single test can entirely rule out publication bias, taken as a whole, the several empirical tests we conduct are consistent with minimal bias.

(1) Studies that find null results may be less likely to be published than those with significant effects (Franco, Malhotra, and Simonovits 2014; Christensen and Miguel 2018). Accordingly, a test for publication bias compares estimates from published studies to those that are not. In online Appendix Table A.15, we compare average estimates of published and unpublished studies and find no difference in impacts.[28]

(2) If there are biases against the publication of certain kinds of studies, they may be most pronounced at the most selective journals (Brodeur et al. 2016). As such, we compare the average impacts of studies published in the most "elite" journals to studies published in other journals (in columns 2 and 4 of online Appendix Table A.15). We find no difference by journal type.

[27] 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.

[28] We cannot observe the unobservable—or those papers that are not shared fully in any form, published or not.

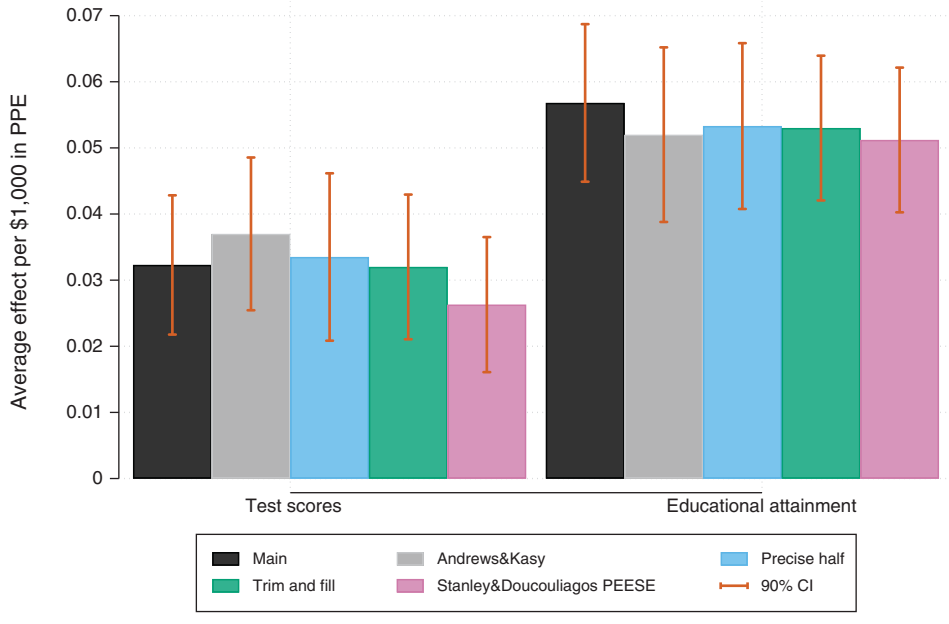

FIGURE 6. FOUR APPROACHES TO PUBLICATION BIAS

Notes: Each bar represents a precision-weighted average estimate for each outcome type using a different approach to account for possible publication bias (detailed in Section VIB). The main estimates are those with no adjustment for publication bias. The Trim and Fill estimates are those that include imputed "missing studies" to ensure that there is no dependence between the study estimate and its standard error. The Andrews and Kasy approach predicts each study's publication probability as a function of its statistical significance and upweighs estimates that are less likely to be published. The PEESE approach models the relationship between precision and estimate and uses the inferred most precise estimate as the truth. The precise half is the meta-analytic average based on only the most precise half of studies for each outcome. The bias-adjusted estimates presented here are reported in tabular form in online Appendix Table A.14. All approaches yield similar estimates.

(3) When there are biases against the publication of insignificant studies, one might observe an overrepresentation of studies right at the significance threshold (in social sciences, this would be a *t*-statistic of 1.96) (Brodeur, Cook, and Heyes 2020). We test for a discontinuous jump in the cumulative density of *t*-statistics at 1.96, and find no consistent evidence of this (see onine Appendix Figure A.23 and Table A.16).

(4) We implement a model that accounts for any selection (should it exist) of significant effects. We show results for the Andrews and Kasy (2019) selection adjustment using their web application in online Appendix Figures A.24 and A.25. They propose estimating the relative publication probabilities (based on the *t*-statistics), and using these to re-weight the distribution of studies to account for differences in publication probability (up-weighting studies that are least likely to be observed). For both outcomes, their model fails to reject the null of no selection, and their adjustment approach yields similar estimates to our preferred model (columns 1 and 5 of online Appendix Table A.14).

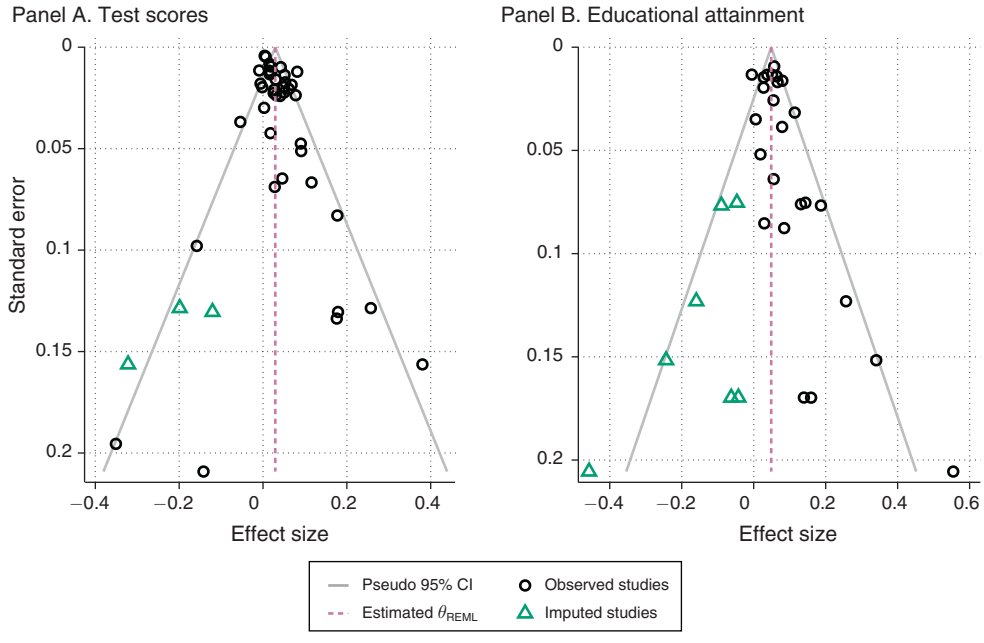

FIGURE 7. FUNNEL PLOTS, MULTIPLE ESTIMATES PER PAPER

Notes: Each panel presents a scatterplot of each estimate $(\hat{\theta}_j)$ against its standard error $(se_{\hat{\theta}_j})$ (black dots). Test scores estimates are in the left panel, and educational attainment estimates are in the right panel. The green triangles are imputed "estimates" designed to make the scatterplot symmetrical around the mean (when the imputed estimates are also included). This imputation is done using the "Trim and Fill" method (Duval and Tweedie 2000).

(5) In a stylized world, with no publication bias, a plot of study impacts against their precision should be symmetric around the grand mean (Borenstein 2009). However, in a stylized world with publication bias, while all precise studies will be published, there may be few published imprecise estimates in the "undesirable" direction. Figure 7 shows the funnel plots for the two outcomes. The black circles indicate the individual study impacts. The distribution of effects is largely symmetrical around the mean for precise studies (at the top of the figures), but there are relatively few imprecise negative study estimates. While this asymmetry needn't be the result of publication bias, to be conservative, we assume that any asymmetry is due to publication bias and assess the impacts of this asymmetry on the estimated pooled average in a few ways:

(i) To create symmetry, the "trim and fill" approach imputes "missing" studies (green triangles)—all of which are negative and very imprecise for both outcomes (Figure 7). The pooled effects that include these imputed studies are 0.032 and 0.053 for test scores and educational attainment, respectively (online Appendix Table A.14 columns 3 and 7)—very similar to our original estimates.

(ii) We also estimate our main model using a "drastic" approach of dropping the majority of the data (suggested in Stanley, Jarrell, and Doucouliagos). Using only those test score and educational attainment studies with estimated standard errors below 0.023 and 0.021, respectively (online Appendix Table A.14, columns 2 and 6), the results are very similar to our main models.

(iii) Finally, we follow Stanley and Doucouliagos (2014) and Ioannidis, Stanley, and Doucouliagos (2017) and implement the precision-effect estimate with standard error (PEESE) approach. Assuming that the most precise estimates will yield the true relationship, one can empirically model the relationship between the estimates and their precision, and infer what the most precise (and therefore, correct) estimate would be. This involves regressing the reported effect on the square of its precision and taking the constant term as the bias-adjusted estimate. For both outcomes, the adjusted estimates are similar to our main estimates—suggesting minimal publication bias (as in all the other tests).

VII. Explaining Heterogeneity

Capital versus Noncapital.—Jackson (2020) and Baron (2022) point out that while impacts of operational spending are consistently positive and significant in the existing literature, impacts of capital spending are less definitive. However, this may be due to many capital studies being underpowered. We assess this using a meta-regression as in (10) on test score estimates only, where C_j connotes a capital spending increase, each estimate is weighted by the inverse of its precision, and errors are clustered among related studies. Note that there are two error terms, a heterogeneity term connoting the difference between the true effect of study j and the grand mean (δ_j), and that due to sampling variability (ϵ_j).

$$\hat{\theta}_j = \alpha_1 + \left(C_j \times \beta_C\right) + \delta_j + \epsilon_j \tag{10}$$

β_C is the difference in the marginal effect of capital relative to other spending increases, and α_1 is the marginal effect for noncapital-specific spending increases. Column 2 of Table 3 shows that effects are larger for noncapital-specific spending (0.0343σ) than for capital (0.027σ), but not statistically significantly so.[29] Indeed, while capital effects are smaller, predicted marginal policy effects for capital and noncapital-specific outlays (of the same value) are similar (Figure 3). From a theoretical perspective, the similarity of marginal impacts is not surprising. That is, if schools seek to maximize outcomes *and money is fully fungible*, the marginal dollar spent on capital should be the same as that spent on labor. As such, in such a stylized world, one might expect that the marginal impact of spending types would be the same for all inputs. While there are likely many frictions such that money is not entirely fungible, our results are consistent with this notion.

[29] Online Appendix Figures A.8 and A.9 for separate forest plots by spending types.

To understand how the distribution of capital spending effects could be similar to that of noncapital-specific spending effects despite capital spending effects being less consistently positive and significant, it is helpful to put the capital spending estimates in perspective. Building a new elementary school typically costs about $27.5 million and houses 624 students (Abramson 2015)—a one-time expense of about $44,000 per pupil. Distributing this over the life of the asset (while accounting for depreciation) yields an average per pupil flow value in the first four years of $2,042. As such, one would expect test scores to increase by about $2.04 \times 0.03 = 0.061\sigma$ six years after the capital outlay. Only four studies examine projects of this magnitude. By way of comparison, a modest $1,000,000 renovation project 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 $0.15 \times 0.03 = 0.0045\sigma$ six years after the capital outlay. This is on the order of magnitude of projects examined in many studies and is smaller than what most individual studies can detect. Our analysis reveals that studies of capital spending on modest building upgrades are unlikely to detect effects. These results reinforce the importance of the statistical precision afforded by formal meta-analysis of multiple studies.

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. Because many policies at least implicitly target low-income populations with larger spending increases, policy effects are often larger for low-income groups. However, this needn't mean that the *marginal* effects of spending differ by income level. We assess this by exploiting the fact that some studies estimate marginal impacts by income status.[30] For both outcomes, we present strong evidence of larger marginal effects for lower-income populations than for more affluent populations.

We quantify the magnitude of these differences using meta-regression. We add the subscript inc such that $\theta_{j,inc}$ is the marginal spending effect for study j (on an outcome y) for population $inc \in \{average, high, low\}$. We then estimate a random effects meta-regression, as described in equation (11) where each study outcome is weighted by the inverse of its precision and errors are adjusted for clustering of dependent estimates:

$$\hat{\theta}_{j,inc} = \alpha_2 + \left(LowIncome_{j,inc} \times \beta_1\right) + \left(NonLowIncome_{j,inc} \times \beta_2\right) + \left(C_j \times \beta_C\right) + \delta_{j,inc} + \epsilon_{j,inc} \tag{11}$$

where $LowIncome_{j,inc}$ and $NonLowIncome_{j,inc}$ are equal to 1 for estimates pertaining to a low-income or non-low-income population (the reference group is the overall population). β_1 and β_2 indicate the *difference* between the effect for the average-resourced student and those from low-income populations and non-low-income populations, respectively. We also control for capital spending.

[30] We detail how specific studies define low income in online Appendix Table A.2. Results are similar if we classify overall Title I studies as low income (see Table A.9).

The marginal test score effect for low-income students is $\hat{\alpha}_2 + \hat{\beta}_1 = 0.0312$, and that for non-low-income students is about half the size, $\hat{\alpha}_2 + \hat{\beta}_2 = 0.0161$. The formal test of equality of effects yields a p-value below 0.01. Consequently, large test score impacts are less likely for more-advantaged groups (Figure 3). That is, a $1,000 increase for four years would improve test scores for non-low-income groups about 90 percent of the time but just over 70 percent of the time for non-low-income groups. Such a policy would increase test scores by 0.04σ among low-income groups one-third of the time, compared to 13 percent for higher-income groups. Also, while one might observe test score impacts as large as 0.06σ nearly 10 percent of the time among low-income students, such an effect would be observed less than 2 percent of the time among higher-income students.

Our results for differential impacts by income status for educational attainment are even more pronounced. The difference across groups is a sizable 0.055σ and is significant at the 0.05 significance level. The marginal effect for low-income groups is 0.0791σ, while that for the non-low-income group is 0.0238σ. That is, the marginal effect on educational attainment is more than three times as large for low-income students than for non-low-income students. Accordingly, sizable policy impacts are much more likely for the less-advantaged than more-advantaged students (Figure 3).

Using our estimates and distributional assumptions, we can predict the likelihood of certain magnitudes of marginal spending impacts for both low- and non-low-income populations. A $1,000 per pupil increase for four years would improve educational attainment for low-income and non-low-income students 99 and 79 percent of the time, respectively. Such a policy would increase college going by 2 pp among low-income students about 90 percent of the time, compared to less than 30 percent of the time for higher-income students. Large college-going effects above 5 pp would occur over one-fifth of the time for low-income students, but almost never for non-low-income students. Importantly, our predicted impacts indicate that effects on college-going larger than 7.5 percentage points per $1,000 of per pupil spending are unlikely, even for disadvantaged populations. The fact that there are several imprecise large estimates in this range underscores that the policy predictions afforded by formal meta-analysis (which takes the precision of estimates into account to measure true heterogeneity) provide valuable information for policymakers that could not have been gleaned from individual studies based on particular policies.

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 (region, urbanicity) and whether the study involved multiple states or a single state. The results, in online Appendix Table A.3, show that none of these dimensions significantly predict differential effects, on average.

Examining Evidence of Diminishing Returns.—Some scholars hypothesize that school spending in the United States is sufficiently high that due to diminishing

returns, the marginal impact of spending is approaching zero. However, across our set of included studies, this is *not* evident. We examine if the marginal spending impacts differ by the baseline spending level in the study context. Per pupil school spending levels have more than doubled in the past 30 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; Lee and Polachek 2018) might have smaller marginal impacts than studies of older policies in the 1960s (e.g., Cascio, Gordon, and Reber 2013) or in lower-spending states such as Texas (e.g., Martorell, Stange, and McFarlin 2016).

To assess this, in the bottom panel of Figure 5, 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 include the precision-weighted linear relationship and its 95 percent confidence interval, separating the scatter plot of marginal test score impacts (left) and educational attainment (right). Both outcomes show little evidence that marginal impacts are smaller at higher baseline spending levels, with a precision-weighted linear regression of the scatterplots yielding a slightly *positive* slope for test scores and a slightly negative slope for educational attainment—both with p-values well above 0.1. There are larger estimates at very low levels of spending, but these estimates are imprecise, and the best estimates of their true effects (i.e., BLUPs) are very close to the pooled average. For both test score and educational attainment outcomes, the *true* marginal impacts of public K-12 school spending are similar across a wide range of per pupil baseline spending levels—suggesting that K-12 education spending in the United States is not yet "*on the flat*." Because education is a labor-intensive field, as wages rise in many sectors, wages for educators may also rise with limited ability to reduce workers (Baumol and De Ferranti 2012). This could explain rising education costs that do not stem from 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 school spending is not allocated to the most productive inputs on the margin, so additional monies go toward bundles of inputs that are generally similarly productive. Whatever the reason, we highlight this important result that our data suggest that current policies will lead to similar true marginal spending effects (after accounting for inflation) as those we report in this study.

VIII. Discussion and Conclusions

Social scientists have long debated the effect of school spending on student outcomes. To deepen our understanding of this, we perform a meta-analysis on a comprehensive set of recent design-based studies of the impact of public school spending on student outcomes in the United States. On average, a $1,000 increase in per pupil school spending (sustained over four years) increases test scores by 0.0316σ, high school graduation by 2.0 percentage points, and college going by 2.8 percentage points. Moreover, by explicitly estimating heterogeneity and modeling the distribution of true policy impacts, we provide relatively precise policy predictions for new settings. We

find little indication that our estimated effects are skewed by confounding biases or publication biases. We highlight that meta-analyses of instrumental variables estimates face an inherent tension between including as many studies as possible (which may include weekly identified studies with unreliable inference) and screening on the strength of the first stage (which introduces a different set of biases). Importantly, our conclusions are robust across more versus less restrictive approaches.

We uncover a variety of new insights: school spending impacts on educational attainment are larger than on test scores when benchmarked against impacts of other interventions; 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; the average marginal effects of capital spending are similar to noncapital-specific spending increases (but most studies are underpowered to detect them); and marginal effects are larger for more-advantaged student populations than for less-advantaged populations (particularly for educational attainment). These insights and policy predictions may inform future school spending policies. Despite these new insights and the fact that school spending effects are *almost* always positive, we document nontrivial unexplained heterogeneity. As such, further research that seeks to explain how and why marginal effects differ across contexts may be fruitful.

REFERENCES

Abott, Carolyn, Vladimir Kogan, Stéphane Lavertu, and Zachary Peskowitz. 2020. "School District Operational Spending and Student Outcomes: Evidence from Tax Elections in Seven States." *Journal of Public Economics* 183: 104142.

Abramson, Paul. 2015. "20th Annual School Construction Report: National Statistics, Building Trends, and Detailed Analysis." *School Planning and Management*.

Andrews, Isaiah, and Maximilian Kasy. 2019. "Identification of and Correction for Publication Bias." *American Economic Review* 109 (8): 2766–94.

Angrist, Joshua, and Michal Kolesár. 2021. "One Instrument to Rule Them All: The Bias and Coverage of Just-ID IV." NBER Working Paper No. 29417.

Angrist, Joshua D., and Jörn-Steffen Pischke. 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, Oriana, Greg Fischer, Andrea Prat, and Erina Ytsma. 2021. "Do Women Respond Less to Performance Pay? Building Evidence from Multiple Experiments." *American Economic Review: Insights* 3 (4): 435–54.

Baron, E. Jason. 2022. "School Spending and Student Outcomes: Evidence from Revenue Limit Elections in Wisconsin." *American Economic Journal: Economic Policy* 14 (1): 1–39.

Baumol, William J., and David M. De Ferranti. 2012. *The Cost Disease: Why Computers Get Cheaper and Health Care Doesn't*. New Haven: Yale University Press.

Beath, Ken J. 2014. "A Finite Mixture Method for Outlier Detection and Robustness in Meta-analysis." *Research Synthesis Methods* 5 (4): 285–93.

Beuermann, Diether W., C. Kirabo Jackson, Laia Navarro-Sola, and Francisco Pardo. 2023. "What is a Good School, and Can Parents Tell? Evidence on the Multidimensionality of School Output." *Review of Economic Studies* 90 (1): 65–101.

Borenstein, Michael. 2009. *Introduction to Meta-analysis*. Chichester, UK: John Wiley and Sons.

Borenstein, Michael, Julian P. T. Higgins, Larry V. Hedges, and Hannah R. Rothstein. 2017. "Basics of Meta-analysis: I Is Not an Absolute Measure of Heterogeneity." *Research Synthesis Methods* 8 (1): 5–18.

Bound, John, David A. Jaeger, and Regina M. Baker. 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): 443–50.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020. "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics." *American Economic Review* 110 (11): 3634–60.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. 2016. "Star Wars: The Empirics Strike Back." *American Economic Journal: Applied Economics* 8 (1): 1–32.

Brunner, Eric, Joshua Hyman, and Andrew Ju. 2020. "School Finance Reforms, Teachers' Unions, and the Allocation of School Resources." *Review of Economics and Statistics* 102 (3): 473–89.

Card, David, and Alan B. Krueger. 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.

Cascio, Elizabeth U., Nora Gordon, and Sarah Reber. 2013. "Local Responses to Federal Grants: Evidence from the Introduction of Title I in the South." *American Economic Journal: Economic Policy* 5 (3): 126–59.

Chaudhary, Latika. 2009. "Education Inputs, Student Performance, and School Finance Reform in Michigan." *Economics of Education Review* 28 (1): 90–98.

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

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane W. Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star." *Quarterly Journal of Economics* 126 (4): 1593–1660.

Christensen, Garret, and Edward Miguel. 2018. "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature* 56 (3): 920–80.

Clark, Melissa A. 2003. "Education Reform, Redistribution, and Student Achievement: Evidence from the Kentucky Education Reform Act." PhD diss. Princeton University.

Clarke, Bronwyn. 2019. *The Evidence Decision-Makers Want: Literature Review*. Washington, DC: Center for the Study of Social Policy.

Conley, Timothy G., Christian B. Hansen, and Peter E. Rossi. 2012. "Plausibly Exogenous." *Review of Economics and Statistics* 94 (1): 260–72.

Conlin, Michael, and Paul N. Thompson. 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.

CPI. 2020. "US Bureau of Labor Statistics, Consumer Price Index for All Urban Consumers: All Items in US City Average [CPIAUCSL]." CPI.

Dehejia, Rajeev, Cristian Pop-Eleches, and Cyrus Samii. 2021. "From Local to Global: External Validity in a Fertility Natural Experiment." *Journal of Business and Economic Statistics* 39 (1): 217–43.

Deke, John. 2003. "A Study of the Impact of Public School Spending on Postsecondary Educational Attainment Using Statewide School District Refinancing in Kansas." *Economics of Education Review* 22 (3): 275–84.

Delaigle, Aurore, Peter Hall, and Alexander Meister. 2008. "On Deconvolution with Repeated Measurements." *Annals of Statistics* 36 (2): 665–85.

Dempster, A. P., and Louise M. Ryan. 1985. "Weighted Normal Plots." *Journal of the American Statistical Association* 80 (392): 845–50.

DerSimonian, R., and N. Laird. 1986. "Meta-analysis in Clinical Trials." *Controlled Clinical Trials* 7 (3): 177–88.

Downes, Thomas A., Richard F. Dye, and Therese J. McGuire. 1998. "Do Limits Matter? Evidence on the Effects of Tax Limitations on Student Performance." *Journal of Urban Economics* 43 (3): 401–17.

Duval, S., and R. Tweedie. 2000. "Trim and Fill: A Simple Funnel-Plot-Based Method of Testing and Adjusting for Publication Bias in Meta-analysis." *Biometrics* 56 (2): 455–63.

Dynarski, Susan, Joshua Hyman, and Diane Whitmore Schanzenbach. 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.

Efron, Bradley. 2016. "Empirical Bayes Deconvolution Estimates." *Biometrika* 103 (1): 1–20.

Efron, Bradley, and Carl Morris. 1973. "Stein's Estimation Rule and its Competitors—An Empirical Bayes Approach." *Journal of the American Statistical Association* 68 (341): 117–30.

Figlio, David N. 1997. "Did the 'Tax Revolt' Reduce School Performance?" *Journal of Public Economics* 65 (3): 245–69.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345 (6203): 1502–05.

Gigliotti, Philip, and Lucy C. Sorensen. 2018. "Educational Resources and Student Achievement: Evidence from the Save Harmless Provision in New York State." *Economics of Education Review* 66: 167–82.

Goncalves, Felipe. 2015. "The Effects of School Construction on Student and District Outcomes: Evidence from a State-Funded Program in Ohio." Unpublished.

Guryan, Jonathan. 2001. "Does Money Matter? Regression-Discontinuity Estimates from Education Finance Reform in Massachusetts." NBER Working Paper No. 8269.

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

Hedberg, Eric, J. Pustejovsky, and E. Tipton. 2017. "robumeta: A Macro for Stata."

Hedges, Larry V. 1983. "A Random Effects Model for Effect Sizes." *Psychological Bulletin* 93 (2): 388–95.

Hedges, Larry V., Elizabeth Tipton, and Matthew C. Johnson. 2010. "Robust Variance Estimation in Meta-regression with Dependent Effect Size Estimates." *Research Synthesis Methods* 1 (1): 39–65.

Higgins, J. P. T, Thomas J., Chandler J., Cumpston M., Li T., Page M. J., Welch V. A. (editors). 2019. *Cochrane Handbook for Systematic Reviews of Interventions.* 2nd Edition. Chichester (UK): John Wiley & Sons.

Hill, J., and L. Zhou. 2006. *Documentation for the NCES Common Core of Data School District Finance Survey (F-33), School Year 1991–92 (Fiscal Year 1992)*. Washington, DC: National Center for Education Statistics.

Holden, Kristian L. 2016. "Buy the Book? Evidence on the Effect of Textbook Funding on School-Level Achievement." *American Economic Journal: Applied Economics* 8 (4): 100–27.

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

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

Hyman, Joshua. 2017. "Does Money Matter in the Long Run? Effects of School Spending on Educational Attainment." *American Economic Journal: Economic Policy* 9 (4): 256–80.

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

Jackson, C. Kirabo. 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. Kirabo. 2020. "Does School Spending Matter? The New Literature on an Old Question." In *Confronting Inequality: How Policies and Practices Shape Children's Opportunities*, edited by Laura Tach, Rachel Dunifon, and Douglas L. Miller, 165–86. Washington, DC: American Psychological Association.

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

Jackson, C. Kirabo, and Claire Mackevicius. 2024. "Replication data for: What Impacts Can We Expect from School Spending Policy? Evidence from Evaluations in the United States." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.38886/E182042V1.

Jackson, C. Kirabo, Shanette C. Porter, John Q. Easton, Alyssa Blanchard, and Sebastián Kiguel. 2020. "School Effects on Socioemotional Development, School-Based Arrests, and Educational Attainment." *American Economic Review: Insights* 2 (4): 491–508.

Jackson, C. Kirabo, Cora Wigger, and Heyu Xiong. 2021. "Do School Spending Cuts Matter? Evidence from the Great Recession." *American Economic Journal: Economic Policy* 13 (2): 304–35.

Johnson, Rucker C. 2015. "Follow the Money: School Spending from Title I to Adult Earnings." *RSF: The Russell Sage Foundation Journal of the Social Sciences* 1 (3): 50–76.

Johnson, Rucker C., and C. Kirabo Jackson. 2019. "Reducing Inequality through Dynamic Complementarity: Evidence from Head Start and Public School Spending." *American Economic Journal: Economic Policy* 11 (4): 310–49.

Kendall, Maurice G., Alan Stuart, and John Keith Ord. 1994. *The Advanced Theory of Statistics*. London: Griffin.

Kline, Patrick, Evan K. Rose, and Christopher R. Walters. 2022. "Systemic Discrimination among Large US Employers." *Quarterly Journal of Economics* 137 (4): 1963–2036.

Kogan, Vladimir, Stéphane Lavertu, and Zachary Peskowitz. 2017. "Direct Democracy and Administrative Disruption." *Journal of Public Administration Research and Theory* 27 (3): 381–99.

Kreisman, Daniel, and Matthew P. Steinberg. 2019. "The Effect of Increased Funding on Student Achievement: Evidence from Texas's Small District Adjustment." *Journal of Public Economics* 176: 118–41.

Krueger, Alan B. 1998. "Reassessing the View that American Schools are Broken." *Economic Policy Review* 4 (1).

Lafortune, Julien, and David Schönholzer. 2022. "The Impact of School Facility Investments on Students and Homeowners: Evidence from Los Angeles." *American Economic Journal: Applied Economics*, 14 (3): 254-89.

Lafortune, Julien, Jesse Rothstein, and Diane Whitmore Schanzenbach. 2018. "School Finance Reform and the Distribution of Student Achievement." *American Economic Journal: Applied Economics* 10 (2): 1–26.

Lecker, Wendy. 2020. "School Finance Litigation Update: Delaware, Illinois, and New Mexico." *Education Law Center*, May 26. https://edlawcenter.org/news/archives/other-states/school-finance-litigation-update-delawre,-illinois,-and-new-mexico.html.

Lee, Katherine J., and Simon G. Thompson. 2008. "Flexible Parametric Models for Random-Effects Distributions." *Statistics in Medicine* 27 (3): 418–34.

Lee, Kyung-Gon, and Solomon W. Polachek. 2018. "Do School Budgets Matter? The Effect of Budget Referenda on Student Dropout Rates." *Education Economics* 26 (2): 129–44.

Martorell, Paco, Kevin Stange, and Isaac McFarlin Jr. 2016. "Investing in Schools: Capital Spending, Facility Conditions, and Student Achievement (Revised and Edited)." *Journal of Public Economics* 140: 13–29.

Matsudaira, Jordan D., Adrienne Hosek, and Elias Walsh. 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, Rachael. 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.

Morris, Carl N. 1983. "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association* 78 (381): 47–55.

Nakajima, Nozomi. 2021. "Evidence-Based Decisions and Education Policymakers." SREE 2021 Conference 44.

Neilson, Christopher A., and Seth D. Zimmerman. 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.* Paris, France: OECD Publishing.

Papke, Leslie E. 2008. "The Effects of Changes in Michigan's School Finance System." *Public Finance Review* 36 (4): 456–74.

Rambachan, Ashesh, and Jonathan Roth. 2020. "Design-Based Uncertainty for Quasi-experiments." *arXiv*: 2008.00602.

Raudenbush, Stephen W., Sean F. Reardon, and Takako Nomi. 2012. "Statistical Analysis for Multisite Trials Using Instrumental Variables with Random Coefficients." *Journal of Research on Educational Effectiveness* 5 (3): 303–32.

Rauscher, Emily. 2020a. "Delayed Benefits: Effects of California School District Bond Elections on Achievement by Socioeconomic Status." *Sociology of Education* 93 (2): 110–31.

Rauscher, Emily. 2020b. "Does Money Matter More in the Country? Education Funding Reductions and Achievement in Kansas, 2010–2018." *AERA Open* 6 (4): 1-38.

Reardon, Sean F., Demetra Kalogrides, and Andrew D. Ho. 2021. "Validation Methods for Aggregate-Level Test Scale Linking: A Case Study Mapping School District Test Score Distributions to a Common Scale." *Journal of Educational and Behavioral Statistics* 46 (2): 138–67.

Roy, Joydeep. 2011. "Impact of School Finance Reform on Resource Equalization and Academic Performance: Evidence from Michigan." *Education Finance and Policy* 6 (2): 137–67.

Schlaffer, James, and Gregory Burge. 2023. "The Asymmetric Effects of School Facilities on Academic Achievement: Evidence from Texas Bond Votes." *Social Science Journal* 60 (2): 235–53.

Stanley, T. D., and Hristos Doucouliagos. 2014. "Meta-regression Approximations to Reduce Publication Selection Bias." *Research Synthesis Methods* 5 (1): 60–78.

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

Tipton, Elizabeth. 2015. "Small Sample Adjustments for Robust Variance Estimation with Meta-regression." *Psychological Methods* 20 (3): 375–93.

Tipton, Elizabeth, and Robert B. Olsen. 2018. "A Review of Statistical Methods for Generalizing from Evaluations of Educational Interventions." *Educational Researcher* 47 (8): 516–24.

Tipton, Elizabeth, Christopher Bryan, and David Yeager. 2020. "To Change the World, Behavioral Intervention Research Will Need to Get Serious about Heterogeneity." Unpublished.

Tyler, John H., and Magnus Lofstrom. 2009. "Finishing High School: Alternative Pathways and Dropout Recovery." *Future of Children* 19 (1): 77–103.

van der Klaauw, Wilbert. 2008. "Breaking the Link between Poverty and Low Student Achievement: An Evaluation of Title I." *Journal of Econometrics* 142 (2): 731–56.

Vivalt, Eva. 2020. "How Much Can We Generalize from Impact Evaluations?" *Journal of the European Economic Association* 18 (6): 3045–89.

Wang, Chia-Chun, and Wen-Chung Lee. 2020. "Evaluation of the Normality Assumption in Meta-analyses." *American Journal of Epidemiology* 189 (3): 235–42.

Wang, Xiao-Feng, and Bin Wang. 2011. "Deconvolution Estimation in Measurement Error Models: The R Package decon." *Journal of Statistical Software* 39 (10): 1–24.

Weinstein, Meryle G., Leanna Stiefel, Amy E. Schwartz, and Luis Chalico. 2009. "Does Title I Increase Spending and Improve Performance? Evidence from New York City." IESP Working Paper No. 09-09.