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

C. Kirabo Jackson
Claire L. Mackevicius

Online Appendix

## A Data Gathering

### A.1 Overall Steps

Figure A.1: Exemplar Connected Papers Graph

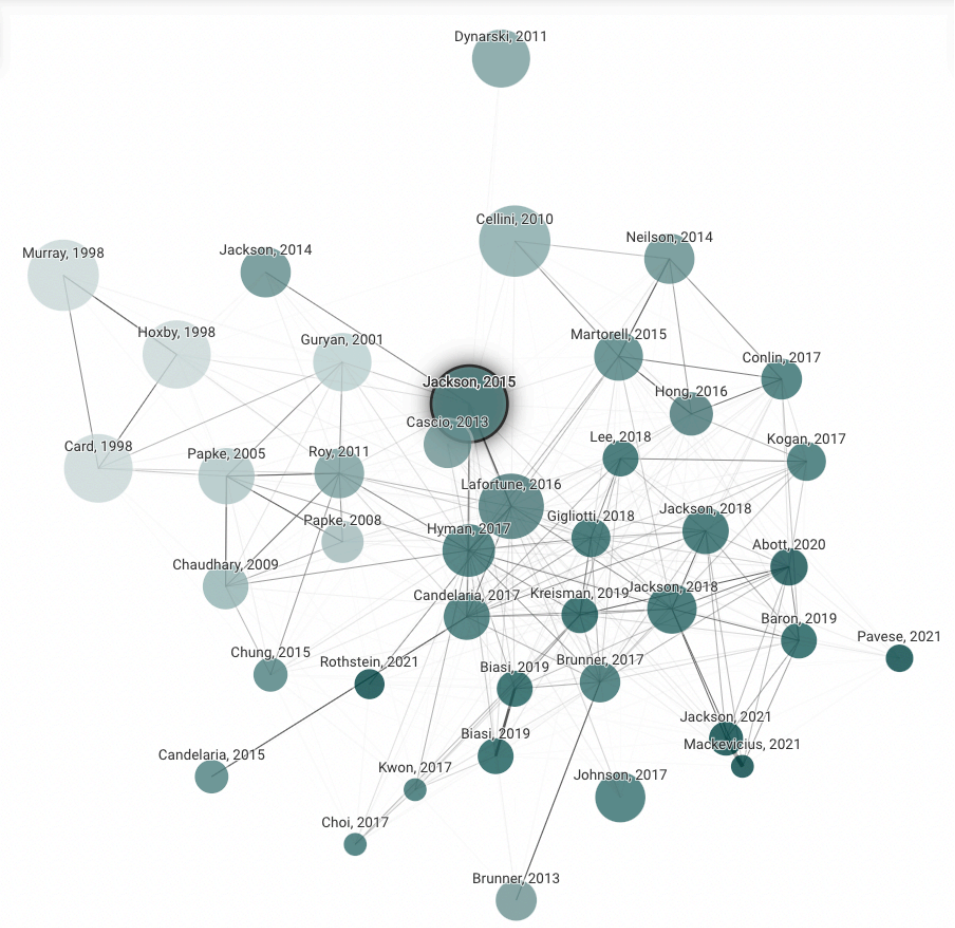


Figure A.2: Count of Included Studies per Year

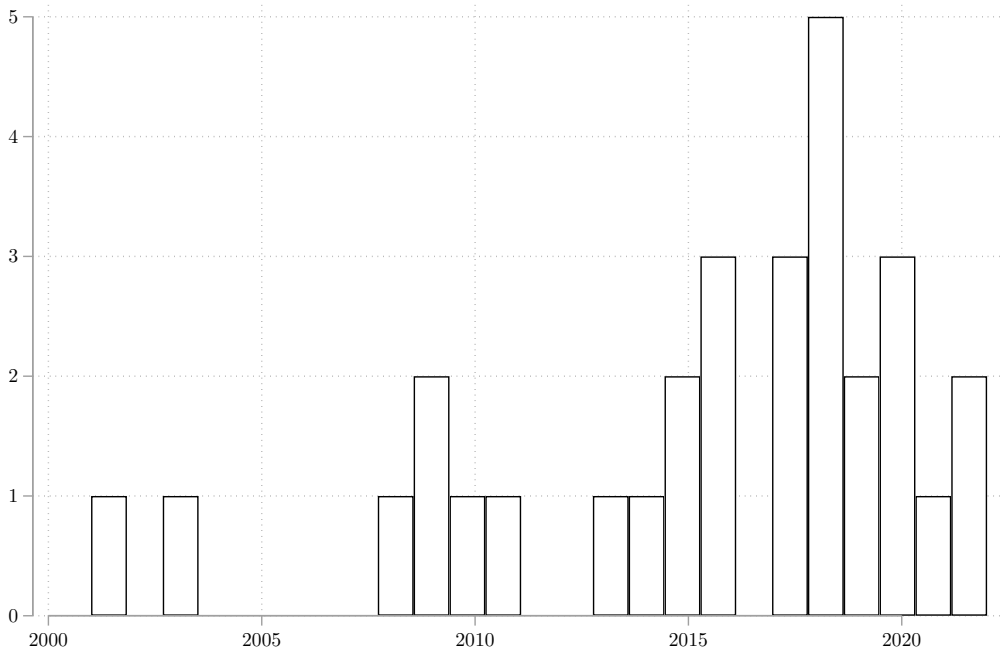


Table A.1: Summary of per-study steps

| study | outcome | effect per $1000 | $ Δ: source | outcome Δ: source |
|---|---|---|---|---|
| Abott Kogan Lavertu Peskowitz (2020) | High school graduation | 0.0850 | $417 (2012$): Table 8 Expend. P.P. Operations, $\leq 5yrs$, Bandwidth $+/-10$ | 0.0174: Table 8 Grad. Rate, $\leq 5yrs$, Bandwidth $+/-10$, standardized (Table 2 Grad. Rate (4yr), Passed); adjusted by factor of 0.8 (5 years to 4 year equivalent) |
| Abott Kogan Lavertu Peskowitz (2020) | Test scores | 0.1160 | $417 (2012$): Table 8 Expend. P.P. Operations, $\leq 5yrs$, Bandwidth $+/-10$ | 0.066: Table 8 Math/ELA (SDs), $\leq 5yrs$, Bandwidth $+/-10$; adjusted by factor of 0.8 (5 years to 4 year equivalent) |
| Baron (2022) | College enrollment | 0.1870 | $289.743 (2010$): Figure 1 (b) Total Operational Expenditures, averaged across 1-10yrs Relative to the Election (exact estimates provided by author) | 0.195: Figure 2 Panel (d) Log(Postsecondary Enrollment) Year 10 relative to election (exact estimates provided by author), multiplied by baseline rate (.39, Table 2), standardized; adjusted by factor of 0.4 (10 years to 4) |
| Baron (2022) | Test scores | -0.1580 | $4400 (2010$): "the median per-pupil bond campaign approved in Wisconsin is only approximately $4,400 per pupil" (24), depreciated over 15 years and averaged over first 6 years | -0.0567: Figure 6 panel (c) Average 10th Grade Math Score, cubic Year 6 relative to election (exact estimates provided by author), divided by student-level SDs (43.2, footnote 28) |
| Baron (2022) | Test scores | 0.1790 | $346 (2010$): Figure 1 (b) Total Operational Expenditures, averaged across 1-4yrs Relative to the Election (exact estimates provided by author) | 3.084: Figure 2 Panel (c) Average 10th Grade Math Score Year 4 relative to election (exact estimates provided by author), divided by student-level SDs (43.2, footnote 28) |
| Brunner Hyman Ju (2020) | Test scores | 0.0530 | $498 (2015$): Table 2 Current Expenditures, State Aid, Expanded controls Yes | 0.007: Table 7 All Districts Years postreform, multiply by 4 (years) |
| Candelaria Shores (2019) | High school graduation | 0.0510 | $795.02 (2010$): .1xbaseline (Table 2 Weighted Mean Total revenues) | 0.197: Table 5, Full log(Rev/Pupil), standardized (Table 2, Graduation rates) |
| Carlson Lavertu (2018) | Test scores | 0.0900 | $2048.79 (2014$): Table 8 Dynamic RD model SIG eligibilty, average Year 1-4 | 0.221, 0.171: Table 5 Dynamic model SIG eligibility Year 4 of SIG, average Reading and Math |
| Cascio Gordon Reber (2013) | High school dropout | 0.5550 | $100 (2009$):"each additional $100 increase in annual current expenditure per pupil..." (pg. 152) | -3.46, 0.66: Table 7 Δ White and Black high school dropout (reverse sign), population weighted (0.9/0.1) and translated to SD units based on baseline (pg 147, population-weighted) |
| Cellini Ferreira Rothstein (2010) | Test scores | 0.1770 | $6300 (2010$): "the average bond proposal in close elections is about $6,300 per pupil" (249), depreciated over 15 years and averaged over first 6 | 0.103, 0.160: Table VII, Academic achievement 6 yrs later Reading and Math, standardized ("the year-six point estimates correspond to effects of roughly 0.067 student-level standard deviations for reading and 0.077 for mathematics" (252) |

| | | | | |
|---|---|---|---|---|
| Chaudhary (2009) | Test scores | 0.0180 | \$5348 (1991\$): From Table 1 baseline | .1765516: Table 3, 4th and 7th scaled scores |
| Clark (2003) | Test scores | 0.0150 | \$1094.28 (2001\$): Table 3 Current expenditures per pupil Post-reform (1=yes) | 0.023: Table 6 Composite, Kentucky x post model (3) |
| Conlin Thompson (2017) | Test proficiency rates | 0.0060 | \$4000 (2013\$): "Capital expenditure and capital stock variables in Panels A and B are listed in \$1000s" (Table 3 note) x4 (years), depreciated 15 years averaged over first 3 | 0.081, 0.07: Table 3 Capital Exp $PP_t$ model (2) Percent Proficient in Math and Reading, relative to time t-3, standardized (Table 1 Percent Proficient in Math and Reading) |
| Gigliotti Sorensen (2018) | Test scores | 0.0420 | \$1000 (2016\$): "models. . . measure the effect of a \$1000 spending increase" (175) | 0.0468, 0.042: Table 4 PPE Math and Reading |
| Goncalves (2015) | Test proficiency rates | -0.0020 | \$23740.4 (2010\$): Table 1 Construction Cost Per Pupil Total, depreciated over 36.875 (weighted between 15 and 50 based on "60-65% of projects are new facilities" (6), averaged across first 6 years | 1.266, -1.442: Table 4 6+ yr. Completion Exposure Math and Reading, standardized (baseline Avg. Proficiency Table 4) |
| Guryan (2001) | Test scores | 0.0280 | \$1000 (1991\$): "median estimate. . . implies that a one standard deviation increase in per-pupil spending (\$1,000). . . " (21) | 0.039, 0.032, -0.034, -0.026: Table V and Table VI Math and Reading, subject-combined and standardized (assumed student-level SD of 100), then precision-weighted across grades |
| Hong Zimmer (2016) | Test proficiency rates | 0.0910 | \$8123 (2000\$): Table 1 Avg. bond amount per pupil, depreciated over 26.9 years (weighted between 15 and 50 based on Table 4 Passed a measure New building) averaged over 6 years | 2.13, 1.44: Table 5 4th7th proficiency Relative year 6, standardized based on Table 3 proficiency baseline |
| Hyman (2017) | College enrollment | 0.0550 | \$1000 (2012\$): "interpretation. . . is that \$1,000 of additional spending during each of grades four through seven. . . " (269) | 0.03: Table 4 model (4) Enroll in postsecondary schooling, standardized (baseline Table 1 All districts and cohorts Enrolls in postsecondary school) |
| Jackson Johnson Persico (2016) | High school graduation | 0.0800 | \$480 (2000\$): Table I All Per pupil spending (avg., ages 5-17) (\$4,800) x0.1 | 0.07053: Table III Prob(High School Graduate) model (7), standardized based on avg. national baseline graduation rate of 0.77; adjusted by factor of 0.33 (12 to 4 years) |
| Jackson Wigger Xiong (2021) | College enrollment | 0.0380 | \$1000 (2015\$): "preferred model, a \$1000 reduction in per-pupil spending. . . " (14) | 0.0201: Table A19 model (8) 4-Year Avg Per-Pupil Spending (thousands), standardized based on Table 1 College Enrollment Rate baseline |
| Jackson Wigger Xiong (2021) | Test scores | 0.0500 | \$1000 (2015\$): "preferred model, a \$1000 reduction in per-pupil spending. . . " (14) | 0.0529: Table A19 model (4) 4-Year Avg Per-Pupil Spending (thousands) |
| Johnson (2015) | High school graduation | 0.1440 | \$85 (2000\$): "results indicate that a \$100 increase in per-pupil Title I funding. . . " (66) times 0.85 passed through in real dollars seen by students (Figure 9) | 0.0225: Table 2 first column County Title I per-pupil spending (00s), average ages five to seventeen, standardized based on avg. national baseline graduation rate of 0.77; adjusted by factor of 0.33 (12 to 4 years) |

| | | | | |
|---|---|---|---|---|
| Kogan Lavertu Peskowitz (2017) | Test scores | 0.0190 | -$303.096 (2010$): Table 3 Total average Election year-3 years after, times 12000 ("District spending per pupil is just under $12,000 annually" (384)) | -0.14: Table 7 3 years after, to student-level SD units based on footnote 34 |
| Kreisman Steinberg (2019) | High school graduation | 0.0280 | $1000 (2011$): specification, abstract | 0.021: Table 8 Graduation, standardized based on Table 1 Graduation rate baseline; adjusted by factor of 0.44 (9 to 4 years) |
| Kreisman Steinberg (2019) | Test scores | 0.0780 | $1000 (2011$): specification, abstract | 0.097, 0.077: Table 5 Reading and Math |
| Lafortune Rothstein Schanzenbach (2018) | Test scores | 0.0160 | $907 (2013$): Table 4 Mean Total exenditures | 0.004: Table 8 Post event x years elapsed times 4 (years) |
| Lafortune Schonholzer (2022) | Test scores | 0.0500 | $87,701 (2013$): correspondence with author | 0.031xyear - 0.016, 0.027xyear - 0.004: Table 3 2SLS New School + Newschool Trend, Math and English Language Arts, 6 years |
| Lee Polachek (2018) | High school dropout | 0.0640 | $169.40 (2018$): Table 2 (percent change) times baseline spend by authors' calculations ($16939.79) | -0.1837: Table 3 9th-12th Grade Cubic, standardized based on baseline dropout rate Table 1 Mean Dropout Rate 9-12th Grade |
| Martorell Stange McFarlin (2016) | Test scores | 0.0250 | $7800 (2010$): "average per-pupil size of capital campaigns in Texas, the state we study in this paper, is about $7800" (14), depreciated over 15 years averaged over first 6 years | 0.016, 0.03: Table 5 Standardized Test Scores 6 years after bond passage Reading and Math |
| Miller (2018) | High school graduation | 0.0660 | $1371.9 (2013$): specification, 0.1 times baseline spend $13,719.24 (pg. 30) | 0.384: Table 4 10th Grade Cohort 1-4 years, standardized based on Table 1 Graduation Rate 4-year lag |
| Miller (2018) | Test scores | 0.0520 | $1371.9 (2013$): specification, 0.1 times baseline spend $13,719.24 (30) | 0.775, 0.879, 0.929, 0.477: Table 5 4th Grade Math and Reading and 8th Grade Math and Reading, subject-combined then precision-weighted across grades |
| Neilson Zimmerman (2014) | Test scores | 0.0310 | $70000 (2005$): "about $70,000 in the New Haven SCP" (25), depreciated over 50 years averaged over first 6 years | 0.153, 0.031: Table 6 > 5 Reading and Math, FE |
| Papke (2008) | Test proficiency rates | 0.0820 | $684.75 (2004$): 0.1 times baseline spend $6847.5 (Table 3 Average Expenditure per Pupil 1992-2004) | 36.77: Table 7 Fixed Effects-Instrumental Variables log(average eral per pupil expend), standardized based on baseline Table 5 average 50th percentile first three years |

| | | | | |
|---|---|---|---|---|
| Rauscher (2020) | Test scores | 0.0070 | $9600 (2014$): average capital outlays years 1-6 post election (Table 5), depreciated over 15 years averaged across first 6 years | 47.77, 12.36: Table 4 models (3) and(6) 6 Years after election Low-SES achievement and High-SES achievement, to student-level standard deviation units extrapolating from "These estimates amount to 0.40 to 0.57 standard deviations..." (119), distributed across estimated students per school (NCES) |
| Rauscher (2020) | Test scores | 0.0160 | -$745, the average of the decrease in spending in rural (-$940) and nonrual (-$550), ($2019) | .016: Tables 4 math and A11 ELA model(3), rural and nonrural, to student-level SD units from author correspondance |
| Roy (2011) | Test scores | 0.3800 | $1000 (2010$): specification, "estimates imply... for every $1,000" (159) | 0.057, 0.061: Table 8 Instrumental variables regressions Lagged spending 1998-2001 Reading and Math, standardized based on baseline SE (Footnote 35) |
| Weinstein Stiefel Schwartz Chalico (2009) | High school graduation | 0.1600 | $391.7 (2003$): Table 6 Direct Expenditure Title I model (2) | 3.59: Table 8 Graduation Rate Title I model (2), standardized based on avg. national baseline graduation rate of 0.77 |
| Weinstein Stiefel Schwartz Chalico (2009) | Test scores | -0.0540 | $284.3 (2003$): Table 5: Direct Expenditure Title I model (2) | -0.011, -.031: Table 7 Title I Math and Reading |

This describes the steps per *overall* study-outcome (and by spending type, relevant for Baron (2022)).

## A.2 Low-Income versus Non-Low-Income

Table A.2: Studies with LI and non-LI estimates

| Study | Outcome | non-LI $ | LI $ | non-LI effect | LI effect | LI definition |
|---|---|---|---|---|---|---|
| Abott Kogan Lavertu Peskowitz (2020) | Test scores | 279.99 | 609.19 | 0.2572 | 0.0460 | "compare spending and educational outcomes between districts that are above or below our sample median in terms of poverty rates among 5–17-year-olds (according to the American Community Survey)" (9) |
| Abott Kogan Lavertu Peskowitz (2020) | High school graduation | 279.99 | 609.19 | 0.1396 | 0.0295 | "compare spending and educational outcomes between districts that are above or below our sample median in terms of poverty rates among 5–17-year-olds (according to the American Community Survey)" (9) |
| Baron (2022) | College enrollment | . | 428.72 | . | 0.2566 | "I classify a school district as having an initially-high share of economically disadvantaged students if its share falls above the median of the Wisconsin 2000-01 school district distribution (the earliest year this variable is made publicly available)." (18) |
| Baron (2022) | Test scores | 329.54 | 392.81 | -0.3509 | -0.1419 | "I classify a school district as having an initially-high share of economically disadvantaged students if its share falls above the median of the Wisconsin 2000-01 school district distribution (the earliest year this variable is made publicly available)." (18) |
| Baron (2022) | Test scores | . | 532.74 | . | 0.1760 | "I classify a school district as having an initially-high share of economically disadvantaged students if its share falls above the median of the Wisconsin 2000-01 school district distribution (the earliest year this variable is made publicly available)." (18) |
| Brunner Hyman Ju (2020) | Test scores | 527.60 | 527.60 | 0.0303 | 0.0682 | "We separate the effects of SFRs by within-state 1980 income terciles because reforms were designed to differentially impact state aid for low- and high-income districts, with the goal of equalizing school funding" (478) |
| Candelaria Shores (2019) | High school graduation | 915.52 | 915.52 | 0.0188 | 0.1313 | "state-specific poverty quartiles, defined using free lunch eligibility status" (39) |

| | | | | | | |
|---|---|---|---|---|---|---|
| Goncalves (2015) | Test proficiency rates | . | 1332.85 | . | 0.0027 | Poorest 25% (Table 3) |
| Hyman (2017) | College enrollment | 1093.70 | 1093.70 | 0.0791 | 0.0055 | "districts with below-median 1995 district-level fraction receiving free lunch" (276) |
| Jackson Johnson Persico (2016) | High school graduation | 710.59 | 686.24 | 0.0275 | 0.1140 | "... a child is defined as low income if parental family income falls below two times the poverty line for any year during childhood" (165) |
| Johnson (2015) | High school graduation | 123.95 | 123.95 | 0.0556 | 0.3406 | |
| Kreisman Steinberg (2019) | Test scores | 1116.33 | 1116.33 | 0.0264 | 0.0618 | tercile of poverty (economically disadvantaged) (Table 6) |
| Kreisman Steinberg (2019) | High school graduation | 1116.33 | 1116.33 | -0.0053 | 0.0571 | tercile of poverty (economically disadvantaged) (Table 6) |
| Lafortune Rothstein Schanzenbach (2018) | Test scores | 672.62 | 1484.28 | -0.0059 | 0.0189 | "bottom or top quintile, respectively, of the state district-level income distribution" (Table 5) |
| Rauscher (2020) | Test scores | 916.53 | 916.53 | 0.0039 | 0.0152 | "The CDE defines low-SES students as those who are eligible for free or reduced-price lunch *or* whose parents both have less than a high school diploma... I refer to the distinction as SES throughout the article" (114) |

This represents all studies included in our meta-analyses which report separate effects for LI and non-LI populations (Except Baron (2022) operational and Goncalves (2015), which report for LI but not non-LI). The studies not included in our analyses, but relevant for identifying whether effects of spending are generally larger for LI populations include: Biasi (2019) on income mobility, Card & Payne (2002) on test score gaps, JJP (2015) on wages and poverty, Johnson (2015) on wages and poverty. These papers all find either a decrease in outcome gaps between LI and non-LI groups, or specifically more pronounced effects for LI individuals exposed to increased spending. This assumes the *same* dollar change for LI and non-LI districts in Hyman (2017). Without additional information about within- and across-district demographic heterogeneity, we are unable to capture (potentially) different spending changes for LI and non-LI students despite evidence in the paper which suggests money was distributed disproportionately to non-LI schools within districts. Analogous to our inclusion criteria for studies, we include only low-income estimates from Baron (2022) and not non-low-income estimates because (estimates provided by author) indicated no detectable spending change associated with operational referendum change for that population.

## A.3 Excluded paper details

We excluded papers unrelated to spending and student outcomes[32] in the United States, and all those that did not satisfy our inclusion criteria. Here, to shed light on how we applied the inclusion criteria, we detail a few well-known papers that were considered but were excluded based in each inclusion criteria.

### No Identifiable Policy (Condition a)

Some studies are excluded based on this criterion. For example, Husted and Kenny (2000) that states "Our preferred resource equalization measure... equals the change in resource inequality since 1972 relative to the predicted change (that is, the unexplained change in inequality). A fall in this variable reflects either the adoption of state policies that have reduced districts' ability to determine how much to spend in their district or an otherwise unmeasured drop in spending inequality" (298).

### No Testing of Exclusion Restriction (Condition b)

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

### No Effect on Spending (Condition c)

This condition corresponds to a first stage F-statistic of 3.85 for the policy instrument on per-pupil school spending. In a two-stage-least-squares (2SLS) framework, the typical threshold would be a first stage F-statistic of 10. We impose a weaker restriction. Still, some well-known studies are excluded based on this criterion. Specifically, van der Klaauw (2008) states that Title I "eligibility does not necessarily lead to a statistically significant increase in average per pupil expenditures" (750). Similarly, Matsudaira et al. (2012) do not find a robust association between the policy (Title I eligibility) and per-pupil spending. Some studies examine the effects of policies that influence school spending, but they do not report the effect of the policies on school spending in a way that allows us to construct a first-stage F-statistic. These include Downes et al. (1998), Figlio (1997), Hoxby (2001) and, more recently, Holden (2016) and Schlaffer and Burge (2020)[33]. Given its prominence, we discuss Hoxby (2001) in more detail: Hoxby (2001) reports that *some* key policy parameters (such as the inverted tax price) do predict differences in school spending but that others do not (such as the income/sales tax rate in support of school spending, which has a t-statistic smaller

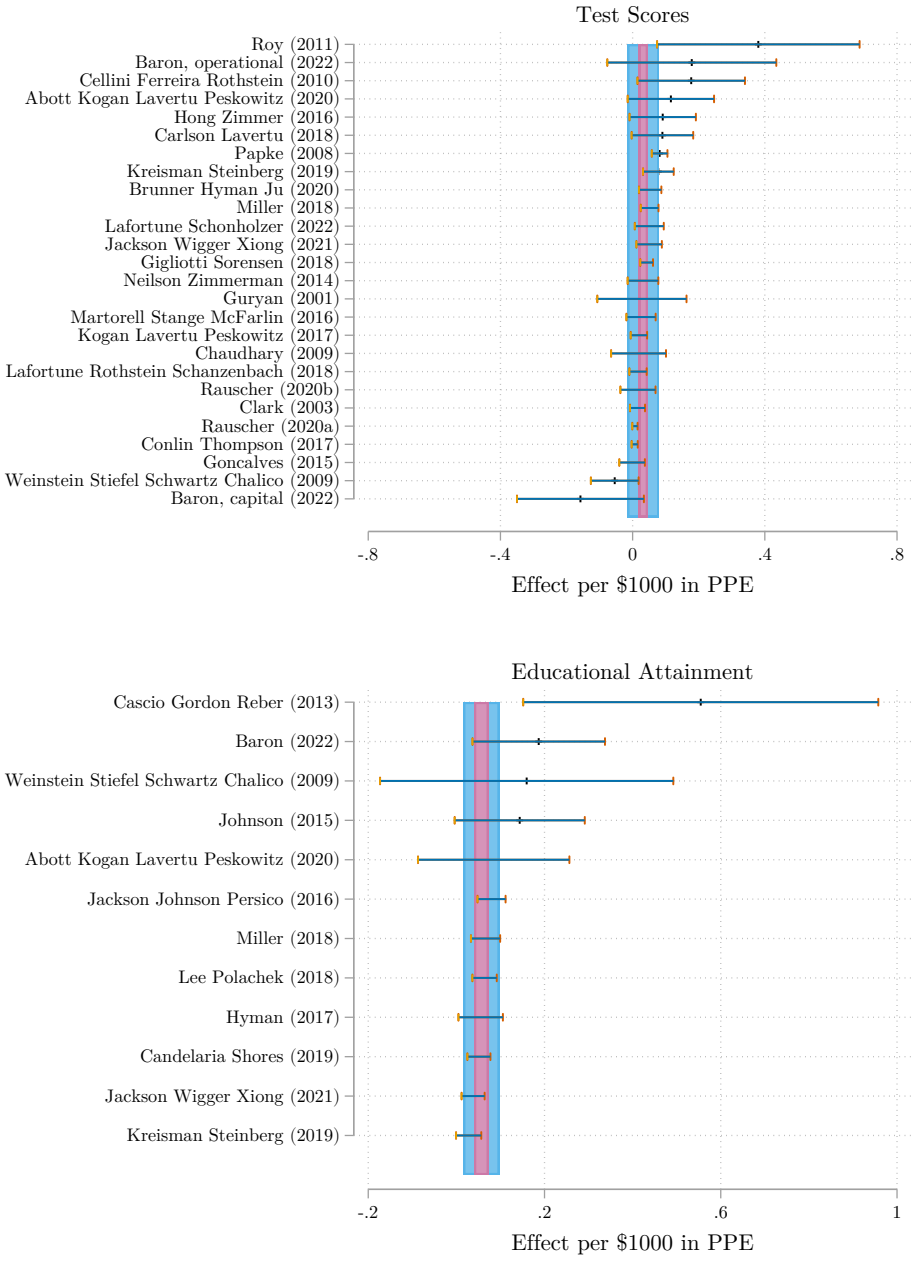
[32] As an additional exemplar, we excluded Deke (2003) as it reports on estimated effects of attending non-four-year postsecondary institutions. On one margin, comparing no postsecondary education to non-four-year postsecondary, this is a positive outcome. However, if people are sorting away from four-year postsecondary and into non-four-year postsecondary attendance, this is not necessarily an improvement. Thus, we do not include this paper as it is not comparable to other educational attainment outcomes reported in other papers.

[33] In particular, this paper only reports estimated effects of capital bonds–and does not specify the change in spending associated with them. Thus, we are unable to associated estimated effects with a dollar change in spending.

than 1 in predicting per-pupil spending). In a 2SLS model, all the policy variables (including the weak predictors) are used and no first stage F-statistic is reported. As such, because a strong first stage is not demonstrated, the 2SLS model predicting spending effects on dropout rates does not satisfy our inclusion criteria. Having said this, two policy variables are *individually* significant at the 5 percent level in most first stage regressions (inverted tax price and the flat grant/median income). In reduced form models, both these variables individually indicate that increased school spending reduces dropout rates. As Hoxby concludes, "*while the estimated effects of equalization on student achievement are generally weak, it does appear that the drop-out rate falls in districts that are constrained to raise spending by the imposition of a per-pupil spending floor*" (p. 1229).

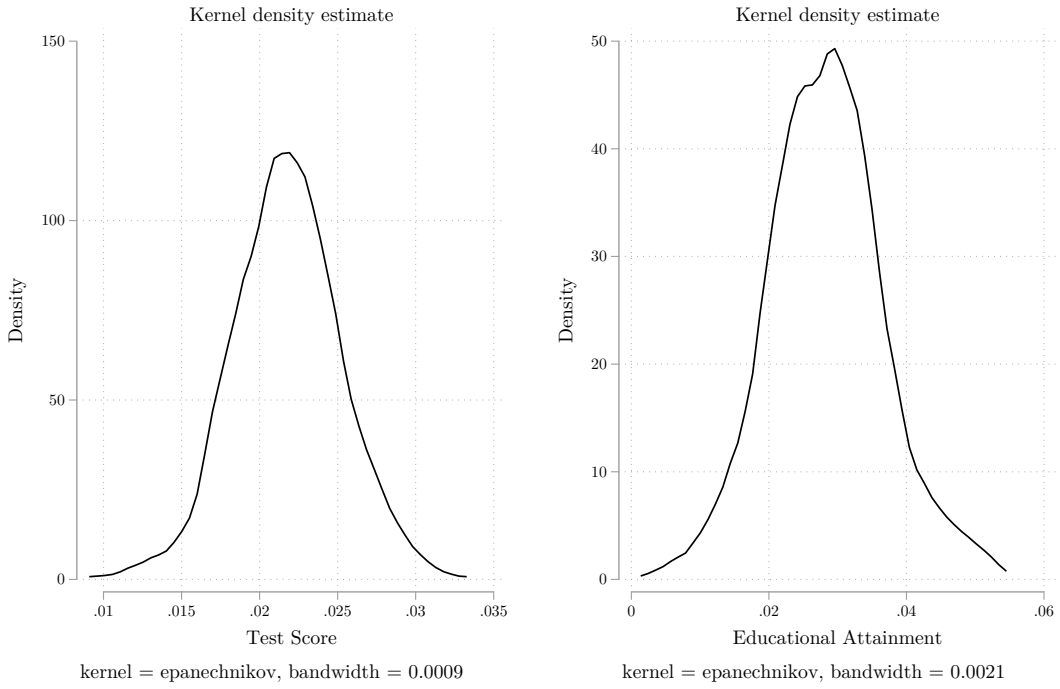
# B Supplemental Figures & Tables

Figure A.3: Forest Plot: One Estimate per Paper



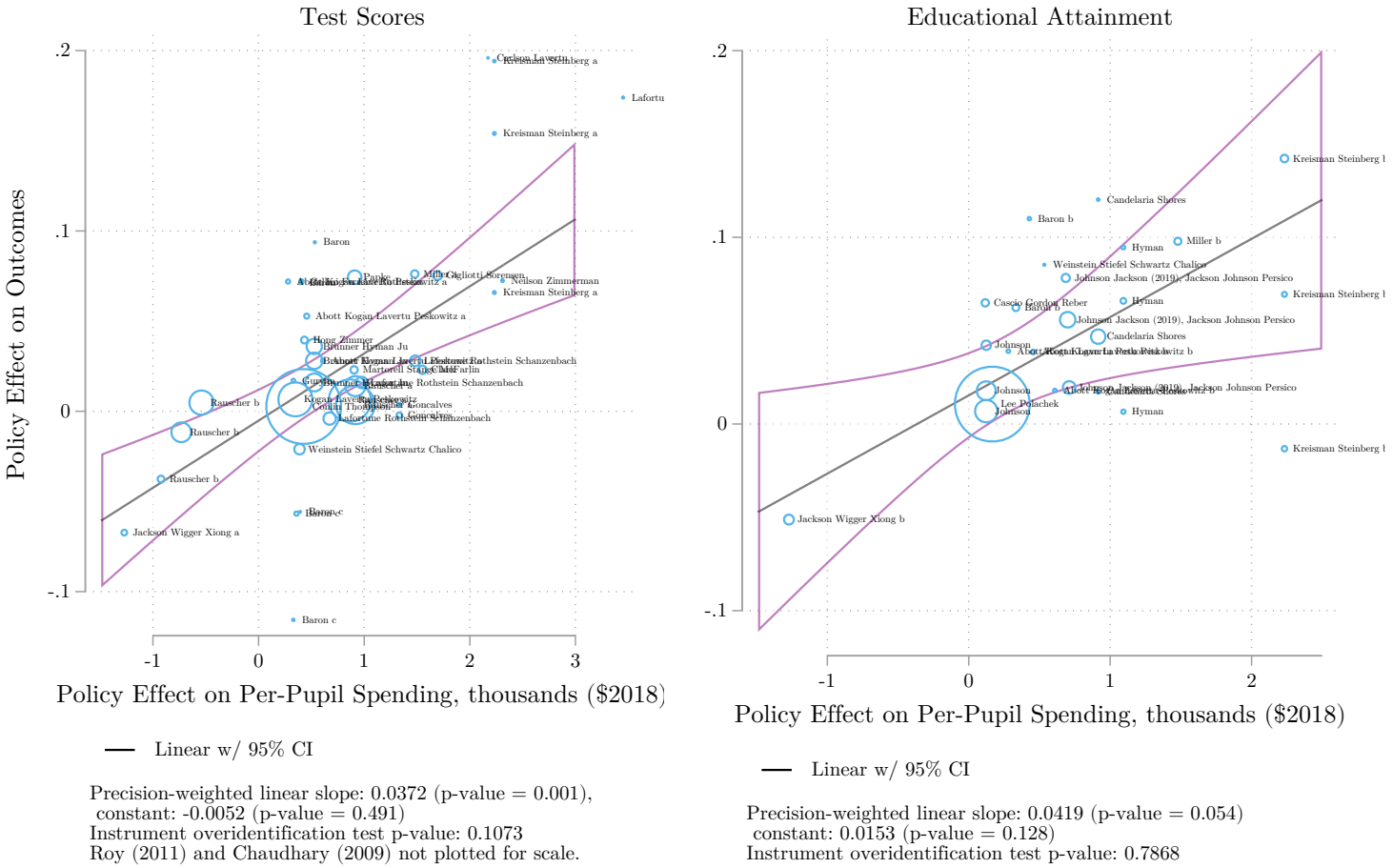

**Note:** The top panel shows papers that examine effects on test scores, and the bottom shows papers that examine spending effects on educational attainment. Each estimate represents the marginal effect of a $1000 per-pupil spending increase sustained over four years on standardized outcomes. The error bars represent the 95% Confidence Interval for each estimate. We show the 95% Confidence Interval for the Pooled Overall effect in pink and the 95% Prediction Interval in blue.

Figure A.4: Density of Positive Boostrap $\tau$ Estimates



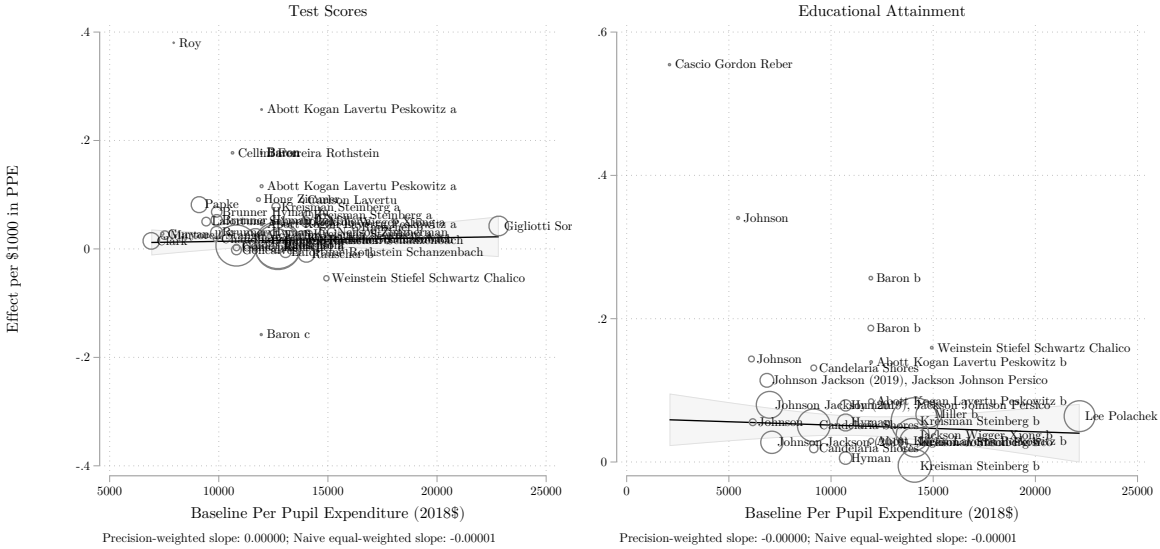

**Note:** These are kernel density plots of the distribution of estimated $\hat{\tau}$ based on 500 bootstrap replications.

Figure A.5: Policy Impacts against Increase in Spending (multiple estimates per study)



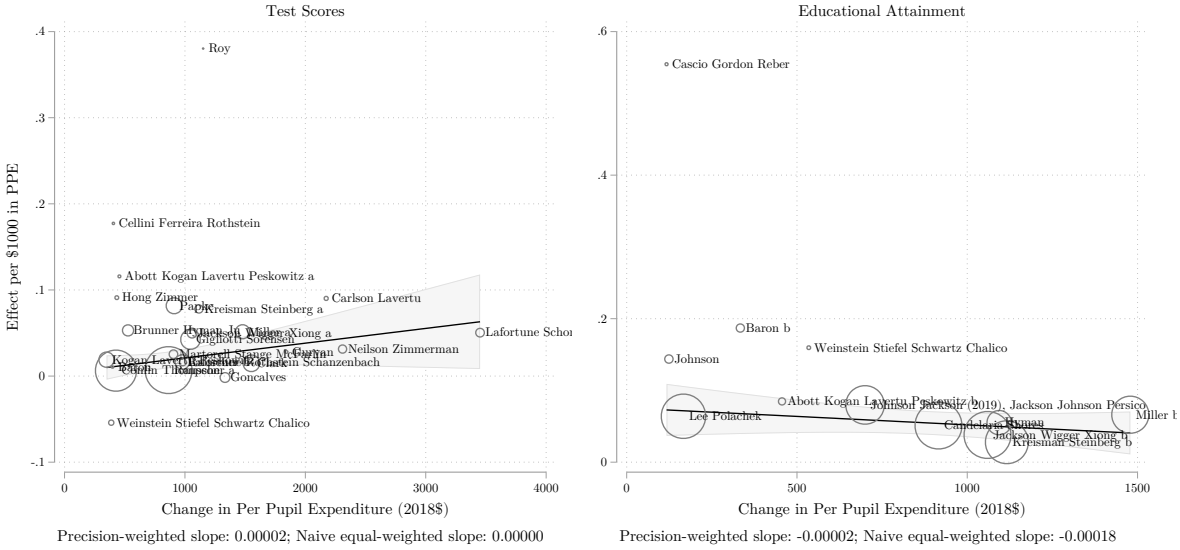

**Note:** This is a scatterplot of each policy's effect on standardized outcomes ($\Delta y_j$) against its effect on spending ($\Delta\$_j$). More precise estimates are depicted with larger circles and we plot the precision-weighted slope and its 95 % Confidence Interval. Note that there are multiple observations per study.

Figure A.6: Marginal Impacts by Baseline Spending Level (multiple estimates per study)



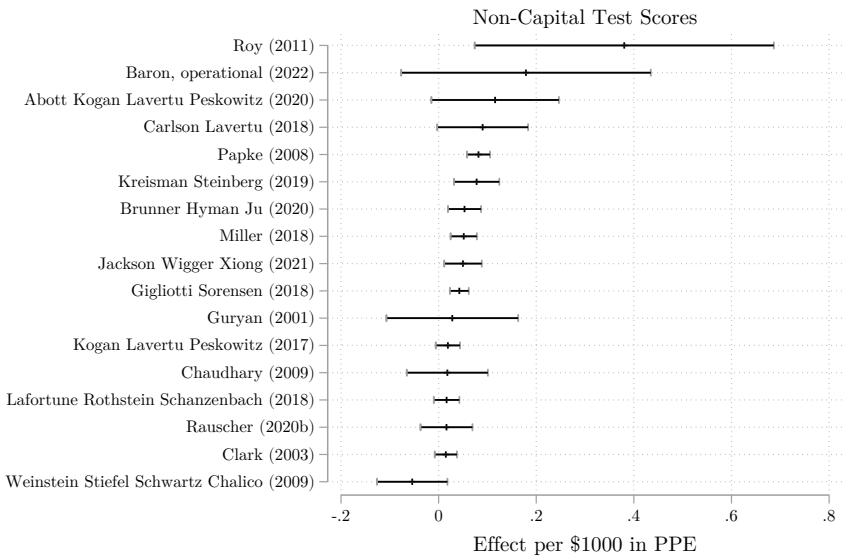

**Note:** This is a scatterplot of each policy's marginal effect on standardized outcomes ($\hat{\theta}_j$) against the baseline spending level in the policy context. More precise estimates are depicted with larger circles, and we plot the precision-weighted slope. Note that there are multiple observations per study

Figure A.7: Marginal Impacts by Change in Spending Level



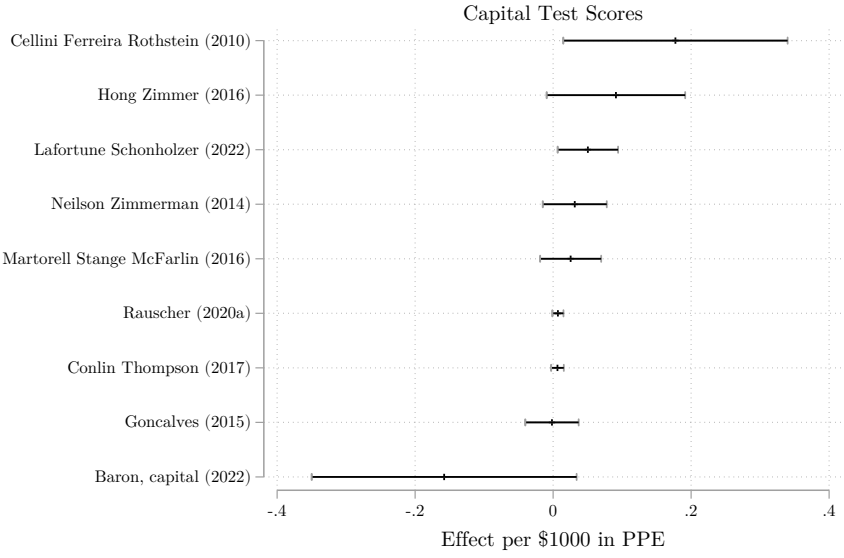

**Note:** This is a scatterplot of each policy's marginal effect on standardized outcomes ($\hat{\theta}_j$) against its effect on spending ($\Delta\$_j$). More precise estimates are depicted with larger circles, and we plot the precision-weighted slope. Note that there are multiple observations per study

Figure A.8: Non-Capital Test Score



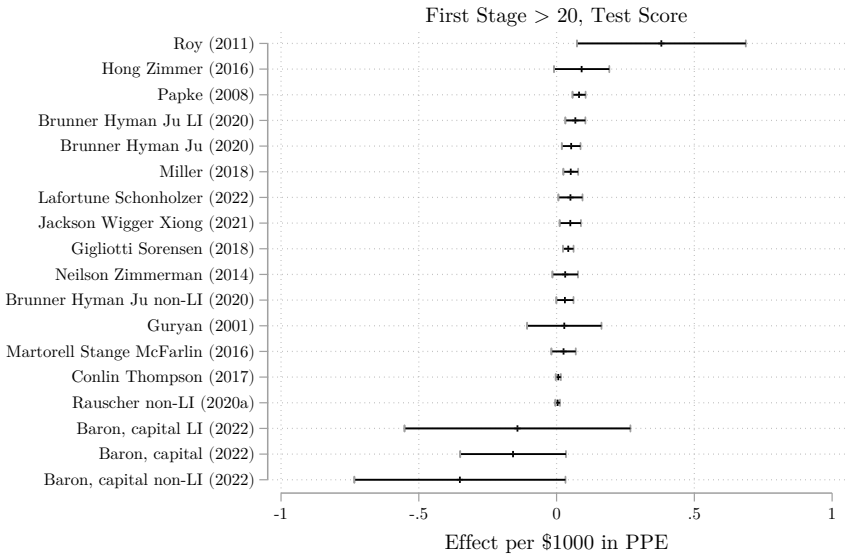

**Note:** Each estimate represents the marginal effect of a $1000 per-pupil spending increase sustained over four years on standardized outcomes. The error bars represent the 95% Confidence Interval for each estimate.

Figure A.9: Capital Test Score



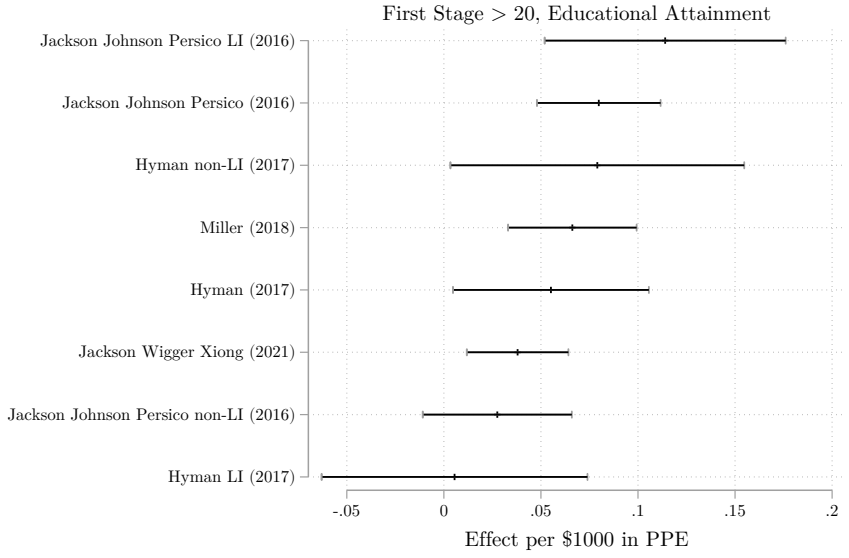

**Note:** Each estimate represents the marginal effect of a $1000 per-pupil spending increase sustained over four years on standardized outcomes. The error bars represent the 95% Confidence Interval for each estimate.

Figure A.10: Test Scores




**Note:** Each estimate represents the marginal effect of a $1000 per-pupil spending increase sustained over four years on standardized outcomes. The error bars represent the 95% Confidence Interval for each estimate.

Figure A.11: Educational Attainment




**Note:** Each estimate represents the marginal effect of a $1000 per-pupil spending increase sustained over four years on standardized outcomes. The error bars represent the 95% Confidence Interval for each estimate.

Table A.3: Meta-Analysis Estimates by Geographic Characteristics

| | (1) Test Scores by Multistate | (2) Test Scores by Region | (3) Test Scores by Urbanicity | (4) Educational Attainment by Multistate | (5) Educational Attainment by Region |
|---|---|---|---|---|---|
| Average Effect | 0.0274*** | 0.0471*** | 0.0402*** | 0.0536*** | 0.0625*** |
| | (0.00624) | (0.00782) | (0.00814) | (0.0154) | (0.00976) |
| Multistate | 0.0187* | | | 0.00985 | |
| | (0.00972) | | | (0.0190) | |
| South | | -0.0207* | | | -0.0249 |
| | | (0.0120) | | | (0.0760) |
| North | | -0.0115 | | | 0.0161 |
| | | (0.0202) | | | (0.0684) |
| Northeast | | -0.0262 | | | 0.00676 |
| | | (0.0219) | | | (0.0239) |
| West | | -0.0131 | | | |
| | | (0.0259) | | | |
| Urban | | | -0.0217 | | |
| | | | (0.0273) | | |
| Rural | | | 0.000489 | | |
| | | | (0.0321) | | |
| N | 40 | 40 | 24 | 25 | 25 |
| $\tau$ | 0.0198 | 0.0266 | 0.0290 | 0.0313 | 0.0398 |

Standard errors in parentheses are adjusted for clustering of related papers.

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

# C The Common Parameter Estimate

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

### Step 1: Choice of outcomes

We report effects on student achievement (measured by test scores or proficiency rates) and educational attainment (measured by dropout rates, high school graduation, or college (postsecondary) enrollment). If multiple test score outcomes are reported (e.g., proficiency rates and raw scores) we use the impacts on raw scores. This allows for standardized test score effects that are more comparable across studies, and avoids comparing impacts across thresholds of differing difficulty (i.e., where some areas have higher proficiency standards than others).[34] For educational attainment outcomes, we capture impacts on high-school completion measures and college enrollment. For studies that report multiple of these measures, we take the highest level reported.[35]

### Step 2: Computing Population Average Treatment Effects

For much of our analysis, we use one estimate per outcome per study. When studies report estimates for multiple specifications, we capture estimates from the authors' preferred specification. When there is a reported overall estimate across all populations (e.g., high-income and low-income), all subjects (e.g., Math and English), and all grade levels (e.g., 8th grade and 4th grade), we take the overall estimate as reported in the study. When studies report effects by subject, grade level, or population, we combine across estimates to generate an overall estimate and standard error for analysis.[36] When we combine test score effects across subjects for the same grade, we assume these stem from the same population and use the simple average as our overall effect.[37] [38] We combine

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

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

[36] Note that we estimate our main models across a range of assumed correlations, displayed visually in Figure A.18. These have little effect on our main results.

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

[38] In the single paper (Baron (2022)) that presents impacts for two separate types of spending (non-capital and capital) on one outcome (test scores), we use the simple average of the impacts of both spending types as our single

test score effects across grade levels using a precision-weighted average.[39] When we combine test score or educational attainment effects across populations (i.e., high- and low-income), we use the population-weighted average (i.e., put greater weight on the larger population) as our overall study effect.[40] This ensures that our overall estimate is as representative as feasible of what the effect would be for the entire population, and facilitates comparison across studies. In Section E, We show that all of our results are remarkably similar to alternative ways to combine estimates.

**Step 3: Standardize the Effect on the Outcome**

Studies report effects on test scores with different scales, and may report impacts on different outcomes (e.g., district proficiency rates or high school graduation). To facilitate comparison across studies, we convert each estimated effect into student-level standardized units if not already reported in these units.[41]

**Step 4: Equalize the Years of Exposure**

Because education is a cumulative process, one would expect larger effects for students exposed to school spending increases for a longer period of time. To account for differential treatment over time, we standardize all effects to reflect (where possible) the effect of being exposed to a spending increase for four years. Several studies report the dynamic effects of a school-spending policy (i.e., the effect over time). For test scores, when the dynamic effects are reported, we take the outcome

overall effect for the coin test analysis; we include both (non-capital and capital) distinct estimates of effects on test score outcomes for our meta-analysis. To compute the standard error of the overall test score effect for Baron (2022) we assume a correlation of zero.

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

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

[41] When effects are not reported in student-level standardized units, we divide the reported raw effect, $\Delta\mathring{y}$, by the student-level standard deviation of the outcome to capture the estimated effect on the outcome in student-level standard deviation units (i.e. $\sigma_{\mathring{y}}$). To perform this standardization, we gather information from each paper on the standard deviation of the outcome of interest. This standard deviation is generally reported in summary statistics. In two cases (Rauscher (2020a) and Kogan et al. (2017)), the standard deviation is reported at the school or district level. In these two exceptional cases, we convert the school- or district-level standard deviation into a student-level standard deviation by dividing the school or district-level standardized estimate impacts by the square root of the school or district size. Our results are robust to excluding these two studies (see Table A.7). For binary outcomes such as proficiency rates, graduation rates, or college-going rates, we use the fact that the standard deviation of a binary variable is $\sqrt{p \times (1-p)}$. In the three studies that report on graduation rates for relatively old samples (Jackson et al. (2016), Johnson (2015) and Weinstein et al. (2009)), we standardize estimated effects using graduation rates that prevailed at that time (77%) from national aggregate statistics, rather than using the baseline reported for the study sample. This choice makes studies more comparable by using the same standardization across studies of the same outcome and time period.

measured four years after the policy change.[42] Some papers do not report dynamic effects, and only report a single change in outcome after a policy-induced change in spending. In such cases, we take the average reported effect.[43] Because high school lasts four years, many papers report the effect on educational attainment of four years of exposure, but not all do.[44] [45] We adjust the captured effects to reflect four years of exposure by dividing the overall effect by the number of years of exposure and then multiplying by four.

To justify this modelling decision, we show empirical evidence that the benefits to increased spending increases approximately linearly with years of exposure. We focus on educational attainment because educational attainment is measured at the same age for all respondents, but there is variation in years of exposure across studies.

That is, some studies of educational attainment outcomes show the effects of four year of exposure to a spending increase, while others present effects of 9 years and 12 years so we can test if our assumption is reasonable. We plot the estimates (not adjusted for exposure) on educational attainment in Figure A.12, with more precise studies represented with larger circles. The pattern indicates larger overall impacts for estimates that relate to more years of exposure (per $1000 per-pupil spending increase). We run a meta-regression on the years-unadjusted effects, and include the years of exposure underlying each estimate as a covariate. The slope of year of exposure is 0.00x ($p-value = 0.03$) and one cannot reject the average four-year effect (the shortest exposure reported) is the same as four times the average impact of an additional year of exposure ($p - value = 0.79$). In sum, the data indicate that the educational attainment impacts increase with years of exposure and that the increase is approximately linear in years of exposure. This is both (a) a substantively important result to inform policy, and (b) validates our modelling assumption.

**Step 5: Equalize the Size of the Spending Change**

Each included study isolates the effect of the policy on spending (and that of the policy on outcomes) from other potential confounding factors and policies. We seek to determine the change in outcomes associated with a particular change in per-pupil spending. To ensure comparability of dollar amounts across time, we adjust reported dollars in each study into 2018 equivalent dollars using the Consumer Price Index (CPI).[46] Because we measure the impacts of exposure to four years of a spending change, we relate this four-year outcome effect to the change in spending during these same four years. For each study $j$ we collect the *average* change in per-pupil spending (in 2018 CPI

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

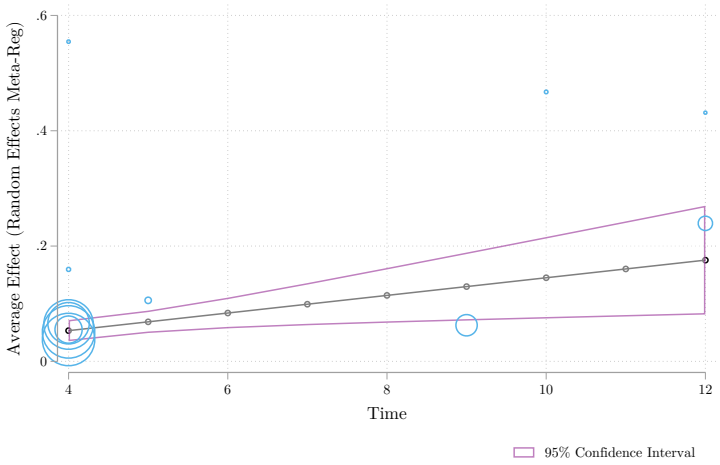
[43] In many cases, the average exposure is less than four years so that (if at all) we may *understate* the magnitude of any school spending effects for these studies.

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

[45] We capture the effect of referendum passage on college enrollment 10 years post-election in the case of Baron (2022) to ensure comparability with other studies which report on the same outcome.

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

Figure A.12: Educational Attainment by Years of Exposure




adjusted dollars) over the four years preceding the observed outcome, $\Delta\$_j$.[47] When the effect of spending on outcomes is directly reported in a study, we record this estimate directly. See Section C.1 for a detailed description of accounting for capital spending.

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

For each study, we obtain an estimate of the change in the standardized outcome per $1000 policy-induced change in school spending (averaged over four years and in 2018 dollars). Our standardized effect on outcome $y$ from study $j$ is $\hat{\theta}_j = (\Delta y_j)/(\Delta\$_j)$. For 5 out of 31 study-outcomes, we compute this ratio manually after standardizing the impact of the policy on both student outcomes and per-pupil spending. For the 26 out of 31 study-outcomes that report marginal spending effects directly, we take the reported marginal effect and adjust it (where needed) for exposure, CPI, and student-level standardization. Importantly, this parameter estimate is comparable across studies.[48] $\hat{\theta}_j$ can be interpreted an Instrumental Variables (IV) estimate of the marginal impacts of school spending

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

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

on outcomes using the exogenous policy-induced variation in school spending as the instrument.[49]

To illustrate the importance of computing the same parameter from each paper, consider the following two papers: Lafortune et al. (2018) report that the "*implied impact is between 0.12 and 0.24 standard deviations per $1,000 per pupil in annual spending*" while Clark (2003) reports that "*the increased spending [...] had no discernable effect on students' test scores*", reporting small, positive, statistically insignificant impacts. At first blush, these two studies suggest very different school spending impacts. However, when compared based on the same empirical relationship, the papers are similar. Specifically, precision aside, $\hat{\theta}_j$ for Clark (2003) is $0.0148\sigma$. By comparison, the large positive impact in Lafortune et al. (2018) is based the change in the test-score gap between high- and low-income groups (a *relative* achievement effect which is an important estimate for distributional analysis) over ten years. Their estimates of absolute overall test score impacts over 4 years yields a $\hat{\theta}_j$ of $0.0164\sigma$.[50]

## C.1 Detailed Approach to Making Capital Comparable to Non-Capital

To account for the difference in timing between when capital spending occurs and when the inputs purchased may affect outcomes, we use the annualized accounting value of the one-time increase in spending as the spending change associated with estimates of student outcomes.

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

In a more realistic scenario with depreciation, during the first year of a building's life, it is more valuable than in its $50^{th}$ year, due to wear and tear and obsolescence. In our example, the building's value in its first year would be greater than $500,000 and in its last year less than $500,000. To account for this, we follow convention in accounting and apply the depreciated value of capital

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

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

spending projects over the life of the asset. We assume annual depreciation of 4.7% and 16.5% for building and non-building projects such that only ten percent of the initial asset value remains after 50 and 15 years, respectively. That is, we assume that expenses that went primarily to new building construction or sizable renovations last 50 years.[51] Similarly, we assume that expenses that went primarily to less durable assets (such as equipment or upgrading electrical wiring for technology) last 15 years. For studies that report the proportion of capital spending that went to new building construction, we depreciate the capital amount proportionally between 50 and 15 years.[52] In Table A.11 we show that our main conclusions are robust to using lower and upper bounds of years depreciated, as well as to assuming no depreciation.

For each study of capital spending, we compute the change in student outcomes for each $1000 in average flow value of the capital spending in the years preceding the measured effect.[53] We illustrate this depreciation in Figure A.13, which shows the 15-year depreciation of a $7,800 per-pupil ($4.7 million per school) expenditure (as in Martorell et al. (2016)) and the 50-year depreciation of a $70,000 per-pupil ($42 million per school) expenditure (as in Neilson and Zimmerman (2014)). This transforms the extraordinarily large one-time expenditure over the projected life of the asset, which falls in value over time. After computing the flow value of the capital outlay for each year after initial payment, we can relate observed student outcomes associated with the average depreciated value of the asset in the years preceding measured outcomes.[54]

**Accounting for Construction Time**

Because the typical capital project does not lead to contemporaneous changes in classroom experiences, it is reasonable to expect any possible student improvements to take several years to materialize after the capital outlay. Indeed, large capital projects that involve entirely new construction or major upgrades to a new wing of a building can take multiple years to complete. Moreover, capital projects often entail some temporary disruption to everyday operations during the renovation/construction period, which may be deleterious to student outcomes. For these reasons, we assign the first two years of a capital spending project to a "construction/adjustment

[51] In 2013-14, the average age of school buildings in since original construction was 44 years (NCES 2016). Studies report on building age, including: Lafortune and Schönholzer (2022) (44.5 years), Martorell et al. (2016) (36 years), and Neilson and Zimmerman (2014) (well over 50 years).

[52] For example, Martorell et al. (2016) report that most of the spending went to renovations, and Cellini et al. (2010) provide an example of specific capital projects funded by a bond referenda, including to "improve student safety conditions, upgrade electrical wiring for technology, install fire doors, replace outdated plumbing/sewer systems, repair leaky rundown roofs/bathrooms, decaying walls, drainage systems, repair, construct, acquire, equip classrooms, libraries, science labs, sites and facilities..." (220). We describe capital paper coding in Table A.4.

[53] Depreciating the asset puts more value on the early years when test scores are measured and less on the years for which outcomes are not measured (many studies do not evaluate what the effect is more than 6 years after the funds are used). Because our parameter includes the spending change in the denominator, this reduces the reported school spending effect relative to not depreciating the asset. Accordingly, our approach may be considered conservative.

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

period" and capture outcomes six years after the increase in capital spending.[55]

To assess whether this temporal decision is reasonable, Figure 1 presents the dynamic effects of the nine studies estimating changes in capital spending on student test score outcomes. The left panel plots the *raw* effects for each study, not the marginal per-$1000 effects, over time as relative to a baseline year zero (t = 0) in which there should be no effect of the policy (the year of the construction or the policy change).[56] Consistent with an initial disruption, in several cases there is an immediate dip in outcomes. Consistent with long-run benefits to capital spending, this initial dip is followed by a gradual increase in outcomes in most studies. By about 5 or 6 years after a capital spending increase, one observes improved outcomes in most cases. To more formally assess the evolution of outcomes over time, we present the average dynamic effect in the right panel of Figure 1. We plot the average (across the nine studies) effects 1 through 6 years after the capital project or construction along with the 90 and 95 percent confidence intervals. This shows the same per-study pattern of no change (or possibly a slight dip) in the first two years and then improving outcomes after about 5 or 6 years. Indeed, one rejects that the effect of capital spending is zero at the 5-percent level by year five. This pattern validates our assigning the first two years of these studies to a "construction/disruption" period and using the six-year effect for capital spending increases as the most comparable to non-capital spending four-year effects. Overall, the pattern indicates that (a) capital spending *does* improve outcomes on average, and (b) these benefits take between 4 and 6 years to materialize. We present more formal statistical tests in Section 5 that quantify the extent to which capital spending may affect outcomes.

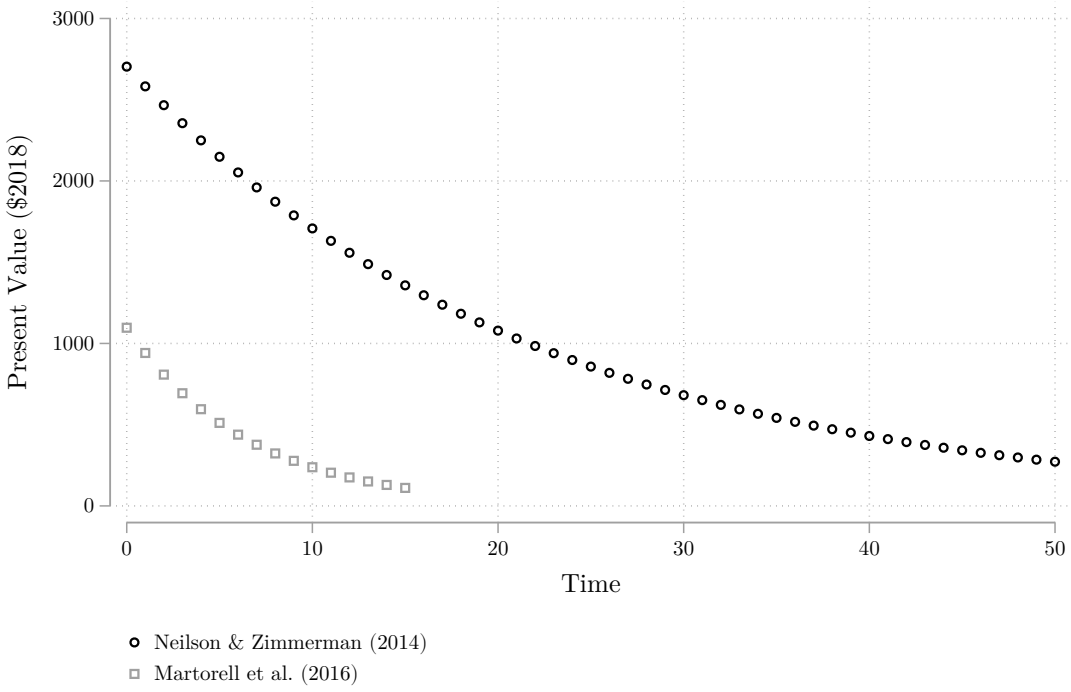
[55]Eight of nine papers report six-year estimates of the effect of capital spending changes on student outcomes. When the six-year effect is not reported, we use the latest year reported. Conlin and Thompson (2017) reports only 3 years after capital spending, so we capture their year-three effect as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

[56]For Lafortune and Schönholzer (2022), Neilson and Zimmerman (2014), and Goncalves (2015), year one (t = 1) represents first year of occupancy at a new or renovated school. In the case of Conlin and Thompson (2017) year one is the first year of program eligibility. For all other studies, year one (t = 1) represents the first year after a capital bond was passed.

Table A.4: Summary of capital depreciation decision

| Study | Depreciate over (years) | Life of project description |
|---|---|---|
| Baron (2022) | 15 | "the median per-pupil bond campaign approved in Wisconsin is only approximately $4,400 per pupil, and bond funds are frequently used to repair, maintain, and modernize existing structures, rather than to build new schools" (24) |
| Cellini Ferreira Rothstein (2010) | 15 | "Anecdotally, bonds are frequently used to build new permanent classrooms that replace temporary buildings (e.g., Sebastian (2006)), although repair, maintenance, and modernization are common uses as well' (220) // Table 1 average amount per pupil is of smaller magnitude than full-building construction |
| Conlin Thompson (2017) | 15 | this paper doesn't specify, and they translate effects into per-$1000 but the OH program was for both new construction and renovations |
| Goncalves (2015) | 36.875 | "I corresponded with an OSFC employee who reported that about 60-65 |
| Hong Zimmer (2016) | 26.9 | for the three years of data they have more detailed spending, percent new building is about 34 |
| Lafortune Schonholzer (2022) | 50 | "We restrict attention only to large new school projecst" (261) |
| Martorell Stange McFarlin (2016) | 15 | "typical capital campaigns deliver only modest facility improvements for the average student" (14) // "evidence is stronger for the claim that capital campaigns increase exposure to renovated schools" (20) |
| Neilson Zimmerman (2014) | 50 | "Of 42 school buildings, 12 had been rebuild completely by 2010, and 18 had been significantly renovated. . . school renovations were generally substantial, incurring costs similar to those of new construction" (20) |
| Rauscher (2020) | 15 | looks at CA bonds, which "can be used only for construction, rehabilitation, equipping school facilities, or acquisition/lease of real property for school facilities" (113) |

Figure A.13: Exemplar Capital Expenditure Depreciation



# D Modelling Assumptions

## D.1 Normality of True Effects

Wang and Lee, 2020 suggests that one can test for the normality of true effects by implementing standard tests of normality on suitably standardized effect sizes. That is, they point out that under the null hypothesis that the $\theta_j$'s are normally distributed, it follows that $\hat{\theta_j^S}$ as defined below follows a standard normal distribution.

$$\hat{\theta_j^S} = \frac{\hat{\theta}_j - \hat{\Theta}_{(-j)}}{(\hat{\tau}^2 + se_j^2 + se^2_{\hat{\theta}_{(-j)}})^{(1/2)}} \tag{12}$$

In 12, all variables are defined as previously, and the subscript -j denoted estimates that exclude estimate j. With the appropriately standardized estimates, they propose implementing standard tests for normality – i.e., the Shapiro–Wilk test and quantile-quantile plots Dempster and Ryan (1985). Implementing their standardization (using the R-code provided), the Shapiro–Wilk test on the standardized estimates yield p-values of 0.78 and 0.99 for test scores and educational attainment, respectively. That is, the tests fail to reject the null hypothesis that the effects are normally distributed. We also present the precision-weighted and equal-weighted quantile-quantile plots plots in figures A.14 and A.15. In both cases, there are no sizable deviations from normality.

Figure A.14: Test Scores Equal Weighted (L) and Weighted (R)

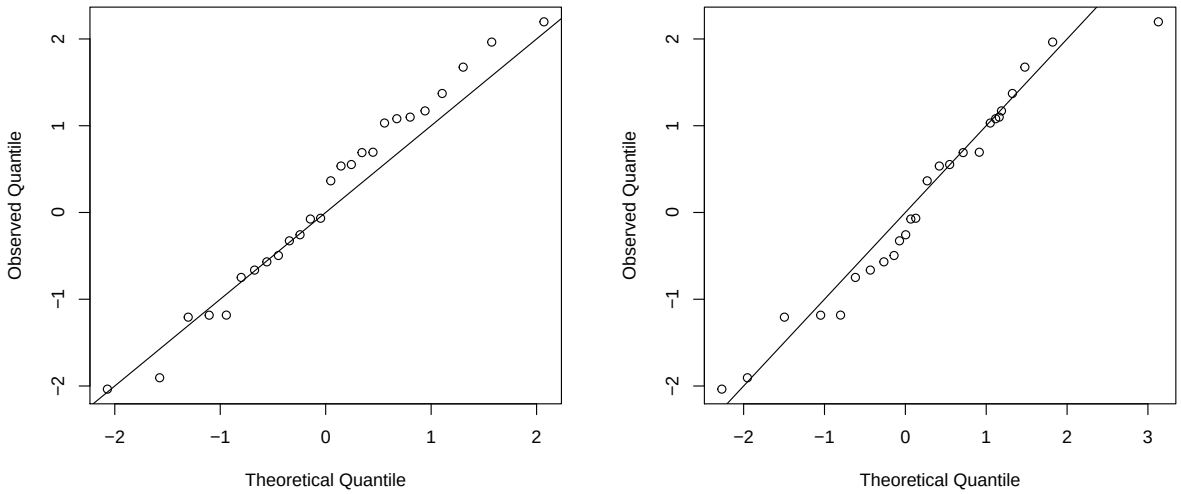
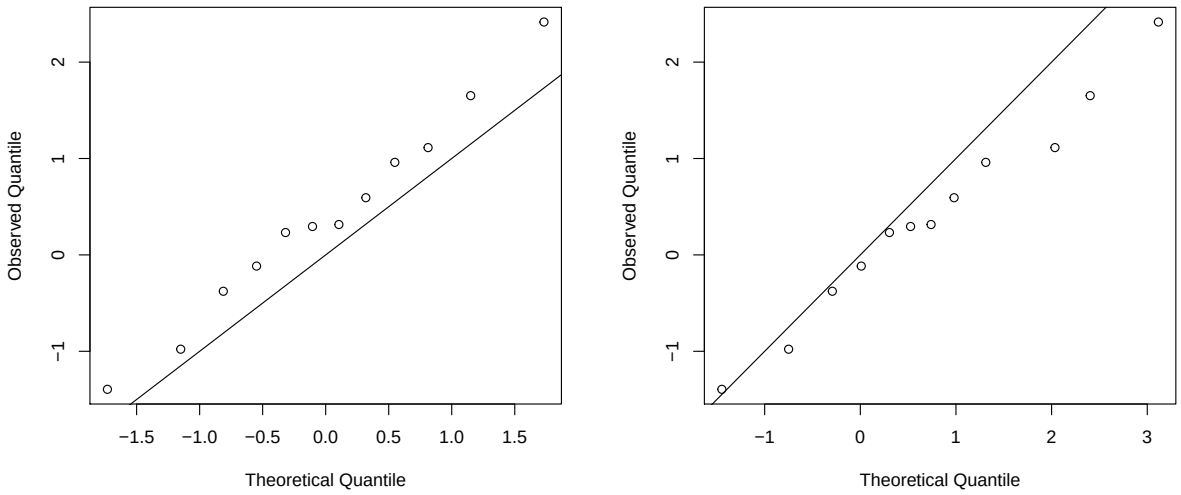


Figure A.15: Educational Attainment Unweighted (L) and Weighted (R)




## Deconvolve Approaches

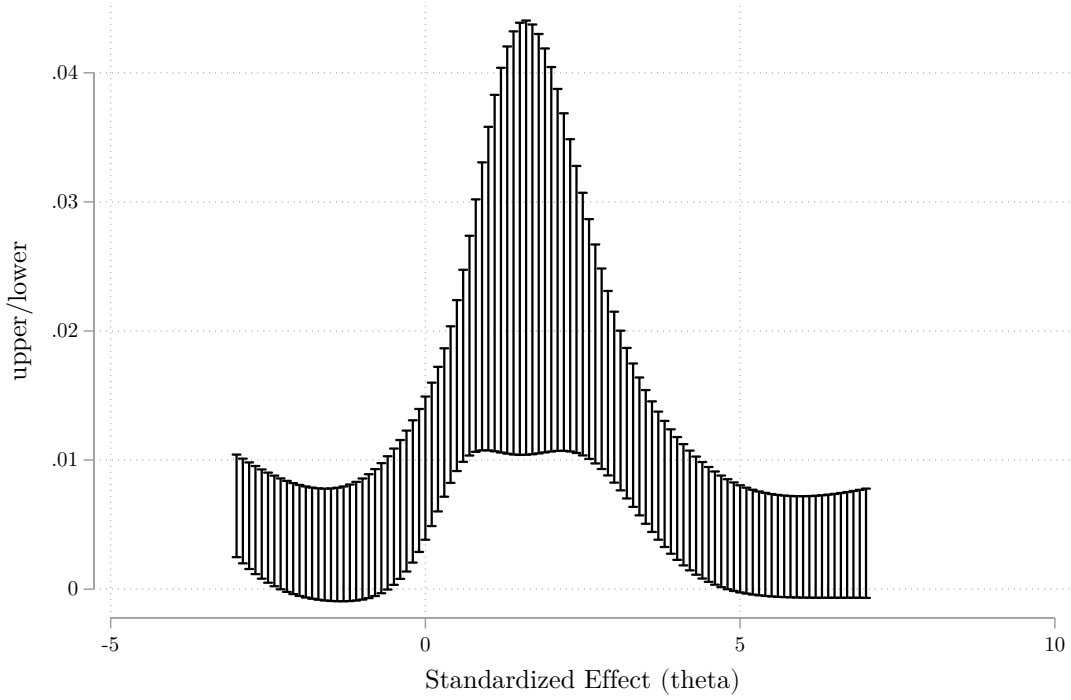
We also implement deconvolution approaches to test for meaningful deviation from normality in our data. The basic deconvolution problem is that we do not observe the distribution of true effects ($\theta_j$), but only the distribution of estimates ($\hat{\theta}_j$) - which is a noisy version of the distribution of true effects $g(\theta)$. Because we make policy predictions about the true effect in other settings, the distribution of true effects is of interest.

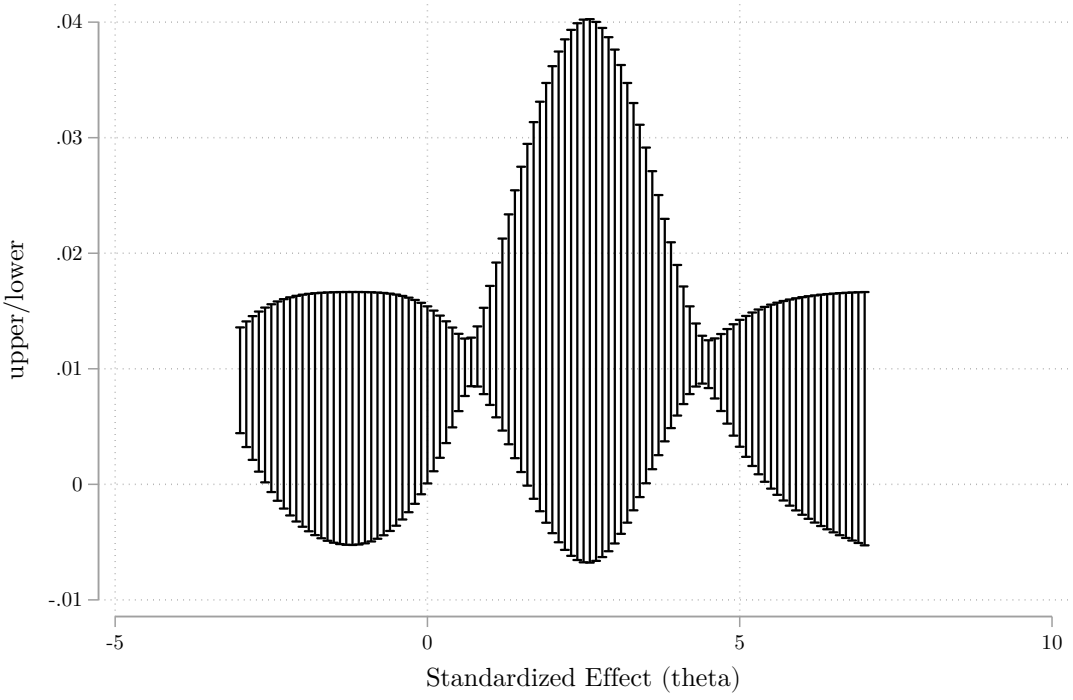
The deconvolution problem is to estimate $g(\theta)$ from the observed estimates $\hat{\theta}_j$'s under the assumption that each estimate $\hat{\theta}_j$ is normally distributed around its true effect $\theta_j$ with precision approximately equal to $se_j$ as in Equation (1). One kernel deconvolution approach is to model density of the distribution flexibly with a Fourier Transform – following Delaigle et al., 2008 as implemented by Wang and Wang (2011). Because using multiple correlated estimates per study could artificially skew the distribution, we estimate the distribution of true effects for the one-per-study sample. Using this approach, as with fitting data nonparametric models, one must choose the tuning parameter. We set the tuning parameter (i.e., set the bandwidth to 0.008) to match the estimated variance of the data – that is estimated using no distributional assumptions. The resulting deconvolved disruptions of true effect for test scores and educational attainment are shown in Figure 4.

One limitation of the approach above is that it does not provide confidence bounds for the estimated densities. To shed light on whether any deviations from the normal distribution are statistically significant, we implement Efron (2016) empirical Bayes (EB) deconvolution of the z-scores that allows for confidence bounds (following Kline et al. (2022)). This approach models the underlying distribution with an exponential family flexibly parameterized by an 8th-order spline. For both outcomes, the deconvoluted distribution of z-scores is approximately normal (Figure A.16).

for both outcomes, the confidence bounds are sufficiently wide that one would not reject the null hypothesis that the distribution of effects follows a normal distribution – bolstering our modeling decision.

Figure A.16: Deconvolve (With Confidence Interval)



### Allowing for Other Functional Forms

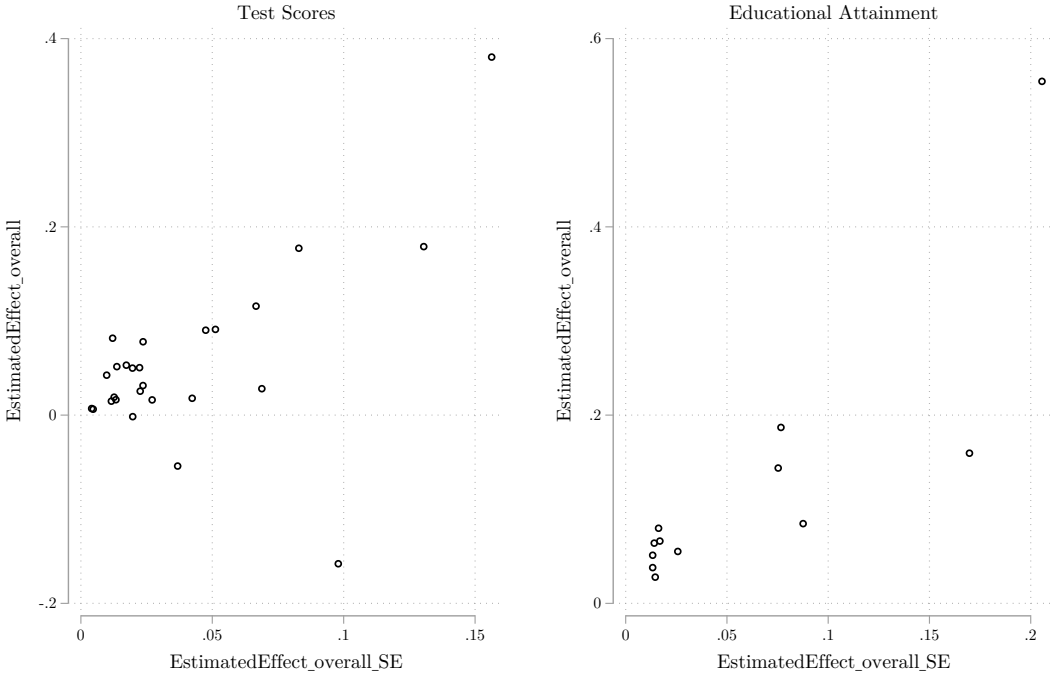
Finally, to assuage concerns that our conclusions are driven by the assumption of normality, we implement meta-analytic models that allow for outliers by modeling the distribution of effects using a t-distribution Baker and Jackson (2008); Lee and Thompson (2008)), or a mixture of normal distributions Beath (2014). These are implemented using the metaplus package in R. Importantly, in addition to reporting estimates of $\Theta$ and $\tau$, these models report likelihood ratio test relative to the standard normal model. We report the meta-analytic results along with the tests of normality from two models in Table A.5. One models allows for the distribution of effect to follow a t-distribution, and the other allows the effect to be a mixture of normal models. These alternative models yield similar results to the normal model and likelihood-ratio tests fail to reject the null of normally distributed effect.

Table A.5: Mixture and Normal Results

| | Test Scores | | Educational Attainment | |
|---|---|---|---|---|
| | t-dist | mixture | t-dist | mixture |
| $\theta$ | 0.0333 | 0.0333 | 0.0557 | 0.0557 |
| | (0.0065) | (0.0065) | (0.0068) | (0.0068) |
| $\tau$ | 0.0218 | 0.0218 | 0.00837 | 0.00838 |
| N | 40 | 40 | 25 | 25 |
| Pr(normal) | >0.999 | >0.999 | >0.999 | >0.999 |

## D.2 Dependence Between Effect and Precision

Figure A.17: Relationship between precision and effect size




**Note:** This is a plot of each marginal effect ($\hat{\theta}_j$) against its standard error ($se_j$).

Table A.6: Relationship between precision and effect size

| | Test Scores | | | | | Educational Attainment | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| | Multiple | Multiple F>20 | One Per | One Per Less Two | One Per F>20 | Multiple | Multiple F>20 | One Per | One Per Less Two | One Per F>20 |
| Avg. Effect | 0.0379** | 0.0273 | 0.0486*** | 0.0353** | 0.0524 | 0.104*** | 0.0579*** | 0.0716* | 0.0581*** | 0.0592** |
| | (0.0183) | (0.0343) | (0.0171) | (0.0129) | (0.0321) | (0.0187) | (0.0130) | (0.0335) | (0.0116) | (0.0110) |
| Centered-SE | 0.0112 | -0.0299 | 0.0421** | 0.00576 | 0.0335 | 0.0839*** | 0.0112 | 0.100*** | 0.0489*** | 0.0105 |
| | (0.0178) | (0.0292) | (0.0183) | (0.0157) | (0.0338) | (0.0195) | (0.0350) | (0.0291) | (0.0132) | (0.0446) |
| N | 40 | 18 | 26 | 24 | 13 | 25 | 8 | 12 | 10 | 4 |

Standard errors in parentheses are adjusted for clustering of related papers.

Reported Centered-SE subtracts the median standard error from the estimate standard error.

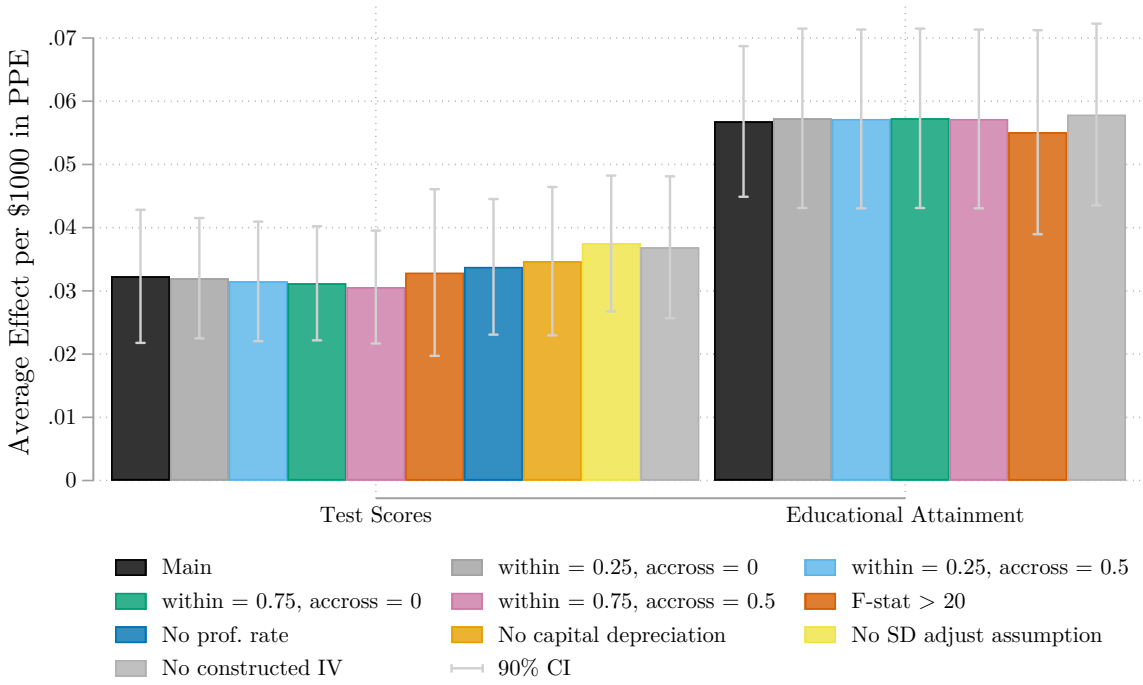
* $p < .1$, ** $p < .05$, *** $p < .01$

# E Robustness to Modelling Assumptions and Sample Restrictions

## E.1 Modelling Assumptions

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

Figure A.18: Modelling Assumptions




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

1. **Strong First Stage:** It is well understood that when the first stage relationship between the treatment and the instrument (in this case the policy) is weak, the resulting estimates may be biased and have unreliable standard errors (Bound et al. (1995) and Conley et al. (2012)). We are relatively liberal in our inclusion of studies, using any study with a first stage F-statistic of 3.85. Because we use precision weighting, and studies with weak first stages are likely to have larger standard errors, our method of moment estimator should be relatively robust to this problem. However, to assuage concerns, Table 4 presents our main specifications for those studies with a first stage F-statistic greater than 20, as constructed by the strength of the policy impact on the change in school spending. The results are very similar to our main results (dark orange bars in Figure A.18).

2. **Coding Proficiency Rates:** To make all test score estimates comparable, we converted reported effects into standardized effects. This is common practice for tests that are given on different scales, but less common for test score outcomes such as proficiency rates. For these outcomes, we reported standardized proficiency rate changes by dividing the effect by the student-level standard deviation of the proficiency rate $\sqrt{p(1-p)}$, where p is the proficiency rate. Improvements in student outcomes above or below the proficiency threshold may lead to very small changes in the proficiency rate, even if they reflect large changes in standardized raw scores. Or conversely, concentrated changes right around the proficiency threshold may appear much larger as proficiency rate increases than they reflect changes in standardized raw scores. As such, one may worry that our modelling of outcomes for these studies may skew our results. To assess this, we estimate test score models that remove the 3 studies that report effects on proficiency rates. We plot this effect and the confidence interval in the blue bar on the left panel of Figure A.18. Dropping these studies has no appreciable effect on our results – indicating that this modelling choice does not affect our conclusions in a meaningful way.

3. **Combining Effects**: For our main analysis we seek to have one single effect per study-outcome. As such, in many cases we combine impacts across subjects, grade levels, and populations making different assumptions about the correlation between effects. To ensure that these assumptions do not drive our conclusions, we re-estimate combined studies under very different assumptions and show that they all yield very similar results. We summarize these alternative approaches below.

   Our main analysis assumes 0 correlation between independent effects (across grades or populations), but these correlations could reasonably lie between 0 and 0.5. Our main analysis assumes 0.5 correlation between dependent effects (math/reading), but the correlations between dependent effects could reasonably range from 0.25 to 0.75. To show the practical importance of these assumptions on our estimates, we estimate our main models assuming all four combinations of the upper and lower bound assumed correlations. We plot the resulting estimates in grey, blue, green, and pink bars in Figure A.18.[57] The stability of our results indicates that our main estimates and conclusions are largely insensitive to reasonable assumptions about the correlations between effect across subjects, grades, and populations.

4. **Capital Depreciation:** To directly compare the effects of operational and capital spending, we depreciate capital expenditures following commonly accepted accounting approaches. To assess the robustness to different assumptions about length of time capital projects depreciated over, we re-run our main specifications with lower and upper bounds on years across which capital investments are depreciated. At a lower bound, we depreciate buildings at 30 and non-buildings at 10 years. At an upper bound, we depreciate buildings at 50 and non-buildings at 30 years. Additionally, one may also worry that the percent depreciation rate

[57] See full results in Table A.10

is too high and that the value of the asset should be more evenly distributed over time. To gauge the importance of this choice, we estimate models that assume the the value is uniform over the life of the asset (or that there is no depreciation). We report the estimated effects in Appendix Table A.11. Irrespective of the assumptions made, our estimates of the marginal effect of capital spending are largely similar and cannot be distinguished from each other nor from our preferred approach using formal statistical tests.

5. **First and Second Stage Standard Errors Correlations:** While many studies report marginal spending effects that we can take directly, for 5 study-outcomes, we must form the IV effects manually using the policy effects on spending and on outcomes.[58] When computing the standard error of this IV estimate, we assume zero correlation between the spending effect and the outcome effect. To provide bounds on the importance of this assumption, we estimate models that assume correlations of -1 and 1 (See Table A.8). The effects are largely unchanged under either assumed upper and lower bound correlations – underscoring the robustness of our meta-analytic average to this assumption.

6. **Student Level Standard Deviations:** For three studies (Kogan et al. (2017), Rauscher (2020a), and Rauscher (2020b)), we convert school- or district-level standard deviations to standardize the effect size at the student standard deviation level. Because this conversion relies on some assumptions, to assuage any concerns that this drives our results, we drop these two studies and re-estimate our model, resulting in very similar effects to including them (see Figure A.18).

[58]These include Brunner et al. (2020), Johnson (2015), Kogan et al. (2017), Lafortune and Schönholzer (2022), and Rauscher (2020a).

## E.2 Sensitivity and Robustness Analyses

Table A.7: Meta-Regression Estimates

| | No Clustering | | No SD Adjustment | No Constructed IV | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| | Test Scores | Educational Attainment | Test Scores | Test Scores | Educational Attainment |
| Average Effect | 0.0303*** | 0.0545*** | 0.0375*** | 0.0369*** | 0.0579*** |
| | (0.00524) | (0.00790) | (0.00651) | (0.00680) | (0.00871) |
| Observations | 40 | 25 | 35 | 30 | 22 |
| $\tau$ | 0.0220 | 0.0248 | 0.0215 | 0.0209 | 0.0280 |
| Average 90% PI | [-0.007,0.068] | [0.013,0.096] | [0.001,0.074] | [0.001,0.073] | [0.011,0.105] |

Standard errors in parentheses. Standard errors in models 3-5 are adjusted for clustering of related papers.

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

Table A.8: Meta-Regression Estimates by Correlation Sensitivity

| | Corr = -1 | | Corr = 1 | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | Test Scores | Educational Attainment | Test Scores | Educational Attainment |
| Average Effect | 0.0322*** | 0.0564*** | 0.0306*** | 0.0596*** |
| | (0.00577) | (0.00845) | (0.00528) | (0.00922) |
| Observations | 40 | 25 | 40 | 25 |
| $\tau$ | 0.0206 | 0.0253 | 0.0203 | 0.0296 |
| Average 90% PI | [-0.003,0.068] | [0.014,0.099] | [-0.004,0.065] | [0.010,0.109] |

Standard errors in parentheses are adjusted for clustering of related papers.

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

Table A.9: Meta-Regression Estimates, Title I Classified as Low-Income

| | Test Scores | Educational Attainment |
|---|---|---|
| | (1) | (2) |
| Overall | 0.0384*** | 0.0555*** |
| | (0.00678) | (0.00762) |
| Low-Income | -0.0244 | 0.0322 |
| | (0.0188) | (0.0241) |
| Non-Low-Income | -0.0237** | -0.0317 |
| | (0.0104) | (0.0194) |
| Capital | -0.00328 | |
| | (0.0112) | |
| LI - Non-LI | -0.001 | 0.064** |
| (se) | (0.019) | (0.029) |
| Observations | 40 | 23 |
| Clusters | 22 | 11 |
| $\tau$ | 0.0203 | 0.0225 |
| Average 90% PI | [0.008,0.069] | [0.004,0.107] |

Standard errors in parentheses are adjusted for clustering of related papers.

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

Table A.10: Meta-Regression Estimates by Within and Across Correlations

| | (w/in pop. low (0.25) across pop. low (0)) | | (w/in pop. low (0.25) across pop. high (0.5)) | | (w/in pop. high (0.75) across pop. low (0)) | | (w/in pop. high (0.75) across pop. high (0.5)) | |
|---|---|---|---|---|---|---|---|---|
| | (1) Test Scores | (2) Educational Attainment | (3) Test Scores | (4) Educational Attainment | (5) Test Scores | (6) Educational Attainment | (7) Test Scores | (8) Educational Attainment |
| Average Effect | 0.0320*** | 0.0573*** | 0.0315*** | 0.0572*** | 0.0312*** | 0.0573*** | 0.0306*** | 0.0572*** |
| | (0.00577) | (0.00860) | (0.00573) | (0.00857) | (0.00547) | (0.00860) | (0.00541) | (0.00857) |
| Observations | 40 | 25 | 40 | 25 | 40 | 25 | 40 | 25 |
| $\tau$ | 0.0217 | 0.0267 | 0.0214 | 0.0265 | 0.0198 | 0.0267 | 0.0193 | 0.0265 |
| Average 90% PI | [-0.005,0.069] | [0.012,0.102] | [-0.005,0.068] | [0.013,0.102] | [-0.003,0.065] | [0.012,0.102] | [-0.003,0.064] | [0.013,0.102] |

Standard errors in parentheses are adjusted for clustering of related papers.

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

Table A.11: Meta-Regression Estimates, by Depreciation Sensitivity

| | (1) | (2) | (3) |
|---|---|---|---|
| | Low Depreciation | High Depreciation | No Depreciation |
| Average Effect | 0.0384*** | 0.0555*** | 0.0347*** |
| | (0.00678) | (0.00762) | (0.00711) |
| Capital | -0.00328 | | 0.00526 |
| | (0.0112) | | (0.0208) |
| EstEffLI | -0.0244 | 0.0322 | |
| | (0.0188) | (0.0241) | |
| EstEffnonLI | -0.0237** | -0.0317 | |
| | (0.0104) | (0.0194) | |
| Capital | 0.035*** | | 0.040** |
| (se) | (0.009) | | (0.019) |
| Observations | 40 | 23 | 40 |
| $\tau$ | | | 0.0241 |
| Average 90% PI | [0.008,0.069] | [0.004,0.107] | [-0.018,0.087] |

Standard errors in parentheses are adjusted for clustering of related papers.

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

# F Assessing Bias in Individual Studies and Publication Bias

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

## F.1 Testing for Bias in Individual Studies

A common criticism of meta-analysis is that the end result is only credible if the studies included are themselves credible. For this reason, we are careful to only include studies that employ methods that may yield credibly causal effects. However, one may reasonably worry that even these individual studies may still suffer from bias – potentially biasing our meta-analytic average. In this section, we formalize a discussion of these biases and discuss when they may bias our meta-analytic average. We also present empirical tests to assess the existence and extent of such possible biases. Finally, we also propose a new meta-analytic approach that is robust to the existence of bias in individual studies under certain reasonable conditions.[59]

### A Framework For Assessing Confounding Bias

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

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

$$\Delta y_j = \underbrace{(\Theta \times \Delta\$_j)}_{Effect of Spending} + \underbrace{v_j}_{Noise} + \underbrace{b_j}_{Bias} \tag{13}$$

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

$$\hat{\theta_j} \equiv \frac{\Delta y_j}{\Delta\$_j} = \Theta + \frac{v_j}{\Delta\$_j} + \frac{b_j}{\Delta\$_j} \tag{14}$$

This is the true average marginal effect, plus the error to treatment ratio, plus the bias to treatment

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

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

$$\hat{\Theta}_{pw} = \frac{\sum(\frac{\Delta y_j + \upsilon_j}{\Delta \$_j})w_{yj}}{\sum w_{yj}} \equiv \underbrace{\Theta}_{\text{True Average}} + \underbrace{\frac{\sum(\frac{\upsilon_j}{\Delta \$_j})w_{yj}}{\sum w_{yj}}}_{\text{Average of Noise Ratio}} + \underbrace{\frac{\sum(\frac{b_j}{\Delta \$_j})w_{yj}}{\sum w_{yj}}}_{\text{Average of Bias Ratio}} \quad (15)$$

The observed average is comprised of three pieces; the true effect, the average of the random noise ratios (noise divided by the change in spending) across all the papers, and the average of the bias ratio terms (bias divided by the change in spending) across all the papers. Equation 15 makes clear that the meta-analytic average is an unbiased estimate of the true average (i.e., $E[\hat{\Theta}_{pw}] = \Theta$) so long as (1) the average of the noise terms is equal to zero in expectation, and (2) the average of the bias terms is equal to zero in expectation. The first condition implies that while some studies' impacts may be overstated due to measurement error or sampling variability, others will be understated for the same reasons so that, on average, the errors cancel each other out. So long as there are enough studies in the pooled sample and the random errors are unrelated to the policy-induced spending change, this condition will be satisfied. The second condition is less straightforward. It would trivially be satisfied if the individual studies are themselves unbiased. However, *even with bias in individual studies*, the second condition would hold if some studies' impacts are biased upward while others are biased downward so that the average bias is zero *and* the bias is unrelated to the policy-induced spending change so that the *average* bias ratio is approximately zero. We will present empirical evidence that this holds in our setting.

**Tests For Bias**

It is known that biases due to violations of the exclusion restriction tend to be more severe when the first stage relationship is weak (Bound et al. (1995) and Conley et al. (2012)). In our context, one can see this clearly because the individual bias-ratio for study $j$ represented by $b_j/\Delta\$_j$ in equation (14) is smaller for policies that generate larger changes in spending. It follows that if biases exist in the included studies, the marginal spending effects should be systematically different (a) among studies that have strong first stages, and (b) among studies based on policies that generate larger versus smaller spending changes.

**Strength of the First Stage:** First, we show the main effects based on studies that have first-stage F-statistics over 20 (Table 4 (columns 3-8)). The results are very similar to models that use estimates with first-stage F-statistics above 3.85 (Table 3) and above 10 (Table 4 (columns 1 and 2)) – suggesting minimal bias in the individual studies.

**Size of the First Stage:** As a second test, we examine if the marginal policy impact varies by the magnitude of the spending change induced by the policy. If there were biases (which one expects to be larger in studies with weaker first stages), then the average marginal effects would be larger for small spending changes than for larger spending changes. We test this by regressing the marginal effect of the study against the magnitude of the spending change (see associated

scatterplot in Figure A.7). Such a model yields a slope of 0.00002 ($p$-value of 0.259) for test scores and a slope of -0.00002 ($p$-value = 0.2390) for educational attainment outcomes – indicating no relationship between the marginal effect and the size of the spending change. While these tests are not dispositive on their own, they suggest that the individual studies included (which were specifically chosen *because* they are credibly causal) are by-and-large not appreciably biased on average.

**Only look at well-powered studies** Some difference-in-difference-based studies may be biased due to a violation of the common trends assumption (Rambachan and Roth (2020)), studies using regression discontinuity designs may have bias due to extrapolation away from the cutoff point, and credible instrumental variables-based studies may have modest violations of the exclusion restriction (Conley et al. (2012)). Because some of our included studies may be underpowered, such violations may not have been detected. This motivates a simple test. If underpowered studies are less able to detect bias, then in the presence of bias, well-powered studies will be less susceptible to bias. We can assess the importance of this bias by seeing how robust our estimates are to the exclusion of underpowered studies. That is, we estimate models only among studies that would have detected (based on the standard error) our main meta-analytic averages at the 5% level. Using this approach, we obtain effects similar to our main estimates (Table A.13).

**Examine Voluntary versus involuntary policies** There is no reason to expect that bias of this sort would be correlated with the policy effect on spending. However, the most plausible cause for concern regarding correlated bias is for policies that involve voluntary adoption. One may expect that places that voluntarily implement policies that lead to larger spending increase also are more likely to do other things that improve student outcomes. Such dynamics would generate bias correlated with the spending increase and would inflate the marginal estimate. While we cannot entirely rule out this form of bias, because we *can* distinguish studies that rely on variation induced by the voluntary adoption of policies, we are able to test for its potential presence. Specifically, we compare the average marginal effect for studies that rely on a new policy implementation (e.g., budget-increasing referenda) versus those that rely on variation conditional on policies being in place (e.g., differential impacts of the recession or fluctuating student enrolment). We find that studies based on a voluntary policy adoption are similar to other studies and (See Table A.13), suggesting little bias of this form.

### An Approach to Testing For and Removing Bias

The test above suggests that the meta-analytic average likely does not suffer from considerable bias. However, taking the possibility of bias seriously, we present a novel approach to estimating a meta-analytic average that is robust to the existence of the bias laid out in Equation (15) *even if the average of the bias is non-zero.* The meta-analytic average in Equation (14) is an estimate of the average marginal effect across all papers. However, the equation predicting the policy effect on outcomes laid out in Equation (13), reveals that one could also estimate the average marginal effect of *differences* in spending increases by estimating the relationship between the policy-induced

changes in outcomes and the policy-induced changes in spending. Equation (13) indicates that a regression of the change in outcomes for a given policy against the change in spending may yield an estimate of the true average under certain conditions. Abstracting from precision-weighting, the simple linear regression of (13) would yield:

$$\Theta_{diff} = \Theta + \frac{cov(v_j, \Delta\$_j)}{var(\Delta\$_j)} + \frac{cov(b_j, \Delta\$_j)}{var(\Delta\$_j)} \tag{16}$$

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

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

To assess this in our setting, in Figure A.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.[61] Each study is represented by a circle, and larger circles indicate more precise outcome estimates. We also plot the fitted values from a precision-weighted regression relating the

[60] 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 $1000 and test scores by $0.05\sigma$ (leading to $\hat{\theta}_A = 0.05$), while Policy B increases per-pupil spending by $2000 and test scores by $0.04\sigma$ (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 on test scores, so that the difference-based relationship is negative (i.e., $\hat{\theta}_{diff} < 0$). While this may seem counter-intuitive, if there is some correlation between the size of the policy and other contextual factors that determine policy efficacy, this could occur.

[61] There are 6 studies that report policy effects on student outcomes translated into $1000-increases, already having made assumptions about the linear relationship between effect size and per-pupil spending change. For these studies, if possible, we capture the reported average policy effect on per-pupil spending, and adjust the reported policy effect on outcomes assuming linearity in the dollars-effect relationship (Gigliotti and Sorensen (2018), Guryan (2001), **?**, Kreisman and Steinberg (2019)). For the two papers that study Michigan's Proposal A (Hyman (2017) and Roy (2011)), there is no one clear policy effect on per-pupil spending, and we rely on effects-per-$1000 as reported. In Figure 5, we plot and report regression results for all studies—adjusted for the four we can adjust. Our results do not change appreciably when we exclude those studies which do not report one average policy effect on per-pupil spending.

two, along with the 95 percent confidence interval. There is a clear positive relationship between the size of the spending increase caused by a policy and the increase in outcomes associated with that policy. Using random effects meta-regression, the slope is $0.0379\sigma/\$1000$ for test scores and $0.0419\sigma/\$1000$ for educational attainment – both significant. Remarkably, for both outcomes, one fails to reject that the averages of the *within-study* relationships are the same as the *across-study* relationships at the 5 percent significance level.[62] This suggests that, for both outcomes, the documented positive causal relationships between school spending and outcomes are robust. For both test scores and educational attainment, those policies that lead to larger spending increases also lead to larger outcome improvements, on average, and the magnitudes of the differences across policies are similar to those documented within studies. To ensure that our finding is robust, we conduct the same tests excluding studies for which we were forced to make assumptions about the size of the policy effect of spending (Table A.12), and our findings are robust to this.

### A Suggestive Test of the Exclusion Restriction

The difference-based model allows for a direct and intuitive test of the exclusion restriction on average. Specifically, the exclusion restriction is that the only mechanism through which the policies examined affect outcomes is through school spending. If this condition holds, the regression line relating the effect of the policy on outcomes to the effect of the policy on spending should go through the origin. That is, the regression model should predict that a policy that has no effect on school spending should have no effect on outcomes. One can see this mathematically by the fact that the constant term in (13) reflects the average of the noise plus the average of the bias. Given that the average of the noise is zero in expectation, this largely reflects the average of the bias. This is a simple test that the constant in the regression is zero. For test scores, the constant is -0.0026 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. Taken together, the data suggest that the exclusion restriction is likely satisfied for both outcomes.

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

Table A.12: Relationship between Size of Policy Effect on Spending and Student Outcomes

| | (1) Test Scores All | (2) Test Scores w/o assumed | (3) Ed Attain All | (4) Ed Attain w/o assumed | (5) IV Model Test | (6) IV Model Ed Attain |
|---|---|---|---|---|---|---|
| Policy on Exp. ($1000s) | 0.0379*** | 0.0295*** | 0.0419*** | 0.0572*** | 0.0379*** | 0.0419*** |
| | (0.00620) | (0.00758) | (0.0110) | (0.0115) | (0.00620) | (0.0110) |
| Constant | -0.00587 | -0.00226 | 0.0153 | 0.0138 | -0.00587 | 0.0153 |
| | (0.00669) | (0.00695) | (0.00740) | (0.0128) | (0.00669) | (0.00740) |
| N | 40 | 33 | 25 | 18 | 40 | 25 |
| Pr(slope = pooled avg.) | 0.454 | 0.900 | 0.294 | 0.243 | | |
| Overidentification p-val | | | | | 0.107 | 0.774 |
| F-Stat | | | | | 23.39 | 17.21 |

Standard errors in parentheses are adjusted for clustering of related papers.

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

**Note:** The excluded instrument in the Instrumental Variables (IV) models (columns 5 and 6) are the individual study indicators.

Table A.13: Meta-Regression Estimates by Power and Policy Categories

| | Power to Detect Main Effect | | By Policy Categories | |
|---|---|---|---|---|
| | (1) Test Score | (2) Educational Attainment | (3) Test Score | (4) Educational Attainment |
| Average Effect | 0.0255*** | 0.0493*** | 0.0363*** | 0.0475*** |
| | (0.00901) | (0.00602) | (0.00927) | (0.0123) |
| Voluntary Policy | | | -0.00805 | 0.0217 |
| | | | (0.0112) | (0.0144) |
| Observations | 12 | 10 | 40 | 25 |
| $\tau$ | 0.0223 | 0.0157 | 0.0193 | 0.0273 |
| Average 90% PI | [-0.014,0.065] | [0.022,0.077] | [-0.001,0.073] | [-0.003,0.098] |

Standard errors in parentheses are adjusted for clustering of related papers.

Voluntary Policy includes: Equalization, Referenda,

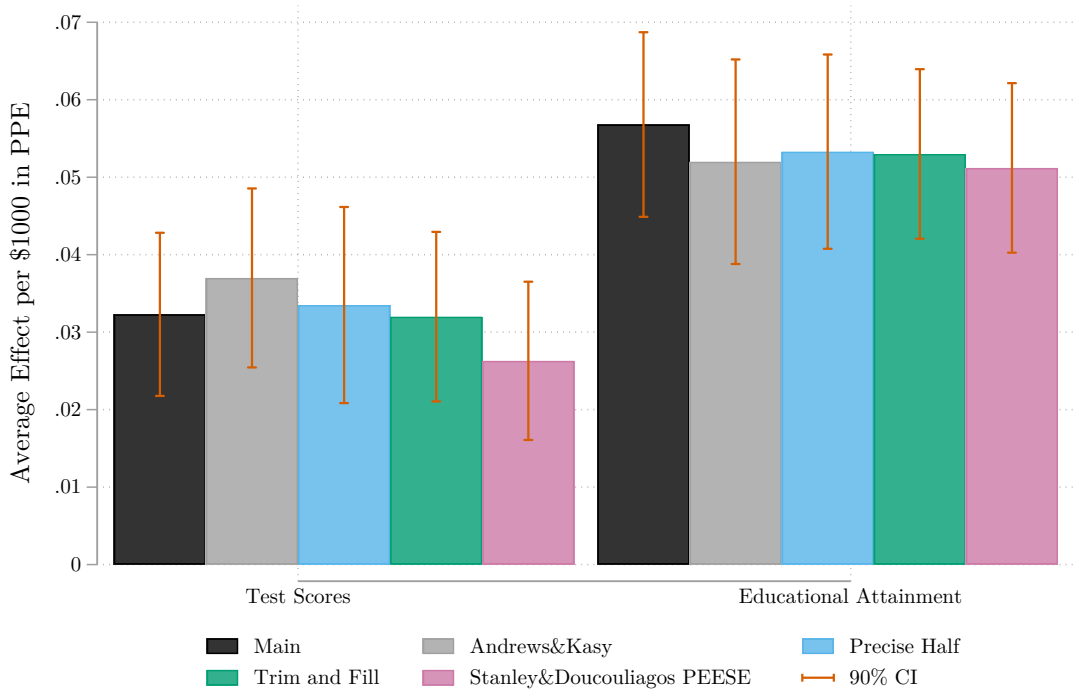
School Finance Reform, New Construction, and School Improvement Grants.

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

## F.2 Publication Biases

Our analysis may be biased if certain kinds of studies – especially those which find no effect of a policy or intervention – are systematically not published. There are two kinds of publication biases that one may worry about in our context. First, journals may be less likely to publish studies that are not statistically significant. If so, assuming that there is an overall positive effect, those studies with larger positive impacts (and therefore larger $t$-statistics) will be more likely to be published – such that the average among published studies may overstate effects. Second, if researchers and journals are more likely to publish results consistent with "preferred" results, precisely estimated impacts of all signs will be published (because they are credible), but imprecise studies (where the results are are more ambiguous) of the non-preferred sign will be disproportionately not published. This would lead to a meta-analytic average biased toward the preferred result. We conduct several tests to assess the extent to which these are a concern in our setting. We visually represent estimates from these approaches in Figure A.19, present regression results in Table A.14, and summarize them here:

Figure A.19: Four Approaches to Publication Bias




1. Studies that find null results may be less likely to be published than studies that find significant effects (Franco et al. (2014), Christensen and Miguel (2018)). If one can observe studies that are not published, a simple test for publication bias compares estimates from published studies to those that are not published. In line with this, we compare average estimates of published and unpublished studies and find no difference in impacts.[63] In Table A.15, the coefficients on the indicator for "Unpublished"

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

show no evidence that there is any difference in average effects reported in published versus unpublished papers for both test scores and educational attainment outcomes.

2. Related to the first test, if there are biases against the publication of certain kinds of studies, one might expect these biases to be most pronounced at the most selective journals (Brodeur et al. (2016)). Informed by this notion, we compare the average impacts of studies published in the most elite journals to studies published in other journals, and similarly find no differences across journal prestige (in columns 2 and 4 of Table A.15, the formal tests of equality across publication type suggest no evidence that publication status or type have any bearing on the estimates reported in studies of effects of school spending.

3. Publication bias is thought to be most prevalent among imprecise studies (Andrews and Kasy (2019)), and when there are biases against the publication of insignificant studies, one might observe an over-representation of studies right at the significance threshold (in social sciences this would be the 5 percent level pertaining to a $t$-statistic of 1.96) and an under-representation of studies right below the significance threshold (Brodeur et al. (2020)). To test for this in our data, we test for a discontinuity in the cumulative density of t-statistics at 1.96. We show that there is no over-representation of studies right at the significance threshold ($t$-statistic = 1.96) in Figure A.22. In Table A.16, we show that there is no significant jump in density, by outcome type or combining across both test score and educational attainment outcomes, at the significance threshold ($t$-stat $> 1.96$) – indeed, for test scores there happens to be a *decline* at that point.

4. Even though we find limited evidence of selection of significant impacts, we implement a model that accounts for any such selection (should it exist). To this aim, we show results for the Andrews and Kasy (2019) selection adjustment using their web application in Figures A.23 and A.24. They propose estimating the publication probabilities (based on the $t$-statistics) for studies, and using these probabilities to produce bias-corrected estimators and confidence sets. More specifically, using the relative publication probabilities, this approach re-weights the distribution of studies to account for differences in publication probability (up-weighting studies that are least likely to be observed). For both test scores and educational attainment, their model fails to reject the null of no selection at the 1.96 $t$-statistic threshold. Reassuringly, their adjustment approach yields similar estimates to our preferred model (columns 1 and 5 of Table A.14).

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

   We visualize the Duval and Tweedie (2000) "trim and fill" approach in Figure A.20, where black circles indicate the individual study impacts. The distribution of effects is largely symmetrical around the mean for very precise studies (at the top of the figures), but the distribution may be asymmetric for studies with larger standard errors (the bottom of the plots). That is, while there is little visual

evidence of publication bias among precisely estimated studies, there is some suggestive evidence that imprecise positive studies with large impacts may be more likely to be published (or written) than imprecise studies with negative or small impacts. To be clear, because (a) our inclusion criteria require that the policy has meaningful impacts on school spending, and (b) one would expect there to be some effect heterogeneity across states and policies, some asymmetry is likely even absent publication bias. Even so, to be conservative, one can assume that any asymmetry is due to publication bias, and assess the impacts of this asymmetry on the estimated pooled average. We follow this approach.

In the left panel of Figure A.20, to create symmetry, the "trim and fill" approach imputes four "missing" studies of test score outcomes (green triangles) – both of which are negative and very imprecise. These imputed studies are outside of the more precise range employed for our first test of bias – validating that approach. The re-estimated pooled effect that includes these four additional imputed studies is 0.032 (Table A.14 column 3) – very similar to our original estimate including all observed estimates. Following this same approach for educational attainment, "trim and fill" imputes five additional negative and relatively imprecise estimates. The re-estimated pooled effect that includes the three additional imputed studies is 0.053 (Table A.14 column 7) – also similar to our original estimate including all observed estimates. The fact that estimates do not change very much with the imputed data also reflects the fact that the evidence of asymmetry is only among very imprecise estimates, which receive lower weight in our precision-weighted pooled average. This suggests that the impacts of any *potential* publication bias on our estimates are small.

When we estimate our main model on all studies using a drastic approach of dropping the majority of the data (Stanley et al. (2010)), specifically those test score studies with an estimated standard error of 0.023 or less (Table A.14 column 2) and educational attainment studies with estimated standard errors of 0.021 or less (Table A.14 column 6), our results are similar to our main models. We indicate these precision levels in the higher horizontal lines in the funnel plot in Figure A.20. Above this cut-off, estimates are very tightly clustered around the pooled average.[64] In this most precise sample (where there is no evidence of asymmetry), the coefficient estimate for test scores is 0.03 Table A.14 column 3). This is very similar to our preferred estimate – indicating minimal bias. Following this same approach for educational attainment, when we restrict our sample to studies with standard errors below 0.021, the Egger's tests indicate no asymmetry, and the regression estimate is 0.0533 (Table A.14 column 6).[65]

Finally, we follow both Stanley and Doucouliagos (2014) and Ioannidis et al. (2017) and implement the precision-effect estimate with standard error (PEESE) approach. This approach estimates the relationship between the precision of the estimates and the estimates reported in each study. Under the assumption that the most precise estimates will yield the true relationship, one can empirically model the relationship between the precision of the estimates and the reported estimates and then infer what the most precise estimate would be. In practice, this involves regressing the reported effect on the square of its precision and taking the constant term as the bias-adjusted estimate. This approach has been found to perform well in simulations. This approach yields a meta-regression estimate which takes into account the influence of publication bias – based on estimate precision. In columns (4) and (8) of Table A.14 we report meta-regression results. For test scores, the PEESE method estimates a precision-weighted pooled average of 0.0263 and for educational attainment of 0.0512.

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

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

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

Table A.14: Meta-Regressions w/ Approaches to Potential Biases

| | Test Scores | | | | Educational Attainment | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Avg. Effect | 0.037*** | 0.0335*** | 0.032*** | 0.0263*** | 0.052*** | 0.0533*** | 0.053*** | 0.0512*** |
| | (0.007) | (0.00767) | (0.00638) | (0.00663) | (0.008) | (0.0076) | (0.00663) | (0.00704) |
| $\tau$ | 0.022 | 0.0242 | . | 0.0195 | 0.010 | 0.0117 | . | 0.00713 |
| Observations | 26 | 13 | 29 | 26 | 12 | 6 | 17 | 12 |

Standard errors in parentheses. When possible, standard errors are adjusted for clustering of related papers.

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

Educational Attainment: (5) Andrews & Kasy (6) SE $<$ .021 (7) Meta Trim&Fill (8) PEESE

$^{*}$ $p < .1$, $^{**}$ $p < .05$, $^{***}$ $p < .01$

Figure A.20: Funnel Plots

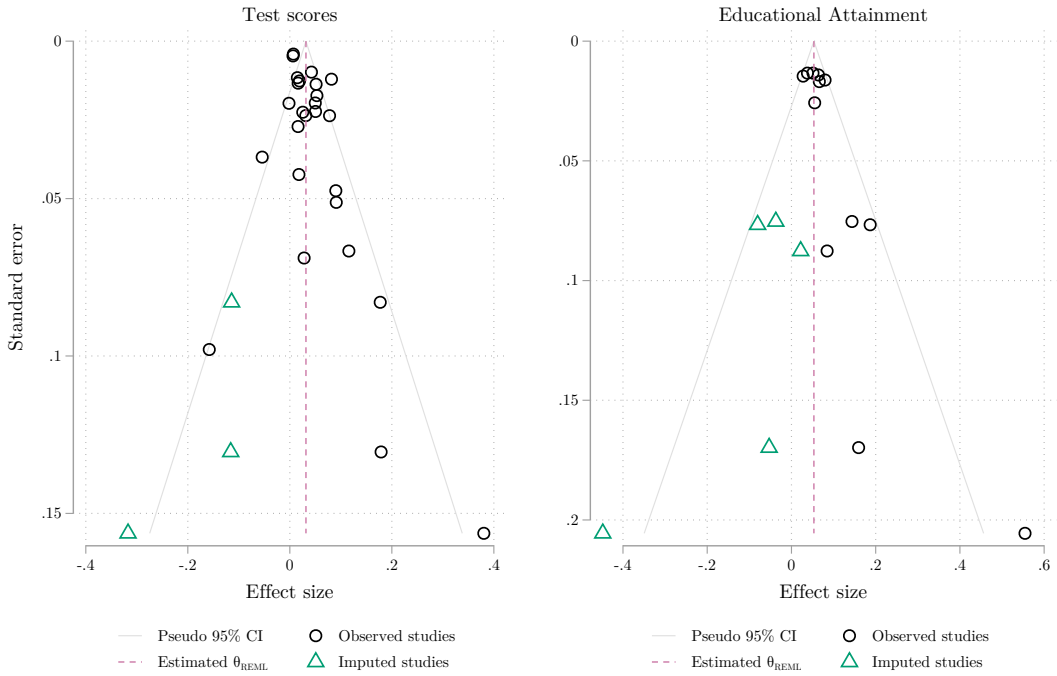



Figure A.21: Funnel Plots, Multiple Estimates per Paper

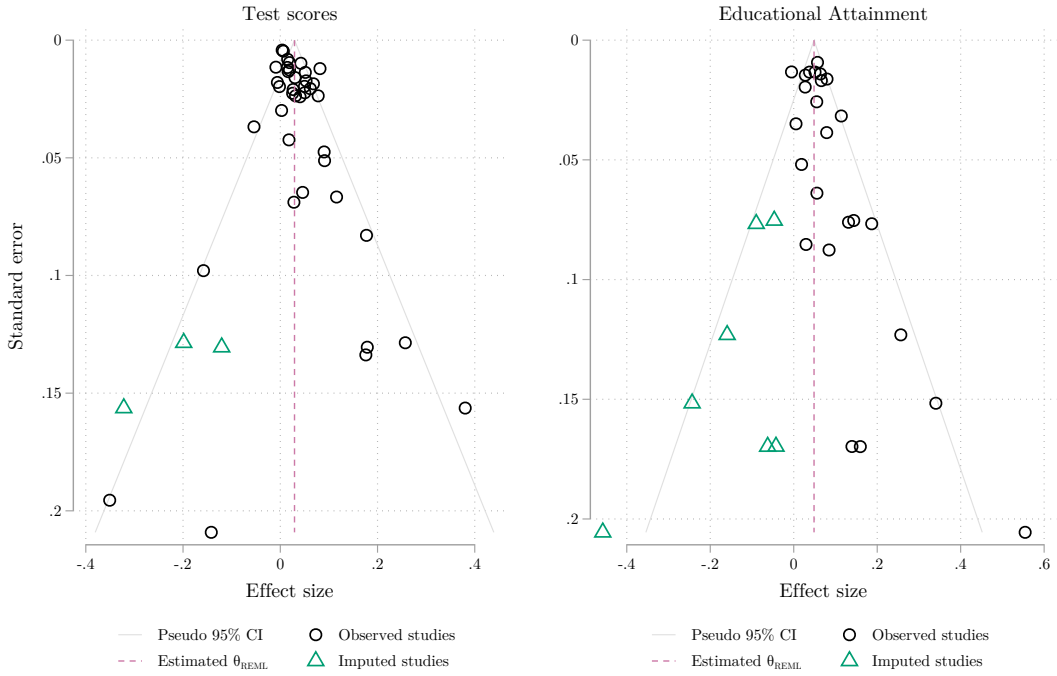


Table A.15: Meta-Regressions w/ Publication Type

| | (1) Test Score | (2) Test Score | (3) Educational Attainment | (4) Educational Attainment |
|---|---|---|---|---|
| Unpublished | -0.0109 | -0.00687 | 0.0137 | 0.00448 |
| | (0.0123) | (0.0175) | (0.0203) | (0.0308) |
| Top Field Journal | | 0.0100 | | -0.0349 |
| | | (0.0170) | | (0.0269) |
| Field Journal | | 0.00611 | | 0.00274 |
| | | (0.0182) | | (0.0202) |
| Average Effect | 0.0363*** | 0.0322** | 0.0560*** | 0.0685*** |
| | (0.00679) | (0.0138) | (0.00998) | (0.0191) |
| N | 40 | 40 | 25 | 25 |
| $\tau$ | 0.0226 | 0.0246 | 0.0294 | 0.0462 |
| Top Field = Field = Unpublished = 0 (p-val) | | 0.687 | | 0.352 |
| Unpublished = 0 (p-val) | 0.376 | 0.694 | 0.500 | 0.884 |

Standard errors in parentheses are adjusted for clustering of related papers.

Reference category High Impact ommitted.

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

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

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

$^{*}$ $p < .1$, $^{**}$ $p < .05$, $^{***}$ $p < .01$

Table A.16: Regressions to test for jump at 5% significance, Outcome: Cumulative T-stat density

| | (1) Test Scores (all tstats) | (2) Test Scores $1 < tstat < 3$ | (3) Ed. Attain (all tstats) | (4) Ed. Attain $1 < tstat < 3$ | (5) All Outcomes (all tstats) | (6) All Outcomes $1 < tstat < 3$ |
|---|---|---|---|---|---|---|
| Sig, 5%-level (ind) | 0.0597 | -0.0151 | 0.110* | -0.00669 | 0.0763* | -0.0148 |
| | (0.0553) | (0.0260) | (0.0572) | (0.0724) | (0.0411) | (0.0272) |
| $N$ | 26 | 15 | 12 | 6 | 38 | 21 |

Standard errors in parentheses

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

In column 5 pooled models (both outcome types) we include an indicator

for the outcome and interact t-stat and t-stat squared with the outcome.

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

Figure A.22: Histogram of *all* effects

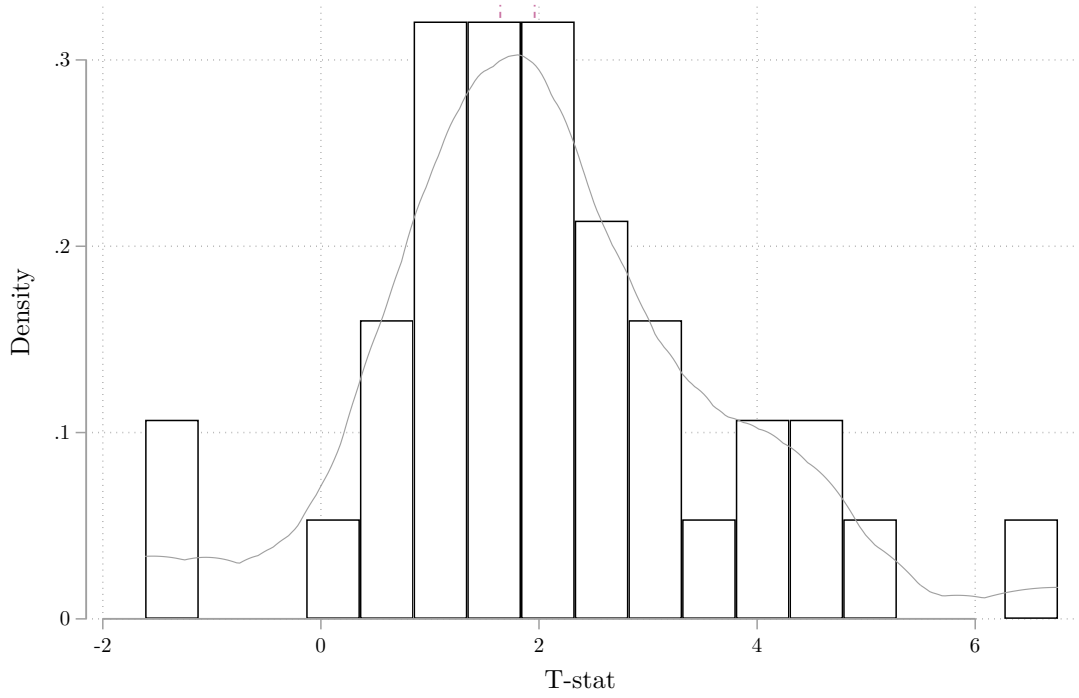


Figure A.23: Test Scores, One Estimate per Study

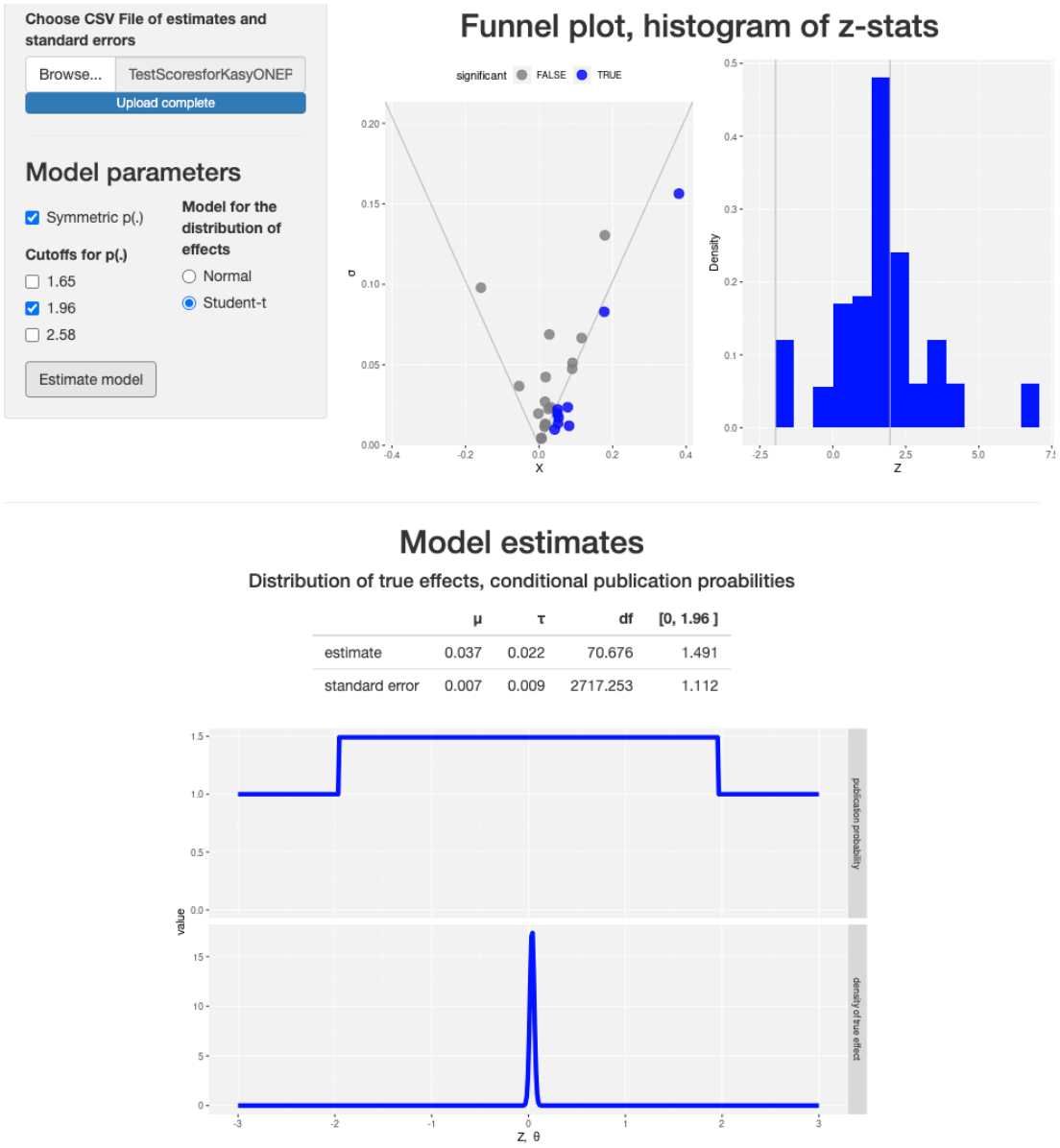
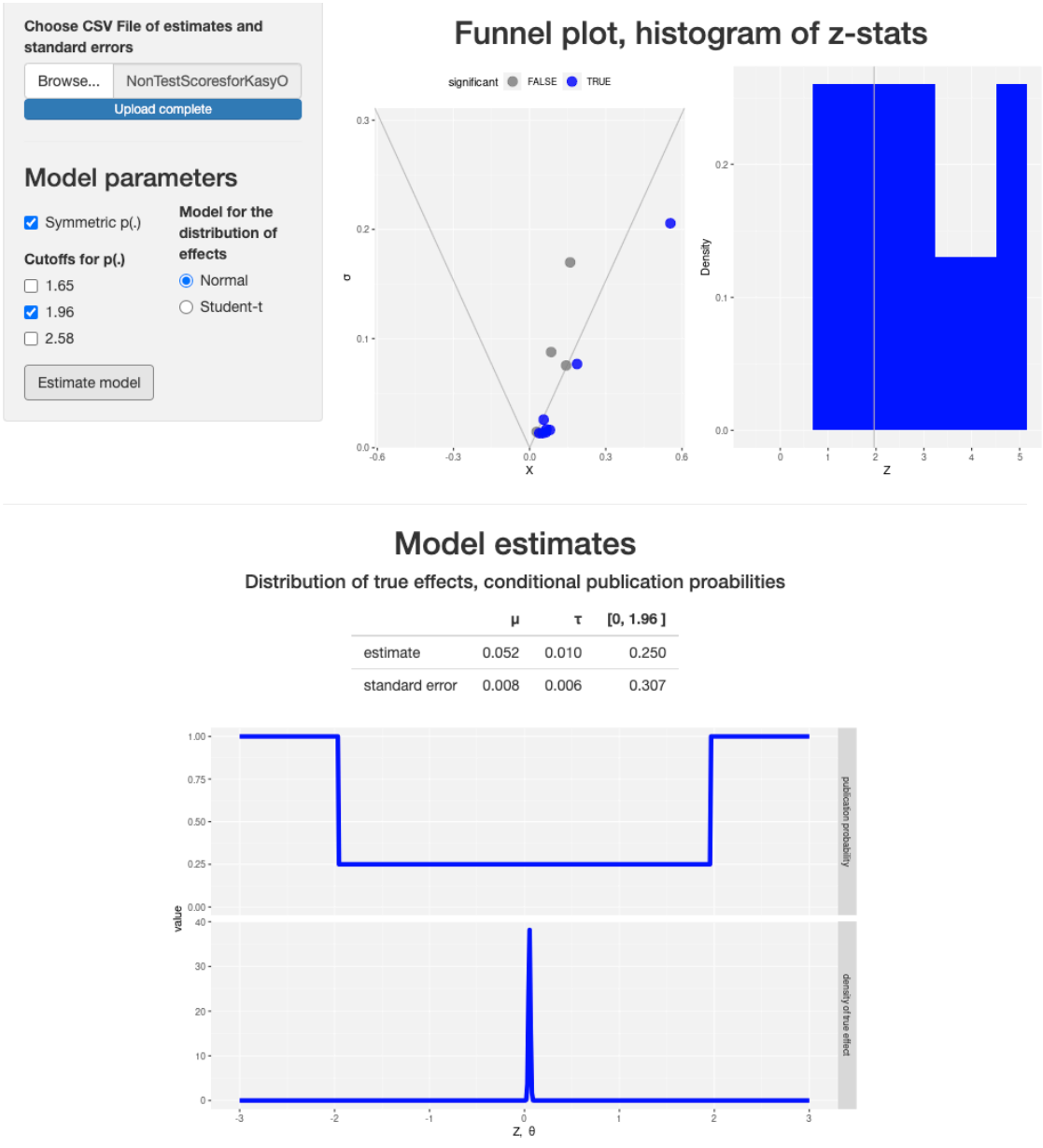


Figure A.24: Non-Test Scores, One Estimate per Study



# G Bayesian Estimates

We implement a specific random effects meta-regression model. To assuage any concerns about our choice of method, we also implement a full Bayesian model. The Bayesian model setup is identical to that of the random effects model (that is, equations (1), (2), ((4)), and (4)). How these models differ is how they estimate $\tau$. By imposing some additional assumptions on the distribution of $\tau$, the Bayesian model obtains estimates of the model parameters that perform well even in small samples.

### Estimating $\tau$ Using a Bayesian Approach

It is helpful to clarify some notation. Let the set of true effects be $\theta = [\theta_1, \theta_2, \theta_3, .....\theta_J]'$. The observed estimates of these true effects are $\hat{\theta} = [\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, .....\hat{\theta}_J]'$. The corresponding sampling standard deviations are $\sigma = [\sigma_1, \sigma_2, \sigma_3, .....\sigma_J]'$ which is approximated by $se = [se_1, se_2, se_3, .....se_J]'$. Because the probability of observing the estimated effects ($\hat{\theta}$) is a function of true effects ($\theta$), the probability of which is determined by $\tau$, the likelihood of observing estimates ($\hat{\theta}$) and sampling standard deviations ($se$) can be computed for any given value of $\tau$, $\Theta$, and $\theta$ – that is, $\mathcal{L}(\tau, \Theta, \theta)$. Frequentist approaches, such as Maximum Likelihood, solve for the values of $\tau$, $\Theta$, and $\theta$ that maximize this likelihood.

Bayes' rule says that the joint posterior probability for the parameters (i.e., $p(\tau, \theta, \Theta|\hat{\theta}, se)$, is proportional to the likelihood of the data given certain parameter values ($\mathcal{L}(\tau, \Theta, \theta)$) multiplied by the prior probability of those parameters ($\pi(\tau, \Theta, \theta)$). As such, using Bayes rule, given some prior distribution, one can compute the posterior distribution of the true effects $\Theta$, $\theta$, and $\tau$. Moments (such as the mean) of the posterior distributions of $\tau$,$\Theta$, and $\theta$ provide information about the values of these parameters. Moreover, the spread of the posterior distributions sheds light on the uncertainty around the values of these parameters.

The Bayesian model works as follows:

1. One chooses a probability density — i.e., prior distribution — that expresses beliefs about the distribution of each parameter *before seeing any data.*

2. One defines a statistical model $p(\hat{\theta}, se|\tau, \theta, \Theta)$ that reflects our beliefs about the data given the parameters.

3. After observing data $\hat{\theta}$ and $se$, the model updates our beliefs using Bayes rule and calculates the joint posterior distribution for the parameters of interest $p(\tau, \theta, \Theta|\hat{\theta}, se)$.

4. The model takes random draws of $\tau$, $\Theta$, and $\theta$ from the posterior distributions and reports moments (in our case, the mean) of the posterior distribution of the parameter estimates. Note that $p(\theta, \Theta, \tau|\hat{\theta}, se)$ can be written as $p(\theta|\Theta, \tau, \hat{\theta}, se)p(\Theta, \tau|\hat{\theta}, se)p(\tau, |\hat{\theta}, se)$. As such, the model will draw the hyperparameters $\tau$, then $\Theta$, from their marginal posterior distributions and then draw $\theta$ from its posterior distribution conditional on the drawn values of $\tau$ and $\Theta$.

Under this approach, we must define the prior distributions for $\tau$ and $\Theta$. To this aim, we assume that the true effect is a random draw from a normal distribution (justified by the central limit theorem), and that the heterogeneity parameter $\tau^2$ follows an inverse Gamma distribution as in (14) and (15).

$$\Theta \sim \mathcal{N}(.) \tag{17}$$

$$\tau^2 \sim InvGamma(.) \tag{18}$$

The inverse Gamma distribution is commonly used to model variance parameters and avoids the non-negative estimates one can obtain from method of moments approaches. Reassuringly, We obtain similar results if we assume a $\chi^2$ distribution.

We estimate this model with starting values such that $\tau^2 \sim InvGamma(0.0001, 0.0001)$ and that $\Theta \sim \mathcal{N}(0, 100)$. The model estimates are reported in Table A.17 and A.18. The model converges well and provides very similar results across simulations and starting values – suggesting that the results are sensible and largely driven by the data (as opposed to the priors). Consistent with this, the resulting $\Theta$ and $\tau$ (and the uncertainty in these estimates) from these models are similar to those using frequentist methods.

Table A.17: Bayes Estimates,

| | One Estimate Per Study | | | | Multiple Estimates Per Study | | | |
|---|---|---|---|---|---|---|---|---|
| | Test Scores | | Educational Attainment | | Test Scores | | Educational Attainment | |
| | RE | Bayes | RE | Bayes | RE | Bayes | RE | Bayes |
| $\theta$ | 0.032 | 0.034 | 0.057 | 0.059 | 0.032 | 0.031 | 0.057 | 0.057 |
| | (0.006) | (0.007) | (0.007) | (0.011) | (0.006) | (0.007) | (0.009) | (0.010) |
| $\tau$ | 0.022 | 0.025 | 0.017 | 0.022 | 0.021 | 0.026 | 0.027 | 0.022 |
| $\tau$ 95% CI | | (0.014, 0.039) | | (0.007, 0.046) | | (0.015, 0.042) | | (0.007, 0.048) |
| N | 26 | 26 | 12 | 12 | 40 | 40 | 25 | 25 |

**Note:** We report random effects estimates in the columns labeled RE. Those with the header labeled Bayes are from the full Bayesian model. $\tau$ 95% CI represents the reported 95% Credibility Interval obtained from the Bayesian model.

Table A.18: Bayes Estimates, , by First Stage Strength

| | F-stat > 10 | | | | F-stat > 20 | | | |
|---|---|---|---|---|---|---|---|---|
| | Test Scores | | Educational Attainment | | Test Scores | | Educational Attainment | |
| | RE | Bayes | RE | Bayes | RE | Bayes | RE | Bayes |
| $\theta$ | 0.033 | 0.035 | 0.054 | 0.056 | 0.033 | 0.039 | 0.055 | 0.060 |
| | (0.006) | (0.008) | (0.009) | (0.017) | (0.008) | (0.009) | (0.010) | (0.047) |
| $\tau$ | 0.022 | 0.024 | 0.028 | 0.032 | 0.020 | 0.027 | 0.019 | 0.078 |
| N | 30 | 30 | 13 | 13 | 18 | 18 | 8 | 8 |

**Note:** We report random effects estimates in the columns labeled RE. Those with the header labeled Bayes are from the full Bayesian model.